David J. Olive

Large Sample Theory

January 15, 2025

Preface

Many statistics departments offer a one semester graduate course in large sample theory. There are several PhD level texts on large sample theory including, in roughly increasing order of difficulty, Lehmann (1999), Ferguson (1996), Sen and Singer (1993), and Serfling (1980). Cramér (1946) is also an important reference, and White (1984) considers asymptotic theory for econometric applications. The online text Hunter (2014) is useful. Also see DasGupta (2008), Davidson (2021), Hall and Oakes (2024), Jiang (2022), Polansky (2011), Sen, Singer, and Pedrosa De Lima (2010), and van der Vaart (1998). A nice review of large sample theory is Chernoff (1956).

More advanced topics for large sample theory can be found in Lukacs (1970, 1975), Petrov (1995), Pollard (1984) and Shorack and Wellner (1986).

For some roughly Master's level large sample theory, see Bickel and Doksum (1977, section 4.4), Casella and Berger (2002, section 5.5), Hoel, Port, and Stone (1971, sections 8.2-8.4), Lehmann (1983, ch. 5), Olive (2014, ch. 8), Rohatgi (1976, ch. 6), Rohatgi (1984, ch. 9), and Woodroofe (1975, ch. 9). For some PhD level large sample theory, see Olive (2023e, ch. 4).

The prerequisite for this text is a Master's level course in Statistics (USA) such as Casella and Berger (2002) or Olive (2014).

Some highlights of this text follow.

- The large sample theory for the elastic net, lasso, and ridge regression is greatly simplified.
- Large sample theory is given for many variable selection estimators, including multiple linear regression, many GLMs, some time series models, and some survival regression models.
- The large sample theory for the one component partial least squares estimator and marginal maximum likelihood estimator is greatly simplified. Some of the hypothesis tests are valid in high dimensions.
- Large sample theory for prediction regions is given, including a prediction region that works in high dimensions. This theory is used to greatly simplify the large sample theory for some bootstrap confidence regions.

- Theory for some robust statistics is given.

Downloading the book's R functions *lsamppack.txt* and data files *lsampdata.txt* into *R*: The commands

```
source("http://parker.ad.siu.edu/Olive/lspack.txt")
source("http://parker.ad.siu.edu/Olive/lsdata.txt")
```

Acknowledgements

Teaching large sample theory, both as a reading course and as Math 582 in 2022 and 2024, at Southern Illinois University was useful.

Contents

Chapter 1
Introduction

This chapter follows Olive (2014, ch. 1-3) closely. Much of the material can be skimmed, and then the reader can refer back to this chapter as needed.

Often large sample theory is taught after a course in probability and measure, and a probability space $(S, \mathcal{B}, P)$ is used where $\mathcal{B}$ is a σ-field. This text will usually ignore measure theoretic probability. Unless told otherwise, the notation $P(A)$ means that A is an event.

Definition 1.1. *Statistics* is the science of extracting useful information from data.

1.1 Probability, Expected Value, CDF

Definition 1.2. The *sample space* S is the set of all possible outcomes of an experiment.

Definition 1.3. Let $\mathcal{B}$ be a special field of subsets of the sample space S forming the class of events. Then A is an *event* if $A \in \mathcal{B}$.

In the definition of an event above, the special field of subsets $\mathcal{B}$ of the sample space S forming the class of events will not be formally given. However, $\mathcal{B}$ contains all "interesting" subsets of S and every subset that is easy to imagine. The point is that not necessarily all subsets of S are events, but every event A is a subset of S.

The *empty set* $\emptyset$ is the set that contains no elements. The set A is a *subset* of B, written $A \subseteq B$, if every element in A is in B. The *union* $A \cup B$ of A with B is the set of all elements in A or B or in both. The *intersection* $A \cap B$ of A with B is the set of all elements in A and B. The *complement* of A, written $\overline{A}$ or A^c, is the set of all elements in S but not in A. If $A = \emptyset$, then A and B are disjoint. In the following definition, disjoint events are often called pairwise disjoint events.

Definition 1.4. If $A \cap B = \emptyset$, then A and B are *mutually exclusive* or *disjoint events.* Events $A_1, A_2, ...$ are *disjoint* or *mutually exclusive* if $A_i \cap A_j = \emptyset$ for $i \neq j$.

Definition 1.5. Let $\mathcal{B}$ be the class of events of the sample space S. A **probability function** $P : \mathcal{B} \to [0, 1]$ is a set function satisfying the following three properties:
P1) $P(A) \geq 0$ for all events A,
P2) $P(S) = 1$, and
P3) if $A_1, A_2, ...$ are disjoint events, then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

If $A_1, ..., A_n$ are disjoint, then $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$. This result follows from Definition 1.5 using $A_i = \emptyset$ for $i > n$.

Theorem 1.1. DeMorgan's Laws:
a) $\overline{A \cup B} = \overline{A} \cap \overline{B}$.
b) $\overline{A \cap B} = \overline{A} \cup \overline{B}$.
c) $(\bigcup_{i=1}^{\infty} A_i)^c = \bigcap_{i=1}^{\infty} A_i^c$.
d) $(\bigcap_{i=1}^{\infty} A_i)^c = \bigcup_{i=1}^{\infty} A_i^c$.

Proof. The proofs of a) and b) are similar to those of c) and d), and "iff" means "if and only if."

c) $(\bigcup_{i=1}^{\infty} A_i)^c$ occurred iff $\bigcup_{i=1}^{\infty} A_i$ did not occur, iff none of the A_i occurred, iff all of the A_i^c occurred, iff $\bigcap_{i=1}^{\infty} A_i^c$ occurred.

d) $(\bigcap_{i=1}^{\infty} A_i)^c$ occurred iff not all of the A_i occurred, iff at least one of the A_i^c occurred, iff $\bigcup_{i=1}^{\infty} A_i^c$ occurred. □

Theorem 1.2. Let A and B be any two events of S. Then
i) $0 \leq P(A) \leq 1$.
ii) $P(\emptyset) = 0$ where $\emptyset$ is the empty set.
iii) **Complement Rule:** $P(A) = 1 - P(\overline{A})$.
iv) **General Addition Rule:** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
v) If $A \subseteq B$, then $P(A) \leq P(B)$.
vi) **Boole's Inequality:** $P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$ for any events $A_1, A_2, ...$
vii) **Bonferroni's Inequality:** $P(\bigcap_{i=1}^{n} A_i) \geq \sum_{i=1}^{n} P(A_i) - (n-1)$ for any events $A_1, A_2, ..., A_n$.

Note that A and $\overline{A}$ are disjoint and $A \cup \overline{A} = S$. Hence $1 = P(S) = P(A \cup \overline{A}) = P(A) + P(\overline{A})$, proving the complement rule. Note that S and $\emptyset$ are disjoint, so $1 = P(S) = P(S \cup \emptyset) = P(S) + P(\emptyset)$. Hence $P(\emptyset) = 0$. If $A \subseteq B$, let $C = \overline{A} \cap B$. Then A and C are disjoint with $A \cup C = B$. Hence $P(A) + P(C) = P(B)$, and $P(A) \leq P(B)$ by i).

Following Casella and Berger (2002, p. 13), $P(\bigcup_{i=1}^{n} A_i^c) = P[(\bigcap_{i=1}^{n} A_i)^c] = 1 - P(\bigcap_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i^c) = \sum_{i=1}^{n} [1 - P(A_i)] = n - \sum_{i=1}^{n} P(A_i)$, where the first equality follows from DeMorgan's Laws, the second equality holds by the complement rule, and the inequality holds by Boole's inequality

$P(\bigcup_{i=1}^{n} A_i^c) \leq \sum_{i=1}^{n} P(A_i^c)$. Hence $P(\bigcap_{i=1}^{n} A_i) \geq \sum_{i=1}^{n} P(A_i) - (n-1)$, and Bonferonni's inequality holds.

If $A_1, A_2, ...$ are disjoint and if $\bigcup_{i=1}^{\infty} A_i = S$, then the collection of sets $A_1, A_2, ...$ is a *partition* of S. By taking $A_j = \emptyset$ for $j > k$, the collection of disjoint sets $A_1, A_2, ..., A_k$ is a partition of S if $\bigcup_{i=1}^{k} A_i = S$. The **conditional probability** of A **given** B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

if $P(B) > 0$.

Theorem 1.3: Law of Total Probability. If $A_1, A_2, ..., A_k$ form a partition of S such that $P(A_i) > 0$ for $i = 1, ..., k$, then

$$P(B) = \sum_{j=1}^{k} P(B \cap A_j) = \sum_{j=1}^{k} P(B|A_j)P(A_j).$$

Definition 1.6. A *random variable* Y is a real valued function with a sample space as a domain: $Y : S \to \mathbb{R}$ where the set of real numbers $\mathbb{R} = (-\infty, \infty)$.

Definition 1.7. The *population* is the entire group of objects from which we want information. The *sample* is the part of the population actually examined.

For the following definition, F is a right continuous function if for every real number x, $\lim_{y \downarrow x} F(y) = F(x)$. Also, $F(\infty) = \lim_{y \to \infty} F(y)$ and $F(-\infty) = \lim_{y \to -\infty} F(y)$.

Definition 1.8. The **cumulative distribution function** (cdf) of any random variable Y is $F(y) = P(Y \leq y)$ for all $y \in \mathbb{R}$.

Cumulative distribution functions are very important for convergence in distribution. See Chapter 2. If $F(y)$ is a cumulative distribution function, then i) $F(-\infty) = \lim_{y \to -\infty} F(y) = 0$, ii) $F(\infty) = \lim_{y \to \infty} F(y) = 1$, iii) F is a nondecreasing function: if $y_1 < y_2$, then $F(y_1) \leq F(y_2)$, iv) F is right continuous: $\lim_{h \downarrow 0} F(y + h) = F(y)$ for all real y. v) Since a cdf is a probability for fixed y, $0 \leq F(y) \leq 1$ for all real y. vi) A cdf $F(y)$ can have at most countably many points of discontinuity, vii) $P(a < Y \leq b) = F(b) - F(a)$.

Definition 1.9. A random variable is **discrete** if it can assume only a finite or countable number of distinct values. The collection of these probabilities is the *probability distribution* of the discrete random variable.

The **probability mass function** (pmf) of a discrete random variable Y is $f(y) = P(Y = y)$ for all $y \in \mathbb{R}$ where $0 \leq f(y) \leq 1$ and $\sum_{y:f(y)>0} f(y) = 1$.

Remark 1.1. The cdf F of a discrete random variable is a step function with a jump of height $f(y)$ at values of y for which $f(y) > 0$.

Definition 1.10. A random variable Y is **continuous** if its distribution function $F(y)$ is absolutely continuous.

The notation $\forall y$ means "for all y."

Definition 1.11. If Y is a continuous random variable, then a **probability density function** (pdf) $f(y)$ of Y is an integrable function such that

$$F(y) = \int_{-\infty}^{y} f(t)dt \tag{1.1}$$

for all $y \in \mathbb{R}$. If $f(y)$ is a pdf, then $f(y)$ is continuous except at most a countable number of points, $f(y) \geq 0 \ \forall y$, and $\int_{-\infty}^{\infty} f(t)dt = 1$.

Theorem 1.4. If Y has pdf $f(y)$, then $f(y) = \frac{d}{dy}F(y) \equiv F'(y)$ wherever the derivative exists (in this text the derivative will exist and be continuous except for at most a finite number of points in any finite interval).

Theorem 1.5. i) $P(a < Y \leq b) = F(b) - F(a)$.
ii) If Y has pdf $f(y)$, then $P(a < Y < b) = P(a < Y \leq b) = P(a \leq Y < b) = P(a \leq Y \leq b) = \int_a^b f(y)dy = F(b) - F(a)$.
iii) If Y has a probability mass function $f(y)$, then Y is discrete and $P(a < Y \leq b) = F(b) - F(a)$, but $P(a \leq Y \leq b) \neq F(b) - F(a)$ if $f(a) > 0$.

Definition 1.12. Let Y be a discrete random variable with probability mass function $f(y)$. Then the *mean* or **expected value** of Y is

$$EY \equiv E(Y) = \sum_{y:f(y)>0} y \ f(y) \tag{1.2}$$

if the sum exists when y is replaced by $|y|$. If $g(Y)$ is a real valued function of Y, then $g(Y)$ is a random variable and

$$E[g(Y)] = \sum_{y:f(y)>0} g(y) \ f(y) \tag{1.3}$$

if the sum exists when $g(y)$ is replaced by $|g(y)|$. If the sums are not absolutely convergent, then $E(Y)$ and $E[g(Y)]$ do not exist.

Definition 1.13. If Y has pdf $f(y)$, then the *mean* or **expected value** of Y is

$$EY \equiv E(Y) = \int_{-\infty}^{\infty} yf(y)dy \tag{1.4}$$

and

$$E[g(Y)] = \int_{-\infty}^{\infty} g(y)f(y)dy \tag{1.5}$$

provided the integrals exist when y and $g(y)$ are replaced by $|y|$ and $|g(y)|$. If the modified integrals do not exist, then $E(Y)$ and $E[g(Y)]$ do not exist.

Definition 1.14. If $E(Y^2)$ exists, then the *variance* of a random variable Y is

$$\text{VAR}(Y) \equiv \text{Var(Y)} \equiv \text{V Y} \equiv \text{V(Y)} = \text{E}[(\text{Y} - \text{E(Y)})^2]$$

and the *standard deviation* of Y is $\text{SD}(Y) = \sqrt{V(Y)}$. If $E(Y^2)$ does not exist, then $V(Y)$ does not exist.

The notation $E(Y) = \infty$ or $V(Y) = \infty$ when the corresponding integral or sum diverges to ∞ can be useful. The following theorem is also used to find $E(Y^2) = V(Y) + (E(Y))^2$. The theorem is valid for all random variables that have a variance, including continuous and discrete random variables. If Y is a Cauchy (μ, σ) random variable, then neither $E(Y)$ nor $V(Y)$ exist.

Theorem 1.6: Short cut formula for variance.

$$V(Y) = E(Y^2) - (E(Y))^2. \tag{1.6}$$

If Y is a discrete random variable with sample space $S_Y = \{y_1, y_2, ..., y_k\}$ then

$$E(Y) = \sum_{i=1}^{k} y_i f(y_i) = y_1 f(y_1) + y_2 f(y_2) + \cdots + y_k f(y_k)$$

and $E[g(Y)] = \sum_{i=1}^{k} g(y_i)f(y_i) = g(y_1)f(y_1) + g(y_2)f(y_2) + \cdots + g(y_k)f(y_k)$.

In particular,

$$E(Y^2) = y_1^2 f(y_1) + y_2^2 f(y_2) + \cdots + y_k^2 f(y_k).$$

Also

$$V(Y) = \sum_{i=1}^{k} (y_i - E(Y))^2 f(y_i) =$$

$$(y_1 - E(Y))^2 f(y_1) + (y_2 - E(Y))^2 f(y_2) + \cdots + (y_k - E(Y))^2 f(y_k).$$

For a continuous random variable Y with pdf $f(y)$, $V(Y) = \int_{-\infty}^{\infty} (y - E[Y])^2 f(y)dy$. Often using $V(Y) = E(Y^2) - (E(Y))^2$ is simpler.

Theorem 1.7. Let a and b be any constants and assume all relevant expectations exist.
i) $E(a) = a$.
ii) $E(aY + b) = aE(Y) + b$.
iii) $E(aX + bY) = aE(X) + bE(Y)$.
iv) $V(aY + b) = a^2V(Y)$.

Definition 1.15. Random variables X and Y are *identically distributed*, written $X \sim Y$, $X \stackrel{D}{=} Y$, or $Y \sim F_X$, if $F_X(y) = F_Y(y)$ for all real y.

Definition 1.16. i) For positive integers k, the *kth moment* of Y is $E[Y^k]$ while the *kth central moment* is $E[(Y - E[Y])^k]$.
ii) The **moment generating function** (mgf) of a random variable Y is

$$m(t) = m_Y(t) = E[e^{tY}] \tag{1.7}$$

if the expectation exists for t in some neighborhood of 0. Otherwise, the mgf does not exist.
iii) The **characteristic function** of a random variable Y is $c(t) = c_Y(t) = E[e^{itY}]$ where the complex number $i = \sqrt{-1}$.

More information about moment generating functions and characteristic functions is given in Section 1.3.

Theorem 1.8. Let X and Y be random variables. Then X and Y are identically distributed, $X \sim Y$, if any of the following conditions hold.
a) $F_X(y) = F_Y(y)$ for all y,
b) $f_X(y) = f_Y(y)$ for all y,
c) $c_X(t) = c_Y(t)$ for all t, or
d) $m_X(t) = m_Y(t)$ for all t in a neighborhood of zero.

Definition 1.17. Let $f(y) \equiv f_Y(y|\boldsymbol{\theta})$ be the pdf or pmf of a random variable Y. Then the set $\mathcal{Y}_{\boldsymbol{\theta}} = \{y|f_Y(y|\boldsymbol{\theta}) > 0\}$ is called the *sample space* or **support** of Y. Let the set Θ be the set of parameter values $\boldsymbol{\theta}$ of interest. Then Θ is the **parameter space** of Y. Use the notation $\mathcal{Y} = \{y|f(y|\boldsymbol{\theta}) > 0\}$ if the support does not depend on $\boldsymbol{\theta}$. So $\mathcal{Y}$ is the support of Y if $\mathcal{Y}_{\boldsymbol{\theta}} \equiv \mathcal{Y}$ $\forall \boldsymbol{\theta} \in \Theta$.

Definition 1.18. The **indicator function** $I_A(x) \equiv I(x \in A) = 1$ if $x \in A$ and $I_A(x) = 0$, otherwise. Sometimes an indicator function such as $I_{(0,\infty)}(y)$ will be denoted by $I(y > 0)$.

1.2 Multivariate Distributions

Often there are n random variables $Y_1, ..., Y_n$ that are of interest. For example, *age, blood pressure, weight, gender* and *cholesterol level* might be some of the random variables of interest for patients suffering from heart disease.

Notation. Let $\mathbb{R}^n$ be the n–dimensional Euclidean space. Then the vector $\boldsymbol{y} = (y_1, ..., y_n)^T \in \mathbb{R}^n$ if y_i is an arbitrary real number for $i = 1, ..., n$. Typically $\boldsymbol{y}$ is a column vector, but when $\boldsymbol{y}$ is the argument of a pdf, pmf, or cdf, then $\boldsymbol{y}$ is often a row vector, e.g., $f(\boldsymbol{y}) = f(y_1, ..., y_n)$. We may say $\boldsymbol{y} \in \mathbb{R}^n$ or $(y_1, ..., y_n) \in \mathbb{R}^n$.

Definition 1.19. If $Y_1, ..., Y_n$ are discrete random variables, then the **joint pmf** (probability mass function) of $Y_1, ..., Y_n$ is

$$f(y_1, ..., y_n) = P(Y_1 = y_1, ..., Y_n = y_n) \tag{1.8}$$

for any $(y_1, ..., y_n) \in \mathbb{R}^n$. A joint pmf f satisfies $f(\boldsymbol{y}) \equiv f(y_1, ..., y_n) \geq 0$ $\forall \boldsymbol{y} \in \mathbb{R}^n$ and

$$\sum_{\boldsymbol{y} \,:\, f(\boldsymbol{y}) > 0} \cdots \sum f(y_1, ..., y_n) = 1.$$

For any event $A \in \mathbb{R}^n$,

$$P[(Y_1, ..., Y_n) \in A] = \sum_{\boldsymbol{y} \,:\, \boldsymbol{y} \in A \text{ and } f(\boldsymbol{y}) > 0} \cdots \sum f(y_1, ..., y_n).$$

Definition 1.20. The **joint cdf** (cumulative distribution function) of $Y_1, ..., Y_n$ is $F(y_1, ..., y_n) = P(Y_1 \leq y_1, ..., Y_n \leq y_n)$ for any $(y_1, ..., y_n) \in \mathbb{R}^n$.

Definition 1.21. If $Y_1, ..., Y_n$ are continuous random variables, then the **joint pdf** (probability density function) of $Y_1, ..., Y_n$ is a function $f(y_1, ..., y_n)$ that satisfies $F(y_1, ..., y_n) = \int_{-\infty}^{y_n} \cdots \int_{-\infty}^{y_1} f(t_1, ..., t_n) dt_1 \cdots dt_n$ where the y_i are any real numbers. A joint pdf f satisfies $f(\boldsymbol{y}) \equiv f(y_1, ..., y_n) \geq 0 \ \ \forall \boldsymbol{y} \in \mathbb{R}^n$ and $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(t_1, ..., t_n) dt_1 \cdots dt_n$
$= 1$. For any event $A \in \mathbb{R}^n$,

$$P[(Y_1, ..., Y_n) \in A] = \int_A \cdots \int f(t_1, ..., t_n) dt_1 \cdots dt_n.$$

Definition 1.22. If $Y_1, ..., Y_n$ has a joint pdf or pmf f, then the *sample space* or **support** of $Y_1, ..., Y_n$ is

$$\mathcal{Y} = \{(y_1, ..., y_n) \in \mathbb{R}^n : f(y_1, ..., y_n) > 0\}.$$

If $\boldsymbol{Y}$ comes from a family of distributions $f(\boldsymbol{y}|\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta$, then the support $\mathcal{Y}_{\boldsymbol{\theta}} = \{\boldsymbol{y} : f(\boldsymbol{y}|\boldsymbol{\theta}) > 0\}$ may depend on $\boldsymbol{\theta}$.

Theorem 1.9. Let $Y_1, ..., Y_n$ have joint cdf $F(y_1, ..., y_n)$ and joint pdf $f(y_1, ..., y_n)$. Then

$$f(y_1, ..., y_n) = \frac{\partial^n}{\partial y_1 \cdots \partial y_n} F(y_1, ..., y_n)$$

wherever the partial derivative exists.

Definition 1.23. The **marginal pmf** of any subset $Y_{i1}, ..., Y_{ik}$ of the coordinates $(Y_1, ..., Y_n)$ is found by summing the joint pmf over all possible values of the other coordinates where the values $y_{i1}, ..., y_{ik}$ are held fixed. For example,

$$f_{Y_1,...,Y_k}(y_1, ..., y_k) = \sum_{y_{k+1}} \cdots \sum_{y_n} f(y_1, ..., y_n)$$

where $y_1, ..., y_k$ are held fixed. In particular, if Y_1 and Y_2 are discrete random variables with joint pmf $f(y_1, y_2)$, then the marginal pmf for Y_1 is

$$f_{Y_1}(y_1) = \sum_{y_2} f(y_1, y_2) \tag{1.9}$$

where y_1 is held fixed. The marginal pmf for Y_2 is

$$f_{Y_2}(y_2) = \sum_{y_1} f(y_1, y_2) \tag{1.10}$$

where y_2 is held fixed.

Remark 1.2. For $n = 2$, double integrals are used to find marginal pdfs (defined below) and to show that the joint pdf integrates to 1. If the region of integration Ω is bounded on top by the function $y_2 = \phi_T(y_1)$, on the bottom by the function $y_2 = \phi_B(y_1)$ and to the left and right by the lines $y_1 = a$ and $y_1 = b$ then $\int\int_\Omega f(y_1, y_2)dy_1dy_2 = \int\int_\Omega f(y_1, y_2)dy_2dy_1 =$

$$\int_a^b \left[\int_{\phi_B(y_1)}^{\phi_T(y_1)} f(y_1, y_2)dy_2 \right] dy_1.$$

Within the inner integral, treat y_2 as the variable, anything else, including y_1, is treated as a constant.

If the region of integration Ω is bounded on the left by the function $y_1 = \psi_L(y_2)$, on the right by the function $y_1 = \psi_R(y_2)$ and to the top and bottom by the lines $y_2 = c$ and $y_2 = d$ then $\int\int_\Omega f(y_1, y_2)dy_1dy_2 = \int\int_\Omega f(y_1, y_2)dy_2dy_1 =$

$$\int_c^d \left[\int_{\psi_L(y_2)}^{\psi_R(y_2)} f(y_1, y_2)dy_1 \right] dy_2.$$

Within the inner integral, treat y_1 as the variable, anything else, including y_2, is treated as a constant.

Definition 1.24. The **marginal pdf** of any subset $Y_{i1}, ..., Y_{ik}$ of the coordinates $(Y_1, ..., Y_n)$ is found by integrating the joint pdf over all possible values of the other coordinates where the values $y_{i1}, ..., y_{ik}$ are held fixed. For example, $f(y_1, ..., y_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(t_1, ..., t_n) dt_{k+1} \cdots dt_n$ where $y_1, ..., y_k$ are held fixed. In particular, if Y_1 and Y_2 are continuous random variables with joint pdf $f(y_1, y_2)$, then the marginal pdf for Y_1 is

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} f(y_1, y_2) dy_2 = \int_{\phi_B(y_1)}^{\phi_T(y_1)} f(y_1, y_2) dy_2 \tag{1.11}$$

where y_1 is held fixed (to get the region of integration, draw a line parallel to the y_2 axis and use the functions $y_2 = \phi_B(y_1)$ and $y_2 = \phi_T(y_1)$ as the lower and upper limits of integration). The marginal pdf for Y_2 is

$$f_{Y_2}(y_2) = \int_{-\infty}^{\infty} f(y_1, y_2) dy_1 = \int_{\psi_L(y_2)}^{\psi_R(y_2)} f(y_1, y_2) dy_1 \tag{1.12}$$

where y_2 is held fixed (to get the region of integration, draw a line parallel to the y_1 axis and use the functions $y_1 = \psi_L(y_2)$ and $y_1 = \psi_R(y_2)$ as the lower and upper limits of integration).

For independent random variables, the joint cdf is the product of the marginal cdfs, the joint pmf is the product of the marginal pmfs, and the joint pdf is the product of the marginal pdfs. Recall that $\forall$ is read "for all."

Definition 1.25. i) The random variables $Y_1, Y_2, ..., Y_n$ are **independent** if $F(y_1, y_2, ..., y_n) = F_{Y_1}(y_1) F_{Y_2}(y_2) \cdots F_{Y_n}(y_n)$ $\forall y_1, y_2, ..., y_n$.
ii) If the random variables have a joint pdf or pmf f then the random variables $Y_1, Y_2, ..., Y_n$ are independent if $f(y_1, y_2, ..., y_n) = f_{Y_1}(y_1) f_{Y_2}(y_2) \cdots f_{Y_n}(y_n)$ $\forall y_1, y_2, ..., y_n$.
If the random variables are not independent, then they are **dependent**.
In particular random variables Y_1 and Y_2 are **independent**, written $Y_1 \perp\!\!\!\perp Y_2$, if either of the following conditions holds.
i) $F(y_1, y_2) = F_{Y_1}(y_1) F_{Y_2}(y_2)$ $\forall y_1, y_2$.
ii) $f(y_1, y_2) = f_{Y_1}(y_1) f_{Y_2}(y_2)$ $\forall y_1, y_2$.
Otherwise, Y_1 and Y_2 are *dependent*.

The following theorem shows that finding the marginal pdfs or pmfs is simple if $Y_1, ..., Y_n$ are independent. Also **subsets of independent random variables are independent**: if $Y_1, ..., Y_n$ are independent and if $\{i_1, ..., i_k\} \subseteq \{1, ..., n\}$ for $k \geq 2$, then $Y_{i_1}, ..., Y_{i_k}$ are independent.

Theorem 1.10. Suppose that $Y_1, ..., Y_n$ are independent random variables with joint pdf or pmf $f(y_1, ..., y_n)$. Then the marginal pdf or pmf of

any subset $Y_{i_1}, ..., Y_{i_k}$ is $f(y_{i_1}, ..., y_{i_k}) = \prod_{j=1}^{k} f_{Y_{i_j}}(y_{i_j})$. Hence $Y_{i_1}, ..., Y_{i_k}$ are independent random variables for $k \geq 2$.

Proof. The proof for a joint pdf is given below. For a joint pmf, replace the integrals by appropriate sums. The marginal

$$f(y_{i_1}, ..., y_{i_k}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[\prod_{j=1}^{n} f_{Y_{i_j}}(y_{i_j}) \right] dy_{i_{k+1}} \cdots dy_{i_n}$$

$$= \left[\prod_{j=1}^{k} f_{Y_{i_j}}(y_{i_j}) \right] \left[\prod_{j=k+1}^{n} \int_{-\infty}^{\infty} f_{Y_{i_j}}(y_{i_j}) \, dy_{i_j} \right]$$

$$= \left[\prod_{j=1}^{k} f_{Y_{i_j}}(y_{i_j}) \right] (1)^{n-k} = \prod_{j=1}^{k} f_{Y_{i_j}}(y_{i_j}). \quad \square$$

Definition 1.26. Suppose that random variables $\boldsymbol{Y} = (Y_1, ..., Y_n)$ have support $\mathcal{Y}$ and joint pdf or pmf f. Then the **expected value** of the real valued function $h(\boldsymbol{Y}) = h(Y_1, ..., Y_n)$ is

$$E[h(\boldsymbol{Y})] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\boldsymbol{y}) f(\boldsymbol{y}) \, d\boldsymbol{y} = \int \cdots \int_{\mathcal{Y}} h(\boldsymbol{y}) f(\boldsymbol{y}) \, d\boldsymbol{y} \qquad (1.13)$$

if f is a joint pdf and if

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |h(\boldsymbol{y})| f(\boldsymbol{y}) \, d\boldsymbol{y}$$

exists. Otherwise the expectation does not exist. The expected value is

$$E[h(\boldsymbol{Y})] = \sum_{y_1} \cdots \sum_{y_n} h(\boldsymbol{y}) f(\boldsymbol{y}) = \sum_{\boldsymbol{y} \in \mathbb{R}^n} h(\boldsymbol{y}) f(\boldsymbol{y}) = \sum_{\boldsymbol{y} \in \mathcal{Y}} h(\boldsymbol{y}) f(\boldsymbol{y}) \qquad (1.14)$$

if f is a joint pmf and if $\sum_{\boldsymbol{y} \in \mathbb{R}^n} |h(\boldsymbol{y})| f(\boldsymbol{y})$ exists. Otherwise the expectation does not exist.

The notation $E[h(\boldsymbol{Y})] = \infty$ can be useful when the corresponding integral or sum diverges to ∞. The following theorem is useful since multiple integrals with smaller dimension are easier to compute than those with higher dimension.

Theorem 1.11. Suppose that $Y_1, ..., Y_n$ are random variables with joint pdf or pmf $f(y_1, ..., y_n)$. Let $\{i_1, ..., i_k\} \subset \{1, ..., n\}$, and let $f(y_{i_1}, ..., y_{i_k})$ be the marginal pdf or pmf of $Y_{i_1}, ..., Y_{i_k}$ with support $\mathcal{Y}_{Y_{i_1}, ..., Y_{i_k}}$. Assume that $E[h(Y_{i_1}, ..., Y_{i_k})]$ exists. Then

$$E[h(Y_{i_1},...,Y_{i_k})] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(y_{i_1},...,y_{i_k})\ f(y_{i_1},...,y_{i_k})\ dy_{i_1} \cdots dy_{i_k} =$$

$$\int \cdots \int_{\mathcal{Y}_{Y_{i_1},...,Y_{i_k}}} h(y_{i_1},...,y_{i_k})\ f(y_{i_1},...,y_{i_k})\ dy_{i_1} \cdots dy_{i_k}$$

if f is a pdf, and

$$E[h(Y_{i_1},...,Y_{i_k})] = \sum_{y_{i_1}} \cdots \sum_{y_{i_k}} h(y_{i_1},...,y_{i_k})\ f(y_{i_1},...,y_{i_k})$$

$$= \sum_{(y_{i_1},...,y_{i_k}) \in \mathcal{Y}_{Y_{i_1},...,Y_{i_k}}} h(y_{i_1},...,y_{i_k})\ f(y_{i_1},...,y_{i_k})$$

if f is a pmf.

Proof. The proof for a joint pdf is given below. For a joint pmf, replace the integrals by appropriate sums. Let $g(Y_1,...,Y_n) = h(Y_{i_1},...,Y_{i_k})$. Then $E[g(\boldsymbol{Y})] =$

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(y_{i_1},...,y_{i_k}) f(y_1,...,y_n)\ dy_1 \cdots dy_n =$$

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(y_{i_1},...,y_{i_k}) \left[\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(y_1,...,y_n)\ dy_{i_{k+1}} \cdots dy_{i_n} \right] dy_{i_1} \cdots dy_{i_k}$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(y_{i_1},...,y_{i_k}) f(y_{i_1},...,y_{i_k})\ dy_{i_1} \cdots dy_{i_k}$$

since the term in the brackets gives the marginal. □

Example 1.1. Typically $E(Y_i), E(Y_i^2)$, and $E(Y_iY_j)$ for $i \neq j$ are of primary interest. Suppose that (Y_1, Y_2) has joint pdf $f(y_1, y_2)$. Then $E[h(Y_1, Y_2)]$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(y_1, y_2) f(y_1, y_2) dy_2 dy_1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(y_1, y_2) f(y_1, y_2) dy_1 dy_2$$

where $-\infty$ to ∞ could be replaced by the limits of integration for dy_i. **In particular**,

$$E(Y_1Y_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y_1 y_2 f(y_1, y_2) dy_2 dy_1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y_1 y_2 f(y_1, y_2) dy_1 dy_2.$$

Since finding the marginal pdf is usually easier than doing the double integral, if h is a function of Y_i but not of Y_j, find the marginal for Y_i : $E[h(Y_1)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(y_1) f(y_1, y_2) dy_2 dy_1 = \int_{-\infty}^{\infty} h(y_1) f_{Y_1}(y_1) dy_1$. Similarly, $E[h(Y_2)] = \int_{-\infty}^{\infty} h(y_2) f_{Y_2}(y_2) dy_2$.

In particular, $E(Y_1) = \int_{-\infty}^{\infty} y_1 f_{Y_1}(y_1) dy_1$, and $E(Y_2) = \int_{-\infty}^{\infty} y_2 f_{Y_2}(y_2) dy_2$.

Suppose that (Y_1, Y_2) have a joint pmf $f(y_1, y_2)$. Then the **expectation**

$$E[h(Y_1, Y_2)] = \sum_{y_2} \sum_{y_1} h(y_1, y_2) f(y_1, y_2) = \sum_{y_1} \sum_{y_2} h(y_1, y_2) f(y_1, y_2).$$

In particular,

$$E[Y_1 Y_2] = \sum_{y_1} \sum_{y_2} y_1 y_2 f(y_1, y_2).$$

Since finding the marginal pmf is usually easier than doing the double summation, if h is a function of Y_i but not of Y_j, find the marginal for pmf for Y_i: $E[h(Y_1)] = \sum_{y_2} \sum_{y_1} h(y_1) f(y_1, y_2) = \sum_{y_1} h(y_1) f_{Y_1}(y_1)$. Similarly, $E[h(Y_2)] = \sum_{y_2} h(y_2) f_{Y_2}(y_2)$. **In particular**, $E(Y_1) = \sum_{y_1} y_1 f_{Y_1}(y_1)$ and $E(Y_2) = \sum_{y_2} y_2 f_{Y_2}(y_2)$.

For pdfs it is sometimes possible to find $E[h(Y_i)]$, but for $k \geq 2$ these expected values tend to be very difficult to compute unless $f(y_1, ..., y_k) = c\ y_1^{i_1} \cdots y_k^{i_k}$ for small integers i_j on rectangular or triangular support. Independence makes finding some expected values simple.

Theorem 1.12. Let $Y_1, ..., Y_n$ be independent random variables. If $h_i(Y_i)$ is a function of Y_i alone and if the relevant expected values exist, then

$$E[h_1(Y_1) h_2(Y_2) \cdots h_n(Y_n)] = E[h_1(Y_1)] \cdots E[h_n(Y_n)].$$

In particular, $E[Y_i Y_j] = E[Y_i] E[Y_j]$ for $i \neq j$.

Proof. The result will be shown for the case where $\boldsymbol{Y} = (Y_1, ..., Y_n)$ has a joint pdf f. For a joint pmf, replace the integrals by appropriate sums. By independence, the support of $\boldsymbol{Y}$ is a cross product: $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n$. Since $f(\boldsymbol{y}) = \prod_{i=1}^n f_{Y_i}(y_i)$, the expectation $E[h_1(Y_1) h_2(Y_2) \cdots h_n(Y_n)] =$

$$\int \cdots \int_{\mathcal{Y}} h_1(y_1) h_2(y_2) \cdots h_n(y_n)\ f(y_1, ..., y_n)\ dy_1 \cdots dy_n$$

$$= \int_{\mathcal{Y}_n} \cdots \int_{\mathcal{Y}_1} \left[\prod_{i=1}^n h_i(y_i) f_{Y_i}(y_i) \right] dy_1 \cdots dy_n$$

$$= \prod_{i=1}^n \left[\int_{\mathcal{Y}_i} h_i(y_i) f_{Y_i}(y_i)\ dy_i \right] = \prod_{i=1}^n E[h_i(Y_i)]. \quad \square$$

Theorem 1.13. Let $Y_1, ..., Y_n$ be independent random variables. If $h_j(Y_{i_j})$ is a function of Y_{i_j} alone and if the relevant expected values exist, then

$$E[h_1(Y_{i_1}) \cdots h_k(Y_{i_k})] = E[h_1(Y_{i_1})] \cdots E[h_k(Y_{i_k})].$$

Proof. Method 1: Take $X_j = Y_{i_j}$ for $j = 1, ..., k$. Then $X_1, ..., X_k$ are independent and Theorem 1.12 applies.

Method 2: Take $h_j(Y_{i_j}) \equiv 1$ for $j = k+1, ..., n$ and apply Theorem 1.12. □

Theorem 1.14. Let $Y_1, ..., Y_n$ be independent random variables. If $h_i(Y_i)$ is a function of Y_i alone and $X_i = h_i(Y_i)$, then the random variables $X_1, ..., X_n$ are independent.

Definition 1.27. The **covariance** of Y_1 and Y_2 is

$$\text{Cov}(Y_1, Y_2) = E[(Y_1 - E(Y_1))(Y_2 - E(Y_2))]$$

provided the expectation exists. Otherwise the covariance does not exist.

Theorem 1.15: Short cut formula. If $\text{Cov}(Y_1, Y_2)$ exists then $\text{Cov}(Y_1, Y_2) = E(Y_1Y_2) - E(Y_1)E(Y_2)$.

The notation $Y_1 \perp\!\!\!\perp Y_2$ means that Y_1 and Y_2 are independent random variables.

Theorem 1.16. a) Let Y_1 and Y_2 be independent random variables. If $\text{Cov}(Y_1, Y_2)$ exists, then $\text{Cov}(Y_1, Y_2) = 0$.

b) **The converse is false**: $\text{Cov}(Y_1, Y_2) = 0$ does not imply $Y_1 \perp\!\!\!\perp Y_2$.

Definition 1.28. $\boldsymbol{Y} = (Y_1, ..., Y_p)^T$ is a $p \times 1$ **random vector** if Y_i is a random variable for $i = 1, ..., p$. $\boldsymbol{Y}$ is a discrete random vector if each Y_i is discrete, and $\boldsymbol{Y}$ is a continuous random vector if each Y_i is continuous. A random variable Y_1 is the special case of a random vector with $p = 1$.

In this section we will consider n random vectors $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$. Often double subscripts will be used: $\boldsymbol{Y}_i = (Y_{i,1}, ..., Y_{i,p_i})^T$ for $i = 1, ..., n$.

Notation. In this text, $\boldsymbol{Y}$ is usually a column vector, and if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are both vectors, a phrase with $\boldsymbol{Y}$ and $\boldsymbol{X}^T$ means that $\boldsymbol{Y}$ is a column vector and $\boldsymbol{X}^T$ is a row vector where T stands for transpose. Arguments of pdfs, pmfs, and cdfs, are usually taken to be row vectors in this text.

Definition 1.29. The *population mean* or **expected value** of a random $p \times 1$ random vector $\boldsymbol{Y} = (Y_1, ..., Y_p)^T$ is

$$E(\boldsymbol{Y}) = (E(Y_1), ..., E(Y_p))^T$$

provided that $E(Y_i)$ exists for $i = 1, ..., p$. Otherwise the expected value does not exist. The $p \times p$ *population covariance matrix*

$$\text{Cov}(\boldsymbol{Y}) = E(\boldsymbol{Y} - E(\boldsymbol{Y}))(\boldsymbol{Y} - E(\boldsymbol{Y}))^T = (\sigma_{i,j})$$

where the ij entry of $\mathrm{Cov}(\boldsymbol{Y})$ is $\mathrm{Cov}(Y_i, Y_j) = \sigma_{i,j}$ provided that each $\sigma_{i,j}$ exists. Otherwise $\mathrm{Cov}(\boldsymbol{Y})$ does not exist.

The covariance matrix is also called the variance–covariance matrix and variance matrix. Sometimes the notation $\mathrm{Var}(\boldsymbol{Y})$ is used. Note that $\mathrm{Cov}(\boldsymbol{Y})$ is a symmetric positive semidefinite matrix. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $p \times 1$ random vectors, $\boldsymbol{a}$ a conformable constant vector and $\boldsymbol{A}$ and $\boldsymbol{B}$ are conformable constant matrices, then

$$E(\boldsymbol{a} + \boldsymbol{X}) = \boldsymbol{a} + E(\boldsymbol{X}) \quad \text{and} \quad E(\boldsymbol{X} + \boldsymbol{Y}) = E(\boldsymbol{X}) + E(\boldsymbol{Y}) \tag{1.15}$$

and

$$E(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}E(\boldsymbol{X}) \quad \text{and} \quad E(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}. \tag{1.16}$$

Thus

$$\mathrm{Cov}(\boldsymbol{a} + \boldsymbol{A}\boldsymbol{X}) = \mathrm{Cov}(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}\mathrm{Cov}(\boldsymbol{X})\boldsymbol{A}^T. \tag{1.17}$$

Definition 1.30. Let $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ be random vectors with joint pdf or pmf $f(\boldsymbol{y}_1, ..., \boldsymbol{y}_n)$. Let $f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i)$ be the marginal pdf or pmf of $\boldsymbol{Y}_i$. Then $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ are **independent random vectors** if

$$f(\boldsymbol{y}_1, ..., \boldsymbol{y}_n) = f_{\boldsymbol{Y}_1}(\boldsymbol{y}_1) \cdots f_{\boldsymbol{Y}_n}(\boldsymbol{y}_n) = \prod_{i=1}^{n} f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i).$$

The following theorem is a useful generalization of Theorem 1.14.

Theorem 1.17. Let $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ be independent random vectors where $\boldsymbol{Y}_i$ is a $p_i \times 1$ vector for $i = 1, ..., n$. and let $\boldsymbol{h}_i : \mathbb{R}^{p_i} \to \mathbb{R}^{p_{j_i}}$ be vector valued functions and suppose that $\boldsymbol{h}_i(\boldsymbol{y}_i)$ is a function of $\boldsymbol{y}_i$ alone for $i = 1, ..., n$. Then the random vectors $\boldsymbol{X}_i = \boldsymbol{h}_i(\boldsymbol{Y}_i)$ are independent. There are three important special cases.
i) If $p_{j_i} = 1$ so that each h_i is a real valued function, then the random variables $X_i = h_i(\boldsymbol{Y}_i)$ are independent.
ii) If $p_i = p_{j_i} = 1$ so that each Y_i and each $X_i = h(Y_i)$ are random variables, then $X_1, ..., X_n$ are independent.
iii) Let $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ and $\boldsymbol{X} = (X_1, .., X_m)^T$ and assume that $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X}$. If $\boldsymbol{h}(\boldsymbol{Y})$ is a vector valued function of $\boldsymbol{Y}$ alone and if $\boldsymbol{g}(\boldsymbol{X})$ is a vector valued function of $\boldsymbol{X}$ alone, then $\boldsymbol{h}(\boldsymbol{Y})$ and $\boldsymbol{g}(\boldsymbol{X})$ are independent random vectors.

1.3 Characteristic Function, MGF, CGF

Definition 1.16 introduced the moment generating function and the characteristic function. This section will give some more details.

Definition 1.31. The **moment generating function** (mgf) of a random variable Y is

$$m(t) = m_Y(t) = E[e^{tY}] \tag{1.18}$$

if the expectation exists for t in some neighborhood of 0. Otherwise, the mgf does not exist. If Y is discrete, then $m(t) = \sum_y e^{ty} f(y)$, and if Y is continuous, then $m(t) = \int_{-\infty}^{\infty} e^{ty} f(y) dy$.

Notation. The natural logarithm of y is $\log(y) = \ln(y)$. If another base is wanted, it will be given, e.g. $\log_{10}(y)$.

Definition 1.32. If the mgf exists, then the **cumulant generating function** (cgf) $k(t) = k_Y(t) = \log(m(t))$ for the values of t where the mgf is defined.

Definition 1.33. The **characteristic function** of a random variable Y is $c(t) = c_Y(t) = E[e^{itY}] = E[\cos(tY)] + iE[\sin(tY)]$ where the complex number $i = \sqrt{-1}$.

Moment generating functions do not necessarily exist in a neighborhood of zero, but a characteristic function always exists. This text does not require much knowledge of theory of complex variables, but know that $i^2 = -1$, $i^3 = -i$ and $i^4 = 1$. Hence $i^{4k-3} = i$, $i^{4k-2} = -1$, $i^{4k-1} = -i$ and $i^{4k} = 1$ for $k = 1, 2, 3, ...$ Let complex number $z = a + ib$. Then the modulus of z is $|z| = |a + ib| = \sqrt{a^2 + b^2}$.

Remark 1.3. a) Suppose that Y is a random variable with an mgf $m(t)$ that exists for $|t| < b$ for some constant $b > 0$. Then often the characteristic function of Y is i) $c(t) = m(it)$ while ii) $m(t) = c(-it)$. If Y has a pmf with $f(y_j) = P(Y = y_j) = p_j)$, then the characteristic function of Y is $c(t) = c_Y(t) = \sum_j e^{ity_j} p_j$ while the mgf $m_Y(t) = \sum_j e^{ty_j} p_j$. Hence the two formulas i) and ii) "hold" if Y has a pmf, at least for t such that the mgf is defined. If Y is nonnegative then the mgf is a scaled Laplace transformation and $c(t)$ is a scaled Fourier transformation, and then the two formulas i) and ii) hold by Laplace and Fourier transformation theory, at least for t such that the mgf is defined. The Taylor series for the mgf is

$$m_Y(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!} E[X^k]$$

for $|t| < b$ while the characteristic function

$$c_Y(t) = \sum_{k=0}^{\infty} \frac{(it)^k}{k!} E[X^k]$$

for all real t if Y has an mgf defined for all real t. Hence if $b = \infty$, the two formulas i) and ii) hold. See Billingsley (1986, pp. 285, 353).

b) If $E[Y^2]$ is finite, then

$$c_Y(t) = 1 + itE(Y) - \frac{1}{2}t^2E[Y^2] + o(t^2) \text{ as } t \to 0.$$

In particular, if $E(Y) = 0$ and $E(Y^2) = V(Y) = \sigma^2$, then

$$c_Y(t) = 1 - \frac{t^2\sigma^2}{2} + o(t^2) \text{ as } t \to 0. \tag{1.19}$$

Here $a(t) = o(t^2)$ as $t \to 0$ if $\lim_{t\to 0} \frac{a(t)}{t^2} = 0$. See Billingsley (1986, p. 354).

c) Properties of $c(t)$: i) $c(0) = 1$, ii) the modulus $|c(t)| \leq 1$ for all real t, iii) $c(t)$ is a continuous function.

d) Let j and k be positive integers. If $E(Y^k)$ is finite, then $E(Y^j)$ is finite for $1 \leq j \leq k$.

e) If Y has mgf $m(t)$, then $E(Y^k)$ is finite for each positive integer k.

f) A complex random variable $Z = X + iY$ where X and Y are ordinary random variables. Then $E(Z) = E(X) + iE(Y)$, and $E(Z)$ exists if $E(|Z|) = E(\sqrt{X^2 + Y^2}) < \infty$. Linearity of expectation and key inequalities such as $|E(Z)| \leq E(|Z|)$ remain valid. Also, if $Z_1 \perp\!\!\!\perp Z_2$ and $g_i(Z_i)$ is a function of the complex random variable Z_i alone, then $E[g_1(Z_1)g_2(Z_2)] = E[g_1(Z_1)]E[g_2(Z_2)]$ if the expectations exist. $Z = e^{itY}$ is the main complex random variable in this book.

g) The formula $\frac{d}{dt}e^{dt} = de^{dt}$ is valid for any complex constant d.

h) The fundamental theorem of calculus holds. Thus $\int_a^b e^{dt}dt = \frac{e^{db} - e^{da}}{d}$ for any nonzero complex constant d.

Remarks 1.3 and 1.4 are often used in proofs of the Central Limit Theorem. Note that by Remark 1.4a), $\lim_{n\to\infty}\left(1 - \frac{c \pm \epsilon}{n}\right)^n = e^{-[c\pm\epsilon]}$ where ϵ is a real number. By Remark 1.4c), this result holds even if ϵ is complex valued. Letting positive $\epsilon \to 0$ proves Remark 1.4b).

Remark 1.4. For a) and b), assume c and c_n are real.

a) $\lim_{n\to\infty}\left(1 - \frac{c}{n}\right)^n = e^{-c}$.

b) If $c_n \to c$ as $n \to \infty$, then $\lim_{n\to\infty}\left(1 + \frac{-c_n}{n}\right)^n = e^{-c}$.

c) If c_n is a sequence of complex numbers such that $c_n \to c$ as $n \to \infty$ where c is real, then $\lim_{n\to\infty}\left(1 - \frac{c_n}{n}\right)^n = e^{-c}$.

In the following theorem, let the kth derivative of $g(t)$ be $g^{(k)}(t)$ with derivative $g^{(1)}(t) = g'(t)$ and second derivative $g^{(2)}(t) = g''(t)$.

Theorem 1.18. Suppose that the mgf $m(t)$ exists for $|t| < b$ for some constant $b > 0$, and suppose that the kth derivative $m^{(k)}(t)$ exists for $|t| < b$. Then $E[Y^k] = m^{(k)}(0)$ for positive integers k. In particular, $E[Y] = m'(0)$ and $E[Y^2] = m''(0)$. For the cumulant generating function $k(t)$, $E(Y) = k'(0)$ and $V(Y) = k''(0)$.

Remark 1.5. Let $h(y)$, $g(y)$, $n(y)$ and $d(y)$ be functions. Review how to find the derivative $g'(y)$ of $g(y)$ and how to find the kth derivative

$$g^{(k)}(y) = \frac{d^k}{dy^k} g(y)$$

for integers $k \geq 2$. Recall that the *product rule* is

$$(h(y)g(y))' = h'(y)g(y) + h(y)g'(y).$$

The **quotient rule** is

$$\left(\frac{n(y)}{d(y)}\right)' = \frac{d(y)n'(y) - n(y)d'(y)}{[d(y)]^2}.$$

The **chain rule** is

$$[h(g(y))]' = [h'(g(y))][g'(y)].$$

Then given the mgf $m(t)$, find $E[Y] = m'(0)$, $E[Y^2] = m''(0)$ and $V(Y) = E[Y^2] - (E[Y])^2$.

Definition 1.34. The **characteristic function** (cf) of a random vector $\boldsymbol{Y}$ is

$$c_{\boldsymbol{Y}}(\boldsymbol{t}) = E(e^{i\boldsymbol{t}^T\boldsymbol{Y}})$$

$\forall \boldsymbol{t} \in \mathbb{R}^n$ where the complex number $i = \sqrt{-1}$.

Definition 1.35. The **moment generating function** (mgf) of a random vector $\boldsymbol{Y}$ is

$$m_{\boldsymbol{Y}}(\boldsymbol{t}) = E(e^{\boldsymbol{t}^T\boldsymbol{Y}})$$

provided that the expectation exists for all $\boldsymbol{t}$ in some neighborhood of the origin $\boldsymbol{0}$.

Theorem 1.19. If $Y_1, ..., Y_n$ have mgf $m(\boldsymbol{t})$, then moments of all orders exist and for any nonnegative integers $k_1, ..., k_j$,

$$E(Y_{i_1}^{k_1} \cdots Y_{i_j}^{k_j}) = \frac{\partial^{k_1+\cdots+k_j}}{\partial t_{i_1}^{k_1} \cdots \partial t_{i_j}^{k_j}} m(\boldsymbol{t})\Big|_{\boldsymbol{t}=\boldsymbol{0}}.$$

In particular,

$$E(Y_i) = \frac{\partial m(\boldsymbol{t})}{\partial t_i}\Big|_{\boldsymbol{t}=\boldsymbol{0}}$$

and

$$E(Y_iY_j) = \left.\frac{\partial^2 m(\boldsymbol{t})}{\partial t_i \partial t_j}\right|_{\boldsymbol{t}=\boldsymbol{0}}.$$

Theorem 1.20. If $Y_1, ..., Y_n$ have a cf $c_{\boldsymbol{Y}}(\boldsymbol{t})$ and mgf $m_{\boldsymbol{Y}}(\boldsymbol{t})$ then the marginal cf and mgf for $Y_{i_1}, ..., Y_{i_k}$ are found from the joint cf and mgf by replacing t_{i_j} by 0 for $j = k+1, ..., n$. In particular, if $\boldsymbol{Y} = (\boldsymbol{Y}_1^T, \boldsymbol{Y}_2^T)^T$ and $\boldsymbol{t} = (\boldsymbol{t}_1^T, \boldsymbol{t}_2^T)^T$, then

$$c_{\boldsymbol{Y}_1}(\boldsymbol{t}_1) = c_{\boldsymbol{Y}}((\boldsymbol{t}_1^T, \boldsymbol{0}^T)^T) \text{ and } \mathrm{m}_{\boldsymbol{Y}_1}(\boldsymbol{t}_1) = \mathrm{m}_{\boldsymbol{Y}}((\boldsymbol{t}_1^{\mathrm{T}}, \boldsymbol{0}^{\mathrm{T}})^{\mathrm{T}}).$$

Proof. Use the definition of the cf and mgf. For example, if $\boldsymbol{Y}_1 = (Y_1, ..., Y_k)^T$ and $\boldsymbol{s} = \boldsymbol{t}_1$, then $m((\boldsymbol{t}_1^T, \boldsymbol{0}^T)^T) =$

$$E[\exp(t_1Y_1 + \cdots + t_kY_k + 0Y_{k+1} + \cdots + 0Y_n)] = E[\exp(t_1Y_1 + \cdots + t_kY_k)] =$$

$E[\exp(\boldsymbol{s}^T\boldsymbol{Y}_1)] = m_{\boldsymbol{Y}_1}(\boldsymbol{s})$, which is the mgf of $\boldsymbol{Y}_1$. □

Theorem 1.21. Partition the $n \times 1$ vectors $\boldsymbol{Y}$ and $\boldsymbol{t}$ as $\boldsymbol{Y} = (\boldsymbol{Y}_1^T, \boldsymbol{Y}_2^T)^T$ and $\boldsymbol{t} = (\boldsymbol{t}_1^T, \boldsymbol{t}_2^T)^T$. Then the random vectors $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are independent iff their joint cf factors into the product of their marginal cfs:

$$c_{\boldsymbol{Y}}(\boldsymbol{t}) = c_{\boldsymbol{Y}_1}(\boldsymbol{t}_1)c_{\boldsymbol{Y}_2}(\boldsymbol{t}_2) \quad \forall \boldsymbol{t} \in \mathbb{R}^n.$$

If the joint mgf exists, then the random vectors $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are independent iff their joint mgf factors into the product of their marginal mgfs:

$$m_{\boldsymbol{Y}}(\boldsymbol{t}) = m_{\boldsymbol{Y}_1}(\boldsymbol{t}_1)m_{\boldsymbol{Y}_2}(\boldsymbol{t}_2)$$

$\forall \boldsymbol{t}$ in some neighborhood of $\boldsymbol{0}$.

Note that if the random vectors $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are independent, written $\boldsymbol{Y}_1 \perp\!\!\!\perp \boldsymbol{Y}_2$, then

$$c_{\boldsymbol{Y}}(\boldsymbol{t}) = E[\exp(i\boldsymbol{t}^T\boldsymbol{Y})] = E[\exp(i\boldsymbol{t}_1^T\boldsymbol{Y}_1 + i\boldsymbol{t}_2^T\boldsymbol{Y}_2)] = E[\exp(i\boldsymbol{t}_1^T\boldsymbol{Y}_1)\exp(i\boldsymbol{t}_2^T\boldsymbol{Y}_2)]$$

$$\stackrel{ind}{=} E[\exp(i\boldsymbol{t}_1^T\boldsymbol{Y}_1)]E[\exp(i\boldsymbol{t}_2^T\boldsymbol{Y}_2)] = c_{\boldsymbol{Y}_1}(\boldsymbol{t}_1)c_{\boldsymbol{Y}_2}(\boldsymbol{t}_2)$$

for any $\boldsymbol{t} = (\boldsymbol{t}_1^T, \boldsymbol{t}_2^T)^T \in \mathbb{R}^n$.

1.4 Sums of Random Variables

The assumption that the data are iid or a random sample is often used. The iid assumption is useful for finding the joint pdf or pmf, and the exact or large sample distribution of many important statistics.

Definition 1.36. $Y_1, ..., Y_n$ are a **random sample** or **iid** if $Y_1, ..., Y_n$ are independent and identically distributed (all of the Y_i have the same distribution).

An important statistic is $\sum_{i=1}^{n} Y_i$. Some properties of sums are given below.

Theorem 1.22. Assume that all relevant expectations exist. Let a, $a_1, ..., a_n$ and $b_1, ..., b_m$ be constants. Let $Y_1, ..., Y_n$, and $X_1, ..., X_m$ be random variables. Let $g_1, ..., g_k$ be functions of $Y_1, ..., Y_n$.

i) $E(a) = a$.

ii) $E[aY] = aE[Y]$

iii) $V(aY) = a^2V(Y)$.

iv) $E[g_1(Y_1, ..., Y_n) + \cdots + g_k(Y_1, ..., Y_n)] = \sum_{i=1}^{k} E[g_i(Y_1, ..., Y_n)]$.

Let $W_1 = \sum_{i=1}^{n} a_i Y_i$ and $W_2 = \sum_{i=1}^{m} b_i X_i$.

v) $E(W_1) = \sum_{i=1}^{n} a_i E(Y_i)$.

vi) $V(W_1) = \text{Cov}(W_1, W_1) = \sum_{i=1}^{n} a_i^2 V(Y_i) + 2\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_i a_j \text{Cov}(Y_i, Y_j)$.

vii) $\text{Cov}(W_1, W_2) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \text{Cov}(Y_i, X_j)$.

viii) $E(\sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n} E(Y_i)$.

ix) If $Y_1, ..., Y_n$ are independent, $V(\sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n} V(Y_i)$.

Let $Y_1, ..., Y_n$ be iid random variables with $E(Y_i) = \mu$ and $V(Y_i) = \sigma^2$, then the

sample mean $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Then

x) $E(\overline{Y}) = \mu$ and

xi) $V(\overline{Y}) = \sigma^2/n$.

Hence the expected value of the sum is the sum of the expected values, the variance of the sum is the sum of the variances for independent random variables, and the covariance of two sums is the double sum of the covariances. Note that ix) follows from vi) with $a_i \equiv 1$, viii) follows from iv) with $g_i(\boldsymbol{Y}) =$

Y_i or from v) with $a_i \equiv 1$, x) follows from v) with $a_i \equiv 1/n$, and xi) can be shown using iii) and ix) using $\overline{Y} = \sum_{i=1}^{n}(Y_i/n)$.

Example 1.2. Let $Y_1, ..., Y_n$ be independent random variables with $E(Y_i) = \mu_i$ and $V(Y_i) = \sigma_i^2$. Let $W = \sum_{i=1}^{n} Y_i$. Then
a) $E(W) = E(\sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n} E(Y_i) = \sum_{i=1}^{n} \mu_i$, and
b) $V(W) = V(\sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n} V(Y_i) = \sum_{i=1}^{n} \sigma_i^2$.

A **statistic** is a function of the data (often a random sample) and known constants. A statistic is a random variable and the **sampling distribution** of a statistic is the distribution of the statistic. Important statistics are $\sum_{i=1}^{n} Y_i$, $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $\sum_{i=1}^{n} a_i Y_i$ where $a_1, ..., a_n$ are constants. The following theorem shows how to find the mgf and characteristic function of such statistics.

Theorem 1.23. a) The characteristic function uniquely determines the distribution.

b) If the moment generating function exists, then it uniquely determines the distribution.

c) Assume that $Y_1, ..., Y_n$ are independent with characteristic functions $c_{Y_i}(t)$. Then the characteristic function of $W = \sum_{i=1}^{n} Y_i$ is

$$c_W(t) = \prod_{i=1}^{n} c_{Y_i}(t). \tag{1.20}$$

d) Assume that $Y_1, ..., Y_n$ are iid with characteristic functions $c_Y(t)$. Then the characteristic function of $W = \sum_{i=1}^{n} Y_i$ is

$$c_W(t) = [c_Y(t)]^n. \tag{1.21}$$

e) Assume that $Y_1, ..., Y_n$ are independent with mgfs $m_{Y_i}(t)$. Then the mgf of $W = \sum_{i=1}^{n} Y_i$ is

$$m_W(t) = \prod_{i=1}^{n} m_{Y_i}(t). \tag{1.22}$$

f) Assume that $Y_1, ..., Y_n$ are iid with mgf $m_Y(t)$. Then the mgf of $W = \sum_{i=1}^{n} Y_i$ is

$$m_W(t) = [m_Y(t)]^n. \tag{1.23}$$

g) Assume that $Y_1, ..., Y_n$ are independent with characteristic functions $c_{Y_i}(t)$. Then the characteristic function of $W = \sum_{j=1}^{n}(a_j + b_j Y_j)$ is

$$c_W(t) = \exp(it\sum_{j=1}^{n} a_j) \prod_{j=1}^{n} c_{Y_j}(b_j t). \tag{1.24}$$

h) Assume that $Y_1, ..., Y_n$ are independent with mgfs $m_{Y_i}(t)$. Then the mgf of $W = \sum_{i=1}^{n}(a_i + b_i Y_i)$ is

$$m_W(t) = \exp(t\sum_{i=1}^{n} a_i) \prod_{i=1}^{n} m_{Y_i}(b_i t). \tag{1.25}$$

Partial Proof:
c)

$$c_{\sum_{j=1}^{n} Y_j}(t) = E[e^{it\sum_{j=1}^{n} Y_j}] = E[e^{itY_1+\cdots+itY_n}] = E\left[\prod_{j=1}^{n} e^{itY_j}\right] \stackrel{ind}{=}$$

$$\prod_{j=1}^{n} E[e^{itY_j}] = \prod_{j=1}^{n} c_{Y_j}(t).$$

The proofs for d), e), and f) are similar, but for mgfs, omit the i's and change c to m.

g) Recall that $\exp(w) = e^w$ and $\exp(\sum_{j=1}^{n} d_j) = \prod_{j=1}^{n} \exp(d_j)$. Now

$$c_W(t) = E(e^{itW}) = E(\exp[it\sum_{j=1}^{n}(a_j + b_j Y_j)])$$

$$= \exp(it\sum_{j=1}^{n} a_j)\ E(\exp[\sum_{j=1}^{n} itb_j Y_j)])$$

$$= \exp(it\sum_{j=1}^{n} a_j)\ E(\prod_{i=1}^{n} \exp[itb_j Y_j)])$$

$$= \exp(it\sum_{j=1}^{n} a_j) \prod_{i=1}^{n} E[\exp(itb_j Y_j)]$$

since the expected value of a product of independent random variables is the product of the expected values of the independent random variables. Now in the definition of a cf, the t is a dummy variable as long as t is real. Hence $c_Y(t) = E[\exp(itY)]$ and $c_Y(s) = E[\exp(isY)]$. Taking $s = tb_j$ gives $E[\exp(itb_j Y_j)] = c_{Y_j}(tb_j)$. Thus

$$c_W(t) = \exp(it\sum_{j=1}^{n} a_j) \prod_{i=1}^{n} c_{Y_j}(tb_j). \quad \square$$

The distribution of $W = \sum_{i=1}^{n} Y_i$ is known as the convolution of $Y_1, ..., Y_n$. Even for $n = 2$, convolution formulas tend to be hard; however, the following

two theorems suggest that to find the distribution of $W = \sum_{i=1}^{n} Y_i$, first find the mgf or characteristic function of W using Theorem 1.23. If the mgf or cf is that of a brand name distribution, then W has that distribution. For example, if the mgf of W is a normal (ν, τ^2) mgf, then W has a normal (ν, τ^2) distribution, written $W \sim N(\nu, \tau^2)$. This technique is useful for several brand name distributions given in Section 1.10.

Theorem 1.24. a) If $Y_1, ..., Y_n$ are independent binomial BIN(k_i, ρ) random variables, then

$$\sum_{i=1}^{n} Y_i \sim \text{BIN}(\sum_{\text{i=1}}^{\text{n}} \text{k}_\text{i}, \rho).$$

Thus if $Y_1, ..., Y_n$ are iid BIN(k, ρ) random variables, then $\sum_{i=1}^{n} Y_i \sim \text{BIN}(\text{nk}, \rho)$.

b) Denote a chi–square χ^2_p random variable by $\chi^2(p)$. If $Y_1, ..., Y_n$ are independent chi–square $\chi^2_{p_i}$, then

$$\sum_{i=1}^{n} Y_i \sim \chi^2(\sum_{i=1}^{n} p_i).$$

Thus if $Y_1, ..., Y_n$ are iid χ^2_p, then

$$\sum_{i=1}^{n} Y_i \sim \chi^2_{np}.$$

c) If $Y_1, ..., Y_n$ are iid exponential EXP(λ), then

$$\sum_{i=1}^{n} Y_i \sim G(n, \lambda).$$

d) If $Y_1, ..., Y_n$ are independent Gamma $G(\nu_i, \lambda)$ then

$$\sum_{i=1}^{n} Y_i \sim G(\sum_{i=1}^{n} \nu_i, \lambda).$$

Thus if $Y_1, ..., Y_n$ are iid $G(\nu, \lambda)$, then

$$\sum_{i=1}^{n} Y_i \sim G(n\nu, \lambda).$$

e) If $Y_1, ..., Y_n$ are independent normal $N(\mu_i, \sigma_i^2)$, then

$$\sum_{i=1}^{n} (a_i + b_i Y_i) \sim N(\sum_{i=1}^{n} (a_i + b_i \mu_i), \sum_{i=1}^{n} b_i^2 \sigma_i^2).$$

Here a_i and b_i are fixed constants. Thus if $Y_1, ..., Y_n$ are iid $N(\mu, \sigma^2)$, then $\overline{Y} \sim N(\mu, \sigma^2/n)$.

f) If $Y_1, ..., Y_n$ are independent Poisson POIS(θ_i), then

$$\sum_{i=1}^{n} Y_i \sim \text{POIS}(\sum_{\text{i=1}}^{\text{n}} \theta_{\text{i}}).$$

Thus if $Y_1, ..., Y_n$ are iid $POIS(\theta)$, then

$$\sum_{i=1}^{n} Y_i \sim \text{POIS}(\text{n}\theta).$$

Theorem 1.25. a) If $Y_1, ..., Y_n$ are independent Cauchy $C(\mu_i, \sigma_i)$, then

$$\sum_{i=1}^{n}(a_i + b_i Y_i) \sim C(\sum_{i=1}^{n}(a_i + b_i\mu_i), \sum_{i=1}^{n} |b_i|\sigma_i).$$

Thus if $Y_1, ..., Y_n$ are iid $C(\mu, \sigma)$, then $\overline{Y} \sim C(\mu, \sigma)$.

b) If $Y_1, ..., Y_n$ are iid geometric geom(p), then

$$\sum_{i=1}^{n} Y_i \sim \text{NB}(\text{n}, \text{p}).$$

c) If $Y_1, ..., Y_n$ are iid inverse Gaussian $IG(\theta, \lambda)$, then

$$\sum_{i=1}^{n} Y_i \sim IG(n\theta, n^2\lambda).$$

Also

$$\overline{Y} \sim IG(\theta, n\lambda).$$

d) If $Y_1, ..., Y_n$ are independent negative binomial NB(r_i, ρ), then

$$\sum_{i=1}^{n} Y_i \sim \text{NB}(\sum_{\text{i=1}}^{\text{n}} \text{r}_\text{i}, \rho).$$

Thus if $Y_1, ..., Y_n$ are iid $NB(r, \rho)$, then

$$\sum_{i=1}^{n} Y_i \sim NB(nr, \rho).$$

Example 1.3. Suppose $Y_1, ..., Y_n$ are iid $IG(\theta, \lambda)$ where the mgf

$$m_{Y_i}(t) = m(t) = \exp\left[\frac{\lambda}{\theta}\left(1 - \sqrt{1 - \frac{2\theta^2 t}{\lambda}}\right)\right]$$

for $t < \lambda/(2\theta^2)$. Then

$$m_{\sum_{i=1}^n Y_i}(t) = \prod_{i=1}^n m_{Y_i}(t) = [m(t)]^n = \exp\left[\frac{n\lambda}{\theta}\left(1 - \sqrt{1 - \frac{2\theta^2 t}{\lambda}}\right)\right]$$

$$= \exp\left[\frac{n^2\lambda}{n\ \theta}\left(1 - \sqrt{1 - \frac{2(n\theta)^2\ t}{n^2\lambda}}\right)\right]$$

which is the mgf of an $IG(n\theta, n^2\lambda)$ random variable. The last equality was obtained by multiplying $\frac{n\lambda}{\theta}$ by $1 = n/n$ and by multiplying $\frac{2\theta^2 t}{\lambda}$ by $1 = n^2/n^2$. Hence $\sum_{i=1}^n Y_i \sim IG(n\theta, n^2\lambda)$.

1.5 The Multivariate Normal Distribution

Definition 1.37: Rao (1965, p. 437). A $p \times 1$ random vector $\boldsymbol{X}$ has a $p-$dimensional *multivariate normal distribution* $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ iff $\boldsymbol{t}^T\boldsymbol{X}$ has a univariate normal distribution for any $p \times 1$ vector $\boldsymbol{t}$.

If $\boldsymbol{\Sigma}$ is positive definite, then $\boldsymbol{X}$ has a joint pdf

$$f(\boldsymbol{z}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(1/2)(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})} \tag{1.26}$$

where $|\boldsymbol{\Sigma}|^{1/2}$ is the square root of the determinant of $\boldsymbol{\Sigma}$. Note that if $p = 1$, then the quadratic form in the exponent is $(z-\mu)(\sigma^2)^{-1}(z-\mu)$ and X has the univariate $N(\mu, \sigma^2)$ pdf. If $\boldsymbol{\Sigma}$ is positive semidefinite but not positive definite, then $\boldsymbol{X}$ has a degenerate distribution. For example, the univariate $N(0, 0^2)$ distribution is degenerate (the point mass at 0).

Some important properties of MVN distributions are given in the following three theorems. These theorems can be proved using results from Johnson and Wichern (1988, p. 127-132).

Theorem 1.26. a) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and

$$\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}.$$

b) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then any linear combination $\boldsymbol{t}^T\boldsymbol{X} = t_1X_1 + \cdots + t_pX_p \sim N_1(\boldsymbol{t}^T\boldsymbol{\mu}, \boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$. Conversely, if $\boldsymbol{t}^T\boldsymbol{X} \sim N_1(\boldsymbol{t}^T\boldsymbol{\mu}, \boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$ for every $p \times 1$ vector $\boldsymbol{t}$, then $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

c) **The joint distribution of independent normal random variables is MVN.** If $X_1, ..., X_p$ are independent univariate normal $N(\mu_i, \sigma_i^2)$ random vectors, then $\boldsymbol{X} = (X_1, ..., X_p)^T$ is $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = (\mu_1, ..., \mu_p)^T$ and $\boldsymbol{\Sigma} = diag(\sigma_1^2, ..., \sigma_p^2)$ (so the off diagonal entries $\sigma_{i,j} = 0$ while the diagonal entries of $\boldsymbol{\Sigma}$ are $\sigma_{i,i} = \sigma_i^2$.)

d) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and if $\boldsymbol{A}$ is a $q \times p$ matrix, then $\boldsymbol{AX} \sim N_q(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^T)$. If $\boldsymbol{a}$ is a $p \times 1$ vector of constants, then $\boldsymbol{a} + \boldsymbol{X} \sim N_p(\boldsymbol{a} + \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

It will be useful to partition $\boldsymbol{X}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$. Let $\boldsymbol{X}_1$ and $\boldsymbol{\mu}_1$ be $q \times 1$ vectors, let $\boldsymbol{X}_2$ and $\boldsymbol{\mu}_2$ be $(p-q) \times 1$ vectors, let $\boldsymbol{\Sigma}_{11}$ be a $q \times q$ matrix, let $\boldsymbol{\Sigma}_{12}$ be a $q \times (p-q)$ matrix, let $\boldsymbol{\Sigma}_{21}$ be a $(p-q) \times q$ matrix, and let $\boldsymbol{\Sigma}_{22}$ be a $(p-q) \times (p-q)$ matrix. Then

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Theorem 1.27. a) **All subsets of a MVN are MVN:** $(X_{k_1}, ..., X_{k_q})^T \sim N_q(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ where $\tilde{\boldsymbol{\mu}}_i = E(X_{k_i})$ and $\tilde{\boldsymbol{\Sigma}}_{ij} = \text{Cov}(X_{k_i}, X_{k_j})$. In particular, $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{X}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$.

b) If $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent, then $\text{Cov}(\boldsymbol{X}_1, \boldsymbol{X}_2) = \boldsymbol{\Sigma}_{12} = E[(\boldsymbol{X}_1 - E(\boldsymbol{X}_1))(\boldsymbol{X}_2 - E(\boldsymbol{X}_2))^T] = \boldsymbol{0}$, a $q \times (p-q)$ matrix of zeroes.

c) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent iff $\boldsymbol{\Sigma}_{12} = \boldsymbol{0}$.

d) If $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{X}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$ are independent, then

$$\begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \sim N_p \left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right).$$

Theorem 1.28. The conditional distribution of a MVN is MVN. If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. That is,

$$\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2 \sim N_q(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

Example 1.4. Let $p = 2$ and let $(Y, X)^T$ have a bivariate normal distribution. That is,

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \text{Cov}(Y, X) \\ \text{Cov}(X, Y) & \sigma_X^2 \end{pmatrix} \right).$$

Also recall that the population correlation between X and Y is given by

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{VAR}(X)}\sqrt{\text{VAR}(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

if $\sigma_X > 0$ and $\sigma_Y > 0$. Then $Y|X = x \sim N(E(Y|X = x), \text{VAR}(Y|X = x))$ where the conditional mean

$$E(Y|X = x) = \mu_Y + \text{Cov}(Y, X)\frac{1}{\sigma_X^2}(x - \mu_X) = \mu_Y + \rho(X, Y)\sqrt{\frac{\sigma_Y^2}{\sigma_X^2}}(x - \mu_X)$$

and the conditional variance

$$\text{VAR}(Y|X = x) = \sigma_Y^2 - \text{Cov}(X, Y)\frac{1}{\sigma_X^2}\text{Cov}(X, Y)$$

$$= \sigma_Y^2 - \rho(X, Y)\sqrt{\frac{\sigma_Y^2}{\sigma_X^2}}\rho(X, Y)\sqrt{\sigma_X^2}\sqrt{\sigma_Y^2}$$

$$= \sigma_Y^2 - \rho^2(X, Y)\sigma_Y^2 = \sigma_Y^2[1 - \rho^2(X, Y)].$$

Also $aX + bY$ is univariate normal with mean $a\mu_X + b\mu_Y$ and variance

$$a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\ \text{Cov}(X, Y).$$

Remark 1.6. There are several common misconceptions. First, **it is not true that every linear combination $\boldsymbol{t}^T\boldsymbol{X}$ of normal random variables is a normal random variable,** and **it is not true that all uncorrelated normal random variables are independent**. The key condition in Theorem 1.26b and Theorem 1.27c is that the joint distribution of $\boldsymbol{X}$ is MVN. It is possible that $X_1, X_2, ..., X_p$ each has a marginal distribution that is univariate normal, but the joint distribution of $\boldsymbol{X}$ is not MVN. Examine the following example from Rohatgi (1976, p. 229). Suppose that the joint pdf of X and Y is a mixture of two bivariate normal distributions both with $EX = EY = 0$ and $\text{VAR}(X) = \text{VAR}(Y) = 1$, but $\text{Cov}(X, Y) = \pm\rho$. Hence

$$f(x, y) = \frac{1}{2}\ \frac{1}{2\pi\sqrt{1 - \rho^2}}\exp(\frac{-1}{2(1 - \rho^2)}(x^2 - 2\rho xy + y^2)) +$$

$$\frac{1}{2}\ \frac{1}{2\pi\sqrt{1 - \rho^2}}\exp(\frac{-1}{2(1 - \rho^2)}(x^2 + 2\rho xy + y^2)) \equiv \frac{1}{2}f_1(x, y) + \frac{1}{2}f_2(x, y)$$

where x and y are real and $0 < \rho < 1$. Since both marginal distributions of $f_i(x, y)$ are N(0,1) for $i = 1$ and 2 by Theorem 1.27a, the marginal distributions of X and Y are N(0,1). Since $\int\int xyf_i(x, y)dxdy = \rho$ for $i = 1$ and $-\rho$ for $i = 2$, X and Y are uncorrelated, but X and Y are not independent since $f(x, y) \neq f_X(x)f_Y(y)$.

Remark 1.7. In Theorem 1.28, suppose that $\boldsymbol{X} = (Y, X_2, ..., X_p)^T$. Let $X_1 = Y$ and $\boldsymbol{X}_2 = (X_2, ..., X_p)^T$. Then $E[Y|\boldsymbol{X}_2] = \beta_1 + \beta_2X_2 + \cdots + \beta_pX_p$ and $\text{VAR}[Y|\boldsymbol{X}_2]$ is a constant that does not depend on $\boldsymbol{X}_2$. Hence $Y|\boldsymbol{X}_2 = \beta_1 + \beta_2X_2 + \cdots + \beta_pX_p + e$ follows the multiple linear regression model.

Example 1.5. Severini (2005, p. 236): Let $W \sim N(\mu_W, \sigma_W^2)$ and let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The characteristic function of W is

$$c_W(y) = E(e^{iyW}) = \exp\left(iy\mu_W - \frac{y^2}{2}\sigma_w^2\right).$$

Prove that the characteristic function of $\boldsymbol{X}$ is

$$c_{\boldsymbol{X}}(\boldsymbol{t}) = \exp\left(i\boldsymbol{t}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}\right).$$

Proof. Let $W = \boldsymbol{t}^T\boldsymbol{X}$. Then $W \sim N(\mu_W, \sigma_W^2)$ where $\mu_W = E(\boldsymbol{t}^T\boldsymbol{X}) = \boldsymbol{t}^T\boldsymbol{\mu}$ and $\sigma_W^2 = V(\boldsymbol{t}^T\boldsymbol{X}) = \mathrm{Cov}(\boldsymbol{t}^T\boldsymbol{X}) = \boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}$. Then

$$c_{\boldsymbol{X}}(\boldsymbol{t}) = E(e^{i\boldsymbol{t}^T\boldsymbol{X}}) = c_W(1) = \exp\left(i\mu_W - \frac{1}{2}\sigma_w^2\right) = \exp\left(i\boldsymbol{t}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}\right).$$

1.6 Exponential Families

Suppose the data is a random sample from some parametric brand name distribution with parameters $\boldsymbol{\theta}$. This brand name distribution comes from a family of distributions parameterized by $\boldsymbol{\theta} \in \Theta$. Each different value of $\boldsymbol{\theta}$ in the parameter space Θ gives a distribution that is a member of the family of distributions. Often the brand name family of distributions is from an exponential family.

Often a "brand name distribution" such as the normal distribution will have three useful parameterizations: the *usual parameterization* with parameter space Θ_U is simply the formula for the probability distribution or mass function (pdf or pmf, respectively) given when the distribution is first defined. The *k-parameter exponential family parameterization* with parameter space Θ, given in Definition 1.38 below, provides a simple way to determine if the distribution is an exponential family while the *natural parameterization* with parameter space Ω, given in Definition 1.39 below, is used for *theory* that requires a complete sufficient statistic.

Definition 1.38. A *family* of joint pdfs or joint pmfs $\{f(\boldsymbol{y}|\boldsymbol{\theta}) : \boldsymbol{\theta} = (\theta_1, ..., \theta_j) \in \Theta\}$ for a random vector $\boldsymbol{Y}$ is an **exponential family** if

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = h(\boldsymbol{y})c(\boldsymbol{\theta})\exp\left[\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(\boldsymbol{y})\right] \tag{1.27}$$

for all $\boldsymbol{y}$ where $c(\boldsymbol{\theta}) \geq 0$ and $h(\boldsymbol{y}) \geq 0$. The functions c, h, t_i, and w_i are real valued functions. The parameter $\boldsymbol{\theta}$ can be a scalar and $\boldsymbol{y}$ can be a scalar. It is crucial that $c, w_1, ..., w_k$ do not depend on $\boldsymbol{y}$ and that $h, t_1, ..., t_k$ do not

depend on $\boldsymbol{\theta}$. The support of the distribution is $\mathcal{Y}$ and the parameter space is Θ. The family is a **k-parameter exponential family** if k is the smallest integer where (1.27) holds.

Notice that the distribution of Y is an exponential family if

$$f(y|\boldsymbol{\theta}) = h(y)c(\boldsymbol{\theta}) \exp\left[\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(y)\right] \tag{1.28}$$

and the distribution is a one parameter exponential family if

$$f(y|\theta) = h(y)c(\theta) \exp[w(\theta)t(y)]. \tag{1.29}$$

The parameterization is not unique since, for example, w_i could be multiplied by a nonzero constant a if t_i is divided by a. Many other parameterizations are possible. If $h(y) = g(y)I_{\mathcal{Y}}(y)$, then usually $c(\boldsymbol{\theta})$ and $g(y)$ are positive, so another parameterization is

$$f(y|\boldsymbol{\theta}) = \exp\left[\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(y) + d(\boldsymbol{\theta}) + S(y)\right] \; I_{\mathcal{Y}}(y) \tag{1.30}$$

where $S(y) = \log(g(y))$, $d(\boldsymbol{\theta}) = \log(c(\boldsymbol{\theta}))$, and $\mathcal{Y}$ does not depend on $\boldsymbol{\theta}$.

To demonstrate that $\{f(\boldsymbol{y}|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ is an exponential family, find $h(\boldsymbol{y}), c(\boldsymbol{\theta}), w_i(\boldsymbol{\theta})$ and $t_i(\boldsymbol{y})$ such that (1.27), (1.28), (1.29), or (1.30) holds.

Theorem 1.29. Suppose that $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ are iid random vectors from an exponential family. Then the joint distribution of $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ follows an exponential family.

Proof. Suppose that $f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i)$ has the form of (1.27). Then by independence,

$$f(\boldsymbol{y}_1, ..., \boldsymbol{y}_n) = \prod_{i=1}^{n} f_{\boldsymbol{Y}_i}(\boldsymbol{y}_i) = \prod_{i=1}^{n} h(\boldsymbol{y}_i)c(\boldsymbol{\theta}) \exp\left[\sum_{j=1}^{k} w_j(\boldsymbol{\theta})t_j(\boldsymbol{y}_i)\right]$$

$$= [\prod_{i=1}^{n} h(\boldsymbol{y}_i)][c(\boldsymbol{\theta})]^n \prod_{i=1}^{n} \exp\left[\sum_{j=1}^{k} w_j(\boldsymbol{\theta})t_j(\boldsymbol{y}_i)\right]$$

$$= [\prod_{i=1}^{n} h(\boldsymbol{y}_i)][c(\boldsymbol{\theta})]^n \exp\left(\sum_{i=1}^{n}\left[\sum_{j=1}^{k} w_j(\boldsymbol{\theta})t_j(\boldsymbol{y}_i)\right]\right)$$

$$= [\prod_{i=1}^{n} h(\boldsymbol{y}_i)][c(\boldsymbol{\theta})]^n \exp\left[\sum_{j=1}^{k} w_j(\boldsymbol{\theta})\left(\sum_{i=1}^{n} t_j(\boldsymbol{y}_i)\right)\right].$$

To see that this has the form (1.27), take $h^*(\boldsymbol{y}_1, ..., \boldsymbol{y}_n) = \prod_{i=1}^n h(\boldsymbol{y}_i)$, $c^*(\boldsymbol{\theta}) = [c(\boldsymbol{\theta})]^n$, $w_j^*(\boldsymbol{\theta}) = w_j(\boldsymbol{\theta})$ and $t_j^*(\boldsymbol{y}_1, ..., \boldsymbol{y}_n) = \sum_{i=1}^n t_j(\boldsymbol{y}_i)$. □

The parameterization that uses the **natural parameter $\boldsymbol{\eta}$** is especially useful for theory. See Definition 1.40 for the natural parameter space Ω.

Definition 1.39. Let Ω be the natural parameter space for $\boldsymbol{\eta}$. The **natural parameterization for an exponential family** is

$$f(\boldsymbol{y}|\boldsymbol{\eta}) = h(\boldsymbol{y})b(\boldsymbol{\eta}) \exp\left[\sum_{i=1}^k \eta_i t_i(\boldsymbol{y})\right] \tag{1.31}$$

where $h(\boldsymbol{y})$ and $t_i(\boldsymbol{y})$ are the same as in Equation (1.27) and $\boldsymbol{\eta} \in \Omega$. The natural parameterization for a random variable Y is

$$f(y|\boldsymbol{\eta}) = h(y)b(\boldsymbol{\eta}) \exp\left[\sum_{i=1}^k \eta_i t_i(y)\right] \tag{1.32}$$

where $h(y)$ and $t_i(y)$ are the same as in Equation (1.27) and $\boldsymbol{\eta} \in \Omega$. Again, the parameterization is not unique. If $a \neq 0$, then $a\eta_i$ and $t_i(y)/a$ would also work.

Notice that the natural parameterization (1.32) has the same form as (1.27) with $\boldsymbol{\theta}^* = \boldsymbol{\eta}$, $c^*(\boldsymbol{\theta}^*) = b(\boldsymbol{\eta})$ and $w_i(\boldsymbol{\theta}^*) = w_i(\boldsymbol{\eta}) = \eta_i$. In applications often $\boldsymbol{\eta}$ and Ω are of interest while $b(\boldsymbol{\eta})$ is not computed.

The next important idea is that of a regular exponential family (and of a full exponential family). Let $d_i(x)$ denote $t_i(y)$, $w_i(\boldsymbol{\theta})$ or η_i. A *linearity constraint* is satisfied by $d_1(x), ..., d_k(x)$ if $\sum_{i=1}^k a_i d_i(x) = c$ for some constants a_i and c and for all x (or η_i) in the sample or parameter space where not all of the $a_i = 0$. If $\sum_{i=1}^k a_i d_i(x) = c$ for all x only if $a_1 = \cdots = a_k = 0$, then the $d_i(x)$ do not satisfy a linearity constraint. In linear algebra, we would say that the $d_i(x)$ are *linearly independent* if they do not satisfy a linearity constraint.

For $k = 2$, a linearity constraint is satisfied if a plot of $d_1(x)$ versus $d_2(x)$ falls on a line as x varies. If the parameter space for the η_1 and η_2 is a nonempty open set, then the plot of η_1 versus η_2 is that nonempty open set, and the η_i can not satisfy a linearity constraint since the plot is not a line.

Let $\tilde{\Omega}$ be the set where the integral of the kernel function is finite:

$$\tilde{\Omega} = \{\boldsymbol{\eta} = (\eta_1, ..., \eta_k) : \frac{1}{b(\boldsymbol{\eta})} \equiv \int_{-\infty}^{\infty} h(y) \exp[\sum_{i=1}^k \eta_i t_i(y)] dy < \infty\}. \tag{1.33}$$

Replace the integral by a sum for a pmf. An interesting fact is that $\tilde{\Omega}$ is a convex set. If the parameter space Θ of the exponential family is not a convex set, then the exponential family can not be regular. Example 1.7 shows that

the χ^2_p distribution is not regular since the set of positive integers is not convex.

Definition 1.40. Condition E1: the natural parameter space $\Omega = \tilde{\Omega}$.
Condition E2: assume that in the natural parameterization, neither the η_i nor the t_i satisfy a linearity constraint.
Condition E3: Ω is a k-dimensional nonempty open set.
If conditions E1), E2) and E3) hold then the exponential family is a k-parameter **regular exponential family** (REF).
If conditions E1) and E2) hold then the exponential family is a k-parameter *full exponential family*.

Notation. A kP–REF is a k-parameter regular exponential family. So a 1P–REF is a 1-parameter REF and a 2P–REF is a 2-parameter REF.

Notice that every REF is full. Any k–dimensional nonempty open set will contain a k–dimensional nonempty rectangle. A k–fold cross product of nonempty open intervals is a k–dimensional nonempty open set. For a one parameter exponential family, a one dimensional rectangle is just an interval, and the only type of function of one variable that satisfies a linearity constraint is a constant function. In the definition of an exponential family and in the usual parameterization, $\boldsymbol{\theta}$ is a $1 \times j$ vector. Typically $j = k$ if the family is a kP–REF. If $j < k$ and k is as small as possible, the family will usually not be regular. For example, a $N(\theta, \theta^2)$ family has $\boldsymbol{\theta} = \theta$ with $j = 1 < 2 = k$, and is not regular.

Some care has to be taken with the definitions of Θ and Ω since formulas (1.27) and (1.32) need to hold for every $\boldsymbol{\theta} \in \Theta$ and for every $\boldsymbol{\eta} \in \Omega$. Let Θ_U be the usual parameter space given for the distribution. For a continuous random variable or vector, the pdf needs to exist. Hence all degenerate distributions need to be deleted from Θ_U to form Θ and Ω. For continuous and discrete distributions, the natural parameter needs to exist (and often does not exist for discrete degenerate distributions). As a rule of thumb, remove values from Θ_U that cause the pmf to have the form 0^0. For example, for the binomial(k, ρ) distribution with k known, the natural parameter $\eta = \log(\rho/(1-\rho))$. Hence instead of using $\Theta_U = [0, 1]$, use $\rho \in \Theta = (0, 1)$, so that $\eta \in \Omega = (-\infty, \infty)$.

These conditions have some redundancy. If Ω contains a k-dimensional rectangle (e.g. if the family is a kP-REF, then Ω is a k-dimensional open set and contains a k-dimensional open ball which contains a k-dimensional rectangle), no η_i is completely determined by the remaining $\eta_j's$. In particular, the η_i cannot satisfy a linearity constraint. If the η_i do satisfy a linearity constraint, then the η_i lie on a hyperplane of dimension at most k, and such a surface cannot contain a k-dimensional rectangle. For example, if $k = 2$, a line cannot contain an open box. If $k = 2$ and $\eta_2 = \eta_1^2$, then the parameter space is not a 2-dimensional open set and does not contain a 2-dimensional

rectangle. Thus the family is not a 2P–REF although η_1 and η_2 do not satisfy a linearity constraint.

The most important 1P–REFs are the binomial (k, ρ) distribution with k known, the exponential (λ) distribution, and the Poisson (θ) distribution. A one parameter exponential family can often be obtained from a k–parameter exponential family by holding $k-1$ of the parameters fixed. Hence a normal (μ, σ^2) distribution is a 1P–REF if σ^2 is known. When data is modeled with an exponential family, often the scale, location and shape parameters are unknown. For example, the mean and standard deviation are usually both unknown.

The most important 2P–REFs are the beta (δ, ν) distribution, the gamma (ν, λ) distribution and the normal (μ, σ^2) distribution. The chi (p, σ) distribution, the inverted gamma (ν, λ) distribution, the log-gamma (ν, λ) distribution and the lognormal (μ, σ^2) distribution are also 2P-REFs. Olive (2014) gives many other examples showing that a distribution is a 1P-REF or 2P-REF. The two parameter Cauchy distribution is not an exponential family because its pdf cannot be put into the form of Equation (1.27).

The natural parameterization can result in a family that is much larger than the family defined by the usual parameterization. See the definition of $\Omega = \tilde{\Omega}$ given by Equation (1.33). Casella and Berger (2002, p. 114) remarks that

$$\{\boldsymbol{\eta} : \boldsymbol{\eta} = (w_1(\boldsymbol{\theta}), ..., w_k(\boldsymbol{\theta})) | \boldsymbol{\theta} \in \Theta\} \subseteq \Omega, \tag{1.34}$$

but often Ω is a strictly larger set.

Remark 1.8. For the families in this text other than the χ^2_p and inverse Gaussian distributions, make the following assumptions if $\dim(\Theta) = k = \dim(\Omega)$. Assume that $\eta_i = w_i(\boldsymbol{\theta})$. Assume the usual parameter space Θ_U is as big as possible (replace the integral by a sum for a pmf):

$$\Theta_U = \{\boldsymbol{\theta} \in \mathbb{R}^k : \int f(y|\boldsymbol{\theta}) dy = 1\},$$

and let

$$\Theta = \{\boldsymbol{\theta} \in \Theta_U : w_1(\boldsymbol{\theta}), ..., w_k(\boldsymbol{\theta}) \text{ are defined } \}.$$

Then assume that the natural parameter space satisfies condition E1) with

$$\Omega = \{(\eta_1, ..., \eta_k) : \eta_i = w_i(\boldsymbol{\theta}) \text{ for } \boldsymbol{\theta} \in \Theta\}.$$

In other words, simply define $\eta_i = w_i(\boldsymbol{\theta})$. For many common distributions, $\boldsymbol{\eta}$ is a one to one function of $\boldsymbol{\theta}$, and the above map is correct, especially if Θ_U is an open interval or cross product of open intervals.

Example 1.6. Let $f(x|\mu, \sigma)$ be the $N(\mu, \sigma^2)$ family of pdfs. Then $\boldsymbol{\theta} = (\mu, \sigma)$ where $-\infty < \mu < \infty$ and $\sigma > 0$. Recall that μ is the mean and σ is the

standard deviation (SD) of the distribution. The usual parameterization is

$$f(x|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp(\frac{-(x-\mu)^2}{2\sigma^2}) I_{\mathbb{R}}(x)$$

where $\mathbb{R} = (-\infty, \infty)$ and the indicator $I_A(x) = 1$ if $x \in A$ and $I_A(x) = 0$ otherwise. Notice that $I_{\mathbb{R}}(x) = 1 \ \ \forall x$. Since

$$f(x|\mu,\sigma) = \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp(\frac{-\mu^2}{2\sigma^2})}_{c(\mu,\sigma)\geq 0} \exp(\underbrace{\frac{-1}{2\sigma^2}}_{w_1(\boldsymbol{\theta})} \underbrace{x^2}_{t_1(x)} + \underbrace{\frac{\mu}{\sigma^2}}_{w_2(\boldsymbol{\theta})} \underbrace{x}_{t_2(x)}) \underbrace{I_{\mathbb{R}}(x)}_{h(x)\geq 0},$$

this family is a 2-parameter exponential family. Hence $\eta_1 = -0.5/\sigma^2$ and $\eta_2 = \mu/\sigma^2$ if $\sigma > 0$, and $\Omega = (-\infty, 0) \times (-\infty, \infty)$. Plotting η_1 on the horizontal axis and η_2 on the vertical axis yields the left half plane which certainly contains a 2-dimensional rectangle. Since t_1 and t_2 lie on a quadratic rather than a line, the family is a 2P–REF. Notice that if $X_1, ..., X_n$ are iid $N(\mu, \sigma^2)$ random variables, then the joint pdf $f(\boldsymbol{x}|\boldsymbol{\theta}) = f(x_1, ..., x_n|\mu, \sigma) =$

$$\underbrace{[\frac{1}{\sqrt{2\pi}\sigma} \exp(\frac{-\mu^2}{2\sigma^2})]^n}_{C(\mu,\sigma)\geq 0} \exp(\underbrace{\frac{-1}{2\sigma^2}}_{w_1(\boldsymbol{\theta})} \underbrace{\sum_{i=1}^{n} x_i^2}_{T_1(\boldsymbol{x})} + \underbrace{\frac{\mu}{\sigma^2}}_{w_2(\boldsymbol{\theta})} \underbrace{\sum_{i=1}^{n} x_i}_{T_2(\boldsymbol{x})}) \underbrace{1}_{h(\boldsymbol{x})\geq 0},$$

and is thus a 2P–REF.

Example 1.7. The χ^2_p distribution is not a 1P-REF since the usual parameter space Θ_U for the χ^2_p distribution is the set of positive integers, which is neither an open set nor a convex set. Nevertheless, the natural parameterization is the gamma($\nu, \lambda = 2$) family which is a 1P-REF. Note that this family has uncountably many members while the χ^2_p family does not.

Example 1.8. The binomial(k, ρ) pmf is

$$f(x|\rho) = \binom{k}{x} \rho^x (1-\rho)^{k-x} \ I_{\{0,...,k\}}(x)$$

$$= \underbrace{\binom{k}{x} I_{\{0,...,k\}}(x)}_{h(x)\geq 0} \underbrace{(1-\rho)^k}_{c(\rho)\geq 0} \exp[\underbrace{\log(\frac{\rho}{1-\rho})}_{w(\rho)} \underbrace{x}_{t(x)}]$$

where $\Theta_U = [0, 1]$. Since the pmf and $\eta = \log(\rho/(1-\rho))$ is undefined for $\rho = 0$ and $\rho = 1$, we have $\Theta = (0, 1)$. Notice that $\Omega = (-\infty, \infty)$.

Example 1.9. The uniform(0,θ) family is not an exponential family since the support $\mathcal{Y}_\theta = (0, \theta)$ depends on the unknown parameter θ.

1.6.1 Properties of $(t_1(\boldsymbol{Y}), ..., t_k(\boldsymbol{Y}))$

Write the *natural parameterization for the exponential family* as

$$f(y|\boldsymbol{\eta}) = h(y)b(\boldsymbol{\eta}) \exp\left[\sum_{i=1}^{k} \eta_i t_i(y)\right]$$

$$= h(y) \exp\left[\sum_{i=1}^{k} \eta_i t_i(y) \quad - \quad a(\boldsymbol{\eta})\right] \tag{1.35}$$

where $a(\boldsymbol{\eta}) = -\log(b(\boldsymbol{\eta}))$.

Theorem 1.30. Suppose that Y comes from an exponential family (1.35) and that $g(y)$ is any function with $E_{\boldsymbol{\eta}}[|g(Y)|] < \infty$. Then for any $\boldsymbol{\eta}$ in the interior of Ω, the integral $\int g(y)f(y|\boldsymbol{\eta})dy$ is continuous and has derivatives of all orders. These derivatives can be obtained by interchanging the derivative and integral operators. If f is a pmf, replace the integral by a sum.

Proof. See Lehmann (1986, p. 59).

Hence

$$\frac{\partial}{\partial \eta_i} \int g(y)f(y|\boldsymbol{\eta})dy = \int g(y) \frac{\partial}{\partial \eta_i} f(y|\boldsymbol{\eta})dy \tag{1.36}$$

if f is a pdf and

$$\frac{\partial}{\partial \eta_i} \sum g(y)f(y|\boldsymbol{\eta}) = \sum g(y) \frac{\partial}{\partial \eta_i} f(y|\boldsymbol{\eta}) \tag{1.37}$$

if f is a pmf.

Remark 1.9. If $\boldsymbol{Y}$ comes from an exponential family (1.27), then the derivative and integral (or sum) operators can be interchanged. Hence

$$\frac{\partial}{\partial \theta_i} \int ... \int g(\boldsymbol{y})f(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{y} = \int ... \int g(\boldsymbol{y}) \frac{\partial}{\partial \theta_i} f(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{y}$$

for any function $g(\boldsymbol{y})$ with $E_\theta|g(\boldsymbol{Y})| < \infty$.

The behavior of $(t_1(Y), ..., t_k(Y))$ will be of considerable interest in later chapters. The following result is in Lehmann (1983, p. 29-30). Also see Johnson, Ladella, and Liu (1979).

Theorem 1.31. Suppose that Y comes from an exponential family (1.35). Then a)

$$E(t_i(Y)) = \frac{\partial}{\partial \eta_i} a(\boldsymbol{\eta}) = - \frac{\partial}{\partial \eta_i} \log(b(\boldsymbol{\eta})) \tag{1.38}$$

and b)

$$\mathrm{Cov}(t_i(Y), t_j(Y)) = \frac{\partial^2}{\partial \eta_i \partial \eta_j} a(\boldsymbol{\eta}) = - \frac{\partial^2}{\partial \eta_i \partial \eta_j} \log(b(\boldsymbol{\eta})). \tag{1.39}$$

Notice that $i = j$ gives the formula for $\mathrm{VAR}(t_i(Y))$.

Proof. The proof will be for pdfs. For pmfs replace the integrals by sums. Use Theorem 1.30 with $g(y) = 1 \ \forall y$. a) Since $1 = \int f(y|\boldsymbol{\eta}) dy$,

$$0 = \frac{\partial}{\partial \eta_i} 1 = \frac{\partial}{\partial \eta_i} \int h(y) \exp\left[\sum_{m=1}^{k} \eta_m t_m(y) \quad - \quad a(\boldsymbol{\eta}) \right] dy$$

$$= \int h(y) \frac{\partial}{\partial \eta_i} \exp\left[\sum_{m=1}^{k} \eta_m t_m(y) \quad - \quad a(\boldsymbol{\eta}) \right] dy$$

$$= \int h(y) \exp\left[\sum_{m=1}^{k} \eta_m t_m(y) \quad - \quad a(\boldsymbol{\eta}) \right] (t_i(y) - \frac{\partial}{\partial \eta_i} a(\boldsymbol{\eta})) dy$$

$$= \int (t_i(y) - \frac{\partial}{\partial \eta_i} a(\boldsymbol{\eta})) f(y|\boldsymbol{\eta}) dy$$

$$= E(t_i(Y)) - \frac{\partial}{\partial \eta_i} a(\boldsymbol{\eta}).$$

b) Similarly,

$$0 = \int h(y) \frac{\partial^2}{\partial \eta_i \partial \eta_j} \exp\left[\sum_{m=1}^{k} \eta_m t_m(y) \quad - \quad a(\boldsymbol{\eta}) \right] dy.$$

From the proof of a),

$$0 = \int h(y) \frac{\partial}{\partial \eta_j} \left[\exp\left[\sum_{m=1}^{k} \eta_m t_m(y) \quad - \quad a(\boldsymbol{\eta}) \right] (t_i(y) - \frac{\partial}{\partial \eta_i} a(\boldsymbol{\eta})) \right] dy$$

$$= \int h(y) \exp\left[\sum_{m=1}^{k} \eta_m t_m(y) \quad - \quad a(\boldsymbol{\eta}) \right] (t_i(y) - \frac{\partial}{\partial \eta_i} a(\boldsymbol{\eta}))(t_j(y) - \frac{\partial}{\partial \eta_j} a(\boldsymbol{\eta})) dy$$

$$- \int h(y) \exp\left[\sum_{m=1}^{k} \eta_m t_m(y) \quad - \quad a(\boldsymbol{\eta}) \right] (\frac{\partial^2}{\partial \eta_i \partial \eta_j} a(\boldsymbol{\eta})) dy$$

$$= \mathrm{Cov}(t_i(Y), t_j(Y)) - \frac{\partial^2}{\partial \eta_i \partial \eta_j} a(\boldsymbol{\eta})$$

since $\frac{\partial}{\partial \eta_j} a(\boldsymbol{\eta}) = E(t_j(Y))$ by a). □

Theorem 1.32. Suppose that Y comes from an exponential family (1.35), and let $\boldsymbol{T} = (t_1(Y), ..., t_k(Y)) = (T_1, ..., T_k)$. Then the distribution of $\boldsymbol{T}$ is an

exponential family with

$$f(\boldsymbol{t}|\boldsymbol{\eta}) = h^*(\boldsymbol{t}) \exp\left[\sum_{i=1}^{k} \eta_i t_i \quad - \quad a(\boldsymbol{\eta})\right].$$

Proof. See Lehmann (1986, p. 58).

The main point of this section is that $\boldsymbol{T}$ is well behaved even if Y is not. For example, if Y follows a one sided stable distribution, then Y is from an exponential family, but $E(Y)$ does not exist. However the mgf of T exists, so all moments of T exist. If $Y_1, ..., Y_n$ are iid from a one parameter exponential family, then $T \equiv T_n = \sum_{i=1}^{n} t(Y_i)$ is from a one parameter exponential family. One way to find the distribution function of T is to find the distribution of $t(Y)$ using the transformation method, then find the distribution of $\sum_{i=1}^{n} t(Y_i)$ using moment generating functions or Theorems 1.24 and 1.25. This technique results in the following two theorems. Notice that T often has a gamma distribution.

1.7 MSE, Information Number, MLE, UMVUE

Definition 1.41. Let the sample $\boldsymbol{Y} = (Y_1, ..., Y_n)$ where $\boldsymbol{Y}$ has a pdf or pmf $f(\boldsymbol{y}|\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta$. Assume all relevant expectations exist. Let $\tau(\boldsymbol{\theta})$ be a real valued function of $\boldsymbol{\theta}$, and let $T \equiv T(Y_1, ..., Y_n)$ be an estimator of $\tau(\boldsymbol{\theta})$. The **bias** of the estimator T for $\tau(\boldsymbol{\theta})$ is

$$B(T) \equiv B_{\tau(\boldsymbol{\theta})}(T) \equiv \text{Bias(T)} \equiv \text{Bias}_{\tau(\boldsymbol{\theta})}(\text{T}) = \text{E}_{\boldsymbol{\theta}}(\text{T}) - \tau(\boldsymbol{\theta}). \tag{1.40}$$

The *mean squared error* (**MSE**) of an estimator T for $\tau(\boldsymbol{\theta})$ is

$$\text{MSE(T)} \equiv \text{MSE}_{\tau(\boldsymbol{\theta})}(\text{T}) = \text{E}_{\boldsymbol{\theta}}[(\text{T} - \tau(\boldsymbol{\theta}))^2]$$

$$= Var_{\boldsymbol{\theta}}(T) + [\text{Bias}_{\tau(\boldsymbol{\theta})}(\text{T})]^2. \tag{1.41}$$

T is an *unbiased estimator* of $\tau(\boldsymbol{\theta})$ if

$$E_{\boldsymbol{\theta}}(T) = \tau(\boldsymbol{\theta}) \tag{1.42}$$

for all $\boldsymbol{\theta} \in \Theta$. Notice that $\text{Bias}_{\tau(\boldsymbol{\theta})}(\text{T}) = 0$ for all $\boldsymbol{\theta} \in \Theta$ if T is an unbiased estimator of $\tau(\boldsymbol{\theta})$.

Notice that the bias and MSE are functions of $\boldsymbol{\theta}$ for $\boldsymbol{\theta} \in \Theta$. If $MSE_{\tau(\boldsymbol{\theta})}(T_1) < MSE_{\tau(\boldsymbol{\theta})}(T_2)$ for all $\boldsymbol{\theta} \in \Theta$, then T_1 is "a better estimator" of $\tau(\boldsymbol{\theta})$ than T_2. So estimators with small MSE are judged to be better than ones with

large MSE. Often T_1 has smaller MSE than T_2 for some $\boldsymbol{\theta}$ but larger MSE for other values of $\boldsymbol{\theta}$. Often θ is real valued.

Example 1.10. Find the bias and MSE (as a function of n and c) of an estimator $T = c\sum_{i=1}^{n} Y_i$ or $(T = b\overline{Y})$ of μ when $Y_1, ..., Y_n$ are iid with $E(Y_1) = \mu = \theta$ and $V(Y_i) = \sigma^2$.
Solution: $E(T) = c\sum_{i=1}^{n} E(Y_i) = nc\mu$, $V(T) = c^2\sum_{i=1}^{n} V(Y_i) = nc^2\sigma^2$, $B(T) = E(T) - \mu$ and $MSE(T) = V(T) + [B(T)]^2$. (For $T = b\overline{Y}$, use $c = b/n$.)

In the class of unbiased estimators, the UMVUE is best since the UMVUE has the smallest variance, hence the smallest MSE. Often the MLE and method of moments estimator are biased but have a smaller MSE than the UMVUE. MLEs and method of moments estimators are widely used because they often have good statistical properties and are relatively easy to compute. Sometimes the UMVUE, MLE and method of moments estimators for θ are the same for a 1P-REF when $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} t(Y_i)$ and $\theta = E(\hat{\theta}) = E[t(Y)]$.

Definition 1.42. Let the sample $\boldsymbol{Y} = (Y_1, ..., Y_n)$ where $\boldsymbol{Y}$ has a pdf or pmf $f(\boldsymbol{y}|\theta)$ for $\theta \in \Theta$. Assume all relevant expectations exist. Let $\tau(\theta)$ be a real valued function of θ, and let $U \equiv U(Y_1, ..., Y_n)$ be an estimator of $\tau(\theta)$. Then U is the *uniformly minimum variance unbiased estimator* (**UMVUE**) of $\tau(\theta)$ if U is an unbiased estimator of $\tau(\theta)$ and if $\mathrm{Var}_\theta(U) \leq \mathrm{Var}_\theta(\mathrm{W})$ for all $\theta \in \Theta$ where W is any other unbiased estimator of $\tau(\theta)$.

Often students will be asked to compute a lower bound on the variance of unbiased estimators of $\eta = \tau(\theta)$ when θ is a scalar. Some preliminary results are needed to define the lower bound, known as the FCRLB. The Fisher information, defined below, is also useful for large sample theory in Chapter 2 since often the asymptotic variance of a good estimator of $\tau(\theta)$ is $1/I_n(\tau(\theta))$. Good estimators tend to have a variance $\geq c/n$, so the FCRLB should be c/n for some positive constant c that may depend on the parameters of the distribution. Often $c = [\tau'(\theta)]^2/I_1(\theta)$.

Definition 1.43. Let $\boldsymbol{Y} = (Y_1, ..., Y_n)$ have a pdf or pmf $f(\boldsymbol{y}|\theta)$. Then the **information number** or **Fisher Information** is

$$I_{\boldsymbol{Y}}(\theta) \equiv I_n(\theta) = E_\theta\left(\left[\frac{\partial}{\partial\theta}\log(f(\boldsymbol{Y}|\theta))\right]^2\right). \tag{1.43}$$

Let $\eta = \tau(\theta)$ where $\tau'(\theta) \neq 0$. Then

$$I_n(\eta) \equiv I_n(\tau(\theta)) = \frac{I_n(\theta)}{[\tau'(\theta)]^2.} \tag{1.44}$$

Theorem 1.33. a) Equations (1.43) and (1.44) agree if $\tau'(\theta)$ is continuous, $\tau'(\theta) \neq 0$, and $\tau(\theta)$ is one to one and onto so that an inverse function exists such that $\theta = \tau^{-1}(\eta)$.

b) If the $Y_1 \equiv Y$ is from a 1P–REF, then the Fisher information in a sample of size one is

$$I_1(\theta) = -E_\theta \left[\frac{\partial^2}{\partial \theta^2} \log(f(Y|\theta)) \right]. \tag{1.45}$$

c) If the $Y_1, ..., Y_n$ are iid from a 1P–REF, then

$$I_n(\theta) = nI_1(\theta). \tag{1.46}$$

Hence if $\tau'(\theta)$ exists and is continuous and if $\tau'(\theta) \neq 0$, then

$$I_n(\tau(\theta)) = \frac{nI_1(\theta)}{[\tau'(\theta)]^2}. \tag{1.47}$$

Proof. a) See Lehmann (1999, p. 467–468).

b) The proof will be for a pdf. For a pmf replace the integrals by sums. By Remark 1.9, the integral and differentiation operators of all orders can be interchanged. Note that

$$0 = E\left[\frac{\partial}{\partial \theta} \log(f(Y|\theta)) \right] \tag{1.48}$$

since

$$\frac{\partial}{\partial \theta} 1 = 0 = \frac{\partial}{\partial \theta} \int f(y|\theta) dy = \int \frac{\partial}{\partial \theta} f(y|\theta) dy = \int \frac{\frac{\partial}{\partial \theta} f(y|\theta)}{f(y|\theta)} f(y|\theta) dy$$

or

$$0 = \frac{\partial}{\partial \theta} \int f(y|\theta) dy = \int \left[\frac{\partial}{\partial \theta} \log(f(y|\theta)) \right] f(y|\theta) dy$$

which is (1.48). Taking 2nd derivatives of the above expression gives

$$0 = \frac{\partial^2}{\partial \theta^2} \int f(y|\theta) dy = \frac{\partial}{\partial \theta} \int \left[\frac{\partial}{\partial \theta} \log(f(y|\theta)) \right] f(y|\theta) dy =$$

$$\int \frac{\partial}{\partial \theta} \left(\left[\frac{\partial}{\partial \theta} \log(f(y|\theta)) \right] f(y|\theta) \right) dy =$$

$$\int \left[\frac{\partial^2}{\partial \theta^2} \log(f(y|\theta)) \right] f(y|\theta) dy + \int \left[\frac{\partial}{\partial \theta} \log(f(y|\theta)) \right] \left[\frac{\partial}{\partial \theta} f(y|\theta) \right] \frac{f(y|\theta)}{f(y|\theta)} dy$$

$$= \int \left[\frac{\partial^2}{\partial \theta^2} \log(f(y|\theta)) \right] f(y|\theta) dy + \int \left[\frac{\partial}{\partial \theta} \log(f(y|\theta)) \right]^2 f(y|\theta) dy$$

or

$$I_1(\theta) = E_\theta[(\frac{\partial}{\partial\theta}\log f(Y|\theta))^2] = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log(f(Y|\theta))\right].$$

c) By independence,

$$I_n(\theta) = E_\theta\left[\left(\frac{\partial}{\partial\theta}\log(\prod_{i=1}^n f(Y_i|\theta))\right)^2\right] = E_\theta\left[\left(\frac{\partial}{\partial\theta}\sum_{i=1}^n \log(f(Y_i|\theta))\right)^2\right] =$$

$$E_\theta\left[\left(\frac{\partial}{\partial\theta}\sum_{i=1}^n \log(f(Y_i|\theta))\right)\left(\frac{\partial}{\partial\theta}\sum_{j=1}^n \log(f(Y_j|\theta))\right)\right] =$$

$$E_\theta\left[\left(\sum_{i=1}^n \frac{\partial}{\partial\theta}\log(f(Y_i|\theta))\right)\left(\sum_{j=1}^n \frac{\partial}{\partial\theta}\log(f(Y_j|\theta))\right)\right] =$$

$$\sum_{i=1}^n E_\theta\left[\left(\frac{\partial}{\partial\theta}\log(f(Y_i|\theta))\right)^2\right] +$$

$$\sum_{i \neq j}\sum E_\theta\left[\left(\frac{\partial}{\partial\theta}\log(f(Y_i|\theta))\right)\left(\frac{\partial}{\partial\theta}\log(f(Y_j|\theta))\right)\right].$$

Hence

$$I_n(\theta) = nI_1(\theta) + \sum_{i \neq j}\sum E_\theta\left[\left(\frac{\partial}{\partial\theta}\log(f(Y_i|\theta))\right)\right] E_\theta\left[\left(\frac{\partial}{\partial\theta}\log(f(Y_j|\theta))\right)\right]$$

by independence. Hence

$$I_n(\theta) = nI_1(\theta) + n(n-1)\left[E_\theta\left(\frac{\partial}{\partial\theta}\log(f(Y_j|\theta))\right)\right]^2$$

since the Y_i are iid. Thus $I_n(\theta) = nI_1(\theta)$ by Equation (1.48) which holds since the Y_i are iid from a 1P–REF. □

Definition 1.44. Let $\boldsymbol{Y} = (Y_1, ..., Y_n)$ be the data, and consider $\tau(\theta)$ where $\tau'(\theta) \neq 0$. The quantity

$$\mathrm{FCRLB_n}(\tau(\theta)) = \frac{[\tau'(\theta)]^2}{\mathrm{I_n}(\theta)}$$

is called the **Fréchet Cramér Rao lower bound** (FCRLB) for the variance of unbiased estimators of $\tau(\theta)$. In particular, if $\tau(\theta) = \theta$, then $\mathrm{FCRLB}_n(\theta) = \frac{1}{I_n(\theta)}$. The FCRLB is often called the Cramér Rao lower bound (CRLB).

Theorem 1.35, Fréchet Cramér Rao Lower Bound or Information Inequality. Let $Y_1, ..., Y_n$ be iid from a 1P–REF with pdf or pmf $f(y|\theta)$. Let $W(Y_1, ..., Y_n) = W(\boldsymbol{Y})$ be any unbiased estimator of $\tau(\theta) \equiv E_\theta W(\boldsymbol{Y})$. Then

$$\mathrm{VAR}_\theta(W(\boldsymbol{Y})) \geq FCRLB_n(\tau(\theta)) = \frac{[\tau'(\theta)]^2}{I_n(\theta)} = \frac{[\tau'(\theta)]^2}{nI_1(\theta)} = \frac{1}{I_n(\tau(\theta))}.$$

Proof. See Olive (2014, pp. 164-166).

Definition 1.45. Let $f(\boldsymbol{y}|\boldsymbol{\theta})$ be the pmf or pdf of a sample $\boldsymbol{Y}$ with parameter space Θ. If $\boldsymbol{Y} = \boldsymbol{y}$ is observed, then the **likelihood function** is $L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta})$. For each sample point $\boldsymbol{y} = (y_1, ..., y_n)$, let $\hat{\boldsymbol{\theta}}(\boldsymbol{y}) \in \Theta$ be a parameter value at which $L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|\boldsymbol{y})$ attains its maximum as a function of $\boldsymbol{\theta}$ with $\boldsymbol{y}$ held fixed. Then a maximum likelihood estimator (**MLE**) of the parameter $\boldsymbol{\theta}$ based on the sample $\boldsymbol{Y}$ is $\hat{\boldsymbol{\theta}}(\boldsymbol{Y})$.

The following remarks are important. I) It is crucial to observe that the likelihood function is a function of $\boldsymbol{\theta}$ (and that $y_1, ..., y_n$ act as fixed constants). Note that the pdf or pmf $f(\boldsymbol{y}|\boldsymbol{\theta})$ is a function of n variables while $L(\boldsymbol{\theta})$ is a function of k variables if $\boldsymbol{\theta}$ is a $1 \times k$ vector. Often $k = 1$ or $k = 2$ while n could be in the hundreds or thousands.

II) If $Y_1, ..., Y_n$ is an independent sample from a population with pdf or pmf $f(y|\boldsymbol{\theta})$, then the likelihood function

$$L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|y_1, ..., y_n) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}). \tag{1.49}$$

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(y_i|\boldsymbol{\theta})$$

if the Y_i are independent but have different pdfs or pmfs.

III) If the MLE $\hat{\boldsymbol{\theta}}$ exists, then $\hat{\boldsymbol{\theta}} \in \Theta$. Hence if $\hat{\boldsymbol{\theta}}$ is not in the parameter space Θ, then $\hat{\boldsymbol{\theta}}$ is not the MLE of $\boldsymbol{\theta}$.

Theorem 1.35: Invariance Principle. If $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, then $h(\hat{\boldsymbol{\theta}})$ is the MLE of $h(\boldsymbol{\theta})$ where h is a function with domain Θ.

Proof. When h is one to one, let $\boldsymbol{\eta} = h(\boldsymbol{\theta})$. Then the inverse function h^{-1} exists and $\boldsymbol{\theta} = h^{-1}(\boldsymbol{\eta})$. Hence

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = f(\boldsymbol{x}|h^{-1}(\boldsymbol{\eta})) \tag{1.50}$$

is the joint pdf or pmf of $\boldsymbol{x}$. So the likelihood function of $h(\boldsymbol{\theta}) = \boldsymbol{\eta}$ is

$$L^*(\boldsymbol{\eta}) \equiv K(\boldsymbol{\eta}) = L(h^{-1}(\boldsymbol{\eta})). \tag{1.51}$$

Also note that

$$\sup_{\boldsymbol{\eta}} K(\boldsymbol{\eta}|\boldsymbol{x}) = \sup_{\boldsymbol{\eta}} L(h^{-1}(\boldsymbol{\eta})|\boldsymbol{x}) = L(\hat{\boldsymbol{\theta}}|\boldsymbol{x}). \tag{1.52}$$

Thus

$$\hat{\boldsymbol{\eta}} = h(\hat{\boldsymbol{\theta}}) \tag{1.53}$$

is the MLE of $\boldsymbol{\eta} = h(\boldsymbol{\theta})$ when h is one to one.

If h is not one to one, then the new parameters $\boldsymbol{\eta} = h(\boldsymbol{\theta})$ do not give enough information to define $f(\boldsymbol{x}|\boldsymbol{\eta})$. Hence we cannot define the likelihood. That is, a $N(\mu, \sigma^2)$ density cannot be defined by the parameter μ/σ alone. Before concluding that the MLE does not exist if h is not one to one, note that if $X_1, ..., X_n$ are iid $N(\mu, \sigma^2)$ then $X_1, ..., X_n$ remain iid $N(\mu, \sigma^2)$ even though the investigator did not rename the parameters wisely or is interested in a function $h(\mu, \sigma) = \mu/\sigma$ that is not one to one. Berk (1967) said that if h is not one to one, define

$$w(\boldsymbol{\theta}) = (h(\boldsymbol{\theta}), u(\boldsymbol{\theta})) = (\boldsymbol{\eta}, \boldsymbol{\gamma}) = \boldsymbol{\xi} \tag{1.54}$$

such that $w(\boldsymbol{\theta})$ is one to one. Note that the choice

$$w(\boldsymbol{\theta}) = (h(\boldsymbol{\theta}), \boldsymbol{\theta})$$

works. In other words, we can always take u to be the identity function.

The choice of w is not unique, but the inverse function

$$w^{-1}(\boldsymbol{\xi}) = \boldsymbol{\theta}$$

is unique. Hence the likelihood is well defined, and $w(\hat{\boldsymbol{\theta}})$ is the MLE of $\boldsymbol{\xi}$. □

There are **four commonly used techniques** for finding the MLE.

- Potential candidates can be found by differentiating log $L(\boldsymbol{\theta})$, the log likelihood.
- Potential candidates can be found by differentiating the likelihood $L(\boldsymbol{\theta})$.
- The MLE can sometimes be found by direct maximization of the likelihood $L(\boldsymbol{\theta})$.
- **Invariance Principle**: If $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, then $h(\hat{\boldsymbol{\theta}})$ is the MLE of $h(\boldsymbol{\theta})$.

The method of moments is another useful way for obtaining point estimators. Let $Y_1, ..., Y_n$ be an iid sample and let

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i^j \text{ and } \mu_j \equiv \mu_j(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(Y^j) \tag{1.55}$$

for $j = 1, ..., k$. So $\hat{\mu}_j$ is the jth sample moment and μ_j is the jth population moment. Fix k and assume that $\mu_j = \mu_j(\theta_1, ..., \theta_k)$. Solve the system

$$\begin{aligned} \hat{\mu}_1 &\stackrel{\text{set}}{=} \mu_1(\theta_1, ..., \theta_k) \\ \vdots \quad & \quad \vdots \\ \hat{\mu}_k &\stackrel{\text{set}}{=} \mu_k(\theta_1, ..., \theta_k) \end{aligned}$$

for $\tilde{\boldsymbol{\theta}}$.

Definition 1.46. The solution $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, ..., \tilde{\theta}_k)^T$ is the **method of moments estimator** of $\boldsymbol{\theta}$. If g is a continuous function of the first k moments and $h(\boldsymbol{\theta}) = g(\mu_1(\boldsymbol{\theta}), ..., \mu_k(\boldsymbol{\theta}))$, then the method of moments estimator of $h(\boldsymbol{\theta})$ is

$$g(\hat{\mu}_1, ..., \hat{\mu}_k).$$

Definition 1.46 is similar to the invariance principle for the MLE, but note that g needs to be a continuous function, and the definition only applies to a function of $(\hat{\mu}_1, ..., \hat{\mu}_k)$ where $k \geq 1$. Hence $\overline{Y}$ is the method of moments estimator of the population mean μ, and $g(\overline{Y})$ is the method of moments estimator of $g(\mu)$ if g is a continuous function. Sometimes the notation $\hat{\theta}_{MLE}$ and $\hat{\theta}_{MM}$ will be used to denote the MLE and method of moments estimators of θ, respectively. As with maximum likelihood estimators, not all method of moments estimators exist in closed form, and then numerical techniques must be used.

1.8 Mixture Distributions

Mixture distributions are useful for model and variable selection since $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a mixture distribution of $\hat{\boldsymbol{\beta}}_{I_j,0}$, and the lasso estimator $\hat{\boldsymbol{\beta}}_L$ is a mixture distribution of $\hat{\boldsymbol{\beta}}_{L,\lambda_i}$ for $i = 1, ..., M$. See Chapter 6. A random vector $\boldsymbol{u}$ has a mixture distribution if $\boldsymbol{u}$ equals a random vector $\boldsymbol{u}_j$ with probability π_j for $j = 1, ..., J$. See Definition 1.29 for the population mean and population covariance matrix of a random vector.

Definition 1.47. The distribution of a $g \times 1$ random vector $\boldsymbol{u}$ is a mixture distribution if the cumulative distribution function (cdf) of $\boldsymbol{u}$ is

$$F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t}) \tag{1.56}$$

where the probabilities π_j satisfy $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{J} \pi_j = 1$, $J \geq 2$, and $F_{\boldsymbol{u}_j}(\boldsymbol{t})$ is the cdf of a $g \times 1$ random vector $\boldsymbol{u}_j$. Then $\boldsymbol{u}$ has a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j.

Theorem 1.36. Suppose $E(h(\boldsymbol{u}))$ and the $E(h(\boldsymbol{u}_j))$ exist. Then

$$E[h(\boldsymbol{u})] = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)]. \tag{1.57}$$

Hence

$$E(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j], \tag{1.58}$$

and $Cov(\boldsymbol{u}) = E(\boldsymbol{u}\boldsymbol{u}^T) - E(\boldsymbol{u})E(\boldsymbol{u}^T) = E(\boldsymbol{u}\boldsymbol{u}^T) - E(\boldsymbol{u})[E(\boldsymbol{u})]^T = \sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j \boldsymbol{u}_j^T] - E(\boldsymbol{u})[E(\boldsymbol{u})]^T =$

$$\sum_{j=1}^{J} \pi_j Cov(\boldsymbol{u}_j) + \sum_{j=1}^{J} \pi_j E(\boldsymbol{u}_j)[E(\boldsymbol{u}_j)]^T - E(\boldsymbol{u})[E(\boldsymbol{u})]^T. \tag{1.59}$$

If $E(\boldsymbol{u}_j) = \boldsymbol{\theta}$ for $j = 1, ..., J$, then $E(\boldsymbol{u}) = \boldsymbol{\theta}$ and

$$Cov(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j Cov(\boldsymbol{u}_j).$$

This theorem is easy to prove if the $\boldsymbol{u}_j$ are continuous random vectors with (joint) probability density functions (pdfs) $f_{\boldsymbol{u}_j}(\boldsymbol{t})$. Then $\boldsymbol{u}$ is a continuous random vector with pdf

$$f_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j f_{\boldsymbol{u}_j}(\boldsymbol{t}), \quad \text{and} \quad \mathrm{E[h(\boldsymbol{u})]} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathrm{h}(\boldsymbol{t}) \mathrm{f}_{\boldsymbol{u}}(\boldsymbol{t}) \mathrm{d}\boldsymbol{t}$$

$$= \sum_{j=1}^{J} \pi_j \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\boldsymbol{t}) f_{\boldsymbol{u}_j}(\boldsymbol{t}) d\boldsymbol{t} = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)]$$

where $E[h(\boldsymbol{u}_j)]$ is the expectation with respect to the random vector $\boldsymbol{u}_j$. Note that

$$E(\boldsymbol{u})[E(\boldsymbol{u})]^T = \sum_{j=1}^{J} \sum_{k=1}^{J} \pi_j \pi_k E(\boldsymbol{u}_j)[E(\boldsymbol{u}_k)]^T. \tag{1.60}$$

Alternatively, with respect to a Riemann Stieltjes integral, $E[h(\boldsymbol{u})] = \int h(\boldsymbol{t}) dF(\boldsymbol{t})$ provided the expected value exists, and the integral is a linear operator with respect to both h and F. Hence for a mixture distribution, $E[h(\boldsymbol{u})] = \int h(\boldsymbol{t}) dF(\boldsymbol{t}) =$

$$\int h(\boldsymbol{t})\ d\left[\sum_{j=1}^{J}\pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})\right] = \sum_{j=1}^{J}\pi_j \int h(\boldsymbol{t}) dF_{\boldsymbol{u}_j}(\boldsymbol{t}) = \sum_{j=1}^{J}\pi_j E[h(\boldsymbol{u}_j)].$$

1.9 Elliptically Contoured Distributions

Definition 1.48: Johnson (1987, pp. 107-108). A $p \times 1$ random vector $\boldsymbol{X}$ has an *elliptically contoured distribution,* also called an *elliptically symmetric distribution,* if $\boldsymbol{X}$ has joint pdf

$$f(\boldsymbol{z}) = k_p |\boldsymbol{\Sigma}|^{-1/2} g[(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})], \tag{1.61}$$

and we say $\boldsymbol{X}$ has an elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution.

If $\boldsymbol{X}$ has an elliptically contoured (EC) distribution, then the characteristic function of $\boldsymbol{X}$ is

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp(i\boldsymbol{t}^T\boldsymbol{\mu})\psi(\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}) \tag{1.62}$$

for some function ψ. If the second moments exist, then

$$E(\boldsymbol{X}) = \boldsymbol{\mu} \tag{1.63}$$

and

$$\text{Cov}(\boldsymbol{X}) = c_X \boldsymbol{\Sigma} \tag{1.64}$$

where

$$c_X = -2\psi'(0).$$

Definition 1.49. The *population squared Mahalanobis distance*

$$U \equiv D^2 = D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{X}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}). \tag{1.65}$$

For elliptically contoured distributions, U has pdf

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2-1} g(u). \tag{1.66}$$

For $c > 0$, an $EC_p(\boldsymbol{\mu}, c\boldsymbol{I}, g)$ distribution is *spherical about* $\boldsymbol{\mu}$ where $\boldsymbol{I}$ is the $p \times p$ identity matrix. The *multivariate normal distribution* $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has $k_p = (2\pi)^{-p/2}$, $\psi(u) = g(u) = \exp(-u/2)$, and $h(u)$ is the χ^2_p pdf.

1.10 Some Useful Distributions

Let the population quantile be y_δ. Then $P(Y \leq y_\delta) = \delta$ if Y has a pdf that is positive at y_δ.

Definition 1.50. The *gamma function* $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ for $x > 0$.

Some properties of the gamma function follow. i) $\Gamma(k) = (k-1)!$ for integer $k \geq 1$. ii) $\Gamma(x+1) = x\ \Gamma(x)$ for $x > 0$. iii) $\Gamma(x) = (x-1)\ \Gamma(x-1)$ for $x > 1$. iv) $\Gamma(0.5) = \sqrt{\pi}$.

1) $Y \sim \text{beta}(\delta, \nu)$

$$f(y) = \frac{\Gamma(\delta+\nu)}{\Gamma(\delta)\Gamma(\nu)} y^{\delta-1}(1-y)^{\nu-1}$$

where $\delta > 0$, $\nu > 0$ and $0 \leq y \leq 1$.

$$E(Y) = \frac{\delta}{\delta+\nu}, \quad V(Y) = \frac{\delta\nu}{(\delta+\nu)^2(\delta+\nu+1)}.$$

2) Bernoulli(ρ) = binomial$(k = 1, \rho)$ $\quad f(y) = \rho^y(1-\rho)^{1-y}$ for $y = 0, 1$. $E(Y) = \rho$, $\quad V(Y) = \rho(1-\rho)$.

$$m(t) = [(1-\rho) + \rho e^t], \quad c(t) = [(1-\rho) + \rho e^{it}].$$

3) binomial(k, ρ), $Y \sim BIN(k, \rho)$,

$$f(y) = \binom{k}{y} \rho^y (1-\rho)^{k-y}$$

for $y = 0, 1, \ldots, k$ where $0 < \rho < 1$. $E(Y) = k\rho$, $\quad V(Y) = k\rho(1-\rho)$. 1P-REF is k is known, and $I_1(\rho) = \dfrac{k}{\rho(1-\rho)}$. $m(t) = [(1-\rho)+\rho e^t]^k$, $\quad c(t) = [(1-\rho) + \rho e^{it}]^k$. If $Y_1, ..., Y_n$ are independent binomial BIN(k_i, ρ) random variables, then

$$\sum_{i=1}^n Y_i \sim \text{BIN}\left(\sum_{\text{i=1}}^{\text{n}} \text{k}_\text{i}, \rho\right).$$

Thus if $Y_1, ..., Y_n$ are iid BIN(k, ρ) random variables, then $\sum_{i=1}^n Y_i \sim$ BIN(nk, ρ).

4) $Y \sim \text{Cauchy}(\mu, \sigma)$,

$$f(y) = \frac{1}{\pi\sigma[1 + (\frac{y-\mu}{\sigma})^2]}$$

where y and μ are real numbers and $\sigma > 0$. $\quad E(Y) = \infty =$ VAR(Y). $E(Y)$ and $V(Y)$ do not exist. $\quad c(t) = \exp(it\mu - |t|\sigma)$.

$$F(y) = \frac{1}{\pi}[\arctan(\frac{y-\mu}{\sigma}) + \pi/2].$$

5) chi-square(p) = gamma($\nu = p/2, \lambda = 2$), $Y \sim \chi^2_p$,

$$f(y) = \frac{y^{\frac{p}{2}-1}e^{-\frac{y}{2}}}{2^{\frac{p}{2}}\Gamma(\frac{p}{2})}$$

where $y > 0$ and p is a positive integer. $E(Y) = p$, $V(Y) = 2p$.

$$m(t) = \left(\frac{1}{1-2t}\right)^{p/2} = (1-2t)^{-p/2} \text{ for } \mathrm{t} < 1/2, \quad \mathrm{c(t)} = \left(\frac{1}{1-\mathrm{i2t}}\right)^{\mathrm{p/2}}.$$

If $Y_1, ..., Y_n$ are independent chi–square $\chi^2_{p_i}$, then

$$\sum_{i=1}^{n} Y_i \sim \chi^2(\sum_{i=1}^{n} p_i).$$

Thus if $Y_1, ..., Y_n$ are iid χ^2_p, then

$$\sum_{i=1}^{n} Y_i \sim \chi^2_{np}.$$

6) exponential(λ)= gamma($\nu = 1, \lambda$), $Y \sim \mathrm{EXP}(\lambda)$

$$f(y) = \frac{1}{\lambda} \exp(-\frac{y}{\lambda})\ I(y \geq 0)$$

where $\lambda > 0$. $E(Y) = \lambda$, $V(Y) = \lambda^2$, and $y_\delta = -\lambda \ln(1-\delta)$. 1P-REF and $I_1(\lambda) = 1/\lambda^2$.

$$m(t) = 1/(1-\lambda t) \text{ for } \mathrm{t} < 1/\lambda, \quad \mathrm{c(t)} = 1/(1-\mathrm{i}\lambda\mathrm{t}).$$

$$F(y) = 1 - \exp(-y/\lambda),\ y \geq 0.$$

If $Y_1, ..., Y_n$ are iid exponential EXP(λ), then

$$\sum_{i=1}^{n} Y_i \sim G(n, \lambda).$$

7) gamma(ν, λ), $Y \sim G(\nu, \lambda)$,

$$f(y) = \frac{y^{\nu-1}e^{-y/\lambda}}{\lambda^\nu \Gamma(\nu)}$$

where ν, λ, and y are positive. $E(Y) = \nu\lambda$, $V(Y) = \nu\lambda^2$. 2P-REF and if ν is known, then $I_1(\lambda) = \nu/\lambda^2$.

$$m(t) = \left(\frac{1}{1-\lambda t}\right)^{\nu} \quad \text{for} \quad \mathrm{t} < 1/\lambda, \quad \mathrm{c(t)} = \left(\frac{1}{1-\mathrm{i}\lambda \mathrm{t}}\right)^{\nu}.$$

If $Y_1, ..., Y_n$ are independent Gamma $G(\nu_i, \lambda)$ then

$$\sum_{i=1}^{n} Y_i \sim G(\sum_{i=1}^{n} \nu_i, \lambda).$$

Thus if $Y_1, ..., Y_n$ are iid $G(\nu, \lambda)$, then $\sum_{i=1}^{n} Y_i \sim G(n\nu, \lambda)$.

8) $Y \sim N(\mu, \sigma^2)$

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)$$

where $\sigma > 0$ and μ and y are real. $E(Y) = \mu$, $V(Y) = \sigma^2$, and $y_\delta = \mu + \sigma z_\delta$. 2P-REF. If σ^2 is known, then $I_1(\mu) = 1/\sigma^2$. If μ is known, then $I_1(\sigma^2) = \frac{1}{2\sigma^4}$.

$$I_1(\mu, \sigma) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}, \quad I_1(\mu, \sigma^2) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}.$$

$$m(t) = \exp(t\mu + t^2\sigma^2/2), \quad c(t) = \exp(it\mu - t^2\sigma^2/2).$$

$$F(y) = \Phi\left(\frac{y-\mu}{\sigma}\right).$$

If $Y_1, ..., Y_n$ are independent normal $N(\mu_i, \sigma_i^2)$, then

$$\sum_{i=1}^{n}(a_i + b_i Y_i) \sim N(\sum_{i=1}^{n}(a_i + b_i\mu_i), \sum_{i=1}^{n} b_i^2\sigma_i^2).$$

Here a_i and b_i are fixed constants. Thus if $Y_1, ..., Y_n$ are iid $N(\mu, \sigma^2)$, then $\overline{Y} \sim N(\mu, \sigma^2/n)$.

9) Poisson(θ), $Y \sim$ POIS(θ)

$$f(y) = \frac{e^{-\theta}\theta^y}{y!}$$

for $y = 0, 1, \ldots$, where $\theta > 0$. $E(Y) = \theta = V(Y)$. 1P-REF and $I_1(\theta) = 1/\theta$.

$$m(t) = \exp(\theta(e^t - 1)), \quad c(t) = \exp(\theta(e^{it} - 1)).$$

If $Y_1, ..., Y_n$ are independent POIS(θ_i), then

$$\sum_{i=1}^{n} Y_i \sim \mathrm{POIS}(\sum_{\mathrm{i=1}}^{\mathrm{n}} \theta_{\mathrm{i}}).$$

Thus if $Y_1, ..., Y_n$ are iid $POIS(\theta)$, then

$$\sum_{i=1}^{n} Y_i \sim \text{POIS}(\text{n}\theta).$$

10) uniform(θ_1, θ_2), $Y \sim U(\theta_1, \theta_2)$.

$$f(y) = \frac{1}{\theta_2 - \theta_1} I(\theta_1 \leq y \leq \theta_2).$$

$F(y) = (y - \theta_1)/(\theta_2 - \theta_1)$ for $\theta_1 \leq y \leq \theta_2$. $E(Y) = (\theta_1 + \theta_2)/2$. $V(Y) = (\theta_2 - \theta_1)^2/12$, and $y_\delta = (\theta_2 - \theta_1)\delta + \theta_1$. By definition, $m(0) = c(0) = 1$. For $t \neq 0$,

$$m(t) = \frac{e^{t\theta_2} - e^{t\theta_1}}{(\theta_2 - \theta_1)t}, \text{ and } \text{c(t)} = \frac{\text{e}^{\text{it}\theta_2} - \text{e}^{\text{it}\theta_1}}{(\theta_2 - \theta_1)\text{it}}.$$

11) point mass at c: The distribution of Y is a point mass at c (or Y is degenerate at c) if $P(Y = c) = 1$ with pmf $f(c) = 1$. Hence $Y \sim N(c, 0)$, $E(Y) = c$, $V(Y) = 0$. $m(t) = e^{tc}$. $c(t) = e^{itc}$. The point mass at 0 has $m(t) \equiv 1$ and $c(t) \equiv 1$.

More Distributions:

12) If Y has a geometric distribution, $Y \sim \text{geom}(\rho)$ then the pmf of Y is

$$f(y) = P(Y = y) = \rho(1 - \rho)^y$$

for $y = 0, 1, 2, ...$ and $0 < \rho < 1$. $E(Y) = (1 - \rho)/\rho$. $V(Y) = (1 - \rho)/\rho^2$. $Y \sim NB(1, \rho)$. Hence the mgf of Y is

$$m(t) = \frac{\rho}{1 - (1 - \rho)e^t}$$

for $t < -\log(1 - \rho)$.

13) If Y has an inverse Gaussian distribution, $Y \sim \text{IG}(\theta, \lambda)$, then the pdf of Y is

$$f(y) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left[\frac{-\lambda(y - \theta)^2}{2\theta^2 y}\right]$$

where $y, \theta, \lambda > 0$. $E(Y) = \theta$ and

$$V(Y) = \frac{\theta^3}{\lambda}.$$

The mgf is

$$m(t) = \exp\left[\frac{\lambda}{\theta}\left(1 - \sqrt{1 - \frac{2\theta^2 t}{\lambda}}\right)\right] \quad t < \lambda/(2\theta^2), \quad c(t) = \exp\left[\frac{\lambda}{\theta}\left(1 - \sqrt{1 - \frac{2\theta^2 it}{\lambda}}\right)\right].$$

14) If Y has a negative binomial distribution, $Y \sim \text{NB}(\text{r}, \rho)$, then the pmf of Y is

$$f(y) = P(Y = y) = \binom{r+y-1}{y} \rho^r (1-\rho)^y$$

for $y = 0, 1, \ldots$ where $0 < \rho < 1$. $E(Y) = r(1-\rho)/\rho$, and

$$V(Y) = \frac{r(1-\rho)}{\rho^2}.$$

The moment generating function

$$m(t) = \left[\frac{\rho}{1-(1-\rho)e^t}\right]^r$$

for $t < -\log(1-\rho)$.

15) If Y has an F distribution, $Y \sim F(\nu_1, \nu_2)$, then the pdf of Y is

$$f(y) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{y^{(\nu_1-2)/2}}{\left(1+(\frac{\nu_1}{\nu_2})y\right)^{(\nu_1+\nu_2)/2}}$$

where $y > 0$ and ν_1 and ν_2 are positive integers.

$$E(Y) = \frac{\nu_2}{\nu_2 - 2}, \quad \nu_2 > 2$$

and

$$V(Y) = 2\left(\frac{\nu_2}{\nu_2-2}\right)^2 \frac{(\nu_1+\nu_2-2)}{\nu_1(\nu_2-4)}, \quad \nu_2 > 4.$$

16) If Y has a Student's t distribution, $Y \sim t_p$, then the pdf of Y is

$$f(y) = \frac{\Gamma(\frac{p+1}{2})}{(p\pi)^{1/2}\Gamma(p/2)}(1+\frac{y^2}{p})^{-(\frac{p+1}{2})}$$

where p is a positive integer and y is real. This family is symmetric about 0. The t_1 distribution is the Cauchy$(0, 1)$ distribution. If Z is $N(0, 1)$ and is independent of $W \sim \chi^2_p$, then

$$\frac{Z}{(\frac{W}{p})^{1/2}}$$

is t_p. $E(Y) = 0$ for $p \geq 2$. $V(Y) = p/(p-2)$ for $p \geq 3$.

Two Multivariate Distributions:

17) point mass at $\boldsymbol{c}$: The distribution of the $p \times 1$ random vector $\boldsymbol{Y}$ is a point mass at $\boldsymbol{c}$ (or $\boldsymbol{Y}$ is degenerate at $\boldsymbol{c}$) if $P(\boldsymbol{Y} = \boldsymbol{c}) = 1$ with pmf $f(\boldsymbol{c}) = 1$. Hence $\boldsymbol{Y} \sim N_p(\boldsymbol{c}, \boldsymbol{0})$, $E(\boldsymbol{Y}) = \boldsymbol{c}$, $\text{Cov}(\boldsymbol{Y}) = \boldsymbol{0}$, $m(\boldsymbol{t}) = e^{\boldsymbol{t}^T\boldsymbol{c}}$, $c(\boldsymbol{t}) = e^{i\boldsymbol{t}^T\boldsymbol{c}}$. The point mass at $\boldsymbol{0}$ has $m(\boldsymbol{t}) \equiv 1$ and $c(\boldsymbol{t}) \equiv 1$.

18) MVN distribution: If $\boldsymbol{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $E(\boldsymbol{Y}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{Y}) = \boldsymbol{\Sigma}$.

$$m(\boldsymbol{t}) = \exp\left(\boldsymbol{t}^T\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}\right), \quad c(\boldsymbol{t}) = \exp\left(i\boldsymbol{t}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}\right).$$

If $\boldsymbol{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and if $\boldsymbol{A}$ is a $q \times p$ matrix, then $\boldsymbol{A}\boldsymbol{Y} \sim N_q(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. If $\boldsymbol{a}$ is a $p \times 1$ vector of constants, then $\boldsymbol{Y} + \boldsymbol{a} \sim N_p(\boldsymbol{\mu} + \boldsymbol{a}, \boldsymbol{\Sigma})$.

$$\text{Let } \boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \text{and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

All subsets of a MVN are MVN: $(Y_{k_1}, ..., Y_{k_q})^T \sim N_q(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ where $\tilde{\boldsymbol{\mu}}_i = E(Y_{k_i})$ and $\tilde{\boldsymbol{\Sigma}}_{ij} = \text{Cov}(Y_{k_i}, Y_{k_j})$. In particular, $\boldsymbol{Y}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{Y}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$. If $\boldsymbol{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are independent iff $\boldsymbol{\Sigma}_{12} = \boldsymbol{0}$.

1.11 Summary

1) See Section 1.10 for some useful distributions.

1.12 Complements

The properties of the multivariate normal distribution and convergence in distribution to a multivariate normal distribution are rather similar, as will be shown in Chapter 3.

1.13 Problems

1.1*. Suppose that

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4\left(\begin{pmatrix} 49 \\ 100 \\ 17 \\ 7 \end{pmatrix}, \begin{pmatrix} 3 & 1 & -1 & 0 \\ 1 & 6 & 1 & -1 \\ -1 & 1 & 4 & 0 \\ 0 & -1 & 0 & 2 \end{pmatrix}\right).$$

a) Find the distribution of X_2.

b) Find the distribution of $(X_1, X_3)^T$.

c) Which pairs of random variables X_i and X_j are independent?

d) Find the correlation $\rho(X_1, X_3)$.

1.2*. Recall that if $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$.

Let $\sigma_{12} = \text{Cov}(Y, X)$ and suppose Y and X follow a bivariate normal distribution

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 49 \\ 100 \end{pmatrix}, \begin{pmatrix} 16 & \sigma_{12} \\ \sigma_{12} & 25 \end{pmatrix} \right).$$

a) If $\sigma_{12} = 0$, find $Y|X$. Explain your reasoning.

b) If $\sigma_{12} = 10$ find $E(Y|X)$.

c) If $\sigma_{12} = 10$, find $\text{Var}(Y|X)$.

1.3. Let $\sigma_{12} = \text{Cov}(Y, X)$ and suppose Y and X follow a bivariate normal distribution

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 15 \\ 20 \end{pmatrix}, \begin{pmatrix} 64 & \sigma_{12} \\ \sigma_{12} & 81 \end{pmatrix} \right).$$

a) If $\sigma_{12} = 10$ find $E(Y|X)$.

b) If $\sigma_{12} = 10$, find $\text{Var}(Y|X)$.

1.4. Suppose that

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4 \left(\begin{pmatrix} 3 \\ 4 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 3 & 2 & 1 & 1 \\ 2 & 4 & 1 & 0 \\ 1 & 1 & 2 & 0 \\ 1 & 0 & 0 & 3 \end{pmatrix} \right).$$

a) Find the distribution of $(X_1, X_3)^T$.

b) Which pairs of random variables X_i and X_j are independent?

c) Find the correlation $\rho(X_1, X_3)$.

Chapter 2
Univariate Limit Theorems

This chapter discusses the central limit theorem, the delta method, asymptotically efficient estimators, convergence in distribution and convergence in probability. This chapter follows Olive (2014, § 8.1-8.5) closely.

Large sample theory, also called asymptotic theory, is used to approximate the distribution of an estimator when the sample size n is large. This theory is extremely useful if the exact sampling distribution of the estimator is complicated or unknown. To use this theory, one must determine what the estimator is estimating, the rate of convergence, the asymptotic distribution, and how large n must be for the approximation to be useful. Moreover, the (asymptotic) standard error (SE), an estimator of the asymptotic standard deviation, must be computable if the estimator is to be useful for inference.

2.1 The CLT and Delta Method

The CLT is also known as the Lindeberg-Lévy CLT.

Theorem 2.1: the Central Limit Theorem (CLT). Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and $V(Y) = \sigma^2$. Let the sample mean $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Then

$$\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Note that the sample mean is estimating the *population mean* μ with a $\sqrt{n}$ convergence rate, the asymptotic distribution is normal, and the SE $= S/\sqrt{n}$ where S is the *sample standard deviation.* For many distributions the central limit theorem provides a good approximation if the sample size $n > 30$. A special case of the CLT is proven at the end of Section 2.4.

Notation. The notation $X \sim Y$ and $X \stackrel{D}{=} Y$ both mean that the random variables X and Y have the same distribution. See Definition 1.15. The notation $Y_n \xrightarrow{D} X$ means that for large n we can approximate the cdf of Y_n by

the cdf of X. See Section 2.3 for more on convergence in distribution. The distribution of X is the limiting distribution or asymptotic distribution of Y_n, and the limiting distribution does not depend on n. For the CLT, notice that

$$Z_n = \sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) = \left(\frac{\overline{Y}_n - \mu}{\sigma/\sqrt{n}}\right) = \left(\frac{\sum_{i=1}^n Y_i - n\mu}{\sqrt{n}\sigma}\right)$$

is the z–score of $\overline{Y}$ and the z–score of $\sum_{i=1}^n Y_i$. Then $Z_n \xrightarrow{D} N(0,1)$. If $Z_n \xrightarrow{D} N(0,1)$, then the notation $Z_n \approx N(0,1)$, also written as $Z_n \sim AN(0,1)$, means approximate the cdf of Z_n by the standard normal cdf. Similarly, the notation

$$\overline{Y}_n \approx N(\mu, \sigma^2/n),$$

also written as $\overline{Y}_n \sim AN(\mu, \sigma^2/n)$, means approximate the cdf of $\overline{Y}_n$ as if $\overline{Y}_n \sim N(\mu, \sigma^2/n)$. Note that the approximate distribution, unlike the limiting distribution, does depend on n. The standard error $S/\sqrt{n}$ approximates the asymptotic standard deviation $\sqrt{\sigma^2/n}$ of $\overline{Y}$.

The two main applications of the CLT are to give the limiting distribution of $\sqrt{n}(\overline{Y}_n - \mu)$ and the limiting distribution of $\sqrt{n}(Y_n/n - \mu_X)$ for a random variable Y_n such that $Y_n = \sum_{i=1}^n X_i$ where the X_i are iid with $E(X) = \mu_X$ and $V(X) = \sigma_X^2$.

Several of the random variables in Theorems 1.24 and 1.25 can be approximated in this way. The CLT says that $\overline{Y}_n \sim AN(\mu, \sigma^2/n)$. The delta method says that if $T_n \sim AN(\theta, \sigma^2/n)$, and if $g'(\theta) \neq 0$, then $g(T_n) \sim AN(g(\theta), \sigma^2[g'(\theta)]^2/n)$. Hence a smooth function $g(T_n)$ of a well behaved statistic T_n tends to be well behaved (asymptotically normal with a $\sqrt{n}$ convergence rate).

Given iid data from some distribution, a common homework problem is to find the limiting distribution of $\sqrt{n}(\overline{Y}_n - \mu)$ using the CLT. You may need to find $E(Y)$, $E(Y^2)$, and $V(Y) = E(Y^2) - [E(Y)]^2$. A variant of this problem gives a formula for $E(Y^r)$. Then find $E(Y) = E(Y^1)$ with $r = 1$ and $E(Y^2)$ with $r = 2$.

Example 2.1. a) Let $Y_1, ..., Y_n$ be iid Ber(ρ). Then $E(Y) = \rho$ and $V(Y) = \rho(1-\rho)$. Hence

$$\sqrt{n}(\overline{Y}_n - \rho) \xrightarrow{D} N(0, \rho(1-\rho))$$

by the CLT.

b) Now suppose that $Y_n \sim BIN(n, \rho)$. Then $Y_n \stackrel{D}{=} \sum_{i=1}^n X_i$ where $X_1, ..., X_n$ are iid Ber(ρ). Hence

$$\sqrt{n}\left(\frac{Y_n}{n} - \rho\right) \xrightarrow{D} N(0, \rho(1-\rho))$$

since

$$\sqrt{n}\left(\frac{Y_n}{n} - \rho\right) \stackrel{D}{=} \sqrt{n}(\overline{X}_n - \rho) \stackrel{D}{\to} N(0, \rho(1-\rho))$$

by a).

c) Now suppose that $Y_n \sim BIN(k_n, \rho)$ where $k_n \to \infty$ as $n \to \infty$. Then

$$\sqrt{k_n}\left(\frac{Y_n}{k_n} - \rho\right) \approx N(0, \rho(1-\rho))$$

or

$$\frac{Y_n}{k_n} \approx N\left(\rho, \frac{\rho(1-\rho)}{k_n}\right) \text{ or } \mathrm{Y_n} \approx \mathrm{N}\left(\mathrm{k_n}\rho, \mathrm{k_n}\rho(1-\rho)\right).$$

Theorem 2.2: the Delta Method. If $g'(\theta) \neq 0$, and

$$\sqrt{n}(T_n - \theta) \stackrel{D}{\to} N(0, \sigma^2),$$

then

$$\sqrt{n}(g(T_n) - g(\theta)) \stackrel{D}{\to} N(0, \sigma^2[g'(\theta)]^2).$$

Example 2.2. Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and $V(Y) = \sigma^2$. Then by the CLT,

$$\sqrt{n}(\overline{Y}_n - \mu) \stackrel{D}{\to} N(0, \sigma^2).$$

Let $g(\mu) = \mu^2$. Then $g'(\mu) = 2\mu \neq 0$ for $\mu \neq 0$. Hence

$$\sqrt{n}((\overline{Y}_n)^2 - \mu^2) \stackrel{D}{\to} N(0, 4\sigma^2\mu^2)$$

for $\mu \neq 0$ by the delta method.

Example 2.3. Let $X \sim$ Binomial(n, p) where the positive integer n is large and $0 < p < 1$. Find the limiting distribution of $\sqrt{n}\left[\left(\frac{X}{n}\right)^2 - p^2\right]$.

Solution. Example 2.1b gives the limiting distribution of $\sqrt{n}(\frac{X}{n} - p)$. Let $g(p) = p^2$. Then $g'(p) = 2p$ and by the delta method,

$$\sqrt{n}\left[\left(\frac{X}{n}\right)^2 - p^2\right] = \sqrt{n}\left(g\left(\frac{X}{n}\right) - g(p)\right) \stackrel{D}{\to}$$

$$N(0, p(1-p)(g'(p))^2) = N(0, p(1-p)4p^2) = N(0, 4p^3(1-p)).$$

Example 2.4. Let $X_n \sim$ Poisson$(n\lambda)$ where the positive integer n is large and $0 < \lambda$.

a) Find the limiting distribution of $\sqrt{n}\left(\frac{X_n}{n} - \lambda\right)$.

b) Find the limiting distribution of $\sqrt{n}\ \left[\ \sqrt{\frac{X_n}{n}} - \ \sqrt{\lambda}\ \right]$.

Solution. a) $X_n \stackrel{D}{=} \sum_{i=1}^{n} Y_i$ where the Y_i are iid Poisson(λ). Hence $E(Y) = \lambda = V(Y)$. Thus by the CLT,

$$\sqrt{n}\ \left(\ \frac{X_n}{n} - \lambda\ \right) \stackrel{D}{=} \sqrt{n}\ \left(\ \frac{\sum_{i=1}^{n} Y_i}{n} - \lambda\ \right) \stackrel{D}{\to} N(0, \lambda).$$

b) Let $g(\lambda) = \sqrt{\lambda}$. Then $g'(\lambda) = \frac{1}{2\sqrt{\lambda}}$ and by the delta method,

$$\sqrt{n}\ \left[\ \sqrt{\frac{X_n}{n}} - \ \sqrt{\lambda}\ \right] = \sqrt{n}\ \left(\ g\left(\frac{X_n}{n}\right) - g(\lambda)\ \right) \stackrel{D}{\to}$$

$$N(0, \lambda\ (g'(\lambda))^2) = N\left(0, \lambda \frac{1}{4\lambda}\right) = N\left(0, \frac{1}{4}\right).$$

Example 2.5. Let $Y_1, ..., Y_n$ be independent and identically distributed (iid) from a Gamma(α, β) distribution.

a) Find the limiting distribution of $\sqrt{n}\ \left(\ \overline{Y} - \ \alpha\beta\ \right)$.

b) Find the limiting distribution of $\sqrt{n}\ \left(\ (\overline{Y})^2 - \ c\ \right)$ for appropriate constant c.

Solution: a) Since $E(Y) = \alpha\beta$ and $V(Y) = \alpha\beta^2$, by the CLT $\sqrt{n}\ \left(\ \overline{Y} - \ \alpha\beta\ \right) \stackrel{D}{\to} N(0, \alpha\beta^2)$.

b) Let $\mu = \alpha\beta$ and $\sigma^2 = \alpha\beta^2$. Let $g(\mu) = \mu^2$ so $g'(\mu) = 2\mu$ and $[g'(\mu)]^2 = 4\mu^2 = 4\alpha^2\beta^2$. Then by the delta method, $\sqrt{n}\ \left(\ (\overline{Y})^2 - \ c\ \right) \stackrel{D}{\to} N(0, \sigma^2[g'(\mu)]^2) = N(0, 4\alpha^3\beta^4)$ where $c = \mu^2 = \alpha^2\beta^2$.

Remark 2.1. a) Note that if $\sqrt{n}(T_n - k) \stackrel{D}{\to} N(0, \sigma^2)$, then evaluate the derivative at k. Thus use $g'(k)$ where $k = \alpha\beta$ in the above example. A common error occurs when k is a simple function of θ, for example $k = \theta/2$ with $g(\mu) = \mu^2$. Thus $g'(\mu) = 2\mu$ so $g'(\theta/2) = 2\theta/2 = \theta$. Then the common delta method error is to plug in $g'(\theta) = 2\theta$ instead of $g'(k) = \theta$. See Problems 2.3, 2.33, 2.35, 2.36, and 2.37.

b) For the delta method, also note that the function g can not depend on n since then there would be a sequence of functions g_n rather than one function g. This fact also applies to several other theorems in this chapter.

The following extension of the delta method is sometimes useful.

Theorem 2.3: the Second Order Delta Method. Suppose that $g'(\theta) = 0$, $g''(\theta) \neq 0$ and

$$\sqrt{n}(T_n - \theta) \stackrel{D}{\to} N(0, \tau^2(\theta)).$$

Then

$$n[g(T_n) - g(\theta)] \xrightarrow{D} \frac{1}{2}\tau^2(\theta)g''(\theta)\chi_1^2.$$

Example 2.6. Let $X_n \sim \text{Binomial}(n, p)$ where the positive integer n is large and $0 < p < 1$. Let $g(\theta) = \theta^3 - \theta$. Find the limiting distribution of $n\left[g\left(\frac{X_n}{n}\right) - c \right]$ for appropriate constant c when $p = \frac{1}{\sqrt{3}}$.

Solution: Since $X_n \stackrel{D}{=} \sum_{i=1}^n Y_i$ where $Y_i \sim BIN(1, p)$,

$$\sqrt{n}\left(\frac{X_n}{n} - p \right) \xrightarrow{D} N(0, p(1-p))$$

by the CLT. Let $\theta = p$. Then $g'(\theta) = 3\theta^2 - 1$ and $g''(\theta) = 6\theta$. Notice that

$$g(1/\sqrt{3}) = (1/\sqrt{3})^3 - 1/\sqrt{3} = (1/\sqrt{3})(\frac{1}{3} - 1) = \frac{-2}{3\sqrt{3}} = c.$$

Also $g'(1/\sqrt{3}) = 0$ and $g''(1/\sqrt{3}) = 6/\sqrt{3}$. Since $\tau^2(p) = p(1-p)$,

$$\tau^2(1/\sqrt{3}) = \frac{1}{\sqrt{3}}(1 - \frac{1}{\sqrt{3}}).$$

Hence

$$n\left[g\left(\frac{X_n}{n}\right) - \left(\frac{-2}{3\sqrt{3}}\right) \right] \xrightarrow{D} \frac{1}{2}\frac{1}{\sqrt{3}}(1 - \frac{1}{\sqrt{3}})\frac{6}{\sqrt{3}}\ \chi_1^2 = (1 - \frac{1}{\sqrt{3}})\ \chi_1^2.$$

Barndorff–Nielsen (1982), Casella and Berger (2002, p. 472, 515), Cox and Hinckley (1974, p. 286), Lehmann and Casella (2003, Section 6.3), Schervish (1995, p. 418), and many others suggest that under regularity conditions if $Y_1, ..., Y_n$ are iid from a one parameter regular exponential family, and if $\hat{\theta}$ is the MLE (maximum likelihood estimator) of θ, then

$$\sqrt{n}(\tau(\hat{\theta}) - \tau(\theta)) \xrightarrow{D} N\left(0, \frac{[\tau'(\theta)]^2}{I_1(\theta)}\right) = N[0, FCRLB_1(\tau(\theta))] \tag{2.1}$$

where the Fréchet Cramér Rao lower bound for $\tau(\theta)$ is

$$FCRLB_1(\tau(\theta)) = \frac{[\tau'(\theta)]^2}{I_1(\theta)}$$

and the Fisher information based on a sample of size one is

$$I_1(\theta) = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log(f(X|\theta))\right].$$

Hence $\tau(\hat{\theta}) \sim AN[\tau(\theta), FCRLB_n(\tau(\theta))]$ where $FCRLB_n(\tau(\theta)) = FCRLB_1(\tau(\theta))/n$. Notice that if

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N\left(0, \frac{1}{I_1(\theta)}\right),$$

then (2.1) follows by the delta method. Also recall that $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$ by the invariance principle and that

$$I_1(\tau(\theta)) = \frac{I_1(\theta)}{[\tau'(\theta)]^2}$$

if $\tau'(\theta) \neq 0$ by Definition 1.43.

For a 1P–REF, $\overline{T}_n = \frac{1}{n}\sum_{i=1}^{n} t(Y_i)$ is the UMVUE (uniformly minimum variance unbiased estimator) and generally the MLE of its expectation $\mu_t \equiv \mu_T = E_\theta(\overline{T}_n) = E_\theta[t(Y)]$. Let $\sigma_t^2 = V_\theta[t(Y)]$. These values can be found by using the distribution of $t(Y)$.

Theorem 2.4. Suppose Y is a 1P–REF with pdf or pmf

$$f(y|\theta) = h(y)c(\theta)\exp[w(\theta)t(y)]$$

and natural parameterization

$$f(y|\eta) = h(y)b(\eta)\exp[\eta t(y)].$$

Then a)

$$\mu_t = E[t(Y)] = \frac{-c'(\theta)}{c(\theta)w'(\theta)} = \frac{-\partial}{\partial\eta}\log(b(\eta)), \tag{2.2}$$

and b)

$$\sigma_t^2 = V[t(Y)] = \frac{\frac{-\partial^2}{\partial\theta^2}\log(c(\theta)) - [w''(\theta)]\mu_t}{[w'(\theta)]^2} = \frac{-\partial^2}{\partial\eta^2}\log(b(\eta)). \tag{2.3}$$

Proof. The proof will be for pdfs. For pmfs replace the integrals by sums.

By Theorem 1.31, only the middle equalities need to be shown. By Remark 1.9 the derivative and integral operators can be interchanged for a 1P–REF.

a) Since $1 = \int f(y|\theta)dy$,

$$0 = \frac{\partial}{\partial\theta}1 = \frac{\partial}{\partial\theta}\int h(y)\exp[w(\theta)t(y) \ + \ \log(c(\theta))]dy$$

$$= \int h(y)\frac{\partial}{\partial\theta}\exp[w(\theta)t(y) \ + \ \log(c(\theta))]dy$$

$$= \int h(y) \exp[w(\theta)t(y) \ + \ \log(c(\theta))] \left(w'(\theta)t(y) + \frac{c'(\theta)}{c(\theta)} \right) dy$$

or

$$E[w'(\theta)t(Y)] = \frac{-c'(\theta)}{c(\theta)}$$

or

$$E[t(Y)] = \frac{-c'(\theta)}{c(\theta)w'(\theta)}.$$

b) Similarly,

$$0 = \int h(y) \frac{\partial^2}{\partial \theta^2} \exp[w(\theta)t(y) \ + \ \log(c(\theta))] dy.$$

From the proof of a) and since $\frac{\partial}{\partial \theta} \log(c(\theta)) = c'(\theta)/c(\theta)$,

$$0 = \int h(y) \frac{\partial}{\partial \theta} \left[\exp[w(\theta)t(y) \ + \ \log(c(\theta))] \left(w'(\theta)t(y) + \frac{\partial}{\partial \theta} \log(c(\theta)) \right) \right] dy$$

$$= \int h(y) \exp[w(\theta)t(y) \ + \ \log(c(\theta))] \ \left(w'(\theta)t(y) + \frac{\partial}{\partial \theta} \log(c(\theta)) \right)^2 dy$$

$$+ \int h(y) \exp[w(\theta)t(y) \ + \ \log(c(\theta))] \ \left(w''(\theta)t(y) + \frac{\partial^2}{\partial \theta^2} \log(c(\theta)) \right) dy.$$

So

$$E \left(w'(\theta)t(Y) + \frac{\partial}{\partial \theta} \log(c(\theta)) \right)^2 = -E \left(w''(\theta)t(Y) + \frac{\partial^2}{\partial \theta^2} \log(c(\theta)) \right). \quad (2.4)$$

Using a) shows that the left hand side of (2.4) equals

$$E \left(w'(\theta) \left(t(Y) + \frac{c'(\theta)}{c(\theta)w'(\theta)} \right) \right)^2 = [w'(\theta)]^2 \ V(t(Y))$$

while the right hand side of (2.4) equals

$$- \left(w''(\theta)\mu_t + \frac{\partial^2}{\partial \theta^2} \log(c(\theta)) \right)$$

and the result follows. □

The simplicity of the following Olive (2014, p. 221) result is rather surprising. When (as is usually the case) $\frac{1}{n} \sum_{i=1}^n t(Y_i)$ is the MLE of μ_t, $\hat{\eta} = g^{-1}(\frac{1}{n} \sum_{i=1}^n t(Y_i))$ is the MLE of η by the invariance principle.

Theorem 2.5. Let $Y_1, ..., Y_n$ be iid from a 1P–REF with pdf or pmf

$$f(y|\theta) = h(y)c(\theta)\exp[w(\theta)t(y)]$$

and natural parameterization

$$f(y|\eta) = h(y)b(\eta)\exp[\eta t(y)].$$

Let

$$E(t(Y)) = \mu_t \equiv g(\eta)$$

and $V(t(Y)) = \sigma_t^2$.

a) Then

$$\sqrt{n}[\frac{1}{n}\sum_{i=1}^{n} t(Y_i) - \mu_t] \xrightarrow{D} N(0, I_1(\eta))$$

where

$$I_1(\eta) = \sigma_t^2 = g'(\eta) = \frac{[g'(\eta)]^2}{I_1(\eta)}.$$

b) If $\eta = g^{-1}(\mu_t)$, $\hat{\eta} = g^{-1}(\frac{1}{n}\sum_{i=1}^{n} t(Y_i))$, and $g^{-1'}(\mu_t) \neq 0$ exists, then

$$\sqrt{n}[\hat{\eta} - \eta] \xrightarrow{D} N\left(0, \frac{1}{I_1(\eta)}\right).$$

c) Suppose the conditions in b) hold. If $\theta = w^{-1}(\eta)$, $\hat{\theta} = w^{-1}(\hat{\eta})$, $w^{-1'}$ exists and is continuous, and $w^{-1'}(\eta) \neq 0$, then

$$\sqrt{n}[\hat{\theta} - \theta] \xrightarrow{D} N\left(0, \frac{1}{I_1(\theta)}\right).$$

d) If the conditions in c) hold, if τ' is continuous and if $\tau'(\theta) \neq 0$, then

$$\sqrt{n}[\tau(\hat{\theta}) - \tau(\theta)] \xrightarrow{D} N\left(0, \frac{[\tau'(\theta)]^2}{I_1(\theta)}\right).$$

Proof: a) The result follows by the central limit theorem if $V(t(Y)) = \sigma_t^2 = I_1(\eta) = g'(\eta)$. Since $\log(f(y|\eta)) = \log(h(y)) + \log(b(\eta)) + \eta t(y)$,

$$\frac{\partial}{\partial\eta}\log(f(y|\eta)) = \frac{\partial}{\partial\eta}\log(b(\eta)) + t(y) = -\mu_t + t(y) = -g(\eta) + t(y)$$

by Theorem 2.4 a). Hence

$$\frac{\partial^2}{\partial\eta^2}\log(f(y|\eta)) = \frac{\partial^2}{\partial\eta^2}\log(b(\eta)) = -g'(\eta),$$

and thus by Theorem 2.4 b)

$$I_1(\eta) = \frac{-\partial^2}{\partial \eta^2} \log(b(\eta)) = \sigma_t^2 = g'(\eta).$$

b) By the delta method,

$$\sqrt{n}(\hat{\eta} - \eta) \xrightarrow{D} N(0, \sigma_t^2 [g^{-1'}(\mu_t)]^2),$$

but

$$g^{-1'}(\mu_t) = \frac{1}{g'(g^{-1}(\mu_t))} = \frac{1}{g'(\eta)}.$$

Since $\sigma_t^2 = I_1(\eta) = g'(\eta)$, it follows that $\sigma_t^2 = [g'(\eta)]^2 / I_1(\eta)$, and

$$\sigma_t^2 [g^{-1'}(\mu_t)]^2 = \frac{[g'(\eta)]^2}{I_1(\eta)} \frac{1}{[g'(\eta)]^2} = \frac{1}{I_1(\eta)}.$$

So

$$\sqrt{n}(\hat{\eta} - \eta) \xrightarrow{D} N\left(0, \frac{1}{I_1(\eta)}\right).$$

c) By the delta method,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N\left(0, \frac{[w^{-1'}(\eta)]^2}{I_1(\eta)}\right),$$

but

$$\frac{[w^{-1'}(\eta)]^2}{I_1(\eta)} = \frac{1}{I_1(\theta)}.$$

The last equality holds since by Theorem 1.33c, if $\theta = g(\eta)$, if g' exists and is continuous, and if $g'(\theta) \neq 0$, then $I_1(\theta) = I_1(\eta)/[g'(\eta)]^2$. Use $\eta = w(\theta)$ so $\theta = g(\eta) = w^{-1}(\eta)$.

d) The result follows by the delta method. □

Remark 2.2. Following DasGupta (2008, p. 241-242), let $\psi(\eta) = -\log(b(\eta))$. Then $E_\eta[t(Y_1)] = \mu_t = \psi'(\eta) = g(\eta)$ by Theorem 2.4a, and the MLE $\hat{\eta}$ is the solution of $\frac{1}{n}\sum_{i=1}^n t(y_i) \stackrel{set}{=} E_\eta[t(Y_1)] = g(\eta)$ if the MLE exists. Now $g(\eta) = E_\eta[t(Y_1)]$ is an increasing function of η since $g'(\eta) = \psi''(\eta) = V_\eta(t(Y)) > 0$ (1P–REFs do not contain degenerate distributions). So for large n, with probability tending to one, the MLE $\hat{\eta}$ exists and $\hat{\eta} = g^{-1}(\frac{1}{n}\sum_{i=1}^n t(Y_i))$. Since $g'(\eta)$ exists, $g(\eta)$ and $g^{-1}(\eta)$ are continuous and the delta method can be applied to $\hat{\eta}$ as in Theorem 2.5b. By the proof of Theorem 2.5a), $\psi''(\eta) = I_1(\eta)$. Notice that if $\hat{\eta}$ is the MLE of η, then $\frac{1}{n}\sum_{i=1}^n t(Y_i)$ is the MLE of $\mu_t = E[t(Y_1)]$ by invariance. Hence if n is large enough, Theorem 2.5ab is for the MLE of $E[t(Y_1)]$ and the MLE of η.

2.2 Asymptotically Efficient Estimators

Definition 2.1. Let $Y_1, ..., Y_n$ be iid random variables. Let $T_n \equiv T_n(Y_1, ..., Y_n)$ be an estimator of a parameter μ_T such that

$$\sqrt{n}(T_n - \mu_T) \xrightarrow{D} N(0, \sigma_A^2).$$

Then the *asymptotic variance* of $\sqrt{n}(T_n - \mu_T)$ is σ_A^2 and the *asymptotic variance* (AV) of T_n is σ_A^2/n. If S_A^2 is a consistent estimator of σ_A^2, then the (asymptotic) *standard error* (SE) of T_n is $S_A/\sqrt{n}$. If $Y_1, ..., Y_n$ are iid with cdf F, then $\sigma_A^2 \equiv \sigma_A^2(F)$ depends on F.

Remark 2.3. Consistent estimators are defined in the following section. The parameter σ_A^2 is a function of both the estimator T_n and the underlying distribution F of Y_1. Frequently $nV(T_n)$ converges in distribution to σ_A^2, but not always. See Staudte and Sheather (1990, p. 51) and Lehmann (1999, p. 232).

Example 2.7. If $Y_1, ..., Y_n$ are iid from a distribution with mean μ and variance σ^2, then by the central limit theorem,

$$\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Recall that $V(\overline{Y}_n) = \sigma^2/n = AV(\overline{Y}_n)$ and that the standard error $SE(\overline{Y}_n) = S_n/\sqrt{n}$ where S_n^2 is the sample variance. Note that $\sigma_A^2(F) = \sigma^2$. If F is a $N(\mu, 1)$ cdf then $\sigma_A^2(F) = 1$, but if F is the $G(\nu = 7, \lambda = 1)$ cdf then $\sigma_A^2(F) = 7$.

Definition 2.2. Let $T_{1,n}$ and $T_{2,n}$ be two estimators of a parameter θ such that

$$n^{\delta}(T_{1,n} - \theta) \xrightarrow{D} N(0, \sigma_1^2(F))$$

and

$$n^{\delta}(T_{2,n} - \theta) \xrightarrow{D} N(0, \sigma_2^2(F)),$$

then the **asymptotic relative efficiency** of $T_{1,n}$ with respect to $T_{2,n}$ is

$$ARE(T_{1,n}, T_{2,n}) = \frac{\sigma_2^2(F)}{\sigma_1^2(F)}.$$

This definition brings up several issues. First, both estimators must have the same convergence rate n^{δ}. Usually $\delta = 0.5$. If $T_{i,n}$ has convergence rate n^{δ_i}, then estimator $T_{1,n}$ is judged to be "better" than $T_{2,n}$ if $\delta_1 > \delta_2$. Secondly, the two estimators need to estimate the same parameter θ. This condition will often not hold unless the distribution is symmetric about μ. Then $\theta = \mu$ is a natural choice. Thirdly, estimators are often judged by their Gaussian efficiency with respect to the sample mean (thus F is the normal distribution).

Since the normal distribution is a location–scale family, it is often enough to compute the ARE for the standard normal distribution. If the data come from a distribution F and the ARE can be computed, then $T_{1,n}$ is judged to be a "better" estimator (for the data distribution F) than $T_{2,n}$ if the $ARE > 1$. Similarly, $T_{1,n}$ is judged to be a "worse" estimator than $T_{2,n}$ if the $ARE < 1$. *Notice that the "better" estimator has the smaller asymptotic variance.*

The *population median* is any value MED(Y) such that

$$P(Y \leq \text{MED}(Y)) \geq 0.5 \text{ and } P(Y \geq \text{MED}(Y)) \geq 0.5. \tag{2.5}$$

In simulation studies, typically the underlying distribution F belongs to a symmetric location–scale family. There are at least two reasons for using such distributions. First, if the distribution is symmetric, then the population median MED(Y) is the point of symmetry and the natural parameter to estimate. Under the symmetry assumption, there are many estimators of MED(Y) that can be compared via their ARE with respect to the sample mean or the maximum likelihood estimator (MLE). Secondly, once the ARE is obtained for one member of the family, it is typically obtained for *all members of the location–scale family.* That is, suppose that $Y_1, ..., Y_n$ are iid from a location–scale family with parameters μ and σ. Then $Y_i = \mu + \sigma Z_i$ where the Z_i are iid from the same family with $\mu = 0$ and $\sigma = 1$. Typically

$$AV[T_{i,n}(\boldsymbol{Y})] = \sigma^2 AV[T_{i,n}(\boldsymbol{Z})],$$

so

$$ARE[T_{1,n}(\boldsymbol{Y}), T_{2,n}(\boldsymbol{Y})] = ARE[T_{1,n}(\boldsymbol{Z}), T_{2,n}(\boldsymbol{Z})].$$

Theorem 2.6. Let $Y_1, ..., Y_n$ be iid with a pdf f that is positive at the population median: $f(\text{MED}(Y)) > 0$. Then

$$\sqrt{n}(MED(n) - MED(Y)) \xrightarrow{D} N\left(0, \frac{1}{4[f(MED(Y))]^2}\right).$$

Example 2.8. Let $Y_1, ..., Y_n$ be iid $N(\mu, \sigma^2)$, $T_{1,n} = \overline{Y}$ and let $T_{2,n} = \text{MED}(n)$ be the sample median. Let $\theta = \mu = E(Y) = \text{MED}(Y)$. Find $ARE(T_{1,n}, T_{2,n})$.

Solution: By the CLT, $\sigma_1^2(F) = \sigma^2$ when F is the $N(\mu, \sigma^2)$ distribution. By Theorem 2.6,

$$\sigma_2^2(F) = \frac{1}{4[f(MED(Y))]^2} = \frac{1}{4[\frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-0}{2\sigma^2})]^2} = \frac{\pi\sigma^2}{2}.$$

Hence

$$ARE(T_{1,n}, T_{2,n}) = \frac{\pi\sigma^2/2}{\sigma^2} = \frac{\pi}{2} \approx 1.571$$

and the sample mean $\overline{Y}$ is a "better" estimator of μ than the sample median MED(n) for the family of normal distributions.

Recall from Definition 1.43 that $I_1(\theta)$ is the information number for θ based on a sample of size 1. Also recall that $I_1(\tau(\theta)) = I_1(\theta)/[\tau'(\theta)]^2 = 1/FCRLB_1[\tau(\theta)]$. See Definition 1.44.

The following definition says that if T_n is an asymptotically efficient estimator of $\tau(\theta)$, then

$$T_n \sim AN[\tau(\theta), FCRLB_n(\tau(\theta))].$$

Definition 2.3. Assume $\tau'(\theta) \neq 0$. Then an estimator T_n of $\tau(\theta)$ is **asymptotically efficient** if

$$\sqrt{n}(T_n - \tau(\theta)) \xrightarrow{D} N\left(0, \frac{[\tau'(\theta)]^2}{I_1(\theta)}\right) \sim N(0, FCRLB_1[\tau(\theta)]). \tag{2.6}$$

In particular, the estimator T_n of θ is asymptotically efficient if

$$\sqrt{n}(T_n - \theta) \xrightarrow{D} N\left(0, \frac{1}{I_1(\theta)}\right) \sim N(0, FCRLB_1[\theta]). \tag{2.7}$$

Following Lehmann (1999, p. 486), if $T_{2,n}$ is an asymptotically efficient estimator of θ, if $I_1(\theta)$ and $v(\theta)$ are continuous functions, and if $T_{1,n}$ is an estimator such that

$$\sqrt{n}(T_{1,n} - \theta) \xrightarrow{D} N(0, v(\theta)),$$

then under regularity conditions, $v(\theta) \geq 1/I_1(\theta)$ and

$$ARE(T_{1,n}, T_{2,n}) = \frac{\frac{1}{I_1(\theta)}}{v(\theta)} = \frac{1}{I_1(\theta)v(\theta)} \leq 1.$$

Hence asymptotically efficient estimators are "better" than estimators of the form $T_{1,n}$. When $T_{2,n}$ is asymptotically efficient,

$$AE(T_{1,n}) = ARE(T_{1,n}, T_{2,n}) = \frac{1}{I_1(\theta)v(\theta)}$$

is sometimes called the asymptotic efficiency of $T_{1,n}$.

Notice that for a 1P–REF, $\overline{T}_n = \frac{1}{n}\sum_{i=1}^n t(Y_i)$ is an asymptotically efficient estimator of $g(\eta) = E(t(Y))$ by Theorem 2.5. $\overline{T}_n$ is the UMVUE of $E(t(Y))$ by the LSU theorem.

The following theorem suggests that MLEs and UMVUEs are often asymptotically efficient. The theorem often holds for location families where the support does not depend on θ. The theorem does not hold for the

uniform $(0, \theta)$ family. For the MLE $\hat{\theta}$. Geisser (2006, pp. 133-134) shows that if i) the Y_i are iid with pdf $f(y|\theta)$ and likelihood function $L(\theta) = \prod_{i=1}^{n} f(y_i|\theta)$, ii) $E_\theta\left[\left(\frac{d\ log(L(\theta))}{d\theta}\right)\right] = 0$, and iii) $E_\theta\left[\left(\frac{d\ log(L(\theta))}{d\theta}\right)^2\right] = -E_\theta\left[\frac{d^2\ log(L(\theta))}{d\theta^2}\right]$ exists and is nonzero for all θ in a neighborhood of the true value θ_0, then

$$\sqrt{n}[\hat{\theta}_n - \theta_0] \xrightarrow{D} N\left(0, \frac{1}{I_1(\theta_0)}\right).$$

Conditions ii) and iii) hold for a 1P-REF with a pdf by Equations (1.45) and (1.48). See Berk (1972) and Wald (1949) for different regularity conditions. Hence the following theorem holds for the MLE $\hat{\theta}_n$ computed from iid $Y_1, ..., Y_n$ if $f(y|\theta)$ is the pdf of a 1P-REF.

Theorem 2.7: a "Standard Limit Theorem": Let $\hat{\theta}_n$ be the MLE or UMVUE of θ. If $\tau'(\theta) \neq 0$, then under strong regularity conditions,

$$\sqrt{n}[\tau(\hat{\theta}_n) - \tau(\theta)] \xrightarrow{D} N\left(0, \frac{[\tau'(\theta)]^2}{I_1(\theta)}\right).$$

2.3 Modes of Convergence and Consistency

Definition 2.4. Let $\{Z_n, n = 1, 2, ...\}$ be a sequence of random variables with cdfs F_n, and let X be a random variable with cdf F. Then Z_n **converges in distribution to** X, written

$$Z_n \xrightarrow{D} X,$$

or Z_n *converges in law to* X, written $Z_n \xrightarrow{L} X$, if

$$\lim_{n\to\infty} F_n(t) = F(t)$$

at each continuity point t of F. The distribution of X is called the **limiting distribution** or the **asymptotic distribution** of Z_n.

Convergence in distribution is also known as weak convergence or X_n converges weakly to X. An important fact is that **the limiting distribution does not depend on the sample size** n. Notice that the CLT, delta method and Theorem 2.5 give the limiting distributions of $Z_n = \sqrt{n}(\overline{Y}_n - \mu)$, $Z_n = \sqrt{n}(g(T_n) - g(\theta))$ and $Z_n = \sqrt{n}[\frac{1}{n}\sum_{i=1}^{n} t(Y_i) - E(t(Y))]$, respectively.

Remark 2.4. i) An important fact is that **the limiting distribution does not depend on the sample size** n.

ii) **Warning:** A common error is to get a "limiting distribution" that does depend on n.
iii) **Know:** If $F_n(t) \to H(t)$ and $H(t)$ is continuous, then for convergence in distribution, $H(t)$ needs to be a cdf: $H(t) = F_X(t)$ if $X_n \xrightarrow{D} X$. If $H(t)$ is a constant: $H(t) = c \in [0,1]$ $\forall t$, then $H(t)$ is not a cdf, and X_n does not converge in distribution to any random variable X.
iv) Since $F(x) = P(X \leq x)$, it follow that $0 \leq F_n(t) \leq 1$. Thus $\lim_{n\to\infty} F_n(t) = H(t)$ has $0 \leq H(t) \leq 1$ if the limit exists. **Warning:** A common error it to get $H(t) < 0$ or $H(t) > 1$.
v) **Warning**: Convergence in distribution says that the cdf $F_n(t)$ of X_n gets close to the cdf of F(t) of X as $n \to \infty$ provided that t is a continuity point of F. Hence for any $\epsilon > 0$ there exists N_t such that if $n > N_t$, then $|F_n(t) - F(t)| < \epsilon$. Notice that N_t depends on the value of t. Convergence in distribution does not imply that the random variables $X_n \equiv X_n(\omega)$ converge to the random variable $X \equiv X(\omega)$ for all ω.
vi) If $F_{X_n}(t) \to F_X(t)$ at all continuity points of $F_X(t)$, then $X_n \xrightarrow{D} X$. If t_0 is a discontinuity point of $F_X(t)$, then the behavior of $F_{X_n}(t_0)$ is not important: could have $\lim_{n\to\infty} F_{X_n}(t_0) = c_{t_0} \in [0,1]$ or that $\lim_{n\to\infty} F_{X_n}(t_0)$ does not exist. Convergence in distribution does not need $c_{t_0} = F_X(t_0)$.
vii) If $F_{X_n}(t) \to H(t)$ except at discontinuity points of $F_X(t)$, still need $H(t) = F_X(t)$ at continuity points of $F_X(t)$ for $X_n \xrightarrow{D} X$.

Convergence in distribution is useful because if the distribution of X_n is unknown or complicated and the distribution of X is easy to use, then for large n we can approximate the probability that X_n is in an interval by the probability that X is in the interval. To see this, notice that if $X_n \xrightarrow{D} X$, then $P(a < X_n \leq b) = F_n(b) - F_n(a) \to F(b) - F(a) = P(a < X \leq b)$ if F is continuous at a and b. Convergence in distribution is useful for constructing large sample confidence intervals and tests of hypotheses. See Chapter 4.

Example 2.9. Suppose that $X_n \sim U(-1/n, 1/n)$. Then the cdf $F_n(x)$ of X_n is

$$F_n(x) = \begin{cases} 0, & x \leq \frac{-1}{n} \\ \frac{nx}{2} + \frac{1}{2}, & \frac{-1}{n} \leq x \leq \frac{1}{n} \\ 1, & x \geq \frac{1}{n}. \end{cases}$$

Sketching $F_n(x)$ shows that it has a line segment rising from 0 at $x = -1/n$ to 1 at $x = 1/n$ and that $F_n(0) = 0.5$ for all $n \geq 1$. Examining the cases $x < 0$, $x = 0$ and $x > 0$ shows that as $n \to \infty$,

$$F_n(x) \to \begin{cases} 0, & x < 0 \\ \frac{1}{2} & x = 0 \\ 1, & x > 0. \end{cases}$$

Notice that if X is a random variable such that $P(X = 0) = 1$, then X has cdf

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0. \end{cases}$$

Since $x = 0$ is the only discontinuity point of $F_X(x)$ and since $F_n(x) \to F_X(x)$ for all continuity points of $F_X(x)$ (i.e. for $x \neq 0$),

$$X_n \xrightarrow{D} X.$$

Example 2.10. Suppose $Y_n \sim U(0, n)$. Then $F_n(t) = t/n$ for $0 < t \leq n$ and $F_n(t) = 0$ for $t \leq 0$. Hence $\lim_{n\to\infty} F_n(t) = 0$ for $t \leq 0$. If $t > 0$ and $n > t$, then $F_n(t) = t/n \to 0$ as $n \to \infty$. Thus $\lim_{n\to\infty} F_n(t) = H(t) = 0$ for all t, and Y_n does not converge in distribution to any random variable Y since $H(t) \equiv 0$ is a continuous function but not a cdf.

Definition 2.5. A sequence of random variables X_n *converges in distribution to a constant* $\tau(\theta)$, written

$$X_n \xrightarrow{D} \tau(\theta), \quad \text{if} \quad \mathrm{X_n} \xrightarrow{\mathrm{D}} \mathrm{X}$$

where $P(X = \tau(\theta)) = 1$. The distribution of the random variable X is said to be *degenerate at* $\tau(\theta)$ or to be a *point mass at* $\tau(\theta)$.

See Section 1.10 for some properties of the point mass distribution, which corresponds to a discrete random variable that only takes on exactly one value. Using characteristic functions, it can be shown that if X has a point mass at $\tau(\theta)$, then $X \sim N(\tau(\theta), 0)$, a normal distribution with mean $\tau(\theta)$ and variance 0. A point mass at 0, where $P(X = 0) = 1$, is a common limiting distribution. See Examples 2.9 and 2.11.

Example 2.11. X has a point mass distribution at c or X is degenerate at c if $P(X = c) = 1$. Thus X has a probability mass function with all of the mass at the point c. Then $F_X(t) = 1$ for $t \geq c$ and $F_X(t) = 0$ for $t < c$. Often $F_{X_n}(t) \to F_X(t)$ for all $t \neq c$ where $P(X = c) = 1$. Then $X_n \xrightarrow{D} X$ where $P(X = c) = 1$. Thus $F_{X_n}(t) \to H(t)$ for all $t \neq c$ where $H(t) = F_X(t)\ \forall t \neq c$. It is possible that $\lim_{n\to\infty} F_{X_n}(c) = H(c) \in [0, 1]$ or that $\lim_{n\to\infty} F_{X_n}(c)$ does not exist.

Example 2.12. Prove whether the following sequences of random variables X_n converge in distribution to some random variable X. If $X_n \xrightarrow{D} X$, find the distribution of X (for example, find $F_X(t)$ or note that $P(X = c) = 1$, so X has the point mass distribution at c).
a) $X_n \sim U(-n - 1, -n)$
b) $X_n \sim U(n, n + 1)$
c) $X_n \sim U(a_n, b_n)$ where $a_n \to a < b$ and $b_n \to b$.
d) $X_n \sim U(a_n, b_n)$ where $a_n \to c$ and $b_n \to c$.

e) $X_n \sim U(-n, n)$
f) $X_n \sim U(c - 1/n, c + 1/n)$

Solution. If $X_n \sim U(a_n, b_n)$ with $a_n < b_n$, then

$$F_{X_n}(t) = \frac{t - a_n}{b_n - a_n}$$

for $a_n \le t \le b_n$, $F_{X_n}(t) = 0$ for $t \le a_n$ and $F_{X_n}(t) = 1$ for $t \ge b_n$. On $[a_n, b_n]$, $F_{X_n}(t)$ is a line segment from $(a_n, 0)$ to $(b_n, 1)$ with slope $\dfrac{1}{b_n - a_n}$.

a) $F_{X_n}(t) \to H(t) \equiv 1 \ \ \forall t \in \mathbb{R}$ since $F_{X_n}(t) = 1$ for $t \ge -n$. Since $H(t)$ is continuous but not a cdf, X_n does not converge in distribution to any RV X.

b) $F_{X_n}(t) \to H(t) \equiv 0 \ \ \forall t \in \mathbb{R}$ since $F_{X_n}(t) = 0$ for $t < n$. Since $H(t)$ is continuous but not a cdf, X_n does not converge in distribution to any RV X.

c)

$$F_{X_n}(t) \to F_X(t) = \begin{cases} 0 & t \le a \\ \frac{t-a}{b-a} & a \le t \le b \\ 1 & t \ge b. \end{cases}$$

Hence $X_n \xrightarrow{D} X \sim U(a, b)$.

d)

$$F_{X_n}(t) \to \begin{cases} 0 & t < c \\ 1 & t > c. \end{cases}$$

Hence $X_n \xrightarrow{D} X$ where $P(X = c) = 1$. Hence X has a point mass distribution at c. (The behavior of $\lim_{n\to\infty} F_{X_n}(c)$ is not important, even if the limit does not exist.)

e)

$$F_{X_n}(t) = \frac{t + n}{2n} = \frac{1}{2} + \frac{t}{2n}$$

for $-n \le t \le n$. Thus $F_{X_n}(t) \to H(t) \equiv 0.5 \ \ \forall t \in \mathbb{R}$. Since $H(t)$ is continuous but not a cdf, X_n does not converge in distribution to any RV X.

f)

$$F_{X_n}(t) = \frac{t - c + \frac{1}{n}}{\frac{2}{n}} = \frac{1}{2} + \frac{n}{2}(t - c)$$

for $c - 1/n \le t \le c + 1/n$. Thus

$$F_{X_n}(t) \to H(t) = \begin{cases} 0 & t < c \\ 1/2 & t = c \\ 1 & t > c. \end{cases}$$

If X has the point mass at c, then

$$F_X(t) = \begin{cases} 0 & t < c \\ 1 & t \ge c. \end{cases}$$

Hence $t = c$ is the only discontinuity point of $F_X(t)$, and $H(t) = F_X(t)$ at all continuity points of $F_X(t)$. Thus $X_n \xrightarrow{D} X$ where $P(X = c) = 1$.

Definition 2.6. a) A sequence of random variables X_n *converges in probability to a constant* $\tau(\theta)$, written

$$X_n \xrightarrow{P} \tau(\theta),$$

if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - \tau(\theta)| < \epsilon) = 1 \text{ or, equivalently, } \lim_{\mathrm{n}\to\infty} \mathrm{P}(|\mathrm{X_n} - \tau(\theta)| \geq \epsilon) = 0.$$

b) The sequence X_n **converges in probability to** X, written

$$X_n \xrightarrow{P} X,$$

if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1, \text{ or, equivalently, } \lim_{\mathrm{n}\to\infty} \mathrm{P}(|\mathrm{X_n} - \mathrm{X}| \geq \epsilon) = 0.$$

Notice that $X_n \xrightarrow{P} X$ if $X_n - X \xrightarrow{P} 0$.

Definition 2.7. A sequence of estimators T_n of $\tau(\theta)$ is **consistent** for $\tau(\theta)$ if

$$T_n \xrightarrow{P} \tau(\theta)$$

for every $\theta \in \Theta$. If T_n is consistent for $\tau(\theta)$, then T_n is a **consistent estimator** of $\tau(\theta)$.

Consistency is a weak property that is usually satisfied by good estimators. T_n is a consistent estimator for $\tau(\theta)$ if the probability that T_n falls in any neighborhood of $\tau(\theta)$ goes to one, regardless of the value of $\theta \in \Theta$. The probability $P \equiv P_\theta$ is the "true" probability distribution or underlying probability that depends on θ.

Definition 2.8. For a real number $r > 0$, Y_n **converges in** r**th mean** to a random variable Y, written $Y_n \xrightarrow{r} Y$, if

$$E(|Y_n - Y|^r) \to 0$$

as $n \to \infty$. In particular, if $r = 2$, Y_n **converges in quadratic mean** to Y, written

$$Y_n \xrightarrow{2} Y \text{ or } \mathrm{Y_n} \xrightarrow{\mathrm{qm}} \mathrm{Y},$$

if $E[(Y_n - Y)^2] \to 0$ as $n \to \infty$. We say that X_n *converges in* r*th mean* to $\tau(\theta)$, written

$$X_n \xrightarrow{r} \tau(\theta),$$

if $E(|Y_n - \tau(\theta)|^r) \to 0$ as $n \to \infty$.

Convergence in quadratic mean is also known as convergence in mean square and as mean square convergence. From Definition 1.41, the mean square error $MSE_{\tau(\theta)}(X_n) = E_\theta[(X_n - \tau(\theta))^2]$. The notations $Y_n \xrightarrow{r} Y$, $Y_n \xrightarrow{L_r} Y$, and $Y_n \xrightarrow{L^r} Y$ are used in the literature, especially for $r \geq 1$.

Theorem 2.8: Generalized Chebyshev's Inequality or Generalized Markov's Inequality: Let $u : \mathbb{R} \to [0, \infty)$ be a nonnegative function. If $E[u(Y)]$ exists then for any $c > 0$,

$$P[u(Y) \geq c] \leq \frac{E[u(Y)]}{c}.$$

If $\mu = E(Y)$ exists, then taking $u(y) = |y - \mu|^r$ and $\tilde{c} = c^r$ gives
Markov's Inequality: for $r > 0$ with $E[|Y - \mu|^r]$ finite and for any $c > 0$,

$$P(|Y - \mu| \geq c] = P(|Y - \mu|^r \geq c^r] \leq \frac{E[|Y - \mu|^r]}{c^r}.$$

If $r = 2$ and $\sigma^2 = V(Y)$ exists, then we obtain
Chebyshev's Inequality:

$$P(|Y - \mu| \geq c] \leq \frac{V(Y)}{c^2}.$$

Proof. The proof is given for pdfs. For pmfs, replace the integrals by sums. Now

$$E[u(Y)] = \int_{\mathbb{R}} u(y)f(y)dy = \int_{\{y:u(y)\geq c\}} u(y)f(y)dy + \int_{\{y:u(y)<c\}} u(y)f(y)dy$$

$$\geq \int_{\{y:u(y)\geq c\}} u(y)f(y)dy$$

since the integrand $u(y)f(y) \geq 0$. Hence

$$E[u(Y)] \geq c \int_{\{y:u(y)\geq c\}} f(y)dy = cP[u(Y) \geq c]. \ \square$$

Note: if $E[|Y - \mu|^k]$ is finite and $k > 0$, then $E[|Y - \mu|^r]$ is finite for $0 < r \leq k$. See Theorem 2.21.

The following theorem gives sufficient conditions for T_n to be a consistent estimator of $\tau(\theta)$, or for T_n to converge in probability to $\tau(\theta)$. Notice that $MSE_{\tau(\theta)}(T_n) \to 0$ for all $\theta \in \Theta$ is equivalent to $T_n \xrightarrow{qm} \tau(\theta)$ for all $\theta \in \Theta$.

Theorem 2.9. a) If

$$\lim_{n\to\infty} MSE_{\tau(\theta)}(T_n) = 0$$

for all $\theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$.

b) If

$$\lim_{n\to\infty} V_\theta(T_n) = 0 \quad \text{and} \quad \lim_{n\to\infty} E_\theta(T_n) = \tau(\theta)$$

for all $\theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$.

c) If

$$\lim_{n\to\infty} MSE_{\tau(\theta)}(T_n) = 0,$$

then $T_n \xrightarrow{P} \tau(\theta)$.

d) If

$$\lim_{n\to\infty} V_\theta(T_n) = 0 \quad \text{and} \quad \lim_{n\to\infty} E_\theta(T_n) = \tau(\theta),$$

then $T_n \xrightarrow{P} \tau(\theta)$.

Proof. a) and c): Using Theorem 2.8 with $Y = T_n$, $u(T_n) = (T_n - \tau(\theta))^2$ and $c = \epsilon^2$ shows that for any $\epsilon > 0$,

$$P_\theta(|T_n - \tau(\theta)| \geq \epsilon) = P_\theta[(T_n - \tau(\theta))^2 \geq \epsilon^2] \leq \frac{E_\theta[(T_n - \tau(\theta))^2]}{\epsilon^2}.$$

Hence

$$\lim_{n\to\infty} E_\theta[(T_n - \tau(\theta))^2] = \lim_{n\to\infty} MSE_{\tau(\theta)}(T_n) \to 0$$

is a sufficient condition for T_n to be a consistent estimator of $\tau(\theta)$, and for $T_n \xrightarrow{P} \tau(\theta)$.

b) and d): Referring to Definition 1.41,

$$MSE_{\tau(\theta)}(T_n) = V_\theta(T_n) + [\text{Bias}_{\tau(\theta)}(\text{T}_\text{n})]^2$$

where $\text{Bias}_{\tau(\theta)}(\text{T}_\text{n}) = \text{E}_\theta(\text{T}_\text{n}) - \tau(\theta)$. Since $MSE_{\tau(\theta)}(T_n) \to 0$ if both $V_\theta(T_n)$ $\to 0$ and $\text{Bias}_{\tau(\theta)}(\text{T}_\text{n}) = \text{E}_\theta(\text{T}_\text{n}) - \tau(\theta) \to 0$, the result follows from a). □

Remark 2.5. We want conditions $A \Rightarrow B$ where B is $X_n \xrightarrow{P} X$. $A \Rightarrow B$ does not mean that if A does not hold, then B does not hold. $A \Rightarrow B$ means that if A holds, then B holds. A **common error** is for the student to say A does not hold, so X_n does not converge in probability to X.

Theorem 2.10. a) Suppose X_n and X are RVs with the same probability space. If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{D} X$.

b) $X_n \xrightarrow{P} \tau(\theta)$ **iff** $X_n \xrightarrow{D} \tau(\theta)$.

Theorem 2.11. If $X_n \xrightarrow{r} X$, then $X_n \xrightarrow{P} X$.

Proof.

$$P[|X_n - X| \geq \epsilon] = P[|X_n - X|^r \geq \epsilon^r] \leq \frac{E[|X_n - X|^r]}{\epsilon^r} \to 0$$

as $n \to \infty$ by the Generalized Chebyshev Inequality. □

The following result shows estimators that converge at a $\sqrt{n}$ rate are consistent. Use this result and the delta method to show that $g(T_n)$ is a consistent estimator of $g(\theta)$. Note that b) follows from a) with $X_\theta \sim N(0, v(\theta))$. The WLLN shows that $\overline{Y}$ is a consistent estimator of $E(Y) = \mu$ if $E(Y)$ exists.

Theorem 2.12. a) Let X_θ be a random variable with a distribution depending on θ, and $0 < \delta \leq 1$. Suppose

$$n^\delta(T_n - \tau(\theta)) \xrightarrow{D} X_\theta$$

for all $\theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$. If the convergence holds for a fixed θ, then $T_n \xrightarrow{P} \tau(\theta)$.

b) If

$$\sqrt{n}(T_n - \tau(\theta)) \xrightarrow{D} N(0, v(\theta))$$

for all $\theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$.

Definition 2.9. a) A sequence of random variables X_n *converges with probability 1* (or *almost surely*, or *almost everywhere*) to X if

$$P(\lim_{n\to\infty} X_n = X) = 1.$$

This type of convergence will be denoted by

$$X_n \xrightarrow{wp1} X.$$

b)

$$X_n \xrightarrow{wp1} \tau(\theta),$$

if $P(\lim_{n\to\infty} X_n = \tau(\theta)) = 1$.

The convergence in Definition 2.9 is also known as *strong convergence*. Notation such as "X_n converges to X wp1" will also be used. Sometimes "wp1" will be replaced with "as" or "ae." The notations $X_n \xrightarrow{ae} X$, $X_n \xrightarrow{as} X$, and $X_n \xrightarrow{wp1} X$ are often used.

Theorem 2.13. Let Y_i be a sequence of iid random variables with $E(Y_i) = \mu$. Then

a) **Strong Law of Large Numbers (SLLN):** $\overline{Y}_n \xrightarrow{wp1} \mu$, and

b) **Weak Law of Large Numbers (WLLN):** $\overline{Y}_n \xrightarrow{P} \mu$.

Proof of WLLN when $V(Y_i) = \sigma^2$: By Chebyshev's inequality, for every $\epsilon > 0$,

$$P(|\overline{Y}_n - \mu| \geq \epsilon) \leq \frac{V(\overline{Y}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$$

as $n \to \infty$. $\square$

Remark 2.6. a) For i) $X_n \xrightarrow{P} X$, ii) $X_n \xrightarrow{r} X$, or iii) $X_n \xrightarrow{wp1} X$, the X_n and X need to be defined on the same probability space.

b) For $X_n \xrightarrow{D} X$, the probability spaces can differ.

c) For i) $X_n \xrightarrow{P} c$, ii) $X_n \xrightarrow{wp1} c$, iii) $X_n \xrightarrow{D} c$, and iv) $X_n \xrightarrow{r} c$, the probability spaces of the X_n can differ.

d) **Warning:** For the SLLN and WLLN, students often forget that $V(Y_i) = \sigma^2$ is not needed. Only need the Y_i iid with $E(Y_i) = \mu$.

Theorem 2.14: a) $T_n \xrightarrow{P} \tau(\theta)$ iff $\mathrm{T_n} \xrightarrow{\mathrm{D}} \tau(\theta)$.

b) If $T_n \xrightarrow{P} \theta$ and τ is continuous at θ, then $\tau(T_n) \xrightarrow{P} \tau(\theta)$. Hence if T_n is a consistent estimator of θ, then $\tau(T_n)$ is a consistent estimator of $\tau(\theta)$ if τ is a continuous function on Θ.

Theorem 2.15: Suppose X_n and X are RVs with the same probability space for b) and c). Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function.

a) If $X_n \xrightarrow{D} X$, then $g(X_n) \xrightarrow{D} g(X)$.

b) If $X_n \xrightarrow{P} X$, then $g(X_n) \xrightarrow{P} g(X)$.

c) If $X_n \xrightarrow{ae} X$, then $g(X_n) \xrightarrow{ae} g(X)$.

Theorem 2.16: Suppose X_n and X are RVs with the same probability space.

a) If $X_n \xrightarrow{wp1} X$, then $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{D} X$.

b) If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{D} X$.

c) If $X_n \xrightarrow{r} X$, then $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{D} X$.

d) $X_n \xrightarrow{P} \tau(\theta)$ **iff** $X_n \xrightarrow{D} \tau(\theta)$ where $c = \tau(\theta)$ is a constant.

Theorem 2.17: a) If $E[(X_n - X)^2] \to 0$ as $n \to \infty$, then $X_n \xrightarrow{P} X$.

b) If $E(X_n) \to E(X)$ and $V(X_n - X) \to 0$ as $n \to \infty$, then $X_n \xrightarrow{P} X$.

Note: Part a) follows from Theorem 2.16 c) with $r = 2$. See Theorem 2.9 if $P(X = \tau(\theta)) = 1$.

Theorem 2.18: Let X_n have pdf $f_{X_n}(x)$, and let X have pdf $f_X(x)$. If $f_{X_n}(x) \to f_X(x)$ for all x (or for x outside of a set of Lebesgue measure 0), then $X_n \xrightarrow{D} X$.

Theorem 2.19 is a special case of Theorem 2.15.

Theorem 2.19: Let $g : \mathbb{R} \to \mathbb{R}$ be continuous at constant c.

a) If $X_n \xrightarrow{D} c$, then $g(X_n) \xrightarrow{D} g(c)$.

b) If $X_n \xrightarrow{P} c$, then $g(X_n) \xrightarrow{P} g(c)$.

c) If $X_n \xrightarrow{ae} c$, then $g(X_n) \xrightarrow{ae} g(c)$.

Theorem 2.20: Suppose X_n and X are integer valued RVs with pmfs $f_{X_n}(x)$ and $f_X(x)$. Then $X_n \xrightarrow{D} X$ iff $P(X_n = k) \to P(X = k)$ for every integer k iff $f_{X_n}(x) \to f_X(x)$ for every real x.

The following theorem uses the fact that $E(W)$ is finite iff $E(|W|)$ is finite.

Theorem 2.21: Let $k > 0$. If $E(X^k)$ is finite, then $E(X^j)$ is finite for $0 < j \leq k$.

Proof. If $|y| \leq 1$, then $|y^j| = |y|^j \leq 1$. If $|y| > 1$ then $|y|^j \leq |y|^k$. Thus $|y|^j \leq |y|^k + 1$ and $|X|^j \leq |X|^k + 1$. Hence $E[|X|^j] \leq E[|X|^k] + 1 < \infty$. □

Theorem 2.22, Jensen's Inequality:

$$g[E(X)] \leq E[g(X)]$$

if the expected values exist and the function g is convex on an interval containing the range of X.

Proof for when g is twice differentiable: Assume that $g''(x) \geq 0$ on the interval containing the range of X. The Taylor's series expansion of $g(x)$ about $\mu = E(X)$ gives (for x in the interval)

$$g(x) = g(\mu) + g'(\mu)(x - \mu) + \frac{g''(\eta)(x - \mu)^2}{2}$$

where η is some value between x and μ. Thus $g(x) \geq g(\mu) + g'(\mu)(x - \mu)$ and $g(X) \geq g(\mu) + g'(\mu)(X - \mu)$. Taking expectations gives

$$E[g(X)] \geq g(\mu) + g'(\mu)E(X - \mu) = g(\mu) = g(E[X]). \quad \square$$

Remark 2.7. a) Let (a, b) be an open interval where $a = -\infty$ and $b = \infty$ are allowed. A sufficient condition for a function g to be convex on an open interval (a, b) is $g''(x) \geq 0$ on (a, b). If $(a, b) = (0, \infty)$ and g is continuous on $[0, \infty)$ and convex on $(0, \infty)$, then g is convex on $[0, \infty)$.
b) If X is a positive RV, then the range of X is $(0, \infty)$.

Theorem 2.23.: If $X_n \xrightarrow{r} X$, then $X_n \xrightarrow{k} X$ where $0 < k < r$.

Proof. Let $U_n = |X_n - X|^r$ and $W_n = |X_n - X|^k$. then $U_n = W_n^t$ where $t = r/k > 1$. The function $g(x) = x^t$ is convex on $[0, \infty)$. By Jensen's inequality,

$$E[|X_n - X|^r] = E[U_n] = E[W_n^t] \geq (E[W_n])^t = (E[|X_n - X|^k])^{r/k}$$

for $r > k$. Thus $\lim_{n\to\infty} E[|X_n - X|^r = 0$ implies that $\lim_{n\to\infty} E[|X_n - X|^k = 0$ for $0 < k < r$. □

Example 2.13. a) Let $P(X_n = n) = 1/n$ and $P(X_n = 0) = 1 - 1/n$. Hence X_n is discrete and takes on two values with $E(X_n) = n\frac{1}{n} = 1$ for all positive integers n. Hence $E[|X_n - 0|] = E(X_n) = 1 \ \ \forall n$ and X_n **does not satisfy** $X_n \xrightarrow{1} 0$. Let $\epsilon > 0$. Then

$$P[|X_n - 0| \geq \epsilon] \leq P(X_n = n) = \frac{1}{n} \to 0$$

as $n \to \infty$. Hence $X_n \xrightarrow{P} 0$ and $X_n \xrightarrow{D} 0$.

b) Let $P(X_n = 0) = 1 - \frac{1}{n}$ and $P(X_n = 1) = 1/n$. Hence X_n is discrete and takes on two values with

$$E[(X_n - 0)^2] = E(X_n^2) = \sum x^2 P(X_n = x) = 0^2(1 - \frac{1}{n}) + 1^2\frac{1}{n} = \frac{1}{n} \to 0$$

as $n \to \infty$. Hence $X_n \xrightarrow{2} 0$, $X_n \xrightarrow{P} 0$, and $X_n \xrightarrow{D} 0$. Note that

$$E[|X_n - 0|] = E(X_n) = \frac{1}{n} \to 0.$$

Hence $X_n \xrightarrow{1} 0$ as expected by Theorem 2.23 since $X_n \xrightarrow{2} 0$.

Theorem 2.24.: Let X_n have pdf $f_{X_n}(x)$, and let X have pdf $f_X(x)$. If $f_{X_n}(x) \to f_X(x)$ for all x (or for x outside of a set of Lebesgue measure 0), then $X_n \xrightarrow{D} X$.

Theorem 2.25.: Suppose X_n and X are integer valued RVs with pmfs $f_{X_n}(x)$ and $f_X(x)$. Then $X_n \xrightarrow{D} X$ iff $P(X_n = k) \to P(X = k)$ for every integer k iff $f_{X_n}(x) \to f_X(x)$ for every real x.

2.4 Slutsky's Theorem and Related Results

Theorem 2.26. Suppose X_n and X are RVs with the same probability space.
a) If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{D} X$.
b) If $X_n \xrightarrow{wp1} X$, then $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{D} X$.
c) If $X_n \xrightarrow{r} X$, then $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{D} X$.
d) $X_n \xrightarrow{P} \tau(\theta)$ **iff** $X_n \xrightarrow{D} \tau(\theta)$.
e) If $X_n \xrightarrow{D} X$ and $X_n \xrightarrow{D} Y$, then $X \stackrel{D}{=} Y$ and $F_X(x) = F_Y(x)$ for all real x.

Partial Proof. a) See Theorem 2.10. c) See Theorem 2.11. d) See Theorem 2.10.

e) Suppose X has cdf F and Y has cdf G. Then F and G agree at their common points of continuity. Hence F and G agree at all but countably many points since F and G are cdfs. Hence F and G agree at all points by right continuity. □

Note: If $X_n \xrightarrow{A} X$ and $X_n \xrightarrow{A} Y$, then $X \stackrel{D}{=} Y$ where A is $wp1$, r, or P. This result holds by Theorem 2.26 e) since if $X_n \xrightarrow{A} X$ and $X_n \xrightarrow{A} Y$, then $X_n \xrightarrow{D} X$ and $X_n \xrightarrow{D} Y$.

Theorem 2.27: Slutsky's Theorem. Suppose $Y_n \xrightarrow{D} Y$ and $W_n \xrightarrow{P} w$ for some constant w. Then

a) $Y_n + W_n \xrightarrow{D} Y + w$,
b) $Y_n W_n \xrightarrow{D} wY$, and
c) $Y_n/W_n \xrightarrow{D} Y/w$ if $w \neq 0$.

Remark 2.8. Note that $Y_n \xrightarrow{A} Y$ implies $Y_n \xrightarrow{D} Y$ where $A = wp1, r$, or P. Also $W_n \xrightarrow{P} w$ iff $W_n \xrightarrow{D} w$. If a sequence of constants $c_n \to c$ as $n \to \infty$ (regular convergence is everywhere convergence), then $c_n \xrightarrow{wp1} c$ and $c_n \xrightarrow{P} c$. So $W_n \xrightarrow{P} w$ can be replaced by $W_n \xrightarrow{B} w$ where $B = D, wp1, r, P$, or regular convergence.
i) So Slutsky's theorem a), b), and c) hold if $Y_n \xrightarrow{A} Y$ and $W_n \xrightarrow{B} w$.
ii) If $Y \equiv y$ where y is a constant, then $Y_n \xrightarrow{A} y$ and $W_n \xrightarrow{B} w$ implies that a), b) and c) hold with Y replaced by y, and $\xrightarrow{D}$ can be replaced by $\xrightarrow{P}$.
iii) If $Y_n \xrightarrow{D} Y$, $a_n \xrightarrow{P} a$, and $b_n \xrightarrow{P} b$, then $a_n + b_n Y_n \xrightarrow{D} a + bY$.

Theorem 2.28. a) If $X_n \xrightarrow{P} \theta$ and τ is continuous at θ, then $\tau(X_n) \xrightarrow{P} \tau(\theta)$.
b) If $X_n \xrightarrow{D} \theta$ and τ is continuous at θ, then $\tau(X_n) \xrightarrow{D} \tau(\theta)$.

Theorem 2.28 is a special case of the continuous mapping theorem. See Theorem 2.30. Suppose that for all $\theta \in \Theta$, $T_n \xrightarrow{D} \tau(\theta)$, $T_n \xrightarrow{r} \tau(\theta)$ or $T_n \xrightarrow{wp1} \tau(\theta)$. Then T_n is a consistent estimator of $\tau(\theta)$ by Theorem 2.26. We are assuming that the function τ does not depend on n since we want a single function $\tau(\theta)$ rather than a sequence of functions $\tau_n(\theta)$. See Remark 2.1 b).

Example 2.14. Let $Y_1, ..., Y_n$ be iid with mean $E(Y_i) = \mu$ and variance $V(Y_i) = \sigma^2$. Then the sample mean $\overline{Y}_n$ is a consistent estimator of μ since i) the SLLN holds (use Theorem 2.13 and 2.26), ii) the WLLN holds and iii) the CLT holds (use Theorem 2.12). Since

$$\lim_{n\to\infty} V_\mu(\overline{Y}_n) = \lim_{n\to\infty} \sigma^2/n = 0 \quad \text{and} \quad \lim_{n\to\infty} E_\mu(\overline{Y}_n) = \mu,$$

$\overline{Y}_n$ is also a consistent estimator of μ by Theorem 2.9b. By the delta method and Theorem 2.12b, $T_n = g(\overline{Y}_n)$ is a consistent estimator of $g(\mu)$ if $g'(\mu) \neq 0$ for all $\mu \in \Theta$. By Theorem 2.28a, $g(\overline{Y}_n)$ is a consistent estimator of $g(\mu)$ if g is continuous at μ for all $\mu \in \Theta$.

Theorem 2.29: Continuity Theorem. Let Y_n be sequence of random variables with characteristic functions $\phi_n(t)$. Let Y be a random variable with cf $\phi(t)$.
a)

$$Y_n \xrightarrow{D} Y \text{ iff } \phi_n(t) \to \phi(t) \ \forall t \in \mathbb{R}.$$

b) Also assume that Y_n has mgf m_n and Y has mgf m. Assume that all of the mgfs m_n and m are defined on $|t| \leq d$ for some $d > 0$. Then if $m_n(t) \to m(t)$ as $n \to \infty$ for all $|t| < c$ where $0 < c < d$, then $Y_n \xrightarrow{D} Y$.

The following theorem is often part of the continuity theorem in the literature, and helps explain why Theorem 2.29 is called the continuity theorem.

Theorem 2.30: If $\lim_{n\to\infty} c_{X_n}(t) = g(t)$ for all t where g is continuous at $t = 0$, then $g(t) = c_X(t)$ is a characteristic function for some RV X, and $X_n \xrightarrow{D} X$.

Remark 2.9. a) Continuity at $t = 0$ implies continuity everywhere since $g(t) = c_X(t)$ is continuous. If $g(t)$ is not continuous at 0, then X_n does not converge in distribution.

b) If $c_{Y_n}(t) \to h(t)$ where $h(t)$ is not continuous, then Y_n does not converge in distribution to any RV Y, by the Continuity Theorem and a).

c) **Warning:** $c_{X_n}(0) \equiv 1$, but $c_{X_n}(0) \to 1$ as $n \to \infty$ does not imply that g is continuous at $t = 0$ if $\lim_{n\to\infty} c_{X_n}(t) = g(t)$ for all real t.

d) Let $X_1, ..., X_n$ be independent RVs with characteristic functions $c_{X_j}(t)$. Then the characteristic function of $\sum_{j=1}^n X_j$ is $c_{\sum_{j=1}^n X_j}(t) = \prod_{j=1}^n c_{X_j}(t)$. If the RVs also have mgfs $m_{X_j}(t)$, then the mgf of $\sum_{j=1}^n X_j$ is $m_{\sum_{j=1}^n X_j}(t) = \prod_{j=1}^n m_{X_j}(t)$.

Theorem 2.31, Helly-Bray-Pormanteau Theorem: $X_n \xrightarrow{D} X$ iff $E[g(X_n)] \to E[g(X)]$ for every bounded, real, continuous function g.

The above theorem is used to prove Theorem 2.32 b).

Theorem 2.32. a) **Generalized Continuous Mapping Theorem:** If $X_n \xrightarrow{D} X$ and the function g is such that $P[X \in C(g)] = 1$ where $C(g)$ is the set of points where g is continuous, then $g(X_n) \xrightarrow{D} g(X)$.

b) **Continuous Mapping Theorem:** If $X_n \xrightarrow{D} X$ and the function g is continuous, then $g(X_n) \xrightarrow{D} g(X)$.

Proof of the Continuous Mapping Theorem: If g is real and continuous, then $\cos[tg(x)]$ and $\sin[tg(x)]$ are bounded real continuous functions. Hence by the Helly-Bray-Pormanteau theorem, for each real t, the characteristic function

$$c_{g(X_n)}(t) = E[e^{itg(X_n)}] = E(\cos[tg(X_n)]) + iE(\sin[tg(X_n)]) \to$$

$$E(\cos[tg(X)]) + iE(\sin[tg(X)]) = E[e^{itg(X)}] = c_{g(X)}(t).$$

Thus $g(X_n) \xrightarrow{D} g(X)$ by the continuity theorem. □

Remark 2.10. For Theorem 2.26, a) follows from Slutsky's Theorem by taking $Y_n \equiv X = Y$ and $W_n = X_n - X$. Then $Y_n \xrightarrow{D} Y = X$ and $W_n \xrightarrow{P} 0$. Hence $X_n = Y_n + W_n \xrightarrow{D} Y + 0 = X$. The convergence in distribution parts of b) and c) follow from a). Theorem 2.28b follows from Theorems 2.26d) and 2.28a). Theorem 2.28a) implies that if T_n is a consistent estimator of θ and τ is a continuous function, then $\tau(T_n)$ is a consistent estimator of $\tau(\theta)$. Theorem 2.32 says that convergence in distribution is preserved by continuous functions, and even some discontinuities are allowed as long as the set of continuity points is assigned probability 1 by the asymptotic distribution. Equivalently, the set of discontinuity points is assigned probability 0.

Example 2.15. (Ferguson 1996, p. 40): If $X_n \xrightarrow{D} X$ then $1/X_n \xrightarrow{D} 1/X$ if X is a continuous random variable since $P(X = 0) = 0$ and $x = 0$ is the only discontinuity point of $g(x) = 1/x$.

Example 2.16. Show that if $Y_n \sim t_n$, a t distribution with n degrees of freedom, then $Y_n \xrightarrow{D} Z$ where $Z \sim N(0, 1)$.

Solution: $Y_n \stackrel{D}{=} Z/\sqrt{V_n/n}$ where $Z \perp\!\!\!\perp V_n \sim \chi^2_n$. If $W_n = \sqrt{V_n/n} \xrightarrow{P} 1$, then the result follows by Slutsky's Theorem. But $V_n \stackrel{D}{=} \sum_{i=1}^n X_i$ where the iid $X_i \sim \chi^2_1$. Hence $V_n/n \xrightarrow{P} 1$ by the WLLN and $\sqrt{V_n/n} \xrightarrow{P} 1$ by Theorem 2.28a.

Before reading the proof for the CLT, review Remarks 1.3 and 1.4.

Remark 2.11, Notes for Proving the CLT. a) Suppose the Y_i are iid with characteristic function $c_Y(t)$. Then $E(Y_i - \mu) = 0$ and $V(Y_i - \mu) = E[(Y_i - \mu)^2] = \sigma^2$. Thus by Remark 1.3,

$$C_{Y-\mu}(t) = 1 - \frac{\sigma^2}{2}t^2 + o(t^2) \text{ and}$$

$$C_{Y-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + o(t^2/n)$$

where

$$\frac{o(t^2/n)}{t^2/n} \to 0$$

as $n \to \infty$. Hence $n \ o(t^2/n) \to 0$ as $n \to 0$.

b) Let the Z-score of $\overline{Y}_n$ be

$$Z_n = \frac{\sqrt{n}(\overline{Y} - \mu)}{\sigma} = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^n Y_i - n\mu}{\sigma\sqrt{n}} = \frac{\sum_{i=1}^n (Y_i - \mu)}{\sigma\sqrt{n}}$$

where the $Y_i - \mu$ are iid with characteristic function $c_{Y-\mu}(t)$. Then the characteristic function of $\frac{Y_i - \mu}{\sigma\sqrt{n}}$ is $c_{Y-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right)$, and the characteristic function of Z_n is

$$c_{Z_n}(t) = \left[c_{Y-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n.$$

If $c_{Z_n}(t) \to c_Z(t)$, the $N(0,1)$ characteristic function, then $\sigma Z_n = \sqrt{n}(\overline{Y}_n - \mu)$ has

$$c_{\sigma Z_n}(t) \to c_{\sigma Z}(t) = c_Z(\sigma_t) = e^{-\sigma^2 t^2/2},$$

the $N(0, \sigma^2)$ characteristic function, and the CLT holds.

Proof of the CLT: Let Z_n be the Z-score of $\overline{Y}_n$. By Remark 2.11,

$$c_{Z_n}(t) = \left[1 - \frac{t^2}{2n} + o(t^2/n)\right]^n =$$

$$\left[1 - \frac{\frac{t^2}{2} - n \;\; o(t^2/n)}{n}\right]^n \to e^{-t^2/2} = c_Z(t)$$

for all t by Remark 1.3b). Thus $Z_n \xrightarrow{D} Z \sim N(0,1)$ and $\sigma Z_n = \sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{D} N(0, \sigma^2)$. $\square$

The next proof does not use characteristic functions, but only applies to iid random variables Y_i that have a moment distribution function. Thus $E(Y_i^j)$ exists for each positive integer j. The CLT only needs $E(Y)$ and $E(Y^2)$ to exist. In the proof, $k(t) = \log(m(t))$ is the cumulant generating function with $k'(0) = E(X)$ and $k''(x) = V(X)$.

L'Hôspital's Rule: Suppose functions $f(x) \to 0$ and $g(x) \to 0$ as $x \downarrow d$, $x \uparrow d$, $x \to d$, $x \to \infty$, or $x \to -\infty$. If

$$\frac{f'(x)}{g'(x)} \to L \;\; \text{then} \;\; \frac{\text{f(x)}}{\text{g(x)}} \to \text{L}$$

as $x \downarrow d$, $x \uparrow d$, $x \to d$, $x \to \infty$, or $x \to -\infty$.

Proof of a Special Case of the CLT. Following Rohatgi (1984, pp. 569-9) and Tardiff (1981), let $Y_1, ..., Y_n$ be iid with mean μ, variance σ^2 and mgf $m_Y(t)$ for $|t| < t_o$. Then

$$Z_i = \frac{Y_i - \mu}{\sigma}$$

has mean 0, variance 1 and mgf $m_Z(t) = \exp(-t\mu/\sigma) m_Y(t/\sigma)$ for $|t| < \sigma t_o$. Want to show that

$$W_n = \sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) \xrightarrow{D} N(0,1).$$

Notice that $W_n =$

$$n^{-1/2}\sum_{i=1}^{n} Z_i = n^{-1/2}\sum_{i=1}^{n}\left(\frac{Y_i-\mu}{\sigma}\right) = n^{-1/2}\frac{\sum_{i=1}^{n} Y_i - n\mu}{\sigma} = \frac{n^{-1/2}}{\frac{1}{n}}\,\frac{\overline{Y}_n-\mu}{\sigma}.$$

Thus

$$m_{W_n}(t) = E(e^{tW_n}) = E[\exp(tn^{-1/2}\sum_{i=1}^{n} Z_i)] = E[\exp(\sum_{i=1}^{n} tZ_i/\sqrt{n})]$$

$$= \prod_{i=1}^{n} E[e^{tZ_i/\sqrt{n}}] = \prod_{i=1}^{n} m_Z(t/\sqrt{n}) = [m_Z(t/\sqrt{n})]^n.$$

The cumulant generating function $k_Z(t) = \log(m_Z(x))$. Then

$$k_{W_n}(t) = \log[m_{W_n}(t)] = n\log[m_Z(t/\sqrt{n})] = nk_Z(t/\sqrt{n}) = \frac{k_Z(t/\sqrt{n})}{\frac{1}{n}}.$$

Now $k_Z(0) = \log[m_Z(0)] = \log(1) = 0$. Thus by L'Hôpital's rule (where the derivative is with respect to n), $\lim_{n\to\infty}\log[m_{W_n}(t)] =$

$$\lim_{n\to\infty}\frac{k_Z(t/\sqrt{n}\,)}{\frac{1}{n}} = \lim_{n\to\infty}\frac{k_Z'(t/\sqrt{n}\,)[\frac{-t/2}{n^{3/2}}]}{(\frac{-1}{n^2})} = \frac{t}{2}\lim_{n\to\infty}\frac{k_Z'(t/\sqrt{n}\,)}{\frac{1}{\sqrt{n}}}.$$

Now $k_Z'(0) = E(Z_i) = 0$, so L'Hôpital's rule can be applied again, giving $\lim_{n\to\infty}\log[m_{W_n}(t)] =$

$$\frac{t}{2}\lim_{n\to\infty}\frac{k_Z''(t/\sqrt{n}\,)[\frac{-t}{2n^{3/2}}]}{(\frac{-1}{2n^{3/2}})} = \frac{t^2}{2}\lim_{n\to\infty}k_Z''(t/\sqrt{n}\,) = \frac{t^2}{2}k_Z''(0).$$

Now $k_Z''(0) = V(Z_i) = 1$. Hence $\lim_{n\to\infty}\log[m_{W_n}(t)] = t^2/2$ and

$$\lim_{n\to\infty} m_{W_n}(t) = \exp(t^2/2)$$

which is the N(0,1) mgf. Thus by the continuity theorem,

$$W_n = \sqrt{n}\left(\frac{\overline{Y}_n-\mu}{\sigma}\right) \xrightarrow{D} N(0,1). \ \square$$

By Theorem 2.34, $d_nF_{g,d_n,1-\delta} \to \chi^2_{g,1-\delta}$ as $d_n \to \infty$. Here $P(X \le \chi^2_{g,1-\delta}) = 1-\delta$ if $X \sim \chi^2_g$, and $P(X \le F_{g,d_n,1-\delta}) = 1-\delta$ if $X \sim F_{g,d_n}$.

Theorem 2.34. If $W_n \sim F_{r,d_n}$ where the positive integer $d_n \to \infty$ as $n \to \infty$, then $rW_n \xrightarrow{D} \chi^2_r$.

Proof. If $X_1 \sim \chi^2_{d_1} \perp\!\!\!\perp X_2 \sim \chi^2_{d_2}$, then

$$\frac{X_1/d_1}{X_2/d_2} \sim F_{d_1,d_2}.$$

If $U_i \sim \chi^2_1$ are iid then $\sum_{i=1}^k U_i \sim \chi^2_k$. Let $d_1 = r$ and $k = d_2 = d_n$. Hence if $X_2 \sim \chi^2_{d_n}$, then

$$\frac{X_2}{d_n} = \frac{\sum_{i=1}^{d_n} U_i}{d_n} = \overline{U} \xrightarrow{P} E(U_i) = 1$$

by the law of large numbers. Hence if $W \sim F_{r,d_n}$, then $rW_n \xrightarrow{D} \chi^2_r$. □

Example 2.17. a) Let $X_n \sim \text{bin}(n, p_n)$ where $np_n = \lambda > 0$ for all positive integers n. Then the mgf $m_{X_n}(t) = (1 - p_n + p_n e^t)^n$ for all t. Thus

$$m_{X_n}(t) = \left(1 - \frac{\lambda}{n} + \frac{\lambda}{n}e^t\right)^n = \left(1 + \frac{\lambda(e^t - 1)}{n}\right)^n \to e^{\lambda(e^t-1)} = m_X(t)$$

for all t where $X \sim POIS(\lambda)$. Hence $X_n \xrightarrow{D} X \sim POIS(\lambda)$ by the continuity theorem.

b) Now let $X_n \sim \text{bin}(n, p_n)$ where $np_n \to \lambda > 0$ as $n \to \infty$. Thus

$$m_{X_n}(t) = \left(1 + \frac{-np_n + np_n e^t}{n}\right)^n \to e^{\lambda(e^t-1)} = m_X(t)$$

for all t since

$$\left(1 + \frac{c_n}{n}\right)^n \to e^c$$

if $c_n \to c$. Here $c = -\lambda + \lambda e^t = \lambda(e^t - 1)$. See Remark 1.4. Hence $X_n \xrightarrow{D} X \sim POIS(\lambda)$ by the continuity theorem.

Note: In the above example, a) is easier, and making assumptions that make the large sample theory easier is a useful techniques.

2.5 Order Relations and Convergence Rates

Definition 2.10. Lehmann (1999, p. 53-54): a) A sequence of random variables W_n is *tight* or *bounded in probability,* written $W_n = O_P(1)$, if for every $\epsilon > 0$ there exist positive constants D_ϵ and N_ϵ such that

$$P(|W_n| \leq D_\epsilon) \geq 1 - \epsilon$$

for all $n \geq N_\epsilon$. Also $W_n = O_P(X_n)$ if $|W_n/X_n| = O_P(1)$.

b) The sequence $W_n = o_P(n^{-\delta})$ if $n^\delta W_n = o_P(1)$ which means that

$$n^{\delta} W_n \xrightarrow{P} 0.$$

c) W_n has the *same order as* X_n *in probability*, written $W_n \asymp_P X_n$, if for every $\epsilon > 0$ there exist positive constants N_ϵ and $0 < d_\epsilon < D_\epsilon$ such that

$$P(d_\epsilon \leq \left| \frac{W_n}{X_n} \right| \leq D_\epsilon) \geq 1 - \epsilon$$

for all $n \geq N_\epsilon$.

d) Similar notation is used for a $k \times r$ matrix $\boldsymbol{A}_n = [a_{i,j}(n)]$ if each element $a_{i,j}(n)$ has the desired property. For example, $\boldsymbol{A}_n = O_P(n^{-1/2})$ if each $a_{i,j}(n) = O_P(n^{-1/2})$.

Definition 2.11. Let $\hat{\boldsymbol{\beta}}_n$ be an estimator of a $p \times 1$ vector $\boldsymbol{\beta}$, and let $W_n = \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\|$.

a) If $W_n \asymp_P n^{-\delta}$ for some $\delta > 0$, then both W_n and $\hat{\boldsymbol{\beta}}_n$ have (tightness) **rate** n^{δ}.

b) If there exists a constant κ such that

$$n^{\delta}(W_n - \kappa) \xrightarrow{D} X$$

for some nondegenerate random variable X, then both W_n and $\hat{\boldsymbol{\beta}}_n$ have *convergence rate* n^{δ}.

Theorem 2.35. Suppose there exists a constant κ such that

$$n^{\delta}(W_n - \kappa) \xrightarrow{D} X.$$

a) Then $W_n = O_P(n^{-\delta})$.

b) If X is not degenerate, then $W_n \asymp_P n^{-\delta}$.

The above result implies that if W_n has convergence rate n^{δ}, then W_n has tightness rate n^{δ}, and the term "tightness" will often be omitted. Part a) is proved, for example, in Lehmann (1999, p. 67).

The following result shows that if $W_n \asymp_P X_n$, then $X_n \asymp_P W_n$, $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$. Notice that if $W_n = O_P(n^{-\delta})$, then n^{δ} is a lower bound on the rate of W_n. As an example, if the CLT holds then $\overline{Y}_n = O_P(n^{-1/3})$, but $\overline{Y}_n \asymp_P n^{-1/2}$.

Theorem 2.36. a) If $W_n \asymp_P X_n$ then $X_n \asymp_P W_n$.

b) If $W_n \asymp_P X_n$ then $W_n = O_P(X_n)$.

c) If $W_n \asymp_P X_n$ then $X_n = O_P(W_n)$.

d) $W_n \asymp_P X_n$ iff $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$.

Proof. a) Since $W_n \asymp_P X_n$,

$$P(d_\epsilon \leq \left|\frac{W_n}{X_n}\right| \leq D_\epsilon) = P(\frac{1}{D_\epsilon} \leq \left|\frac{X_n}{W_n}\right| \leq \frac{1}{d_\epsilon}) \geq 1-\epsilon$$

for all $n \geq N_\epsilon$. Hence $X_n \asymp_P W_n$.

b) Since $W_n \asymp_P X_n$,

$$P(|W_n| \leq |X_n D_\epsilon|) \geq P(d_\epsilon \leq \left|\frac{W_n}{X_n}\right| \leq D_\epsilon) \geq 1-\epsilon$$

for all $n \geq N_\epsilon$. Hence $W_n = O_P(X_n)$.

c) Follows by a) and b).

d) If $W_n \asymp_P X_n$, then $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$ by b) and c). Now suppose $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$. Then

$$P(|W_n| \leq |X_n| D_{\epsilon/2}) \geq 1-\epsilon/2$$

for all $n \geq N_1$, and

$$P(|X_n| \leq |W_n| 1/d_{\epsilon/2}) \geq 1-\epsilon/2$$

for all $n \geq N_2$. Hence

$$P(A) \equiv P(\left|\frac{W_n}{X_n}\right| \leq D_{\epsilon/2}) \geq 1-\epsilon/2$$

and

$$P(B) \equiv P(d_{\epsilon/2} \leq \left|\frac{W_n}{X_n}\right|) \geq 1-\epsilon/2$$

for all $n \geq N = \max(N_1, N_2)$. Since $P(A \cap B) = P(A) + P(B) - P(A \cup B) \geq P(A) + P(B) - 1$,

$$P(A \cap B) = P(d_{\epsilon/2} \leq \left|\frac{W_n}{X_n}\right| \leq D_{\epsilon/2}) \geq 1-\epsilon/2+1-\epsilon/2-1 = 1-\epsilon$$

for all $n \geq N$. Hence $W_n \asymp_P X_n$. □

The following result is used to prove the following Theorem 2.38 which says that if there are K estimators $T_{j,n}$ of a parameter $\boldsymbol{\beta}$, such that $\|T_{j,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ where $0 < \delta \leq 1$, and if T_n^* picks one of these estimators, then $\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta})$.

Theorem 2.37: Pratt (1959). Let $X_{1,n}, ..., X_{K,n}$ each be $O_P(1)$ where K is fixed. Suppose $W_n = X_{i_n,n}$ for some $i_n \in \{1, ..., K\}$. Then

$$W_n = O_P(1). \tag{2.8}$$

Proof.

$$P(\max\{X_{1,n},...,X_{K,n}\} \leq x) = P(X_{1,n} \leq x,...,X_{K,n} \leq x) \leq$$

$$F_{W_n}(x) \leq P(\min\{X_{1,n},...,X_{K,n}\} \leq x) = 1 - P(X_{1,n} > x,...,X_{K,n} > x).$$

Since K is finite, there exists $B > 0$ and N such that $P(X_{i,n} \leq B) > 1-\epsilon/2K$ and $P(X_{i,n} > -B) > 1 - \epsilon/2K$ for all $n > N$ and $i = 1,...,K$. Bonferroni's inequality states that $P(\cap_{i=1}^{K} A_i) \geq \sum_{i=1}^{K} P(A_i) \ - \ (K-1)$. Thus

$$F_{W_n}(B) \geq P(X_{1,n} \leq B,...,X_{K,n} \leq B) \geq$$

$$K(1-\epsilon/2K) - (K-1) = K - \epsilon/2 - K + 1 = 1 - \epsilon/2$$

and

$$-F_{W_n}(-B) \geq -1 + P(X_{1,n} > -B,...,X_{K,n} > -B) \geq$$

$$-1 + K(1-\epsilon/2K) - (K-1) = -1 + K - \epsilon/2 - K + 1 = -\epsilon/2.$$

Hence

$$F_{W_n}(B) - F_{W_n}(-B) \geq 1 - \epsilon \ \text{ for } \ \text{n} > \text{N}. \ \square$$

Theorem 2.38. Suppose $\|T_{j,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ for $j = 1,...,K$ where $0 < \delta \leq 1$. Let $T_n^* = T_{i_n,n}$ for some $i_n \in \{1,...,K\}$ where, for example, $T_{i_n,n}$ is the $T_{j,n}$ that minimized some criterion function. Then

$$\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta}). \tag{2.9}$$

Proof. Let $X_{j,n} = n^{\delta}\|T_{j,n} - \boldsymbol{\beta}\|$. Then $X_{j,n} = O_P(1)$ so by Theorem 2.37, $n^{\delta}\|T_n^* - \boldsymbol{\beta}\| = O_P(1)$. Hence $\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta})$. $\square$

2.6 More CLTs

Remark 2.12. For each positive integer n, let $W_{n1},...,W_{nr_n}$ be independent. The probability space may change with n, giving a triangular array of random variables. Let $E[W_{nk}] = 0$, $V(W_{nk}) = E[W_{nk}^2] = \sigma_{nk}^2$, and $s_n^2 = \sum_{k=1}^{r_n} \sigma_{nk}^2 = V[\sum_{k=1}^{r_n} W_{nk}]$. Then

$$Z_n = \frac{\sum_{k=1}^{r_n} W_{nk}}{s_n}$$

is the z-score of $\sum_{k=1}^{r_n} W_{nk}$.

For the above remark, let $r_n = n$. Then the triangular array is shown below.
W_{11}
W_{21}, W_{22}

W_{31}, W_{32}, W_{33}
$\vdots$
$W_{n1}, W_{n2}, W_{n3}, ..., W_{nn}$
$\vdots$

Theorem 2.39, Lyapounov's CLT: Under Remark 2.12, assume the $|W_{nk}|^{2+\delta}$ are integrable for some $\delta > 0$. Assume Lyapounov's condition:

$$\lim_{n\to\infty} \sum_{k=1}^{r_n} \frac{E[|W_{nk}|^{2+\delta}]}{s_n^{2+\delta}} = 0. \tag{2.10}$$

Then

$$Z_n = \frac{\sum_{k=1}^{r_n} W_{nk}}{s_n} \xrightarrow{D} N(0,1).$$

Theorem 2.39 can be proved using Theorem 2.40. Note that Z_n is the Z-score of $\sum_{k=1}^{r_n} W_{nk}$.

Example 2.18. Special cases: i) $r_n = n$ and $W_{nk} = W_k$ has $W_1, ..., W_n, ...$ independent with $s_n^2 = \sum_{k=1}^n \sigma_k^2$.
ii) $W_{nk} = X_{nk} - E(X_{nk}) = X_{nk} - \mu_{nk}$ has

$$\frac{\sum_{k=1}^{r_n}(X_{nk} - \mu_{nk})}{s_n} \xrightarrow{D} N(0,1).$$

iii) Suppose $X_1, X_2, ...$ are independent with $E(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$. Let

$$Z_n = \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i}{\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2}}$$

be the z-score of $\sum_{i=1}^n X_i$. Assume $E[|X_i - \mu_i|^3] < \infty$ for all $n \in \mathbb{N}$ and

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n E[|X_i - \mu_i|^3]}{\left(\sum_{i=1}^n \sigma_i^2\right)^{3/2}} = 0. \tag{2.11}$$

Then $Z_n \xrightarrow{D} N(0,1)$.

Proof of iii): Take $W_{nk} = X_k - \mu_k$, $\delta = 1$, $s_n^2 = \sum_{k=1}^n \sigma_k^2$, and apply Lyapounov's CLT. Note that

$$\left(\sum_{k=1}^n \sigma_k^2\right)^{3/2} = (s_n^2)^{3/2} = s_n^3 = s_n^{2+1}.$$

□

The (Lindeberg-Lévy) CLT has the X_i iid with $V(X_i) = \sigma^2 < \infty$. The Lyapounov CLT in Example 2.18 iii) has the X_i independent (not necessar-

ily identically distributed), but needs stronger moment conditions to satisfy Equation (2.11) or (2.12).

Theorem 2.40, Lindeberg CLT: Let the W_{nk} satisfy Remark 2.12 and Lindeberg's condition

$$\lim_{n\to\infty} \sum_{k=1}^{r_n} \frac{E(W_{nk}^2\ I[|W_{nk}| \geq \epsilon s_n])}{s_n^2} = 0 \tag{2.12}$$

for any $\epsilon > 0$. Then

$$Z_n = \frac{\sum_{k=1}^{r_n} W_{nk}}{s_n} \xrightarrow{D} N(0,1).$$

Note: The Lindeberg CLT is sometimes called the Lindeberg-Feller CLT. Lindeberg's condition is nearly necessary for $Z_n = \frac{\sum_{k=1}^{r_n} W_{nk}}{s_n} \xrightarrow{D} N(0,1)$.

Example 2.19. a) Special case of the Lindeberg CLT: Let $r_n = n$ and let the $W_{nk} = W_k$ be independent. If

$$\lim_{n\to\infty} \sum_{k=1}^{n} \frac{E(W_k^2\ I[|W_k| \geq \epsilon s_n])}{s_n^2} = 0$$

for any $\epsilon > 0$. Then

$$Z_n = \frac{\sum_{k=1}^{n} W_k}{s_n} \xrightarrow{D} N(0,1).$$

b) **uniformly bounded sequence**: Let $r_n = n$ and $W_{nk} = W_k$. If there is a constant $c > 0$ such that $P(|W_k| < c) = 1\ \forall k$, and if $s_n \to \infty$ as $n \to \infty$, then Lindeberg's CLT holds.

c) Let $r_n = n$ and let the $W_{nk} = W_k$ be **iid** with $V(W_k) = \sigma^2 \in (0, \infty)$. Then Lindeberg's CLT holds. (Taking $W_i = X_i - \mu$ proves the usual CLT with the Lindeberg CLT.)

d) If Lyapunov's condition holds, then Lindeberg's condition holds. Hence the Lindeberg CLT proves the Lyapounov CLT.

Example 2.20. DeGroot (1975, pp. 229-230): Suppose the X_i are independent $\text{Ber}(p_i) \sim \text{bin}(m = 1, p_i)$ random variables with $E(X_i) = p_i$, $V(X_i) = p_i q_i$, $q_i = 1 - p_i$, and $\sum_{i=1}^{\infty} p_i q_i = \infty$. Prove that

$$Z_n = \frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} p_i}{(\sum_{i=1}^{n} p_i q_i)^{1/2}} \xrightarrow{D} N(0,1)$$

as $n \to \infty$.

Proof. Let $Y_i = |W_i| = |X_i - p_i|$. Then $P(Y_i = 1 - p_i) = p_i$ and $P(Y_i = q_i) = q_i$. Thus

$$E[|X_i - p_i|^3] = E[|W_i|^3] = \sum_y y^3 f(y) = (1-p_i)^3 p_i + p_i^3 q_i = q_i^3 p_i + p_i^3 q_i$$

$$= p_i q_i (p_i^2 + q_i^2) \leq p_i q_i$$

since $p_i^2 + q_i^2 \leq (P_i + q_i)^2 = 1$. Thus $\sum_{i=1}^n E[|X_i - p_i|^3] \leq \sum_{i=1}^n p_i q_i$. Dividing both sides by $(\sum_{i=1}^n p_i q_i)^{3/2}$ gives

$$\frac{\sum_{i=1}^n E[|X_i - p_i|^3]}{(\sum_{i=1}^n p_i q_i)^{3/2}} \leq \frac{1}{(\sum_{i=1}^n p_i q_i)^{1/2}} \to 0$$

as $n \to \infty$. Thus Equation (2.12) holds and $Z_n \xrightarrow{D} N(0,1)$. $\square$

Theorem 2.41, Hájek Šidak CLT: Let $X_1, ..., X_n$ be iid with $E(X_i) = \mu$ and $V(X_i) = \sigma^2$. Let $\boldsymbol{c}_n = (c_{n1}, ..., c_{nn})^T$ be a vector of constants such that

$$\max_{1 \leq i \leq n} \frac{c_{ni}^2}{\sum_{j=1}^n c_{nj}^2} \to 0 \text{ as } n \to \infty.$$

Then

$$Z_n = \frac{\sum_{i=1}^n c_{ni}(X_i - \mu)}{\sigma \sqrt{\sum_{j=1}^n c_{nj}^2}} \xrightarrow{D} N(0,1).$$

Note: $c_{ni} = 1/n$ gives the usual CLT.

Example 2.21, Simple Linear Regression. Let $Y_i = \alpha + \beta x_i + e_i$ for $i = 1, ..., n$ where α and β are unknown constants, the x_i are treated as constants, the e_i are unobserved random variables with mean $E(e_i) = 0$ and variance $V(e_i) = \sigma^2$. Then

$$\hat{\boldsymbol{\beta}} = \frac{\sum_{i=1}^n (Y_i - \overline{Y}_n)(x_i - \overline{x}_n)}{\sum_{i=1}^n (x_i - \overline{x}_n)^2} = \frac{\sum_{i=1}^n [\beta(x_i - \overline{x}_n) + e_i - \overline{e}_n](x_i - \overline{x}_n)}{\sum_{i=1}^n (x_i - \overline{x}_n)^2}$$

$$= \beta + \frac{\sum_{i=1}^n e_i (x_i - \overline{x}_n)}{\sum_{i=1}^n (x_i - \overline{x}_n)^2}.$$

So

$$\hat{\beta} - \beta = \frac{\sum_{i=1}^n c_{ni} e_i}{\sum_{j=1}^n c_{nj}^2} = \frac{\sum_{i=1}^n c_{ni} e_i}{\sqrt{\sum_{j=1}^n c_{nj}^2}} \quad \frac{\sigma}{\sigma} \quad \frac{1}{\sqrt{\sum_{j=1}^n c_{nj}^2}}$$

where $c_{ni} = x_i - \overline{x}_n$. So by Theorem 2.41,

$$\sqrt{\sum_{i=1}^n (x_i - \overline{x}_n)^2} \; \frac{\hat{\beta} - \beta}{\sigma} = \frac{\sum_{i=1}^n c_{ni} e_i}{\sigma \sqrt{\sum_{j=1}^n c_{nj}^2}} \xrightarrow{D} N(0,1)$$

if

$$\max_{1\le i\le n} \frac{(x_i - \overline{x}_n)^2}{\sum_{j=1}^n (x_j - \overline{x}_n)^2} \to 0$$

as $n \to \infty$. Thus

$$\sqrt{n}\sqrt{\frac{\sum_{i=1}^n (x_i - \overline{x}_n)^2}{n}}\,\frac{\sigma}{\sigma}(\hat{\beta} - \beta) \xrightarrow{D} N(0, \sigma^2),$$

or

$$\hat{\beta} \sim AN\left(\beta, \frac{MSE}{n}\frac{S_x^2}{}\right).$$

Note that we do not need the sample variance of the x_i to satisfy $S_x^2 \xrightarrow{P} \sigma_x^2$ where $V(x_i) = \sigma_x^2$ for all i.

2.7 The Plug-In Principle

Suppose that $X_n \xrightarrow{D} X = X_{\boldsymbol{\tau}} \sim D(\boldsymbol{\tau})$ where the distribution of X depends on unknown parameters $\boldsymbol{\tau}$. The plug-in principle says approximate the distribution of $X_{\boldsymbol{\tau}}$ by $Z_n = X_{\hat{\boldsymbol{\tau}}} \sim D(\hat{\boldsymbol{\tau}})$ where $\hat{\boldsymbol{\tau}}$ is a consistent estimator of $\boldsymbol{\tau}$. Then Z_n is often used to make large sample confidence intervals and for large sample tests of hypotheses. For example, let $X_n = \sqrt{n}(T_n - \theta)$.

The plug-in principle is also often used to get an asymptotic normal approximation for a statistic, and often the bootstrap confidence regions are closely related to the plug-in principle. For the CLT, $X \sim N(0, \sigma^2)$ and $Z_n \sim N(0, S_n^2)$. For the MLE, $X \sim N(0, 1/I_1(\theta))$ and $Z_n \sim N(0, 1/I_1(\hat{\theta}_n))$ where $\hat{\theta}_n$ is the MLE of θ.

It is not clear whether Z_n converge in distribution to X. To see this, consider $X = X_\theta \sim U(0, \theta)$ distribution. a) Let $X_1, ..., X_n$ be iid $U(0, \theta)$ and use the plug-in estimator "$Z_n \sim U(0, X_{(n)})$" where $X_{(n)} = \max(X_1, ..., X_n)$.For this uniform distribution, the cdf $F_X(t) = 0$ for $t < 0$, $F_X(t) = t/\theta$ for $0 \le t \le \theta$ and $F_X(t) = 1$ for $t > \theta$. Now $F_{Z_n}(t) = t/X_{(n)}$ for $0 \le t \le X_{(n)}$. Suppose $\theta = 10$. For $F_{Z_n}(10)$ to converge to $F_X(10) = 1$, for any $\epsilon > 0$, there must exist a positive integer N_ϵ such that for $n \ge N_\epsilon$, we have $|F_{Z_n}(10) - 1| < \epsilon$. For $0 < \epsilon \le 1$, this result requires $X_{(n)} > 1 - \epsilon$, but $X_{(n)} < 1 - \epsilon$ with nonzero probability $(10 - \epsilon)^n$ = probability that all n of the $X_i < 1 - \epsilon$. Thus $U(0, X_{(n)})$ can't converge in distribution to X. Note that $F_{Z_n}(t)$ is a consistent estimator of $F_X(t)$ for any t. b) On the other hand, the $U(0, \theta)$ distribution is a scale family. If we interpret the "$U(0, X_{(n)})$" distribution as the $X_{(n)}U(0, 1)$ distribution, then $Z_n = X_{(n)}U(0, 1) \xrightarrow{D} \theta U(0, 1) \sim U(0, \theta)$.

The problem with a) is how to interpret a distribution when a parameter is replaced by a random variable. For a scale family, the interpretation in b)

makes more sense, but not all distributions are scale families. Note that the $N(0, \sigma^2)$ distribution is a scale family with scale parameter $\sigma > 0$.

Thus the plug-in principle approximation $Z_n = X_{\hat{\boldsymbol{\tau}}_n} \sim D(\hat{\boldsymbol{\tau}}_n)$ for $X = X_{\boldsymbol{\tau}} \sim D(\boldsymbol{\tau})$ appears to be weaker that convergence in distribution. We may use the notation $Z_n \xrightarrow{C} X$ when $\hat{\boldsymbol{\tau}}_n$ is a consistent estimator of $\boldsymbol{\tau}$.

2.8 Summary

1) **CLT**: Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and $V(Y) = \sigma^2$. Then $\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{D} N(0, \sigma^2)$.

2 a) $Z_n = \sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) = \left(\frac{\overline{Y}_n - \mu}{\sigma/\sqrt{n}}\right) = \left(\frac{\sum_{i=1}^n Y_i - n\mu}{\sqrt{n}\sigma}\right)$ is the z–score of $\overline{X}_n$ (and the z-score of $\sum_{i=1}^n Y_i$), and $Z_n \xrightarrow{D} N(0, 1)$. b) Two applications of the CLT are to give the limiting distribution of $\sqrt{n}(\overline{Y}_n - \mu)$ and the limiting distribution of $\sqrt{n}(Y_n/n - \mu_Y)$ for a random variable Y_n such that $Y_n = \sum_{i=1}^n X_i$ where the X_i are iid with $E(X) = \mu_X$ and $V(X) = \sigma_X^2$. See Section 1.4. c) The CLT is the Lindeberg-Lévy CLT.

3) **Delta Method**: If $g'(\theta) \neq 0$ and $\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma^2)$, then $\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{D} N(0, \sigma^2[g'(\theta)]^2)$.

4) **Second Order Delta Method**: Suppose that $g'(\theta) = 0$, $g''(\theta) \neq 0$ and $\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \tau^2(\theta))$. Then $n[g(T_n) - g(\theta)] \xrightarrow{D} \frac{1}{2}\tau^2(\theta)g''(\theta)\chi_1^2$.

5) **1P–REF Limit Theorem**: Let $Y_1, ..., Y_n$ be iid from a 1P–REF with pdf or pmf $f(y|\theta) = h(y)c(\theta)\exp[w(\theta)t(y)]$ and natural parameterization $f(y|\eta) = h(y)b(\eta)\exp[\eta t(y)]$. Let $E(t(Y)) = \mu_t \equiv g(\eta)$ and $V(t(Y)) = \sigma_t^2$. Then $\sqrt{n}[\frac{1}{n}\sum_{i=1}^n t(Y_i) - \mu_t] \xrightarrow{D} N(0, I_1(\eta))$ where $I_1(\eta) = \sigma_t^2 = g'(\eta)$.

6) **Limit theorem for the Sample Median:**
$\sqrt{n}(MED(n) - MED(Y)) \xrightarrow{D} N\left(0, \frac{1}{4f^2(MED(Y))}\right)$.

7) If $n^\delta(T_{1,n} - \theta) \xrightarrow{D} N(0, \sigma_1^2(F))$ and $n^\delta(T_{2,n} - \theta) \xrightarrow{D} N(0, \sigma_2^2(F))$, then the **asymptotic relative efficiency** of $T_{1,n}$ with respect to $T_{2,n}$ is

$$ARE(T_{1,n}, T_{2,n}) = \frac{\sigma_2^2(F)}{\sigma_1^2(F)}.$$

The "better" estimator has the smaller asymptotic variance or $\sigma_i^2(F)$.

8) An estimator T_n of $\tau(\theta)$ is **asymptotically efficient** if

$$\sqrt{n}(T_n - \tau(\theta)) \xrightarrow{D} N\left(0, \frac{[\tau'(\theta)]^2}{I_1(\theta)}\right).$$

9) For a 1P–REF, $\frac{1}{n}\sum_{i=1}^{n} t(Y_i)$ is an asymptotically efficient estimator of $g(\eta) = E(t(Y))$.

10) Standard Limit Theorem: Under strong regularity conditions, if $\hat{\theta}_n$ is the MLE or UMVUE of θ, then $T_n = \tau(\hat{\theta}_n)$ is an asymptotically efficient estimator of $\tau(\theta)$. Hence if $\tau'(\theta) \neq 0$, then

$$\sqrt{n}[\tau(\hat{\theta}_n) - \tau(\theta)] \xrightarrow{D} N\left(0, \frac{[\tau'(\theta)]^2}{I_1(\theta)}\right).$$

11) $X_n \xrightarrow{D} X$ if

$$\lim_{n\to\infty} F_n(t) = F(t)$$

at each continuity point t of F. **Convergence in distribution** is also known as weak convergence and convergence in law. X is the limiting distribution or asymptotic distribution of X_n. **The limiting distribution does not depend on** the sample size n. $X_n \xrightarrow{D} \tau(\theta)$ if $X_n \xrightarrow{D} X$ where $P(X = \tau(\theta)) = 1$: hence X is *degenerate at* $\tau(\theta)$ or the distribution of X is a *point mass at* $\tau(\theta)$.

12) If $X_n \xrightarrow{D} X$ and $X_n \xrightarrow{D} Y$, then i) $X \stackrel{D}{=} Y$ and ii) $F_X(x) = F_Y(x)$ for all real x.

13) **Convergence in probability**: a) $X_n \xrightarrow{P} \tau(\theta)$ if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - \tau(\theta)| < \epsilon) = 1 \text{ or, equivalently, } \lim_{\mathrm{n}\to\infty} \mathrm{P}(|\mathrm{X_n} - \tau(\theta)| \geq \epsilon) = 0.$$

b) $X_n \xrightarrow{P} X$ if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1, \text{ or, equivalently, } \lim_{\mathrm{n}\to\infty} \mathrm{P}(|\mathrm{X_n} - \mathrm{X}| \geq \epsilon) = 0.$$

14) T_n is a **consistent estimator** of $\tau(\theta)$ if $T_n \xrightarrow{P} \tau(\theta)$ for every $\theta \in \Theta$.

15) Theorem: T_n is a **consistent estimator** of $\tau(\theta)$ if any of the following 2 conditions holds:

i) $\lim_{n\to\infty} V_\theta(T_n) = 0$ and $\lim_{n\to\infty} E_\theta(T_n) = \tau(\theta)$ for all $\theta \in \Theta$.

ii) $MSE_{\tau(\theta)}(T_n) = E[(T_n - \tau(\theta))^2] \to 0$ for all $\theta \in \Theta$.

Here

$$MSE_{\tau(\theta)}(T_n) = V_\theta(T_n) + [\mathrm{Bias}_{\tau(\theta)}(\mathrm{T_n})]^2$$

where $\mathrm{Bias}_{\tau(\theta)}(\mathrm{T_n}) = \mathrm{E}_\theta(\mathrm{T_n}) - \tau(\theta)$.

16) Theorem: a) Let X_θ be a random variable with a distribution depending on θ, and $0 < \delta \leq 1$. If

$$n^\delta(T_n - \tau(\theta)) \xrightarrow{D} X_\theta$$

for all $\theta \in \Theta$, then $T_n \xrightarrow{P} \tau(\theta)$.

b) If

$$\sqrt{n}(T_n - \tau(\theta)) \xrightarrow{D} N(0, v(\theta))$$

for all $\theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$.

Note: If $\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma^2)$, then $T_n \xrightarrow{P} \theta$. Often $X_\theta \sim N(0, v(\theta))$.

17) **WLLN:** Let $Y_1, ..., Y_n, ...$ be a sequence of iid random variables with $E(Y_i) = \mu$. Then $\overline{Y}_n \xrightarrow{P} \mu$. Hence $\overline{Y}_n$ is a consistent estimator of μ.

18) Y_n **converges in rth mean** to a random variable Y, $Y_n \xrightarrow{r} Y$, if

$$E(|Y_n - Y|^r) \to 0$$

as $n \to \infty$. In particular, if $r = 2$, Y_n **converges in quadratic mean** to Y, written

$$Y_n \xrightarrow{2} Y \text{ or } \mathrm{Y_n} \xrightarrow{\mathrm{qm}} \mathrm{Y},$$

if $E[(Y_n - Y)^2] \to 0$ as $n \to \infty$. $Y_n \xrightarrow{r} \tau(\theta)$ if $E(|Y_n - \tau(\theta)|^r) \to 0$ as $n \to \infty$. If $r \geq 1$, $Y_n \xrightarrow{r} Y$ is often written as $Y_n \xrightarrow{L^r} Y$ or $Y_n \xrightarrow{L_r} Y$.

19) A sequence of random variables X_n *converges with probability 1* (or *almost surely*, or *almost everywhere*, or *strong convergence*) to X if

$$P(\lim_{n\to\infty} X_n = X) = 1.$$

This type of convergence will be denoted by $X_n \xrightarrow{wp1} X$. Notation such as "X_n converges to X wp1" will also be used. Sometimes "wp1" will be replaced with "as" or "ae."

$$X_n \xrightarrow{wp1} \tau(\theta),$$

if $P(\lim_{n\to\infty} X_n = \tau(\theta)) = 1$.

20) **SLLN**: If $X_1, ..., X_n$ are iid with $E(X_i) = \mu$ finite, then $\overline{X}_n \xrightarrow{wp1} \mu$.

21) a) For i) $X_n \xrightarrow{P} X$, ii) $X_n \xrightarrow{r} X$, or iii) $X_n \xrightarrow{wp1} X$, the X_n and X need to be defined on the same probability space.

b) For $X_n \xrightarrow{D} X$, the probability spaces can differ.

c) For i) $X_n \xrightarrow{P} c$, ii) $X_n \xrightarrow{wp1} c$, iii) $X_n \xrightarrow{D} c$, and iv) $X_n \xrightarrow{r} c$, the probability spaces of the X_n can differ.

22) Theorem: i) $T_n \xrightarrow{P} \tau(\theta)$ iff $\mathrm{T_n} \xrightarrow{\mathrm{D}} \tau(\theta)$.

ii) If $T_n \xrightarrow{P} \theta$ and τ is continuous at θ, then $\tau(T_n) \xrightarrow{P} \tau(\theta)$. Hence if T_n is a consistent estimator of θ, then $\tau(T_n)$ is a consistent estimator of $\tau(\theta)$ if τ is a continuous function on Θ.

23) Theorem: Suppose X_n and X are RVs with the same probability space for b) and c). Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function.

a) If $X_n \xrightarrow{D} X$, then $g(X_n) \xrightarrow{D} g(X)$.

b) If $X_n \xrightarrow{P} X$, then $g(X_n) \xrightarrow{P} g(X)$.

c) If $X_n \xrightarrow{ae} X$, then $g(X_n) \xrightarrow{wp1} g(X)$.

24) Theorem: Suppose X_n and X are RVs with the same probability space.

a) If $X_n \xrightarrow{wp1} X$, then $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{D} X$.

b) If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{D} X$.

c) If $X_n \xrightarrow{r} X$, then $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{D} X$.

d) $X_n \xrightarrow{P} \tau(\theta)$ **iff** $X_n \xrightarrow{D} \tau(\theta)$ where c is a constant.

25) Theorem: a) If $E[(X_n - X)^2] \to 0$ as $n \to \infty$, then $X_n \xrightarrow{P} X$.

b) If $E(X_n) \to E(X)$ and $V(X_n - X) \to 0$ as $n \to \infty$, then $X_n \xrightarrow{P} X$.

Note: See 15) if $P(X = \tau(\theta)) = 1$.

26) Theorem: If $X_n \xrightarrow{r} X$, then $X_n \xrightarrow{k} X$ where $0 < k < r$.

27) Theorem: Let X_n have pdf $f_{X_n}(x)$, and let X have pdf $f_X(x)$. If $f_{X_n}(x) \to f_X(x)$ for all x (or for x outside of a set of Lebesgue measure 0), then $X_n \xrightarrow{D} X$.

28) Theorem: Let $g : \mathbb{R} \to \mathbb{R}$ be continuous at constant c.

a) If $X_n \xrightarrow{D} c$, then $g(X_n) \xrightarrow{D} g(c)$.

b) If $X_n \xrightarrow{P} c$, then $g(X_n) \xrightarrow{P} g(c)$.

c) If $X_n \xrightarrow{wp1} c$, then $g(X_n) \xrightarrow{wp1} g(c)$.

Note: If $X_n \xrightarrow{r} c$, then $X_n \xrightarrow{P} c$ and $g(X_n) \xrightarrow{P} g(c)$.

29) Theorem: Suppose X_n and X are integer valued RVs with pmfs $f_{X_n}(x)$ and $f_X(x)$. Then $X_n \xrightarrow{D} X$ iff $P(X_n = k) \to P(X = k)$ for every integer k iff $f_{X_n}(x) \to f_X(x)$ for every real x.

30) **Slutsky's Theorem**: If $Y_n \xrightarrow{D} Y$ and $W_n \xrightarrow{P} w$ for some constant w, then i) $Y_n W_n \xrightarrow{D} wY$, ii) $Y_n + W_n \xrightarrow{D} Y + w$ and iii) $Y_n / W_n \xrightarrow{D} Y/w$ for $w \neq 0$.

Note that $Y_n \xrightarrow{B} Y$ implies $Y_n \xrightarrow{D} Y$ where $B = wp1, r$, or P. Also $W_n \xrightarrow{P} c$ iff $W_n \xrightarrow{D} c$. If a sequence of constants $c_n \to c$ as $n \to \infty$ (everywhere convergence), then $c_n \xrightarrow{wp1} c$ and $c_n \xrightarrow{P} c$. (So everywhere convergence is a special case of almost everywhere convergence.)

31) The **cumulative distribution function** (cdf) of any random variable Y is $F(y) = P(Y \leq y)$ for all $y \in \mathbb{R}$. If $F(y)$ is a cumulative distribution function, then i) $F(-\infty) = \lim_{y \to -\infty} F(y) = 0$, ii) $F(\infty) = \lim_{y \to \infty} F(y) = 1$, iii) F is a nondecreasing function: if $y_1 < y_2$, then $F(y_1) \leq F(y_2)$, iv) F is right continuous: $\lim_{h \downarrow 0} F(y+h) = F(y)$ for all real y. v) Since a cdf is a probability for fixed y, $0 \leq F(y) \leq 1$ for all real y. vi) A cdf $F(y)$ can have at most countably many points of discontinuity, vii) $P(a < Y \leq b) = F(b) - F(a)$. viii) If Y is a random variable, then $F_Y(y)$ completely determines the distribution of Y.

32) The **moment generating function** (mgf) of a random variable Y is

$$m(t) = E[e^{tY}] \tag{2.13}$$

if the expectation exists for t in some neighborhood of 0. Otherwise, the mgf does not exist. If Y is discrete, then $m(t) = \sum_y e^{ty} f(y)$, and if Y is continuous, then $m(t) = \int_{-\infty}^{\infty} e^{ty} f(y) dy$. If Y is a random variable and $m_Y(t)$ exists, then $m_Y(t)$ completely determines the distribution of Y.

Notes: a) If X has mgf $m_X(t)$, then $E(X^k)$ exists for all positive integers k.

b) Let j and k be positive integers. If $E(X^k)$ is finite, then $E(X^j)$ is finite for $1 \leq j \leq k$.

33) The **characteristic function** of a random variable Y is $c(t) = E[e^{itY}] = E[\cos(tY)] + iE[\sin(tY)]$ where the complex number $i = \sqrt{-1}$. i) $c(0) = 1$, ii) the modulus $|c(t)| \leq 1$ for all real t, iii) $c(t)$ is a continuous function. iv) If $E(Y) = 0$ and $E(Y^2) = V(Y) = \sigma^2$, then

$$c_Y(t) = 1 + \frac{t^2\sigma^2}{2} + o(t^2) \text{ as } t \to 0.$$

Here $a(t) = o(t^2)$ as $t \to 0$ if $\lim_{t \to 0} \frac{a(t)}{t^2} = 0$. v) If Y is discrete with pmf $f_Y(y)$, then $c_Y(t) = \sum_y e^{ity} f_y(y)$. vi) If Y is a random variable, then $c_Y(t)$ always exists, and completely determines the distribution of Y.

34) **Continuity Theorem:** Let Y_n be sequence of random variables with characteristic functions $c_{Y_n}(t)$. Let Y be a random variable with cf $c_Y(t)$.

a)

$$Y_n \xrightarrow{D} Y \text{ iff } c_{Y_n}(t) \to c_Y(t) \ \forall t \in \mathbb{R}.$$

b) Also assume that Y_n has mgf m_{Y_n} and Y has mgf m_Y. Assume that all of the mgfs m_{Y_n} and m_Y are defined on $|t| \leq d$ for some $d > 0$. Then if $m_{Y_n}(t) \to m_Y(t)$ as $n \to \infty$ for all $|t| < c$ where $0 < c < d$, then $Y_n \xrightarrow{D} Y$.

35) Theorem: If $\lim_{n \to \infty} c_{X_n}(t) = g(t)$ for all t where g is continuous at $t = 0$, then $g(t) = c_X(t)$ is a characteristic function for some RV X, and $X_n \xrightarrow{D} X$.

Note: Hence continuity at $t = 0$ implies continuity everywhere since $g(t) = \varphi_X(t)$ is continuous. If $g(t)$ is not continuous at 0, then X_n does not converge in distribution.

36) If $c_{Y_n}(t) \to h(t)$ where $h(t)$ is not continuous, then Y_n does not converge in distribution to any RV Y, by the Continuity Theorem and 35).

37) Let $X_1, ..., X_n$ be independent RVs with characteristic functions $c_{X_j}(t)$. Then the characteristic function of $\sum_{j=1}^n X_j$ is $c_{\sum_{j=1}^n X_j}(t) = \prod_{j=1}^n c_{X_j}(t)$. If the RVs also have mgfs $m_{X_j}(t)$, then the mgf of $\sum_{j=1}^n X_j$ is $m_{\sum_{j=1}^n X_j}(t) = \prod_{j=1}^n m_{X_j}(t)$.

38) **Helly-Bray-Pormanteau Theorem**: $X_n \xrightarrow{D} X$ iff $E[g(X_n)] \to E[g(X)]$ for every bounded, real, continuous function g.

Note: 38) is used to prove 39 b).

39) a) **Generalized Continuous Mapping Theorem:** If $X_n \xrightarrow{D} X$ and the function g is such that $P[X \in C(g)] = 1$ where $C(g)$ is the set of points where g is continuous, then $g(X_n) \xrightarrow{D} g(X)$.

Note: $P[X \in C(g)] = 1$ can be replaced by $P[X \in D(g)] = 0$ where $D(g)$ is the set of points where g is not continuous.

b) **Continuous Mapping Theorem:** If $X_n \xrightarrow{D} X$ and the function g is continuous, then $g(X_n) \xrightarrow{D} g(X)$.

Note: the function g can not depend on n since g_n is a seqeuce of functions rather than a single function.

40) Generalized Chebyshev's Inequality or Generalized Markov's Inequality: Let $u : \mathbb{R} \to [0, \infty)$ be a nonnegative function. If $E[u(Y)]$ exists then for any $c > 0$,

$$P[u(Y) \geq c] \leq \frac{E[u(Y)]}{c}.$$

If $\mu = E(Y)$ exists, then taking $u(y) = |y - \mu|^r$ and $\tilde{c} = c^r$ gives
Markov's Inequality: for $r > 0$ and any $c > 0$,

$$P(|Y - \mu| \geq c] = P(|Y - \mu|^r \geq c^r] \leq \frac{E[|Y - \mu|^r]}{c^r}.$$

If $r = 2$ and $\sigma^2 = V(Y)$ exists, then we obtain
Chebyshev's Inequality:

$$P(|Y - \mu| \geq c] \leq \frac{V(Y)}{c^2}.$$

41) a) $\lim_{n\to\infty} \left(1 - \frac{c}{n}\right)^n = e^{-c}$.

b) If $c_n \to c$ as $n \to \infty$, then $\lim_{n\to\infty} \left(1 + \frac{-c_n}{n}\right)^n = e^{-c}$.

c) If c_n is a sequence of complex numbers such that $c_n \to c$ as $n \to \infty$ where c is real, then $\lim_{n\to\infty} \left(1 - \frac{c_n}{n}\right)^n = e^{-c}$.

42) For each positive integer n, let $W_{n1}, ..., W_{nr_n}$ be independent. The probability space may change with n, giving a triangular array of RVs. Let $E[W_{nk}] = 0$, $V(W_{nk}) = E[W_{nk}^2] = \sigma_{nk}^2$, and $s_n^2 = \sum_{k=1}^{r_n} \sigma_{nk}^2 = V[\sum_{k=1}^{r_n} W_{nk}]$. Then

$$Z_n = \frac{\sum_{k=1}^{r_n} W_{nk}}{s_n}$$

is the z-score of $\sum_{k=1}^{r_n} W_{nk}$.

43) **Lyapounov's CLT**: Under 42), assume the $|W_{nk}|^{2+\delta}$ are integrable for some $\delta > 0$. Assume Lyapounov's condition:

$$\lim_{n\to\infty} \sum_{k=1}^{r_n} \frac{E[|W_{nk}|^{2+\delta}]}{s_n^{2+\delta}} = 0.$$

Then

$$Z_n = \frac{\sum_{k=1}^{r_n} W_{nk}}{s_n} \xrightarrow{D} N(0,1).$$

44) Special cases: i) $r_n = n$ and $W_{nk} = W_k$ has $W_1, ..., W_n, ...$ independent. ii) $W_{nk} = X_{nk} - E(X_{nk}) = X_{nk} - \mu_{nk}$ has

$$\frac{\sum_{k=1}^{r_n}(X_{nk} - \mu_{nk})}{s_n} \xrightarrow{D} N(0,1).$$

iii) Suppose $X_1, X_2, ...$ are independent with $E(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$. Let

$$Z_n = \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i}{\left(\sum_{i=1}^n \sigma_i^2\right)^{1/2}}$$

be the z-score of $\sum_{i=1}^n X_i$. Assume $E[|X_i - \mu_i|^3] < \infty$ for all $n \in \mathbb{N}$ and

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n E[|X_i - \mu_i|^3]}{\left(\sum_{i=1}^n \sigma_i^2\right)^{3/2}} = 0. \qquad (*)$$

Then $Z_n \xrightarrow{D} N(0,1)$.

45) The (Lindeberg-Lévy) CLT has the X_i iid with $V(X_i) = \sigma^2 < \infty$. The Lyapounov CLT in 43 iii) has the X_i independent (not necessarily identically distributed), but needs stronger moment conditions to satisfy $(*)$.

46) **Lindeberg CLT**: Let the W_{nk} satisfy 42) and Lindeberg's condition

$$\lim_{n\to\infty} \sum_{k=1}^{r_n} \frac{E(W_{nk}^2\ I[|W_{nk}| \geq \epsilon s_n])}{s_n^2} = 0$$

for any $\epsilon > 0$. Then

$$Z_n = \frac{\sum_{k=1}^{r_n} W_{nk}}{s_n} \xrightarrow{D} N(0,1).$$

Notes: The Lindeberg CLT is sometimes called the Lindeberg-Feller CLT. Lindeberg's condition is nearly necessary for $Z_n = \frac{\sum_{k=1}^{r_n} W_{nk}}{s_n} \xrightarrow{D} N(0,1)$.

47) Special case of the Lindeberg CLT: Let $r_n = n$ and let the $W_{nk} = W_k$ be independent. If

$$\lim_{n\to\infty} \sum_{k=1}^{n} \frac{E(W_k^2\ I[|W_k| \geq \epsilon s_n])}{s_n^2} = 0$$

for any $\epsilon > 0$. Then

$$Z_n = \frac{\sum_{k=1}^{n} W_k}{s_n} \xrightarrow{D} N(0,1).$$

48) a) **uniformly bounded sequence**: Let $r_n = n$ and $W_{nk} = W_k$. If there is a constant $c > 0$ such that $P(|W_k| < c) = 1\ \forall k$, and if $s_n \to \infty$ as $n \to \infty$, then Lindeberg's CLT 46) holds.

b) Let $r_n = n$ and let the $W_{nk} = W_k$ be **iid** with $V(W_k) = \sigma^2 \in (0, \infty)$. Then Lindeberg's CLT 46) holds. (Taking $W_i = X_i - \mu$ proves the usual CLT with the Lindeberg CLT.)

c) If Lyapunov's condition holds, then Lindeberg's condition holds. Hence the Lindeberg CLT proves the Lyapounov CLT.

2.9 Complements

In analysis, convergence in probability is a special case of convergence in measure and convergence in distribution is a special case of weak convergence. See Ash (1972, p. 322) and Sen and Singer (1993, p. 39). Since $\overline{Y} \xrightarrow{P} \mu$ iff $\overline{Y} \xrightarrow{D} \mu$, the WLLN refers to weak convergence. Almost sure convergence is also called strong convergence. Hence the SLLN refers to strong convergence. Technically the X_n and X need to share a common probability space for convergence in probability and almost sure convergence.

Perlman (1972) and Wald (1949) give general results on the consistency of the MLE while Berk (1972), Lehmann (1980), and Schervish (1995, p. 418) discuss the asymptotic normality of the MLE in exponential families. Theorem 2.5 appears in Olive (2014). Portnoy (1977) gives large sample theory for unbiased estimators in exponential families. Although $\overline{T}_n$ is the UMVUE of $E(t(Y)) = \mu_t$, asymptotic efficiency of UMVUEs is not simple in general. See Pfanzagl (1993).

Casella and Berger (2002, p. 112, 133) give results similar to Theorem 2.4. Some of the order relations of Section 2.5 are discussed in Mann and Wald (1943a). See Ver Hoef (2012) for history of the delta method.

Bickel and Doksum (1977, pp. 135-137) has a useful theorem for method of moments estimators based on iid $Y_1, ..., Y_n$: let $m_k = m_k(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(Y^k)$ and let $T_n = g(\hat{m}_1, ..., \hat{m}_r)$ be the method of moments estimator of $q(\boldsymbol{\theta}) = g(\boldsymbol{m}) = g(m_1, ..., m_r)$ where $g : \mathbb{R}^r \to \mathbb{R}$. If $m_{2r} < \infty$, then

$$\sqrt{n}(T_n - g(\boldsymbol{m})) \xrightarrow{D} N(0, \sigma^2_{MM})$$

where

$$\sigma^2_{MM} = V(\sum_{k=1}^{r} \frac{\partial}{\partial m_k} g(\boldsymbol{m}) Y^k).$$

For the following theorem, see Proschan and Shaw (2016, p. 189). Note that $a_n = \sqrt{n} = n^{1/2}$ is common. If $X \sim N(0, \sigma^2)$ and $g'(\theta) = 0$, then $a_n[g(\hat{\theta}) - g(\theta)] \xrightarrow{D} 0 \sim N(0, 0)$, the point mass at 0. Note that $g'(\theta)N(0, \sigma^2) \sim N(0, [g'(\theta)]^2\sigma^2)$.

Theorem 2.42, Generalized Delta Method: Let a_n be a sequence of constants such that $a_n \to \infty$ as $n \to \infty$. Suppose that $X_n = a_n(\hat{\theta} - \theta) \xrightarrow{D} X$. Let $g(x)$ be a function with derivative $g'(\theta)$ at $x = \theta$. Then
$a_n[g(\hat{\theta}) - g(\theta)] \xrightarrow{D} g'(\theta)X$.

There are many variants of the WLLN. The following theorem gives some examples.

Theorem 2.43. Suppose $X_1, ..., X_n$ are jointly distributed random variables.

a) If $E(X_i) \equiv \mu$ and $V(\overline{X}_n) \to 0$ as $n \to \infty$, then $\overline{X}_n \xrightarrow{P} \mu$ as $n \to \infty$.

b) Suppose $X_1, ..., X_n$ are uncorrelated random variables with $E(X_i) \equiv \mu$ and $V(X_i) = \sigma_i^2$. If $\sum_{i=1}^{n} \sigma_i^2/n^2 \to 0$ as $n \to \infty$, then $\overline{X}_n \xrightarrow{P} \mu$ as $n \to \infty$.

c) If $E(\overline{X}_n) \to \mu$ and $V(\overline{X}_n) \to 0$ as $n \to \infty$, then $\overline{X}_n \xrightarrow{P} \mu$ as $n \to \infty$.

Proof. By Chebyshev's inequality, $P(|\overline{X}_n - \mu| \geq \epsilon) \leq V(\overline{X}_n)/\epsilon^2$ for any $\epsilon > 0$. Hence the result follows if $V(\overline{X}_n) \to 0$ as $n \to 0$. Thus a) holds by assumption.

b) Now $V(\overline{X}_n) = V(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2} V(\sum_{i=1}^{n} X_i) = \frac{\sum_{i=1}^{n} \sigma_i^2}{n^2} \to 0$ as $n \to \infty$.

c) This result follows by Theorem 2.9 b). Note that this result also implies that $\overline{X}_n \xrightarrow{2} \mu$, and note that a) and b) follow from c).

2.10 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

Problems with a Q have appeared on Statistical Inference or Probability and Measure qualifying exams.

Refer to Section 1.10 for the pdf or pmf of the distributions in the problems below.

2.1*. a) Enter the following *R* function that is used to illustrate the central limit theorem when the data $Y_1, ..., Y_n$ are iid from an exponential distribution. The function generates a data set of size n and computes $\overline{Y}_1$ from the data set. This step is repeated $nruns = 100$ times. The output is a vector

$(\overline{Y}_1, \overline{Y}_2, ..., \overline{Y}_{100})$. A histogram of these means should resemble a symmetric normal density once n is large enough.

```
cltsim <- function(n=100, nruns=100){
ybar <- 1:nruns
for(i in 1:nruns){
  ybar[i] <- mean(rexp(n))}
list(ybar=ybar)}
```

b) The following commands will plot 4 histograms with $n = 1, 5, 25$ and 200. Save the plot in *Word*.

```
> z1 <- cltsim(n=1)
> z5 <- cltsim(n=5)
> z25 <- cltsim(n=25)
> z200 <- cltsim(n=200)
> par(mfrow=c(2,2))
> hist(z1$ybar)
> hist(z5$ybar)
> hist(z25$ybar)
> hist(z200$ybar)
```

c) Explain how your plot illustrates the central limit theorem.

d) Repeat parts a), b) and c), but in part a), change *rexp(n)* to *rnorm(n)*. Then $Y_1, ..., Y_n$ are iid N(0,1) and $\overline{Y} \sim N(0, 1/n)$.

2.2*. Let $X_1, ..., X_n$ be iid from a normal distribution with unknown mean μ and known variance σ^2. Let

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Find the limiting distribution of $\sqrt{n}((\overline{X})^3 - c)$ for an appropriate constant c.

2.3^{*Q}. Let $X_1, ..., X_n$ be a random sample from a population with pdf

$$f(x) = \begin{cases} \frac{\theta x^{\theta-1}}{3^\theta} & 0 < x < 3 \\ 0 & \text{elsewhere} \end{cases}$$

The method of moments estimator for θ is $T_n = \dfrac{\overline{X}}{3 - \overline{X}}$.

a) Find the limiting distribution of $\sqrt{n}(T_n - \theta)$ as $n \to \infty$.

b) Is T_n asymptotically efficient? Why?

c) Find a consistent estimator for θ and show that it is consistent.

2.4*. From Theorems 1.24 and 1.25,

if $Y_n = \sum_{i=1}^{n} X_i$ where the X_i are iid from a nice distribution, then Y_n also has a nice distribution. If $E(X) = \mu$ and $V(X) = \sigma^2$ then by the CLT

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Hence

$$\sqrt{n}\left(\frac{Y_n}{n} - \mu\right) \xrightarrow{D} N(0, \sigma^2).$$

Find μ, σ^2 and the distribution of X_i if

i) $Y_n \sim \text{BIN}(n, \rho)$ where BIN stands for binomial.

ii) $Y_n \sim \chi^2_n$.

iii) $Y_n \sim G(n\nu, \lambda)$ where G stands for gamma.

iv) $Y_n \sim NB(n, \rho)$ where NB stands for negative binomial.

v) $Y_n \sim POIS(n\theta)$ where POIS stands for Poisson.

vi) $Y_n \sim N(n\mu, n\sigma^2)$.

2.5***.** Suppose that $X_n \sim U(-1/n, 1/n)$.

a) What is the cdf $F_n(x)$ of X_n?

b) What does $F_n(x)$ converge to?

(Hint: consider $x < 0$, $x = 0$ and $x > 0$.)

c) $X_n \xrightarrow{D} X$. What is X?

2.6. Continuity Theorem problem: Let X_n be sequence of random variables with cdfs F_n and mgfs m_n. Let X be a random variable with cdf F and mgf m. Assume that all of the mgfs m_n and m are defined if $|t| \leq d$ for some $d > 0$. Thus if $m_n(t) \to m(t)$ as $n \to \infty$ for all $|t| < c$ where $0 < c < d$, then $X_n \xrightarrow{D} X$.

Let

$$m_n(t) = \frac{1}{[1 - (\lambda + \frac{1}{n})t]}$$

for $t < 1/(\lambda + 1/n)$. Then what is $m(t)$ and what is X?

2.7. Let $Y_1, ..., Y_n$ be iid, $T_{1,n} = \overline{Y}$ and let $T_{2,n} = \text{MED}(n)$ be the sample median. Let $\theta = \mu$.

Then

$$\sqrt{n}(MED(n) - MED(Y)) \xrightarrow{D} N\left(0, \frac{1}{4f^2(MED(Y))}\right)$$

where the population median is MED(Y) (and MED(Y) $= \mu = \theta$ for a) and b) below).

a) Find $ARE(T_{1,n}, T_{2,n})$ if F is the cdf of the normal $N(\mu, \sigma^2)$ distribution.

b) Find $ARE(T_{1,n}, T_{2,n})$ if F is the cdf of the double exponential $DE(\theta, \lambda)$ distribution.

2.8^Q**.** Let $X_1, ..., X_n$ be independent identically distributed random variables with probability density function

$$f(x) = \theta x^{\theta-1}, \ 0 < x < 1, \ \ \theta > 0.$$

a) Find the MLE of $\frac{1}{\theta}$. Is it unbiased? Does it achieve the information inequality lower bound?

b) Find the asymptotic distribution of the MLE of $\frac{1}{\theta}$.

c) Show that $\overline{X}_n$ is unbiased for $\frac{\theta}{\theta+1}$. Does $\overline{X}_n$ achieve the information inequality lower bound?

d) Find an estimator of $\frac{1}{\theta}$ from part (c) above using $\overline{X}_n$ which is different from the MLE in (a). Find the asymptotic distribution of your estimator using the delta method.

e) Find the asymptotic relative efficiency of your estimator in (d) with respect to the MLE in (b).

Problems from old quizzes and exams. Problems from old qualifying exams are marked with a Q.

2.9. Let $X_1, ..., X_n$ be iid Bernoulli(p) random variables.

a) Find $I_1(p)$.

b) Find the FCRLB for estimating p.

c) Find the limiting distribution of $\sqrt{n}(\ \overline{X}_n - p\ )$.

d) Find the limiting distribution of $\sqrt{n}\ [\ (\overline{X}_n)^2 - c\ ]$ for an appropriate constant c.

2.10. Let $X_1, ..., X_n$ be iid Exponential(β) random variables.

a) Find the FCRLB for estimating β.

b) Find the limiting distribution of $\sqrt{n}(\ \overline{X}_n - \beta\ )$.

c) Find the limiting distribution of $\sqrt{n}\ [\ (\overline{X}_n)^2 - c\ ]$ for an appropriate constant c.

2.11. Let $Y_1, ..., Y_n$ be iid Poisson (λ) random variables.

a) Find the limiting distribution of $\sqrt{n}(\ \overline{Y}_n - \lambda\ )$.

b) Find the limiting distribution of $\sqrt{n}\ [\ (\overline{Y}_n)^2 - c\ ]$ for an appropriate constant c.

2.12. Let $Y_n \sim \chi^2_n$.

a) Find the limiting distribution of $\sqrt{n}\ \left(\ \frac{Y_n}{n} -\ 1\ \right)$.

b) Find the limiting distribution of $\sqrt{n}\ \left[\ \left(\frac{Y_n}{n}\right)^3 -\ 1\ \right]$.

2.13. Let $X_1, ..., X_n$ be iid with cdf $F(x) = P(X \leq x)$. Let $Y_i = I(X_i \leq x)$ where the indicator equals 1 if $X_i \leq x$ and 0, otherwise.

a) Find $E(Y_i)$.

b) Find $V(Y_i)$.

c) Let $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$ for some fixed real number x. Find the limiting distribution of $\sqrt{n}\ \left(\hat{F}_n(x)\ -\ c_x\ \right)$ for an appropriate constant c_x.

2.14. Suppose X_n has cdf

$$F_n(x) = 1 - \left(1 - \frac{x}{\theta n}\right)^n$$

for $x \geq 0$ and $F_n(x) = 0$ for $x < 0$. Show that $X_n \xrightarrow{D} X$ by finding the cdf of X.

2.15. Let X_n be a sequence of random variables such that $P(X_n = 1/n) = 1$. Does X_n converge in distribution? If yes, prove it by finding X and the cdf of X. If no, prove it.

2.16. Suppose that $Y_1, ..., Y_n$ are iid with $E(Y) = (1-\rho)/\rho$ and $V(Y) = (1-\rho)/\rho^2$ where $0 < \rho < 1$.

a) Find the limiting distribution of

$$\sqrt{n}\ \left(\overline{Y}_n -\ \frac{1-\rho}{\rho}\ \right).$$

b) Find the limiting distribution of $\sqrt{n}\ \left[\ g(\overline{Y}_n) -\ \rho\right]$ for appropriate function g.

2.17. Let $X_n \sim$ Binomial(n, p) where the positive integer n is large and $0 < p < 1$.

a) Find the limiting distribution of $\sqrt{n}\ \left(\ \frac{X_n}{n} -\ p\ \right)$.

b) Find the limiting distribution of $\sqrt{n}\ \left[\ \left(\frac{X_n}{n}\right)^2 -\ p^2\ \right]$.

c) Let $g(\theta) = \theta^3 - \theta$. Find the limiting distribution of $n\left[g\left(\frac{X_n}{n}\right) - c \right]$ for appropriate constant c when $p = \frac{1}{\sqrt{3}}$. Hint: Use the Second Order Delta Method.

2.18. Let $Y_1, ..., Y_n$ be iid exponential (λ) so that $E(Y) = \lambda$ and $\text{MED}(Y) = \log(2)\lambda$.

a) Let $T_{1,n} = \log(2)\overline{Y}$ and find the limiting distribution of $\sqrt{n}(T_{1,n} - \log(2)\lambda)$.

b) Let $T_{2,n} = \text{MED}(n)$ be the sample median and find the limiting distribution of $\sqrt{n}(T_{2,n} - \log(2)\lambda)$.

c) Find $ARE(T_{1,n}, T_{2,n})$.

2.19. Suppose that $\eta = g(\theta)$, $\theta = g^{-1}(\eta)$ and $g'(\theta) > 0$ exists. If X has pdf or pmf $f(x|\theta)$, then in terms of η, the pdf or pmf is $f^*(x|\eta) = f(x|g^{-1}(\eta))$. Now

$$A = \frac{\partial}{\partial \eta} \log[f(x|g^{-1}(\eta))] = \frac{1}{f(x|g^{-1}(\eta))} \frac{\partial}{\partial \eta} f(x|g^{-1}(\eta)) =$$

$$\left[\frac{1}{f(x|g^{-1}(\eta))}\right] \left[\frac{\partial}{\partial \theta} f(x|\theta)\Big|_{\theta = g^{-1}(\eta)}\right] \left[\frac{\partial}{\partial \eta} g^{-1}(\eta)\right]$$

using the chain rule twice. Since $\theta = g^{-1}(\eta)$,

$$A = \left[\frac{1}{f(x|\theta)}\right] \left[\frac{\partial}{\partial \theta} f(x|\theta)\right] \left[\frac{\partial}{\partial \eta} g^{-1}(\eta)\right].$$

Hence

$$A = \frac{\partial}{\partial \eta} \log[f(x|g^{-1}(\eta))] = \left[\frac{\partial}{\partial \theta} \log[f(x|\theta)]\right] \left[\frac{\partial}{\partial \eta} g^{-1}(\eta)\right].$$

Now show that

$$I_1^*(\eta) = \frac{I_1(\theta)}{[g'(\theta)]^2}.$$

2.20. Let $Y_1, ..., Y_n$ be iid exponential (1) so that $P(Y \leq y) = F(y) = 1 - e^{-y}$ for $y \geq 0$. Let $Y_{(n)} = \max(Y_1, ..., Y_n)$.

a) Show that $F_{Y_{(n)}}(t) = P(Y_{(n)} \leq t) = [1 - e^{-t}]^n$ for $t \geq 0$.

b) Show that $P(Y_{(n)} - \log(n) \leq t) \to \exp(-e^{-t})$ (for all $t \in (-\infty, \infty)$ since $t + \log(n) > 0$ implies $t \in \mathbb{R}$ as $n \to \infty$).

2.21. Let $Y_1, ..., Y_n$ be iid uniform $(0, 2\theta)$.

a) Let $T_{1,n} = \overline{Y}$ and find the limiting distribution of $\sqrt{n}(T_{1,n} - \theta)$.

b) Let $T_{2,n} = \text{MED}(n)$ be the sample median and find the limiting distribution of $\sqrt{n}(T_{2,n} - \theta)$.

c) Find $ARE(T_{1,n}, T_{2,n})$. Which estimator is better, asymptotically?

2.22. Suppose that $Y_1, ..., Y_n$ are iid from a distribution with pdf $f(y|\theta)$ and that the integral and differentiation operators of all orders can be interchanged (e.g. the data is from a one parameter exponential family).

a) Show that $0 = E\left[\frac{\partial}{\partial\theta}\log(f(Y|\theta))\right]$ by showing that

$$\frac{\partial}{\partial\theta}1 = 0 = \frac{\partial}{\partial\theta}\int f(y|\theta)dy = \int \left[\frac{\partial}{\partial\theta}\log(f(y|\theta))\right] f(y|\theta)dy. \quad (*)$$

b) Take 2nd derivatives of $(*)$ to show that

$$I_1(\theta) = E_\theta[(\frac{\partial}{\partial\theta}\log f(Y|\theta))^2] = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log(f(Y|\theta))\right].$$

2.23. Suppose that $X_1, ..., X_n$ are iid $N(\mu, \sigma^2)$.

a) Find the limiting distribution of $\sqrt{n}\ \left(\overline{X}_n - \ \mu\ \right)$.

b) Let $g(\theta) = [\log(1+\theta)]^2$. Find the limiting distribution of $\sqrt{n}\ \left(g(\overline{X}_n) - \ g(\mu)\ \right)$ for $\mu > 0$.

c) Let $g(\theta) = [\log(1+\theta)]^2$. Find the limiting distribution of $n\ \left(g(\overline{X}_n) - \ g(\mu)\ \right)$ for $\mu = 0$. Hint: Use the Second Order Delta Method: Theorem 2.3.

2.24. Note that $E(X) = E(X1) = E[X(I(A) + I(A^c))] = E[XI(A)] + E[XI(A^c)] \geq E[XI(A)]$ if X is a nonegative random variable since then $XI(A^c)$ is a nonnegative random variable and $E[XI(A^c)] \geq 0$.

Let $W_n = X_n - X$ and let $r > 0$. Notice that for any $\epsilon > 0$,

$$E|X_n - X|^r \geq E[|X_n - X|^r\ I(|X_n - X| \geq \epsilon)] \geq \epsilon^r P(|X_n - X| \geq \epsilon).$$

Show that $W_n \xrightarrow{P} 0$ if $E|X_n - X|^r \to 0$ as $n \to \infty$.

2.25. Let $X_1, ..., X_n$ be iid with $E(X) = \mu$ and $V(X) = \sigma^2$. What is the limiting distribution of $n[(\overline{X})^2 - \mu^2]$ for the value or values of μ where the delta method does not apply? Hint: use Theorem 2.3.

2.26^Q**.** Let $X \sim \text{Binomial}(n, p)$ where the positive integer n is large and $0 < p < 1$.

a) Find the limiting distribution of $\sqrt{n}\ \left(\ \frac{X}{n} - \ p\ \right)$.

b) Find the limiting distribution of $\sqrt{n}\ \left[\ \left(\frac{X}{n}\right)^2 - \ p^2\ \right]$.

c) Show how to find the limiting distribution of $\left[\left(\frac{X}{n}\right)^3 - \frac{X}{n} \right]$ when $p = \frac{1}{\sqrt{3}}$.

(Actually want the limiting distribution of

$$n \left(\left[\left(\frac{X}{n}\right)^3 - \frac{X}{n} \right] - g(p) \right)$$

where $g(\theta) = \theta^3 - \theta$.)

2.27^Q**.** Let $X_1, ..., X_n$ be independent and identically distributed (iid) from a Poisson(λ) distribution.

a) Find the limiting distribution of $\sqrt{n}\ (\ \overline{X} - \lambda\ )$.

b) Find the limiting distribution of $\sqrt{n}\ [\ (\overline{X})^3 - (\lambda)^3\ ]$.

2.28^Q**.** Let $X_1, ..., X_n$ be iid from a normal distribution with unknown mean μ and known variance σ^2. Let $\overline{X} = \frac{\sum_{i=1}^n X_i}{n}$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$.

a) Show that $\overline{X}$ and S^2 are independent.

b) Find the limiting distribution of $\sqrt{n}((\overline{X})^3 - c)$ for an appropriate constant c.

2.29. Suppose that $Y_1, ..., Y_n$ are iid logistic$(\theta, 1)$ with pdf

$$f(y) = \frac{\exp(-(y-\theta))}{[1+\exp(-(y-\theta))]^2}$$

where and y and θ are real.

a) $I_1(\theta) = 1/3$ and the family is regular so the "standard limit theorem" for the MLE $\hat{\theta}_n$ holds. Using this standard theorem, what is the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$?

b) Find the limiting distribution of $\sqrt{n}(\overline{Y}_n - \theta)$.

c) Find the limiting distribution of $\sqrt{n}(MED(n) - \theta)$.

d) Consider the estimators $\hat{\theta}_n$, $\overline{Y}_n$ and MED(n). Which is the best estimator and which is the worst?

2.30. Let $Y_n \sim$ binomial(n, p). Find the limiting distribution of

$$\sqrt{n}\left(\arcsin\left(\sqrt{\frac{\mathrm{Y_n}}{\mathrm{n}}}\right) - \arcsin(\sqrt{\mathrm{p}})\right).$$

(Hint:

$$\frac{d}{dx}\arcsin(\mathrm{x}) = \frac{1}{\sqrt{1-\mathrm{x}^2}}.)$$

2.31. Suppose $Y_n \sim$ uniform$(-n, n)$. Let $F_n(y)$ be the cdf of Y_n.
a) Find $F(y)$ such that $F_n(y) \to F(y)$ for all y as $n \to \infty$.

b) Does $Y_n \xrightarrow{D} Y$? Explain briefly.

2.32. Suppose $Y_n \sim$ uniform$(0, n)$. Let $F_n(y)$ be the cdf of Y_n.
a) Find $F(y)$ such that $F_n(y) \to F(y)$ for all y as $n \to \infty$.

b) Does $Y_n \xrightarrow{D} Y$? Explain briefly.

2.33^Q**.** Let $Y_1, ..., Y_n$ be independent and identically distributed (iid) from a distribution with probability mass function $f(y) = \rho(1-\rho)^y$ for $y = 0, 1, 2, ...$ and $0 < \rho < 1$. Then $E(Y) = (1-\rho)/\rho$ and $V(Y) = (1-\rho)/\rho^2$.

a) Find the limiting distribution of $\sqrt{n}\left(\overline{Y} - \frac{1-\rho}{\rho}\right)$.

b) Show how to find the limiting distribution of $g(\overline{Y}) = \frac{1}{1+\overline{Y}}$. Deduce it completely. (This bad notation means find the limiting distribution of $\sqrt{n}(g(\overline{Y}) - c)$ for some constant c.)

c) Find the method of moments estimator of ρ.

d) Find the limiting distribution of $\sqrt{n}\ (\ (1+\overline{Y}) - \ d\ )$ for appropriate constant d.

e) Note that $1 + E(Y) = 1/\rho$. Find the method of moments estimator of $1/\rho$.

2.34^Q**.** Let $X_1, ..., X_n$ be independent identically distributed random variables from a normal distribution with mean μ and variance σ^2.

a) Find the approximate distribution of $1/\bar{X}$. Is this valid for all values of μ?

b) Show that $1/\bar{X}$ is asymptotically efficient for $1/\mu$, provided $\mu \neq \mu^*$. Identify μ^*.

2.35^Q**.** Let $Y_1, ..., Y_n$ be independent and identically distributed (iid) from a distribution with probability density function

$$f(y) = \frac{2y}{\theta^2}$$

for $0 < y \leq \theta$ and $f(y) = 0$, otherwise.

a) Find the limiting distribution of $\sqrt{n}\ (\ \overline{Y} - c\ )$ for appropriate constant c.

b) Find the limiting distribution of $\sqrt{n}\ (\ \log(\ \overline{Y}\ ) - \ d\ )$ for appropriate constant d.

c) Find the method of moments estimator of θ^k.

2.36^Q**.** Let $Y_1, ..., Y_n$ be independent identically distributed discrete random variables with probability mass function

$$f(y) = P(Y = y) = \binom{r+y-1}{y} \rho^r (1-\rho)^y$$

for $y = 0, 1, \ldots$ where positive integer r **is known** and $0 < \rho < 1$. Then $E(Y) = r(1-\rho)/\rho$, and $V(Y) = r(1-\rho)/\rho^2$.

a) Find the limiting distribution of $\sqrt{n}\ \left(\ \overline{Y} - \ \frac{r(1-\rho)}{\rho} \ \right)$.

b) Let $g(\overline{Y}) = \frac{r}{r+\overline{Y}}$. Find the limiting distribution of $\sqrt{n}\ \left(\ g(\overline{Y}) - \ c \ \right)$ for appropriate constant c.

c) Find the method of moments estimator of ρ.

2.37^Q**.** Let $X_1, ..., X_n$ be independent identically distributed uniform $(0, \theta)$ random variables where $\theta > 0$.

a) Find the limiting distribution of $\sqrt{n}(\overline{X} - c_\theta)$ for an appropriate constant c_θ that may depend on θ.

b) Find the limiting distribution of $\sqrt{n}[(\overline{X})^2 - k_\theta]$ for an appropriate constant k_θ that may depend on θ.

2.38^Q**.** Let $X_1, ..., X_n$ be independent identically distributed (iid) beta(β, β) random variables.

a) Find the limiting distribution of $\sqrt{n}(\ \overline{X}_n - \theta\ )$, for appropriate constant θ.

b) Find the limiting distribution of $\sqrt{n}(\ \log(\overline{X}_n) - d\ )$, for appropriate constant d.

2.39. Suppose that $X_1, ..., X_n$ are iid and $\mathrm{V}(X_1) = \sigma^2$. Given that

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \xrightarrow{P} \sigma^2,$$

give a very short proof that the sample variance

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \xrightarrow{P} \sigma^2.$$

2.40. Suppose

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{D} N(0,1)$$

and $s_n^2 \xrightarrow{P} \sigma^2$ where $\sigma > 0$. Prove that

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{s_n} \xrightarrow{D} N(0,1).$$

2.41. If $Y_n \xrightarrow{D} Y$, $a_n \xrightarrow{P} a$, and $b_n \xrightarrow{P} b$, then $a_n + b_n Y_n \xrightarrow{D} X$. Find X.

2.42. What theorem can be used to prove both the (usual) central limit theorem and the Lyapounov CLT?

2.43. Let $X_1, ..., X_n$ be iid with mean $E(X) = \mu$ and variance $V(X) = \sigma^2 > 0$. Then $n(\overline{X} - \mu)^2 = [\sqrt{n}(\overline{X} - \mu)]^2 \xrightarrow{D} W$. What is W?

2.44. Let $Y_1, ..., Y_n$ be iid gamma($\nu = 2, \lambda$) with $E(Y) = 2\lambda$, $V(Y) = 2\lambda^2$, and $I_1(\lambda) = \dfrac{2}{\lambda^2}$. The gamma$(2, \lambda)$ distribution is a 1PREF. Let $\hat{\lambda}_n$ be the MLE of λ. Find the limiting distribution of $\sqrt{n}(\hat{\lambda}_n - \lambda)$.

2.45. Let $Y_1, ..., Y_n$ be iid double exponential $DE(\theta, \lambda)$ with $E(Y) = \theta$ and $V(Y) = 2\lambda^2$ where θ and y are real and $\lambda > 0$.

a) Find the limiting distribution of $\sqrt{n}[\,\overline{Y} - c]$ for an appropriate constant c.

b) Find the limiting distribution of $\sqrt{n}\left[(\overline{Y})^2 - d\right]$ for appropriate constant d for the values of θ where the delta method applies.

c) What is the limiting distribution of $n\left[(\overline{Y})^2 - d\right]$ for the value or values of θ where the delta method does not apply?

2.46. Let $Y_1, ..., Y_n$ be iid with $E(Y^r) = \exp(r\mu + r^2\sigma^2/2)$ for any real r. Find the limiting distribution of $\sqrt{n}(\overline{Y}_n - c)$ for appropriate constant c.

2.47. Let $Y_n \sim$ Poisson$(n\theta)$. Find the limiting distribution of $\sqrt{n}\left(\dfrac{Y_n}{n} - c\right)$ for appropriate constant c.

More Problems:

2.48. Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and $V(Y) = \sigma^2$. Let $g(\mu) = \mu^2$. For $\mu = 0$, find the limiting distribution of $n[(\overline{Y}_n)^2 - 0^2] = n(\overline{Y}_n)^2$ by using the Second Order Delta Method.

2.49. Rohatgi (1971, p. 248): Let $P(X_n = 0) = 1 - 1/n^r$ and $P(X_n = n) = 1/n^r$ where $r > 0$.

a) Prove that X_n does not converge in rth mean to 0. Hint: Find $E[|X_n|^r]$.

b) Does $X_n \xrightarrow{D} X$ for some random variable X? Prove or disprove.

2.50. Suppose $Y_n \sim EXP(1/n)$ with cdf $F_{Y_n}(y) = 1 - exp(-ny)$ for $y \geq 0$, and $F_{Y_n}(y) = 0$ for $y < 0$. Does $Y_n \xrightarrow{D} Y$ for some random variable Y? Prove or disprove. If $Y_n \xrightarrow{D} Y$, find Y.

2.51. Suppose $X_1, ..., X_n$ are iid from a distribution with mean μ and variance σ^2. $\dfrac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{P} c$. What is c? Hint: Use WLLN on $W_i = X_i^2$.

2.52. Rohatgi (1971, p. 248): Let $P(X_n = 0) = 1 - 1/n^r$ and $P(X_n = n) = 1/n^r$ where $r > 0$.

a) Prove that X_n does not converge in rth mean to 0. Hint: Find $E[|X_n|^r]$.

b) Does $X_n \xrightarrow{D} X$ for some random variable X? Prove or disprove.
Hint: $P(|X_n - 0| \geq \epsilon) \leq P(X_n = n)$.

2.53. Suppose $Y_1, ..., Y_n$ are iid EXP(λ). Let $T_n = Y_{(1)} = Y_{1:n} = \min(Y_1, ..., Y_n)$. It can be shown that the mgf of T_n is

$$m_{T_n}(t) = \frac{1}{1 - \frac{\lambda t}{n}}$$

for $t < n/\lambda$. Show that $T_n \xrightarrow{D} X$ and give the distribution of X.

2.54. Let $Y_1, ..., Y_n$ be iid with

$$E(Y^r) = 2^{r/2}\sigma^r \frac{\Gamma(\frac{r+p}{2})}{\Gamma(p/2)}$$

for $r > -p$ where $\sigma, p > 0$. Find the limiting distribution of $\sqrt{n}(\overline{Y}_n - c)$ for appropriate constant c.

2.55. Suppose

$$F_{X_n}(x) = \begin{cases} 0, & x \leq c - \frac{1}{n} \\ \frac{n}{2}(x - c + \frac{1}{n}), & c - \frac{1}{n} < x < c + \frac{1}{n} \\ 1, & x \geq c + \frac{1}{n}. \end{cases}$$

Does $X_n \xrightarrow{D} X$ for some random variable X? Prove or disprove. If $X_n \xrightarrow{D} X$, find X.

2.56. Suppose $Y_n \sim EXP(n)$ with cdf $F_{Y_n}(y) = 1 - exp(-y/n)$ for $y \geq 0$ and $F_{Y_n}(y) = 0$ for $y < 0$. Does $Y_n \xrightarrow{D} Y$ for some random variable Y? Prove or disprove. If $Y_n \xrightarrow{D} Y$, find Y.

2.57. Suppose $Y_1, ..., Y_n$ are iid $POIS(\theta)$. Then the MLE of θ is $\hat{\theta}_n = \overline{Y}_n$.

a) Find the limiting distribution of $\sqrt{n}(\overline{Y}_n - c)$ for appropriate constant c.

b) Let $\tau(\theta) = \theta^2$. Find the limiting distribution of $\sqrt{n}[\tau(\hat{\theta}_n) - \tau(\theta)]$ using the Delta Method.

2.58. Let X_n be sequence of random variables with cdfs F_n and mgfs m_n. Let X be a random variable with cdf F and mgf m. Assume that all of the mgfs m_n and m are defined to $|t| \leq d$ for some $d > 0$. Let

$$m_n(t) = \frac{1}{[1 - (\lambda + \frac{1}{n})t]}$$

for $t < 1/(\lambda + 1/n)$. Show that $m_n(t) \to m(t)$ by finding $m(t)$.
(Then $X_n \xrightarrow{D} X$ where $X \sim EXP(\lambda)$ with $E(X) = \lambda$ by the continuity theorem for mgfs.)

2.59. Suppose $Y_n \xrightarrow{P} Y$. Then $W_n = Y_n - Y \xrightarrow{P} 0$. Define $X_n = Y$ for all n. Then $X_n \xrightarrow{D} Y$. Then $Y_n = X_n + W_n \xrightarrow{D} Z$ by Slutsky's Theorem. What is Z?

2.60. The method of moments estimator for $Cov(X, Y) = \sigma_{X,Y}$ is $\hat{\sigma}_{X,Y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$. Another common estimator is

$S_{X,Y} = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \dfrac{n}{n-1}\hat{\sigma}_{X,Y}$. Using the fact that $\hat{\sigma}_{X,Y} \xrightarrow{P} \sigma_{X,Y}$ when the covariance exists, prove that $S_{X,Y} \xrightarrow{P} \sigma_{X,Y}$ with Slutsky's Theorem. Hint: $Z_n \xrightarrow{P} c$ iff $Z_n \xrightarrow{D} c$ if c is a constant, and usual convergence $a_n \to a$ of a sequence of constants implies $a_n \xrightarrow{P} a$.

2.61. Suppose that the characteristic function of $\overline{X}_n$ is

$$c_{\overline{X}}(t) = \exp(-\frac{t^2\sigma^2}{2n}).$$

Then the characteristic function of $\sqrt{n}\ \overline{X}_n$ is $c_{\sqrt{n}\ \overline{X}}(t) = c_{\overline{X}}(\sqrt{n}\ t)$. Does $\sqrt{n}\ \overline{X}_n \xrightarrow{D} W$ for some random variable W? Explain.

2.62. Let $X_1, ..., X_n$ be iid with mean $E(X) = \mu$ and variance $V(X) = \sigma^2 > 0$. Then $\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = \sum_{i=1}^{n}(X_i - \mu + \mu - \overline{X}_n)^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{x} - \mu)^2$.

a) $\dfrac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \xrightarrow{P} \theta$. What is θ?

b) Also, $n(\overline{X} - \mu)^2 = [\sqrt{n}(\overline{X} - \mu)]^2 \xrightarrow{D} W$. What is W? Hint: use the continuous mapping theorem. Note that $Z \sim N(0, \sigma^2) \sim \sigma N(0, 1)$.

2.63. Let $X_1, ..., X_n$ be independent and identically distributed (iid) from a Poisson(λ) distribution with $E(X) = \lambda$. Let $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$.

a) Find the limiting distribution of $\sqrt{n}\ (\ \overline{X} - \lambda\ )$.

b) Find the limiting distribution of $\sqrt{n}\ [\ (\overline{X})^3 - (\lambda)^3\ ]$.

2.64. Let $X_1, ..., X_n$ be iid from a normal distribution with unknown mean μ and known variance σ^2. Find the limiting distribution of $\sqrt{n}(\overline{X}^3 - c)$ for an appropriate constant c.

2.65. Let $Y_n \sim \chi^2_n$.

a) Find the limiting distribution of $\sqrt{n}\ \left(\ \dfrac{Y_n}{n} -\ 1\ \right)$.

b) Find the limiting distribution of $\sqrt{n}\ \left[\ \left(\dfrac{Y_n}{n}\right)^3 -\ 1\ \right]$.

2.66. Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and $V(Y) = \sigma^2$. Let $g(\mu) = \mu^2$. For $\mu = 0$, find the limiting distribution of $n[(\overline{Y}_n)^2 - 0^2] = n(\overline{Y}_n)^2$ by using the Second Order Delta Method.

2.67. In earlier courses, you should have used moment generating functions to show that if $Y_n = \sum_{i=1}^{n} X_i$ where the X_i are iid from a nice distribution, then Y_n has a nice distribution where the nice distributions are the binomial, chi–square, gamma, negative binomial, normal, and Poisson distributions. If $E(X) = \mu$ and $\mathrm{V}(X) = \sigma^2$ then by the CLT

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Since $\sqrt{n}(\frac{Y_n}{n} - \mu)$ and $\sqrt{n}(\overline{X}_n - \mu)$ have the same distribution,

$$\sqrt{n}\left(\frac{Y_n}{n} - \mu\right) \xrightarrow{D} N(0, \sigma^2)$$

For example, if $Y_n \sim N(n\mu, n\sigma^2)$ then $Y_n \sim \sum_{i=1}^{n} X_i$ where the X_i are iid $N(\mu, \sigma^2)$. Hence

$$\sqrt{n}\left(\frac{Y_n}{n} - \mu\right) \sim \sqrt{n}(\overline{X}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

which should not be surprising since

$$\sqrt{n}\left(\frac{Y_n}{n} - \mu\right) \sim N(0, \sigma^2).$$

Write down the distribution of X_i if

i) $Y_n \sim \text{BIN}(n, p)$ where BIN stands for binomial.

ii) $Y_n \sim \chi^2_n$.

iii) $Y_n \sim G(n\alpha, \beta)$ where G stands for gamma.

iv) $Y_n \sim NB(n, p)$ where NB stands for negative binomial.

v) $Y_n \sim POIS(n\theta)$ where POIS stands for Poisson.

(Write down the distribution if you know it or can find it. Do not use mgfs unless you can not find the distribution.)

2.68. Suppose that $X_n \sim U(-1/n, 1/n)$.

a) What is the cdf $F_n(x)$ of X_n?

b) What does $F_n(x)$ converge to? (Hint: consider $x < 0$, $x = 0$ and $x > 0$.)

c) $X_n \xrightarrow{D} X$. What is X?

2.69. Suppose X_n is a discrete random variable with $P(X_n = n) = 1/n$ and $P(X_n = 0) = (n-1)/n$.

a) Show $X_n \xrightarrow{D} X$.

b) Does $E(X_n) \to E(X)$? Explain briefly.

2.70. Suppose X_n has cdf

$$F_n(x) = 1 - \left(1 - \frac{x}{\theta n}\right)^n$$

for $x \geq 0$ and $F_n(x) = 0$ for $x < 0$. Show that $X_n \xrightarrow{D} X$ by finding the cdf of X.

2.71. Let $Y_1, ..., Y_n$ be iid $N(\mu, \sigma^2)$ with μ known. Let $\hat{\sigma}^2_n$ be the MLE of σ^2 with $I_1(\sigma^2) = \dfrac{1}{2\sigma^4}$.

a) Find the limiting distribution of $\sqrt{n}(\hat{\sigma}^2_n - \sigma^2)$.

b) Find the limiting distribution of $\sqrt{n}[\sqrt{\hat{\sigma}_n^2} - \sigma]$. Note that $\tau(\sigma^2) = \sqrt{\sigma^2}$. Taking $\theta = \sigma^2$ could be useful.

2.72. Rohatgi (1971, p. 248): Let $P(X_n = 0) = 1 - 1/n^r$ and $P(X_n = n) = 1/n^r$ where $r > 0$.

a) Prove that X_n does not converge in rth mean to 0. Hint: Find $E[|X_n|^r]$.

b) Does $X_n \xrightarrow{D} X$ for some random variable X? Prove or disprove.

2.73. Suppose $X_1, ..., X_n$ are iid $C(\mu, \sigma)$ with characteristic function $c_X(t) = \exp(it\mu - |t|\sigma)$ where $\exp(a) = e^a$.

a) Find the characteristic function $c_{T_n}(t)$ of $T_n = \sum_{i=1}^n X_i$.

b) Find the characteristic function of $\overline{X}_n = T_n/n$.

c) Does $\overline{X}_n \xrightarrow{D} W$ for some RV W? Explain.

2.74. Suppose $X_1, ..., X_n$ are iid from a distribution with mean μ and variance σ^2. The method of moments estimator for σ^2 is

$$S_M^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 \quad - \quad (\overline{X}_n)^2.$$

a) $\dfrac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{P} c$. What is c? Hint: Use WLLN on $W_i = X_i^2$.

b) $(\overline{X}_n)^2 \xrightarrow{P} d$. What is d? Hint: $g(x) = x^2$ is continuous, so if $Z_n \xrightarrow{P} \theta$, then $g(Z_n) \xrightarrow{P} g(\theta)$.

c) Show $S_m^2 \xrightarrow{P} \sigma^2$.

d) $S^2 = \dfrac{n}{n-1} S_M^2 = \dfrac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. Prove $S^2 \xrightarrow{P} \sigma^2$.

2.75. Suppose X_n are random variables with characteristic functions $c_{X_n}(t)$, and that $c_{X_n}(t) \to e^{itc}$ for every $t \in \mathbb{R}$ where c is a constant. Does $X_n \xrightarrow{D} X$ for some random variable X? Explain briefly. Hint: Is the function $g(t) = e^{itc}$ continuous as $t = 0$? Is there a random variable that has characteristic function $g(t)$?

2.76. The characteristic function for $Y \sim N(\mu, \sigma^2)$ is $c_Y(t) = \exp(it\mu - t^2\sigma^2/2)$. Let $X_n \sim N(0, n)$.

a) Prove $c_{X_n}(t) \to h(t)\ \forall t$ by finding $h(t)$.

b) Use a) to prove whether X_n converges in distribution.

2.77. Suppose

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{D} N(0, 1)$$

and $s_n^2 \xrightarrow{P} \sigma^2$ where $\sigma > 0$. Prove that

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{s_n} \xrightarrow{D} N(0, 1).$$

2.78. It is true that W_n has the same order as X_n in probability, written $W_n \asymp_P X_n$, iff for every $\epsilon > 0$ there exist positive constants N_ϵ and $0 < d_\epsilon < D_\epsilon$ such that

$$P(d_\epsilon \leq \left| \frac{W_n}{X_n} \right| \leq D_\epsilon) \geq 1 - \epsilon$$

for all $n \geq N_\epsilon$.

a) Show that if $W_n \asymp_P X_n$ then $X_n \asymp_P W_n$.

b) Show that if $W_n \asymp_P X_n$ then $W_n = O_P(X_n)$.

c) Show that if $W_n \asymp_P X_n$ then $X_n = O_P(W_n)$.

d) Show that if $W_n = O_P(X_n)$ and if $X_n = O_P(W_n)$, then $W_n \asymp_P X_n$.

2.79. This problem will prove the following Theorem which says that if there are K estimators $T_{j,n}$ of a parameter $\boldsymbol{\beta}$, such that $\|T_{j,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ where $0 < \delta \leq 1$, and if T_n^* picks one of these estimators, then $\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta})$.

Lemma: Pratt (1959). Let $X_{1,n}, ..., X_{K,n}$ each be $O_P(1)$ where K is fixed. Suppose $W_n = X_{i_n,n}$ for some $i_n \in \{1, ..., K\}$. Then

$$W_n = O_P(1). \tag{2.14}$$

Proof.

$$P(\max\{X_{1,n}, ..., X_{K,n}\} \leq x) = P(X_{1,n} \leq x, ..., X_{K,n} \leq x) \leq$$

$$F_{W_n}(x) \leq P(\min\{X_{1,n}, ..., X_{K,n}\} \leq x) = 1 - P(X_{1,n} > x, ..., X_{K,n} > x).$$

Since K is finite, there exists $B > 0$ and N such that $P(X_{i,n} \leq B) > 1 - \epsilon/2K$ and $P(X_{i,n} > -B) > 1 - \epsilon/2K$ for all $n > N$ and $i = 1, ..., K$. Bonferroni's inequality states that $P(\cap_{i=1}^K A_i) \geq \sum_{i=1}^K P(A_i) \ - \ (K-1)$. Thus

$$F_{W_n}(B) \geq P(X_{1,n} \leq B, ..., X_{K,n} \leq B) \geq K(1 - \epsilon/2K) - (K-1) = K - \epsilon/2 - K + 1 = 1 - \epsilon/2$$

and

$$-F_{W_n}(-B) \geq -1 + P(X_{1,n} > -B, ..., X_{K,n} > -B) \geq$$

$$-1 + K(1 - \epsilon/2K) - (K-1) = -1 + K - \epsilon/2 - K + 1 = -\epsilon/2.$$

Hence

$$F_{W_n}(B) - F_{W_n}(-B) \geq 1 - \epsilon \ \text{ for n > N. QED}$$

Theorem. Suppose $\|T_{j,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ for $j = 1, ..., K$ where $0 < \delta \leq 1$. Let $T_n^* = T_{i_n,n}$ for some $i_n \in \{1, ..., K\}$ where, for example, $T_{i_n,n}$ is the $T_{j,n}$ that minimized some criterion function. Then

$$\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta}). \tag{2.15}$$

Prove the above theorem using the Lemma with an appropriate $X_{j,n}$.

2.80. Suppose

$$F_{X_n}(x) = \begin{cases} 0, & x \le c - \frac{1}{n} \\ \frac{n}{2}(x - c + \frac{1}{n}), & c - \frac{1}{n} < x < c + \frac{1}{n} \\ 1, & x \ge c + \frac{1}{n}. \end{cases}$$

Does $X_n \xrightarrow{D} X$ for some random variable X? Prove or disprove. If $X_n \xrightarrow{D} X$, find X.

2.81. Suppose $X_1, ..., X_n$ are iid from a distribution with $E(X^k) = \Gamma(3-k)/6\lambda^k$ for integer $k < 4$. Recall that $\Gamma(n) = (n-1)!$ for integers $n \ge 1$. Find the limiting distribution of $\sqrt{n}(\ \overline{X}_n - c\ )$ for appropriate constant c.

2.82. Suppose X_n is a discrete random variable with $P(X_n = n) = 1/n$ and $P(X_n = 0) = (n-1)/n$. Does $X_n \xrightarrow{D} X$? Explain.

2.83. Let $X_n \sim \text{Poisson}(n\theta)$. Find the limiting distribution of $\sqrt{n}\left(\ \frac{X_n}{n} - \theta\ \right)$.

2.84. Let $Y_1, ..., Y_n$ be iid Gamma(θ, θ) random variables with $E(Y_i) = \theta^2$ and $V(Y_i) = \theta^3$ where $\theta > 0$. Find the limiting distribution of $\sqrt{n}(\ \overline{Y}_n - c\ )$ for appropriate constant c.

2.85. Let $X_n = \sqrt{n}$ with probability $1/n$ and $X_n = 0$ with probability $1 - 1/n$.
($X_n = \sqrt{n} I_{[0,1/n]}$ wrt U(0,1) probability.)

a) Prove that $X_n \xrightarrow{1} 0$.

b) Does $X_n \xrightarrow{2} 0$? Prove or disprove.

2.86. Suppose $X_n \sim U(c - 1/n, c + 1/n)$. Does $X_n \xrightarrow{D} X$ for some random variable X? Prove or disprove. (If $Y \sim U(\theta_1, \theta_2)$, then the cdf of Y is $F(y) = (y - \theta_1)/(\theta_2 - \theta_1)$ for $\theta_1 \le y \le \theta_2$.)

2.87. Let $X_n \sim N(0, \sigma_n^2)$ where $\sigma_n^2 \to \infty$ as $n \to \infty$. Let $\Phi(x)$ be the cdf of a $N(0,1)$ RV. Then the cdf of X_n is $F_n(x) = \Phi(x/\sigma_n)$.

a) Find $F(x)$ such that $F_n(x) \to F(x)$ for all real x.

b) Does $X_n \xrightarrow{D} X$? Explain briefly.

2.88. Suppose $X_1, ..., X_n$ are iid $C(\mu, \sigma)$ with characteristic function $\varphi_X(t) = \exp(it\mu - |t|\sigma)$ where $\exp(a) = e^a$.

a) Find the characteristic function $\varphi_{T_n}(t)$ of $T_n = \sum_{i=1}^n X_i$.

b) Find the characteristic function of $\overline{X}_n = T_n/n$.

c) Does $\overline{X}_n \xrightarrow{D} W$ for some RV W? Explain.

2.89. Let $P(X_n = 1) = 1/n$ and $P(X_n = 0) = 1 - 1/n$.

a) Find $P(|X_n| \ge \epsilon)$ for $0 < \epsilon \le 1$.
(Note that $P(|X_n| \ge \epsilon) = 0$ for $\epsilon > 1$.)

b) Does X_n converge in probability? Explain.

2.90. Let $P(X_n = 0) = 1 - 1/n$ and $P(X_n = 1) = 1/n$. Prove $X_n \xrightarrow{2} 0$ by showing $E[(X_n - 0)^2] \to 0$ as $n \to \infty$.

2.91. Let Y_n and Y be random variables such that $Y_n = Y$ with probability $1 - p_n$ and $Y_n = n$ with probability p_n where $p_n \to 0$. Prove or disprove: $Y_n \xrightarrow{D} Y$.

2.92^Q. a) Suppose that $X_n \sim U(-1/n, 1/n)$. Prove whether or not X_n converges in distribution to a random variable X.

b) Suppose $Y_n \sim U(0, n)$. Prove whether or not X_n converges in distribution to a random variable X.

2.93^Q. Prove whether the following sequences of random variables X_n converge in distribution to some random variable X. If $X_n \xrightarrow{D} X$, find the distribution of X (for example, find $F_X(t)$ or note that $P(X = c) = 1$, so X has the point mass distribution at c).

a) $X_n \sim U(-n - 1, -n)$

b) $X_n \sim U(n, n + 1)$

c) $X_n \sim U(a_n, b_n)$ where $a_n \to a < b$ and $b_n \to b$.

d) $X_n \sim U(a_n, b_n)$ where $a_n \to c$ and $b_n \to c$.

e) $X_n \sim U(-n, n)$

f) $X_n \sim U(c - 1/n, c + 1/n)$

2.94^Q. a) Let $P(X_n = n) = 1/n$ and $P(X_n = 0) = 1 - 1/n$.

i) Determine whether $X_n \xrightarrow{1} 0$.

ii) Determine whether $X_n \xrightarrow{P} 0$.

iii) Determine whether $X_n \xrightarrow{D} 0$.

b) Let $P(X_n = 0) = 1 - \dfrac{1}{n}$ and $P(X_n = 1) = 1/n$.

i) Determine whether $X_n \xrightarrow{2} 0$.

ii) Determine whether $X_n \xrightarrow{1} 0$.

iii) Determine whether $X_n \xrightarrow{P} 0$.

iv) Determine whether $X_n \xrightarrow{D} 0$.

2.95^Q. Let $X_1, ..., X_n$ be independent identically distributed (iid) beta(β, β) random variables.

a) Find the limiting distribution of $\sqrt{n}(\ \overline{X}_n - \theta\ )$, for appropriate constant θ.

b) Find the limiting distribution of $\sqrt{n}(\ \log(\overline{X}_n) - d\ )$, for appropriate constant d.

2.96. Let $X_1, ..., X_n$ be a random sample of size n from $U(\theta, 2\theta)$.

a) Find the limiting distribution of $\sqrt{n}(\overline{X} - c)$ for an appropriate constant c.

b) Find the limiting distribution of $\sqrt{n}(\log(\overline{X}) - d)$ for an appropriate constant d.

2.97. Let $Y_n \sim$ Poisson(n).

a) Find the limiting distribution of $\sqrt{n}\ \left(\ \dfrac{Y_n}{n} -\ 1\ \right)$.

b) Find the limiting distribution of $\sqrt{n}\ \left[\left(\frac{Y_n}{n} \right)^2 - 1 \right]$.

2.98*. Let $Y_1, ..., Y_n$ be iid uniform $U(\theta, 2\theta)$ for $\theta > 0$ and iid $U(2\theta, \theta)$ for $\theta < 0$.

a) Find the limiting distribution of $\sqrt{n}[\ \overline{Y} - c]$ for appropriate constant c.

b) Find the limiting distribution of $\sqrt{n}[\ (\overline{Y})^2 - d]$ for appropriate constant d.

2.99*. Let $\boldsymbol{x}_1, ..., \boldsymbol{x}_k$ be iid with $E(\boldsymbol{x}) = \boldsymbol{\mu}$ where $\boldsymbol{x}$ is $p \times 1$. Let $n =$ floor(k/2) $= \lfloor k/2 \rfloor$ be the integer part of $k/2$. So floor(100/2) = floor(101/2) = 50. Let the iid random variables $W_i = \boldsymbol{x}_{2i-1}^T \boldsymbol{x}_{2i}$ for $i = 1, ..., n$. Hence $W_1, W_2, ..., W_n = \boldsymbol{x}_1^T \boldsymbol{x}_2, \boldsymbol{x}_3^T \boldsymbol{x}_4, ..., \boldsymbol{x}_{2n-1}^T \boldsymbol{x}_{2n}$. Then $E(W_i) = \boldsymbol{\mu}^T \boldsymbol{\mu} = \theta \geq 0$ and $V(W_i) = \sigma_W^2$.

a) Find the limiting distribution of $\sqrt{n}(\overline{W} - \theta)$.

b) If $\theta > 0$, find the limiting distribution of $\sqrt{n}\left(\sqrt{\overline{W}} - \sqrt{\theta}\right)$.

2.100. Suppose X_n is a sequence of random variables with $P(X_n = 1/n) = 0.5$ and $P(X_n = -1/n) = 0.5$.

a) Show whether or not $X_n \xrightarrow{1} 0$ (convergence in rth mean with $r = 1$).

b) Does $X_n \xrightarrow{D} X$ for some random variable X? Prove or disprove.

2.101. Let $Y_1, ..., Y_n$ be iid beta$(\delta = \theta, \nu = 1)$ with $E(Y) = \dfrac{\theta}{\theta + 1}$, $V(Y) = \dfrac{\theta}{(\theta + 1)^2(\theta + 2)}$, and $I_1(\theta) = \dfrac{1}{\theta^2}$. The beta$(\theta, 1)$ distribution is a 1PREF. Let $\hat{\theta}_n$ be the MLE of θ. Find the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$.

2.102. Let $Y_1, ..., Y_n$ be iid $C(\mu, \sigma)$. Then the pdf of Y_i is

$$f(y) = \frac{1}{\pi\sigma[1 + (\frac{y-\mu}{\sigma})^2]}$$

where y and μ are real numbers and $\sigma > 0$. Then $MED(Y) = \mu$. Find the limiting distribution of $\sqrt{n}(MED(n) - \mu)$.

2.103. Let $Y_1, ..., Y_n$ be independent and identically distributed (iid) from a Gamma(α, β) distribution.

a) Find the limiting distribution of $\sqrt{n}\ \left(\overline{Y} - \alpha\beta \right)$.

b) Find the limiting distribution of $\sqrt{n}\ \left((\overline{Y})^2 - c \right)$ for appropriate constant c.

2.104. Let $X_1, ..., X_n$ be independent and identically distributed (iid) from a $N(\mu, \sigma^2)$ distribution. Let $\overline{X} = \dfrac{1}{n} \sum_{i=1}^{n} X_i$.

a) Find the limiting distribution of $\sqrt{n}\ (\ \overline{X} - \mu\ )$.

b) Find the limiting distribution of

$$\sqrt{n}\ \left[\frac{1}{\overline{X}} - c \right]$$

for appropriate constant c. You may assume $\mu \neq 0$.

2.105. Suppose $Y_1, ..., Y_n$ are iid gamma(ν, λ), $Y \sim G(\nu, \lambda)$, where ν is known. Then $I_1(\lambda) = \nu/\lambda^2$. Is $\hat{\lambda}_n = \overline{Y}_n/\nu$ an asymptotically efficient estimator of λ? Hint: determine if

$$\sqrt{n}(\overline{Y}_n/\nu - \lambda) \xrightarrow{D} N\left(0, \frac{1}{I_1(\lambda)}\right).$$

2.106. Let $W_1, \ldots, W_n$ be iid random variables with probability density function (pdf)

$$f(w) = \frac{3w^2}{\lambda} e^{-w^3/\lambda}$$

if $w > 0$, and $f(w) = 0$, elsewhere, where $\lambda > 0$, and $I_1(\lambda) = 1/\lambda^2$. This distribution is a 1PREF. Let $\hat{\lambda}$ be the MLE of λ, Find the limiting distribution of $\sqrt{n}(\hat{\lambda} - \lambda)$.

2.107. Suppose $X_1, ..., X_n$ are iid from a distribution with $E(X^k) = 2\theta^k/(k+2)$. Find the limiting distribution of $\sqrt{n}(\ \overline{X}_n - c\ )$ for appropriate constant c.

2.108. Let $Y_1, ..., Y_n$ be iid from a distribution with pdf

$$f(y) = \frac{\theta}{y^2} \exp\left(\frac{-\theta}{y}\right)$$

where $y > 0$ and $\theta > 0$. Then $MED(Y) = \theta/\log(2)$. Find the limiting distribution of $\sqrt{n}(MED(n) - MED(Y))$.

2.109. Let $Y_1, ..., Y_n$ be iid from a 1PREF with parameter θ. Let $\hat{\theta}$ be the MLE of θ with $I_1(\theta) = \dfrac{1}{\theta^2}$.

a) Find the limiting distribution of $\sqrt{n}(\hat{\theta} - \theta)$.

b) Find the limiting distribution of $\sqrt{n}[\hat{\theta}^2 - \theta^2]$.

2.110. Let X_k be sequence of independent Poisson $(1/2^k)$ random variables for k =0, 1, 2, ... Let $S_n = \sum_{k=0}^{n} X_k$. Then the characteristic function of S_n is

$$c_{S_n}(t) = \prod_{k=0}^{n} c_{X_k}(t) = \prod_{k=0}^{n} \exp\left[\frac{1}{2^k}(e^{it} - 1)\right]$$

$$= \exp\left[(\sum_{k=0}^{n} \frac{1}{2^k})(e^{it} - 1)\right] \sim Pois(\sum_{k=0}^{n} \frac{1}{2^k}).$$

Using $\sum_{k=0}^{\infty} \frac{1}{2^k} = 2$ and the continuity theorem for characteristic functions, find $\lim_{n\to\infty} c_{S_n}(t) = c_S(t)$, and thus prove $S_n \xrightarrow{D} S$. Identify the distribution of the random variable S.

2.111. Suppose $Y_1, ..., Y_n$ are iid $POIS(\theta)$. Then $I_1(\theta) = 1/\theta$. Is $\hat{\theta}_n = \overline{Y}_n$ an asymptotically efficient estimator of θ? Hint: determine if

$$\sqrt{n}(\overline{Y}_n - \theta) \xrightarrow{D} N\left(0, \frac{1}{I_1(\theta)}\right).$$

2.112. Suppose that the characteristic function of $X_n \sim N(\mu_n, \sigma_n^2)$ is

$$c_{X_n}(t) = \exp(it\mu_n - t^2\sigma_n^2/2).$$

Suppose that $\mu_n \to \mu$ and $\sigma_n^2 \to \sigma^2$ as $n \to \infty$. Does $X_n \xrightarrow{D} X$ for some random variable X? Explain. (Hint: Does $c_{X_n}(t) \to c_X(t)$ as $n \to \infty$?)

2.113. Suppose the Z_i are iid with $E(Z_i) = \mu$ and $V(Z_i) = \sigma^2$. Let $X_i = (Z_i + Z_{i+1})/2$. Using Slutsky's theorem ane the work below, show $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{D} W$ and find the distribution of W.

It can be shown that

$$\sqrt{n}(\overline{X}_n - \mu) = \sqrt{n-1}\left(\frac{Z_2 + \cdots + Z_n}{n-1} - \mu\right)\sqrt{\frac{n-1}{n}} + \frac{Z_1 + Z_{n+1}}{2\sqrt{n}} - \frac{\mu}{\sqrt{n}}$$

$$= \sqrt{n-1}(\overline{Z}_{n-1} - \mu)\sqrt{\frac{n-1}{n}} + \frac{Z_1 + Z_{n+1}}{2\sqrt{n}} - \frac{\mu}{\sqrt{n}}.$$

2.114. Suppose $r_n \xrightarrow{P} 0$ and $W_n \xrightarrow{D} W$. Let $Z_n = W_n + r_n$. Then $Z_n = W_n + r_n \xrightarrow{D} Z$ by Slutsky's Theorem. What is Z?

2.115. Suppose $Y_1, ..., Y_n$ are iid $EXP(\lambda)$. Then the MLE of λ is $\hat{\lambda}_n = \overline{Y}_n$, and $I_1(\lambda) = 1/\lambda^2$.

a) Find the limiting distribution of $\sqrt{n}(\overline{Y}_n - c)$ for appropriate constant c.

b) The Standard Limit Theorem for the MLE $\hat{\lambda}_n$ says

$$\sqrt{n}(\hat{\lambda}_n - \lambda) \xrightarrow{D} N\left(0, \frac{1}{I_1(\lambda)}\right).$$

Using a), prove that the Standard Limit Theorem holds for Y_i iid $EXP(\lambda)$.

2.116. Suppose $Y_1, ..., Y_n$ are iid and $W_i = t(Y_i)$ for a function t such that $E(W_i) = \mu_W$ and $V(W_i) = \sigma_W^2$. a) Find the limiting distribution of $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} t(Y_i) - c\right)$ for appropriate constant c.

Repeat a) if $W_i = t(Y_i) = Y_i^k$ for positive integer k, assuming that $E(W_i)$ and $E(W_i^2)$ are finite. This part gives a limit theorem for the sample kth moment. So give simple formulas for c, μ_W, and σ_W^2.

2.117. Let $X_1, ..., X_n$ be iid with mean $E(X) = \mu$ and variance $V(X) = \sigma^2 > 0$. Then $\exp[\sqrt{n}(\overline{X} - \mu)] \xrightarrow{D} W$. What is W? Hint: use the continuous mapping theorem: if $Z_n \xrightarrow{D} Z$ and g is continuous, then $g(Z_n) \xrightarrow{D} g(Z)$.

Chapter 3
Multivariate Limit Theorems

This chapter discusses multivariate limit theorems, and follows Olive (2014, § 8.6, 8.7) closely. Review Section 1.3 on characteristic functions and moment generating functions.

3.1 Multivariate Limit Theorems

Many of the univariate results from Chapter 2 can be extended to random vectors. For the limit theorems, the vector $\boldsymbol{X}$ is typically a $k \times 1$ column vector and $\boldsymbol{X}^T$ is a row vector. Let $\|\boldsymbol{x}\| = \sqrt{x_1^2 + \cdots + x_k^2}$ be the Euclidean norm of $\boldsymbol{x}$.

Definition 3.1. Let $\boldsymbol{X}_n$ be a sequence of random vectors with joint cdfs $F_n(\boldsymbol{x})$ and let $\boldsymbol{X}$ be a random vector with joint cdf $F(\boldsymbol{x})$.

a) $\boldsymbol{X}_n$ **converges in distribution** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$, if $F_n(\boldsymbol{x}) \to F(\boldsymbol{x})$ as $n \to \infty$ for all points $\boldsymbol{x}$ at which $F(\boldsymbol{x})$ is continuous. The distribution of $\boldsymbol{X}$ is the **limiting distribution** or **asymptotic distribution** of $\boldsymbol{X}_n$.

b) $\boldsymbol{X}_n$ **converges in probability** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, if for every $\epsilon > 0$, $P(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \epsilon) \to 0$ as $n \to \infty$.

c) Let $r > 0$ be a real number. Then $\boldsymbol{X}_n$ **converges in rth mean** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{r} \boldsymbol{X}$, if $E(\|\boldsymbol{X}_n - \boldsymbol{X}\|^r) \to 0$ as $n \to \infty$.

d) $\boldsymbol{X}_n$ **converges almost everywhere** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{ae} \boldsymbol{X}$, if $P(\lim_{n\to\infty} \boldsymbol{X}_n = \boldsymbol{X}) = 1$.

Theorems 3.1, 3.2, and 3.3 below are the multivariate extensions of the limit theorems in Section 2.1. When the limiting distribution of $\boldsymbol{Z}_n = \sqrt{n}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta}))$ is multivariate normal $N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$, approximate the joint cdf of $\boldsymbol{Z}_n$ with the joint cdf of the $N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$ distribution. Thus to find probabilities, manipulate $\boldsymbol{Z}_n$ as if $\boldsymbol{Z}_n \approx N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. To see that the CLT is a special case of the MCLT below, let $k = 1$, $E(X) = \mu$ and $V(X) = \boldsymbol{\Sigma} = \sigma^2$.

Theorem 3.1: the Multivariate Central Limit Theorem (MCLT). If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $k \times 1$ random vectors with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$, then

$$\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$$

where the sample mean

$$\overline{\boldsymbol{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i.$$

The MCLT is proven after Theorem 3.8.

Remark 3.1. The behavior of convergence in distribution to a MVN distribution in B) is much like the behavior of the MVN distributions in A). The results in B) can be proven using the multivariate delta method. Let $\boldsymbol{A}$ be a $q \times k$ constant matrix, b a constant, $\boldsymbol{a}$ a $k \times 1$ constant vector, and $\boldsymbol{d}$ a $q \times 1$ constant vector. Note that $\boldsymbol{a} + b\boldsymbol{X}_n = a + \boldsymbol{A}\boldsymbol{X}_n$ with $\boldsymbol{A} = b\boldsymbol{I}$. Thus i) and ii) follow from iii).

A) Suppose $\boldsymbol{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then
i) $\boldsymbol{A}\boldsymbol{X} \sim N_q(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.
ii) $\boldsymbol{a} + b\boldsymbol{X} \sim N_k(\boldsymbol{a} + b\boldsymbol{\mu}, b^2\boldsymbol{\Sigma})$.
iii) $\boldsymbol{A}\boldsymbol{X} + \boldsymbol{d} \sim N_q(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{d}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.
(Find the mean and covariance matrix of the left hand side and plug in those values for the right hand side. **Be careful with the dimension** k or q.)

B) Suppose $\boldsymbol{X}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then
i) $\boldsymbol{A}\boldsymbol{X}_n \xrightarrow{D} N_q(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.
ii) $\boldsymbol{a} + b\boldsymbol{X}_n \xrightarrow{D} N_k(\boldsymbol{a} + b\boldsymbol{\mu}, b^2\boldsymbol{\Sigma})$.
iii) $\boldsymbol{A}\boldsymbol{X}_n + \boldsymbol{d} \xrightarrow{D} N_q(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{d}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

To see that the delta method is a special case of the multivariate delta method, note that if T_n and parameter θ are real valued, then $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} = g'(\theta)$.

Theorem 3.2: the Multivariate Delta Method. If

$$\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{\theta}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

then

$$\sqrt{n}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta})) \xrightarrow{D} N_d(\boldsymbol{0}, \boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}\boldsymbol{\Sigma}\boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\theta})})$$

if $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}\boldsymbol{\Sigma}\boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\theta})}$ is nonsingular, where the $d \times k$ Jacobian matrix of partial derivatives

$$\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} = \begin{bmatrix} \frac{\partial}{\partial\theta_1} g_1(\boldsymbol{\theta}) & \dots & \frac{\partial}{\partial\theta_k} g_1(\boldsymbol{\theta}) \\ \vdots & & \vdots \\ \frac{\partial}{\partial\theta_1} g_d(\boldsymbol{\theta}) & \dots & \frac{\partial}{\partial\theta_k} g_d(\boldsymbol{\theta}) \end{bmatrix}.$$

Here the mapping $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^d$ needs to be differentiable in a neighborhood of $\boldsymbol{\theta} \in \mathbb{R}^k$.

Example 3.1. If Y has a Weibull distribution, $Y \sim W(\phi, \lambda)$, then the pdf of Y is

$$f(y) = \frac{\phi}{\lambda} y^{\phi-1} e^{-\frac{y^\phi}{\lambda}}$$

where λ, y, and ϕ are all positive. If $\mu = \lambda^{1/\phi}$ so $\mu^\phi = \lambda$, then the Weibull pdf

$$f(y) = \frac{\phi}{\mu}\left(\frac{y}{\mu}\right)^{\phi-1} \exp\left[-\left(\frac{y}{\mu}\right)^\phi\right].$$

Let $(\hat{\mu}, \hat{\phi})$ be the MLE of (μ, ϕ). According to Bain (1978, p. 215),

$$\sqrt{n}\left(\begin{pmatrix}\hat{\mu}\\ \hat{\phi}\end{pmatrix} - \begin{pmatrix}\mu\\ \phi\end{pmatrix}\right) \xrightarrow{D} N\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}1.109\frac{\mu^2}{\phi^2} & 0.257\mu\\ 0.257\mu & 0.608\phi^2\end{pmatrix}\right)$$

$= N_2(\mathbf{0}, \boldsymbol{I}^{-1}(\boldsymbol{\theta}))$ where $\boldsymbol{I}(\boldsymbol{\theta})$ is given in Definition 3.2.

Let column vectors $\boldsymbol{\theta} = (\mu \ \ \phi)^T$ and $\boldsymbol{\eta} = (\lambda \ \ \phi)^T$. Then

$$\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\theta}) = \begin{pmatrix}\lambda\\ \phi\end{pmatrix} = \begin{pmatrix}\mu^\phi\\ \phi\end{pmatrix} = \begin{pmatrix}g_1(\boldsymbol{\theta})\\ g_2(\boldsymbol{\theta})\end{pmatrix}.$$

So $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} =$

$$\begin{bmatrix}\frac{\partial}{\partial\theta_1}g_1(\boldsymbol{\theta}) & \frac{\partial}{\partial\theta_2}g_1(\boldsymbol{\theta})\\ \frac{\partial}{\partial\theta_1}g_2(\boldsymbol{\theta}) & \frac{\partial}{\partial\theta_2}g_2(\boldsymbol{\theta})\end{bmatrix} = \begin{bmatrix}\frac{\partial}{\partial\mu}\mu^\phi & \frac{\partial}{\partial\phi}\mu^\phi\\ \frac{\partial}{\partial\mu}\phi & \frac{\partial}{\partial\phi}\phi\end{bmatrix} = \begin{bmatrix}\phi\mu^{\phi-1} & \mu^\phi\log(\mu)\\ 0 & 1\end{bmatrix}.$$

Thus by the multivariate delta method,

$$\sqrt{n}\left(\begin{pmatrix}\hat{\lambda}\\ \hat{\phi}\end{pmatrix} - \begin{pmatrix}\lambda\\ \phi\end{pmatrix}\right) \xrightarrow{D} N_2(\mathbf{0}, \boldsymbol{\Sigma})$$

where (see Definition 3.4 below)

$$\boldsymbol{\Sigma} = \boldsymbol{I}(\boldsymbol{\eta})^{-1} = [\boldsymbol{I}(\boldsymbol{g}(\boldsymbol{\theta}))]^{-1} = \boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}\boldsymbol{I}^{-1}(\boldsymbol{\theta})\boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\theta})} =$$

$$\begin{bmatrix}1.109\lambda^2(1 + 0.4635\log(\lambda) + 0.5482(\log(\lambda))^2) & 0.257\phi\lambda + 0.608\lambda\phi\log(\lambda)\\ 0.257\phi\lambda + 0.608\lambda\phi\log(\lambda) & 0.608\phi^2\end{bmatrix}.$$

Definition 3.2. Let X be a random variable with pdf or pmf $f(x|\boldsymbol{\theta})$. Then the **information matrix**

$$\boldsymbol{I}(\boldsymbol{\theta}) = [\boldsymbol{I}_{i,j}]$$

where

$$\boldsymbol{I}_{i,j} = E\left[\frac{\partial}{\partial\theta_i}\log(f(X|\boldsymbol{\theta}))\frac{\partial}{\partial\theta_j}\log(f(X|\boldsymbol{\theta}))\right].$$

Definition 3.3. An estimator $\boldsymbol{T}_n$ of $\boldsymbol{\theta}$ is **asymptotically efficient** if

$$\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{\theta}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{I}^{-1}(\boldsymbol{\theta})).$$

Following Lehmann (1999, p. 511), if $\boldsymbol{T}_n$ is asymptotically efficient and if the estimator $\boldsymbol{W}_n$ satisfies

$$\sqrt{n}(\boldsymbol{W}_n - \boldsymbol{\theta}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{J}(\boldsymbol{\theta}))$$

where $\boldsymbol{J}(\boldsymbol{\theta})$ and $\boldsymbol{I}^{-1}(\boldsymbol{\theta})$ are continuous functions of $\boldsymbol{\theta}$, then under regularity conditions, $\boldsymbol{J}(\boldsymbol{\theta}) - \boldsymbol{I}^{-1}(\boldsymbol{\theta})$ is a positive semidefinite matrix, and $\boldsymbol{T}_n$ is "better" than $\boldsymbol{W}_n$.

Definition 3.4. Assume that $\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\theta})$. Then

$$\boldsymbol{I}(\boldsymbol{\eta}) = \boldsymbol{I}(\boldsymbol{g}(\boldsymbol{\theta})) = [\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}\boldsymbol{I}^{-1}(\boldsymbol{\theta})\boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\theta})}]^{-1}.$$

Notice that this definition agrees with the multivariate delta method if

$$\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{\theta}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma} = \boldsymbol{I}^{-1}(\boldsymbol{\theta})$.

Now suppose that $X_1, ..., X_n$ are iid random variables from a k-parameter REF

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta})\exp\left[\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(x)\right] \tag{3.1}$$

with natural parameterization

$$f(x|\boldsymbol{\eta}) = h(x)b(\boldsymbol{\eta})\exp\left[\sum_{i=1}^{k}\eta_i t_i(x)\right]. \tag{3.2}$$

Then the complete minimal sufficient statistic is

$$\overline{\boldsymbol{T}}_n = \frac{1}{n}(\sum_{i=1}^{n} t_1(X_i), ..., \sum_{i=1}^{n} t_k(X_i))^T.$$

Let $\boldsymbol{\mu}_T = (E(t_1(X), ..., E(t_k(X)))^T$. From Theorem 1.31, for $\boldsymbol{\eta} \in \Omega$,

$$E(t_i(X)) = \frac{-\partial}{\partial \eta_i} \log(b(\boldsymbol{\eta})),$$

and

$$\mathrm{Cov}(\mathrm{t_i(X)}, \mathrm{t_j(X)}) \equiv \sigma_{\mathrm{i,j}} = \frac{-\partial^2}{\partial \eta_{\mathrm{i}} \partial \eta_{\mathrm{j}}} \log(\mathrm{b}(\boldsymbol{\eta})).$$

Theorem 3.3. If the random variable X is a kP–REF with pmf or pdf (3.2), then the information matrix

$$\boldsymbol{I}(\boldsymbol{\eta}) = [\boldsymbol{I}_{i,j}]$$

where

$$\boldsymbol{I}_{i,j} = E\left[\frac{\partial}{\partial \eta_i} \log(f(X|\boldsymbol{\eta})) \frac{\partial}{\partial \eta_j} \log(f(X|\boldsymbol{\eta}))\right] = -E\left[\frac{\partial^2}{\partial \eta_i \partial \eta_j} \log(f(X|\boldsymbol{\eta}))\right].$$

Several authors, including Barndorff–Nielsen (1982), have noted that the multivariate CLT can be used to show that $\sqrt{n}(\overline{\boldsymbol{T}}_n - \boldsymbol{\mu}_T) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. The fact that $\boldsymbol{\Sigma} = \boldsymbol{I}(\boldsymbol{\eta})$ appears in Lehmann (1983, p. 127).

Theorem 3.4. If $X_1, ..., X_n$ are iid from a k-parameter regular exponential family, then

$$\sqrt{n}(\overline{\boldsymbol{T}}_n - \boldsymbol{\mu}_T) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{I}(\boldsymbol{\eta})).$$

Proof. By the multivariate central limit theorem,

$$\sqrt{n}(\overline{\boldsymbol{T}}_n - \boldsymbol{\mu}_T) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma} = [\sigma_{i,j}]$. Hence the result follows if $\sigma_{i,j} = \boldsymbol{I}_{i,j}$. Since

$$\log(f(x|\boldsymbol{\eta})) = \log(h(x)) + \log(b(\boldsymbol{\eta})) + \sum_{l=1}^{k} \eta_l t_l(x),$$

$$\frac{\partial}{\partial \eta_i} \log(f(x|\boldsymbol{\eta})) = \frac{\partial}{\partial \eta_i} \log(b(\boldsymbol{\eta})) + t_i(X).$$

Hence

$$-\boldsymbol{I}_{i,j} = E\left[\frac{\partial^2}{\partial \eta_i \partial \eta_j} \log(f(X|\boldsymbol{\eta}))\right] = \frac{\partial^2}{\partial \eta_i \partial \eta_j} \log(b(\boldsymbol{\eta})) = -\sigma_{i,j}. \quad \square$$

To obtain standard results, use the multivariate delta method, assume that both $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are $k \times 1$ vectors, and assume that $\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\theta})$ is a one to one mapping so that the inverse mapping is $\boldsymbol{\theta} = \boldsymbol{g}^{-1}(\boldsymbol{\eta})$. If $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}$ is nonsingular, then

$$\boldsymbol{D}^{-1}_{\boldsymbol{g}(\boldsymbol{\theta})} = \boldsymbol{D}_{\boldsymbol{g}^{-1}(\boldsymbol{\eta})} \tag{3.3}$$

(see Searle 1982, p. 339), and

$$\boldsymbol{I}(\boldsymbol{\eta}) = [\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}\boldsymbol{I}^{-1}(\boldsymbol{\theta})\boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\theta})}]^{-1}$$

$$= [\boldsymbol{D}^{-1}_{\boldsymbol{g}(\boldsymbol{\theta})}]^T\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{D}^{-1}_{\boldsymbol{g}(\boldsymbol{\theta})} = \boldsymbol{D}^T_{\boldsymbol{g}^{-1}(\boldsymbol{\eta})}\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{D}_{\boldsymbol{g}^{-1}(\boldsymbol{\eta})}. \tag{3.4}$$

Compare Lehmann (1999, p. 500) and Lehmann (1983, p. 127).

For example, suppose that $\boldsymbol{\mu}_T$ and $\boldsymbol{\eta}$ are $k \times 1$ vectors, and

$$\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{I}^{-1}(\boldsymbol{\eta}))$$

where $\boldsymbol{\mu}_T = \boldsymbol{g}(\boldsymbol{\eta})$ and $\boldsymbol{\eta} = \boldsymbol{g}^{-1}(\boldsymbol{\mu}_T)$. Also assume that $\overline{\boldsymbol{T}}_n = \boldsymbol{g}(\hat{\boldsymbol{\eta}})$ and $\hat{\boldsymbol{\eta}} = \boldsymbol{g}^{-1}(\overline{\boldsymbol{T}}_n)$. Then by the multivariate delta method and Theorem 3.4,

$$\sqrt{n}(\overline{\boldsymbol{T}}_n - \boldsymbol{\mu}_T) = \sqrt{n}(\boldsymbol{g}(\hat{\boldsymbol{\eta}}) - \boldsymbol{g}(\boldsymbol{\eta})) \xrightarrow{D} N_k[\boldsymbol{0}, \boldsymbol{I}(\boldsymbol{\eta})] = N_k[\boldsymbol{0}, \boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\eta})}\boldsymbol{I}^{-1}(\boldsymbol{\eta})\boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\eta})}].$$

Hence

$$\boldsymbol{I}(\boldsymbol{\eta}) = \boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\eta})}\boldsymbol{I}^{-1}(\boldsymbol{\eta})\boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\eta})}.$$

Similarly,

$$\sqrt{n}(\boldsymbol{g}^{-1}(\overline{\boldsymbol{T}}_n) - \boldsymbol{g}^{-1}(\boldsymbol{\mu}_T)) = \sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \xrightarrow{D} N_k[\boldsymbol{0}, \boldsymbol{I}^{-1}(\boldsymbol{\eta})] =$$

$$N_k[\boldsymbol{0}, \boldsymbol{D}_{\boldsymbol{g}^{-1}(\boldsymbol{\mu}_T)}\boldsymbol{I}(\boldsymbol{\eta})\boldsymbol{D}^T_{\boldsymbol{g}^{-1}(\boldsymbol{\mu}_T)}].$$

Thus

$$\boldsymbol{I}^{-1}(\boldsymbol{\eta}) = \boldsymbol{D}_{\boldsymbol{g}^{-1}(\boldsymbol{\mu}_T)}\boldsymbol{I}(\boldsymbol{\eta})\boldsymbol{D}^T_{\boldsymbol{g}^{-1}(\boldsymbol{\mu}_T)} = \boldsymbol{D}_{\boldsymbol{g}^{-1}(\boldsymbol{\mu}_T)}\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\eta})}\boldsymbol{I}^{-1}(\boldsymbol{\eta})\boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\eta})}\boldsymbol{D}^T_{\boldsymbol{g}^{-1}(\boldsymbol{\mu}_T)}$$

as expected by Equation (3.4). Typically $\hat{\boldsymbol{\theta}}$ is a function of the sufficient statistic $\boldsymbol{T}_n$ and is the unique MLE of $\boldsymbol{\theta}$. Replacing $\boldsymbol{\eta}$ by $\boldsymbol{\theta}$ in the above discussion shows that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{I}^{-1}(\boldsymbol{\theta}))$ is equivalent to $\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{\mu}_T) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{I}(\boldsymbol{\theta}))$ provided that $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}$ is nonsingular.

3.2 More Multivariate Results

Definition 3.5. If the estimator $\boldsymbol{g}(\boldsymbol{T}_n) \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$, then $\boldsymbol{g}(\boldsymbol{T}_n)$ is a **consistent estimator** of $\boldsymbol{g}(\boldsymbol{\theta})$.

Theorem 3.5. If $0 < \delta \leq 1$, $\boldsymbol{X}$ is a random vector, and

$$n^{\delta}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta})) \xrightarrow{D} \boldsymbol{X},$$

then $\boldsymbol{g}(\boldsymbol{T}_n) \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta})$.

Theorem 3.6. If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid, $E(\|\boldsymbol{X}\|) < \infty$ and $E(\boldsymbol{X}) = \boldsymbol{\mu}$, then

a) WLLN: $\overline{\boldsymbol{X}}_n \xrightarrow{P} \boldsymbol{\mu}$ and

b) SLLN: $\overline{\boldsymbol{X}}_n \xrightarrow{ae} \boldsymbol{\mu}$.

Theorem 3.7: Continuity Theorem. Let $\boldsymbol{X}_n$ be a sequence of $k \times 1$ random vectors with characteristic function $c_n(\boldsymbol{t})$ and let $\boldsymbol{X}$ be a $k \times 1$ random vector with cf $c(\boldsymbol{t})$. Then

$$\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X} \text{ iff } \mathrm{c_n}(\boldsymbol{t}) \to \mathrm{c}(\boldsymbol{t})$$

for all $\boldsymbol{t} \in \mathbb{R}^k$.

Theorem 3.8: Cramér Wold Device. Let $\boldsymbol{X}_n$ be a sequence of $k \times 1$ random vectors and let $\boldsymbol{X}$ be a $k \times 1$ random vector. Then

$$\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X} \text{ iff } \boldsymbol{t}^{\mathrm{T}}\boldsymbol{X}_{\mathrm{n}} \xrightarrow{\mathrm{D}} \boldsymbol{t}^{\mathrm{T}}\boldsymbol{X}$$

for all $\boldsymbol{t} \in \mathbb{R}^k$.

Proof. (Serverini (2005, p. 337)): Let $W_n = \boldsymbol{t}^T\boldsymbol{X}_n$ and $W = \boldsymbol{t}^T\boldsymbol{X}$. Note that

$$c_{W_n}(y) = c_{\boldsymbol{t}^T\boldsymbol{X}_n}(y) = E\left[e^{iy\boldsymbol{t}^T\boldsymbol{X}_n}\right] = c_{\boldsymbol{X}_n}(y\boldsymbol{t})$$

where $y \in \mathbb{R}$, and similarly

$$c_W(y) = c_{\boldsymbol{t}^T\boldsymbol{X}}(y) = c_{\boldsymbol{X}}(y\boldsymbol{t})$$

where $y \in \mathbb{R}$.

If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$, then $c_{\boldsymbol{X}_n}(\boldsymbol{t}) \to c_{\boldsymbol{X}}(\boldsymbol{t}) \ \forall \ \boldsymbol{t} \in \mathbb{R}^k$. Fix $\boldsymbol{t}$. Then $c_{\boldsymbol{X}_n}(y\boldsymbol{t}) \to c_{\boldsymbol{X}}(y\boldsymbol{t}) \ \forall \ y \in \mathbb{R}$. Thus $\boldsymbol{t}^T\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{t}^T\boldsymbol{X}$.

Now assume $\boldsymbol{t}^T\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{t}^T\boldsymbol{X} \ \forall \ \boldsymbol{t} \in \mathbb{R}^k$. Then $c_{\boldsymbol{X}_n}(y\boldsymbol{t}) \to c_{\boldsymbol{X}}(y\boldsymbol{t}) \ \forall \ y \in \mathbb{R}$ and $\forall \ \boldsymbol{t} \in \mathbb{R}^k$. Take $y = 1$ to get $c_{\boldsymbol{X}_n}(\boldsymbol{t}) \to c_{\boldsymbol{X}}(\boldsymbol{t}) \ \forall \ \boldsymbol{t} \in \mathbb{R}^k$. Hence $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ by the Continuity Theorem. □

Application: Proof of the MCLT Theorem 3.1. Note that for fixed $\boldsymbol{t}$, the $\boldsymbol{t}^T\boldsymbol{X}_i$ are iid random variables with mean $\boldsymbol{t}^T\boldsymbol{\mu}$ and variance $\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}$. Hence by the CLT, $\boldsymbol{t}^T\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N(0, \boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$. The right hand side has distribution $\boldsymbol{t}^T\boldsymbol{X}$ where $\boldsymbol{X} \sim N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. Hence by the Cramér Wold Device, $\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. □

Theorem 3.9. a) If $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, then $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$.

b)

$$\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta}) \text{ iff } \boldsymbol{X}_{\mathrm{n}} \xrightarrow{\mathrm{D}} \boldsymbol{g}(\boldsymbol{\theta}).$$

Let $g(n) \geq 1$ be an increasing function of the sample size n: $g(n) \uparrow \infty$, e.g. $g(n) = \sqrt{n}$. See White (1984, p. 15). If a $k \times 1$ random vector $\boldsymbol{T}_n - \boldsymbol{\mu}$ converges to a nondegenerate multivariate normal distribution with convergence rate $\sqrt{n}$, then $\boldsymbol{T}_n$ has (tightness) rate $\sqrt{n}$.

Definition 3.6. Let $\boldsymbol{A}_n = [a_{i,j}(n)]$ be an $r \times c$ random matrix.
a) $\boldsymbol{A}_n = O_P(X_n)$ if $a_{i,j}(n) = O_P(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
b) $\boldsymbol{A}_n = o_p(X_n)$ if $a_{i,j}(n) = o_p(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
c) $\boldsymbol{A}_n \asymp_P (1/(g(n))$ if $a_{i,j}(n) \asymp_P (1/(g(n))$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
d) Let $\boldsymbol{A}_{1,n} = \boldsymbol{T}_n - \boldsymbol{\mu}$ and $\boldsymbol{A}_{2,n} = \boldsymbol{C}_n - c\boldsymbol{\Sigma}$ for some constant $c > 0$. If $\boldsymbol{A}_{1,n} \asymp_P (1/(g(n))$ and $\boldsymbol{A}_{2,n} \asymp_P (1/(g(n))$, then $(\boldsymbol{T}_n, \boldsymbol{C}_n)$ has (tightness) rate $g(n)$.

Theorem 3.10. Let W_n, X_n, Y_n and Z_n be sequences of random variables such that $Y_n > 0$ and $Z_n > 0$. (Often Y_n and Z_n are deterministic, e.g. $Y_n = n^{-1/2}$.)

a) If $W_n = O_P(1)$ and $X_n = O_P(1)$, then $W_n + X_n = O_P(1)$ and $W_n X_n = O_P(1)$, thus $O_P(1) + O_P(1) = O_P(1)$ and $O_P(1)O_P(1) = O_P(1)$.

b) If $W_n = O_P(1)$ and $X_n = o_P(1)$, then $W_n + X_n = O_P(1)$ and $W_n X_n = o_P(1)$, thus $O_P(1) + o_P(1) = O_P(1)$ and $O_P(1)o_P(1) = o_P(1)$.

c) If $W_n = O_P(Y_n)$ and $X_n = O_P(Z_n)$, then $W_n + X_n = O_P(\max(Y_n, Z_n))$ and $W_n X_n = O_P(Y_n Z_n)$, thus $O_P(Y_n) + O_P(Z_n) = O_P(\max(Y_n, Z_n))$ and $O_P(Y_n)O_P(Z_n) = O_P(Y_n Z_n)$.

Recall that the smallest integer function $\lceil x \rceil$ rounds up, e.g. $\lceil 7.7 \rceil = 8$.

Definition 3.7. The *sample ρ quantile* $\hat{y}_{n,\rho} = \hat{\xi}_{n,\rho} = Y_{(\lceil n\rho \rceil)}$. The *population quantile* $y_\rho = \xi_\rho = Q(\rho) = \inf\{y : F(y) \geq \rho\}$.

There are many other ways to define sample quantiles, and the different estimators tend to be asymptotically equivalent. If the inverse F^{-1} of the cdf exists, then $Q(u) = F^{-1}(u)$. $Q(u) \leq x$ iff $u \leq F(x)$. $F(y_\rho) = P(Y \leq y_\rho) \geq \rho$ and $P(Y \geq y_\rho) \geq 1 - \rho$. Let the observed data be $Y_1, ..., Y_n$, and let $\hat{F}(y) = \frac{1}{n} \sum_{i=1}^{n} I(Y_i \leq y)$. Then $\hat{Q}(\rho) = \inf\{y : \hat{F}(y) \geq \rho\} = \hat{y}_{n,\rho} = Y_{(\lceil n\rho \rceil)}$. (An alternative definition of the population quantile that is often used is that y_ρ is any real number satisfying $P(Y \leq y_\rho) \geq \rho$ and $P(Y \geq y_\rho) \geq 1 - \rho$. Then y_ρ is not necessarily unique. Definition 3.7 makes the population quantile unique. The regularity conditions in Theorem 3.11 make y_ρ unique if the alternative definition is used.)

Theorem 3.11: Serfling (1980, p. 80). Let $0 < \rho_1 < \rho_2 < \cdots < \rho_k < 1$. Suppose that F has a density f that is positive and continuous in neighborhoods of $\xi_{\rho_1}, ..., \xi_{\rho_k}$. Then

$$\sqrt{n}[(\hat{\xi}_{n,\rho_1}, ..., \hat{\xi}_{n,\rho_k})^T - (\xi_{\rho_1}, ..., \xi_{\rho_k})^T] \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma} = (\sigma_{ij})$ and

$$\sigma_{ij} = \frac{\rho_i(1-\rho_j)}{f(\xi_{\rho_i})f(\xi_{\rho_j})}$$

for $i \leq j$ and $\sigma_{ij} = \sigma_{ji}$ for $i > j$.

Theorem 3.12: Continuous Mapping Theorem. Let $\boldsymbol{X}_n \in \mathbb{R}^k$. If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and if the function $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^j$ is continuous and does not depend on n, then $\boldsymbol{g}(\boldsymbol{X}_n) \xrightarrow{D} \boldsymbol{g}(\boldsymbol{X})$.

The following theorem is an extension of Theorem 2.8.

Theorem 3.13: Generalized Chebyshev's Inequality or Generalized Markov's Inequality: Let $u : \mathbb{R}^k \to [0, \infty)$ be a nonnegative function. If $E[u(\boldsymbol{X})]$ exists, then for any $\epsilon > 0$,

$$P[u(\boldsymbol{X}) \geq \epsilon] \leq \frac{E[u(\boldsymbol{X})]}{\epsilon}.$$

Proof Sketch. The proof is nearly identical to that of Theorem 2.8.

Example 3.2. Let $u(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{c}\|^r$ for some $r > 0$. Often $\boldsymbol{c} = \boldsymbol{0}$ or $\boldsymbol{a} = E(\boldsymbol{X}) = \boldsymbol{\mu}$. If $E[u(\boldsymbol{X})]$ exists, then for any $\epsilon > 0$,

$$P(\|\boldsymbol{X} - \boldsymbol{c}\| \geq \epsilon] = P(\|\boldsymbol{X} - \boldsymbol{c}\|^r \geq \epsilon^r] \leq \frac{E[\|\boldsymbol{X} - \boldsymbol{c}\|^r]}{\epsilon^r}.$$

Theorem 3.14. Suppose $\boldsymbol{x}_n$ and $\boldsymbol{x}$ are random vectors with the same probability space.
a) If $\boldsymbol{x}_n \xrightarrow{P} \boldsymbol{x}$, then $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x}$.
b) If $\boldsymbol{x}_n \xrightarrow{wp1} \boldsymbol{x}$, then $\boldsymbol{x}_n \xrightarrow{P} \boldsymbol{x}$ and $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x}$.
c) If $\boldsymbol{x}_n \xrightarrow{r} \boldsymbol{x}$ for some $r > 0$, then $\boldsymbol{x}_n \xrightarrow{P} \boldsymbol{x}$ and $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x}$.
d) $\boldsymbol{x}_n \xrightarrow{P} \boldsymbol{c}$ **iff** $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{c}$ where $\boldsymbol{c}$ is a constant vector.

The proof of c) follows from the Generalized Chebyshev inequality. See Example 3.2.

Remark 3.2. Let $\boldsymbol{W}_n$ be a sequence of $m \times m$ random matrices and let $\boldsymbol{C}$ be an $m \times m$ constant matrix.
a) $\boldsymbol{W}_n \xrightarrow{P} \boldsymbol{X}$ iff $\boldsymbol{a}^T\boldsymbol{W}_n\boldsymbol{b} \xrightarrow{P} \boldsymbol{a}^T\boldsymbol{C}\boldsymbol{b}$ for all constant vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^m$.
b) If $\boldsymbol{W}_n \xrightarrow{P} \boldsymbol{C}$, then the determinant $det(\boldsymbol{W}_n) = |\boldsymbol{W}_n| \xrightarrow{P} |\boldsymbol{C}| = det(\boldsymbol{C})$.
c) If $\boldsymbol{W}_n^{-1}$ exists for each n and $\boldsymbol{C}^{-1}$ exists, then If $\boldsymbol{W}_n \xrightarrow{P} \boldsymbol{C}$ iff $\boldsymbol{W}_n^{-1} \xrightarrow{P} \boldsymbol{C}^{-1}$.

The following theorem is taken from Severini (2005, pp. 345-349).

Theorem 3.15. Let $\boldsymbol{X}_n = (X_{1n}, ..., X_{kn})^T$ be a sequence of $k \times 1$ random vectors, let $\boldsymbol{Y}_n$ be a sequence of $k \times 1$ random vectors, and let $\boldsymbol{X} = (X_1, ..., X_k)^T$ be a $k \times 1$ random vector. Let $\boldsymbol{W}_n$ be a sequence of $k \times k$

nonsingular random matrices, and let $\boldsymbol{C}$ be a $k \times k$ constant nonsingular matrix.

a) $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$ iff $X_{in} \xrightarrow{P} X_i$ for $i = 1, ..., k$.

b) **Slutsky's Theorem**: If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and $\boldsymbol{Y}_n \xrightarrow{P} \boldsymbol{c}$ for some constant $k \times 1$ vector $\boldsymbol{c}$, then i) $\boldsymbol{X}_n + \boldsymbol{Y}_n \xrightarrow{D} \boldsymbol{X} + \boldsymbol{c}$ and

ii) $\boldsymbol{Y}_n^T \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{c}^T \boldsymbol{X}$.

c) If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and $\boldsymbol{W}_n \xrightarrow{P} \boldsymbol{C}$, then $\boldsymbol{W}_n \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{C}\boldsymbol{X}$, $\boldsymbol{X}_n^T \boldsymbol{W}_n \xrightarrow{D} \boldsymbol{X}^T \boldsymbol{C}$, $\boldsymbol{W}_n^{-1} \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{C}^{-1} \boldsymbol{X}$, and $\boldsymbol{X}_n^T \boldsymbol{W}_n^{-1} \xrightarrow{D} \boldsymbol{X}^T \boldsymbol{C}^{-1}$.

Theorem 3.16. i) Suppose $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}$ be a $q \times p$ constant matrix. Then $\boldsymbol{A}\sqrt{n}(T_n - \boldsymbol{\mu}) = \sqrt{n}(\boldsymbol{A}T_n - \boldsymbol{A}\boldsymbol{\mu}) \xrightarrow{D} N_q(\boldsymbol{A}\boldsymbol{\theta}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

ii) Let $\boldsymbol{\Sigma} > 0$. If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ where $s > 0$ is some constant, then $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + o_P(1)$, so $D^2_{\boldsymbol{x}}(T, \boldsymbol{C})$ is a consistent estimator of $s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

iii) Let $\boldsymbol{\Sigma} > 0$. If $\sqrt{n}(T - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ and if $\boldsymbol{C}$ is a consistent estimator of $\boldsymbol{\Sigma}$, then $n(T - \boldsymbol{\mu})^T \boldsymbol{C}^{-1} (T - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$. In particular, $n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$.

Proof: ii) $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) =$
$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T [\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1} + s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)$
$= (\boldsymbol{x} - \boldsymbol{\mu})^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu}) + (\boldsymbol{x} - T)^T [\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - T)$
$+(\boldsymbol{x} - \boldsymbol{\mu})^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{\mu} - T) + (\boldsymbol{\mu} - T)^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu})$
$+(\boldsymbol{\mu} - T)^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{\mu} - T) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(1)$.

(Note that $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(n^{-\delta})$ if $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ with rate n^{δ} where $0 < \delta \leq 0.5$ if $[\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1}] = O_P(n^{-\delta})$.)

Alternatively, $D^2_{\boldsymbol{x}}(T, \boldsymbol{C})$ is a continuous function of $(T, \boldsymbol{C})$ if $\boldsymbol{C} > 0$ for $n > 10p$. Hence $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) \xrightarrow{P} D^2_{\boldsymbol{x}}(\mu, s\boldsymbol{\Sigma})$.

iii) Note that $\boldsymbol{Z}_n = \sqrt{n}\ \boldsymbol{\Sigma}^{-1/2}(T - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{I}_p)$. Thus $\boldsymbol{Z}_n^T \boldsymbol{Z}_n = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (T - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$. Now $n(T - \boldsymbol{\mu})^T \boldsymbol{C}^{-1} (T - \boldsymbol{\mu}) =$
$n(T - \boldsymbol{\mu})^T [\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}](T - \boldsymbol{\mu}) = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (T - \boldsymbol{\mu}) +$
$n(T - \boldsymbol{\mu})^T [\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1}](T - \boldsymbol{\mu}) = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (T - \boldsymbol{\mu}) + o_P(1) \xrightarrow{D} \chi^2_p$ since $\sqrt{n}(T - \boldsymbol{\mu})^T [\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1}] \sqrt{n}(T - \boldsymbol{\mu}) = O_P(1) o_P(1) O_P(1) = o_P(1)$. □

Theorem 3.17. Let $\boldsymbol{x}_n = (x_{1n}, ..., x_{kn})^T$ and $\boldsymbol{x} = (x_1, ..., x_k)^T$ be random vectors. Then $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x}$ implies $x_{in} \xrightarrow{D} x_i$ for $i = 1, ..., k$.

Proof. Use the Cramér Wold device with $\boldsymbol{t}_i = (0, ..., 0, 1, 0, ...0)^T$ where the 1 is in the ith position. Thus

$$\boldsymbol{t}_i^T \boldsymbol{x}_n = x_{in} \xrightarrow{D} x_i = \boldsymbol{t}_i^T \boldsymbol{x}.$$

□

Joint convergence in distribution implies marginal convergence in distribution by Theorem 3.16. Typically marginal convergence in distribution $\boldsymbol{x}_{in} \xrightarrow{D} \boldsymbol{x}_i$ for $i = 1, ..., m$ **does not imply** that

$$\begin{pmatrix} \boldsymbol{x}_{1n} \\ \vdots \\ \boldsymbol{x}_{mn} \end{pmatrix} \xrightarrow{D} \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_m \end{pmatrix}.$$

That is marginal convergence in distribution does not imply joint convergence in distribution. An exception is when the marginal random vectors are independent.

Example 3.3. Suppose that $\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{y}_n$ for $n = 1, 2, ...$ Suppose $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x}$, and $\boldsymbol{y}_n \xrightarrow{D} \boldsymbol{y}$ where $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$. Then

$$\begin{bmatrix} \boldsymbol{x}_n \\ \boldsymbol{y}_n \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}$$

by the continuity theorem. To see this, let $\boldsymbol{t} = (\boldsymbol{t}_1^T, \boldsymbol{t}_2^T)^T$, $\boldsymbol{z}_n = (\boldsymbol{x}_n^T, \boldsymbol{y}_n^T)^T$, and $\boldsymbol{z} = (\boldsymbol{x}^T, \boldsymbol{y}^T)^T$. Since $\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{y}_n$ and $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$, the characteristic function

$$\phi_{\boldsymbol{z}_n}(\boldsymbol{t}) = \phi_{\boldsymbol{x}_n}(\boldsymbol{t}_1)\phi_{\boldsymbol{y}_n}(\boldsymbol{t}_2) \to \phi_{\boldsymbol{x}}(\boldsymbol{t}_1)\phi_{\boldsymbol{y}}(\boldsymbol{t}_2) = \phi_{\boldsymbol{z}}(\boldsymbol{t}).$$

Hence $\boldsymbol{z}_n \xrightarrow{D} \boldsymbol{z}$ and $\boldsymbol{g}(\boldsymbol{z}_n) \xrightarrow{D} \boldsymbol{g}(\boldsymbol{z})$ if $\boldsymbol{g}$ is continuous by the continuous mapping theorem.

Remark 3.3. a) In the above example, we can show $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$ instead of assuming $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$. See Ferguson (1996, p. 42).

b) If $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x}$ and $\boldsymbol{y}_n \xrightarrow{P} \boldsymbol{c}$, a constant vector, then

$$\begin{bmatrix} \boldsymbol{x}_n \\ \boldsymbol{y}_n \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{c} \end{bmatrix}.$$

Note that a constant vector $\boldsymbol{c} \perp\!\!\!\perp \boldsymbol{x}$ for any random vector $\boldsymbol{x}$.

Example 3.4. a) Let $X \sim N(0, 1)$. Let $X_n = X \ \forall n$. Let

$$Y_n = \begin{cases} X, & n \text{ even} \\ -X, & n \text{ odd.} \end{cases}$$

Thus $Y_n \sim N(0, 1)$, $X_n \xrightarrow{D} X$, and $Y_n \xrightarrow{D} X$. Then

$$(1 \quad 1)\begin{pmatrix} X_n \\ Y_n \end{pmatrix} = X_n + Y_n = \begin{cases} 2X, & n \text{ even} \\ 0, & n \text{ odd} \end{cases}$$

does not converge in distribution as $n \to \infty$ by the Cramér Wold Device with $\boldsymbol{t} = (1 \quad 1)^T$. Thus

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix}$$

does not converge in distribution.

b) Let $X \sim N(0,1)$ and $W \sim N(0,1)$. Let $X_n = X \ \forall n$ and $Y_n = -X \ \forall n$. Then

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} = \begin{pmatrix} X \\ -X \end{pmatrix} \ \forall n, \text{ and } \begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{\text{D}} \begin{pmatrix} X \\ -X \end{pmatrix}.$$

Now $X_n \xrightarrow{D} W$ and $Y_n \xrightarrow{D} W$. Since

$$(1 \quad 1)\begin{pmatrix} X_n \\ Y_n \end{pmatrix} = X_n + Y_n = 0 \ \forall n, \quad \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$$

does not converge in distribution to

$$\begin{pmatrix} W \\ W \end{pmatrix}$$

as $n \to \infty$.

Example 3.5. a) Let $\boldsymbol{x} = (x_1, ..., x_k)^T$ and $\boldsymbol{x}_n = (x_{1n}, ..., x_{kn})^T$ be $k \times 1$ random vectors. By Theorem 3.14 c), $\boldsymbol{x}_n \xrightarrow{2} \boldsymbol{x}$ implies that $\boldsymbol{x}_n \xrightarrow{P} \boldsymbol{x}$. Now $\boldsymbol{x}_n \xrightarrow{2} \boldsymbol{x}$ iff $E(\|\boldsymbol{x}_n - \boldsymbol{x}\|^2) \to 0$ as $n \to \infty$. Thus $\boldsymbol{x}_n \xrightarrow{2} \boldsymbol{x}$ iff $E[(\boldsymbol{x}_n - \boldsymbol{x})^T(\boldsymbol{x}_n - \boldsymbol{x})] = \sum_{i=1}^k E[(x_{in} - x_i)^2] \to 0$ as $n \to \infty$.

b) Let $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ be iid with mean $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$ and covariance matrix $\text{Cov}(\boldsymbol{x}_i) = \boldsymbol{\Sigma}$. Assume the $\boldsymbol{x}_i$ have the same distribution as $(x_1, ..., x_k)^T$ with $E(x_i) = \mu_i$ and $V(x_i) = \sigma_i^2$. Then $E[\|\overline{\boldsymbol{x}}_n - \boldsymbol{\mu}\|^2] = E[(\overline{\boldsymbol{x}}_n - \boldsymbol{\mu})^T(\overline{\boldsymbol{x}}_n - \boldsymbol{\mu})] = \sum_{i=1}^k E[(\overline{x}_{in} - \mu_i)^2] = \sum_{i=1}^k V(\overline{x}_{in}) = \sum_{i=1}^k \sigma_i^2/n = \frac{1}{n} tr(\boldsymbol{\Sigma}) \to 0$ as $n \to \infty$. Thus $\overline{\boldsymbol{x}}_n \xrightarrow{2} \boldsymbol{\mu}$ by a), and hence $\overline{\boldsymbol{x}}_n \xrightarrow{P} \boldsymbol{\mu}$. This result proves a special case of the WLLN.

3.3 The Plug-In Principle

Suppose that $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x} = \boldsymbol{x}_{\boldsymbol{\tau}} \sim D(\boldsymbol{\tau})$ where the distribution of $\boldsymbol{x}$ depends on unknown parameters $\boldsymbol{\tau}$. The plug-in principle says approximate the distribution of $\boldsymbol{x}_{\boldsymbol{\tau}}$ by $\boldsymbol{z}_n = \boldsymbol{x}_{\hat{\boldsymbol{\tau}}} \sim D(\hat{\boldsymbol{\tau}})$ where $\hat{\boldsymbol{\tau}}$ is a consistent estimator of $\boldsymbol{\tau}$. Then $\boldsymbol{z}_n$ is often used to make large sample confidence intervals and for large sample tests of hypotheses. For example, let $\boldsymbol{x}_n = \sqrt{n}(T_n - \boldsymbol{\theta})$.

The plug-in principle is also often used to get an asymptotic normal approximation for a statistic, and often the bootstrap confidence regions are closely related to the plug-in principle. For the MCLT, $\boldsymbol{x} \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ and

$\boldsymbol{z}_n \sim N_p(\boldsymbol{0}, \boldsymbol{S}_n)$. For the MLE, $\boldsymbol{x} \sim N(0, [\boldsymbol{I}(\boldsymbol{\theta})]^{-1})$ and $\boldsymbol{z}_n \sim N(0, [\boldsymbol{I}(\hat{\boldsymbol{\theta}}_n)]^{-1})$ where $\hat{\boldsymbol{\theta}}_n$ is the MLE of $\boldsymbol{\theta}$.

It is not clear whether "$\boldsymbol{z}_n \sim D(\hat{\boldsymbol{\tau}})$" converges in distribution to $\boldsymbol{x} \sim D(\boldsymbol{\tau})$. See Section 2.7. Thus the plug-in principle approximation $\boldsymbol{z}_n = \boldsymbol{x}_{\hat{\boldsymbol{\tau}}_n} \sim D(\hat{\boldsymbol{\tau}}_n)$ for $\boldsymbol{x} = \boldsymbol{x}_{\boldsymbol{\tau}} \sim D(\boldsymbol{\tau})$ appears to weaker than convergence in distribution. We may use the notation $\boldsymbol{z}_n \xrightarrow{C} \boldsymbol{x}$ when $\hat{\boldsymbol{\tau}}_n$ is a consistent estimator of $\boldsymbol{\tau}$.

There are some exceptions. For example, interpret "$\boldsymbol{z}_n \sim N_p(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n)$" as $\boldsymbol{z}_n \sim \hat{\boldsymbol{\mu}}_n + \hat{\boldsymbol{\Sigma}}_n^{1/2} N_p(\boldsymbol{0}, \boldsymbol{I}_p) \xrightarrow{D} \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2} N_p(\boldsymbol{0}, \boldsymbol{I}_p) \sim \boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

If $\boldsymbol{x}_n = \sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{x} \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\hat{\boldsymbol{\Sigma}}_n$ is a consistent estimator of $\boldsymbol{\Sigma}$ where $\boldsymbol{\Sigma}$ and the $\hat{\boldsymbol{\Sigma}}_n$ are nonsingular, then it can be shown that $\sqrt{n}(T_n - \boldsymbol{\theta})^T \hat{\boldsymbol{\Sigma}}_n^{-1} \sqrt{n}(T_n - \boldsymbol{\theta})^T \xrightarrow{D} \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} \sim \chi^2_p$. Hence the consistent estimator $\hat{\boldsymbol{\Sigma}}_n$ is useful for constructing large sample confidence regions and large sample tests of hypotheses for $\boldsymbol{\theta}$.

3.4 Summary

1) Let $\boldsymbol{X}_n \in \mathbb{R}^k$ be a sequence of random vectors with joint cdfs $F_{\boldsymbol{X}_n}(\boldsymbol{x})$ and let $\boldsymbol{X} \in \mathbb{R}^k$ be a random vector with joint cdf $F_{\boldsymbol{X}}(\boldsymbol{x})$.

a) $\boldsymbol{X}_n$ **converges in distribution** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$, if $F_{\boldsymbol{X}_n}(\boldsymbol{x}) \to F_{\boldsymbol{X}}(\boldsymbol{x})$ as $n \to \infty$ for all points $\boldsymbol{x}$ at which $F_{\boldsymbol{X}}(\boldsymbol{x})$ is continuous. The distribution of $\boldsymbol{X}$ is the **limiting distribution** or **asymptotic distribution** of $\boldsymbol{X}_n$, and the limiting distribution does not depend on n.

b) $\boldsymbol{X}_n$ **converges in probability** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, if for every $\epsilon > 0$, $P(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \epsilon) \to 0$ as $n \to \infty$.

c) Let $r > 0$ be a real number. Then $\boldsymbol{X}_n$ **converges in rth mean** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{r} \boldsymbol{X}$, if $E(\|\boldsymbol{X}_n - \boldsymbol{X}\|^r) \to 0$ as $n \to \infty$.

d) $\boldsymbol{X}_n$ **converges with probability one** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{wp1} \boldsymbol{X}$, if $P(\lim_{n\to\infty} \boldsymbol{X}_n = \boldsymbol{X}) = 1$.

e) Replace $\boldsymbol{X}$ by $\boldsymbol{c}$ for $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{c}$, $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{c}$, $\boldsymbol{X}_n \xrightarrow{r} \boldsymbol{c}$, or $\boldsymbol{X}_n \xrightarrow{wp1} \boldsymbol{c}$.

f) $\xrightarrow{D} = \xrightarrow{L}$ and $\boldsymbol{X}_n \xrightarrow{wp1} \boldsymbol{X} = \boldsymbol{X}_n \xrightarrow{as} \boldsymbol{X} = \boldsymbol{X}_n \xrightarrow{ae} \boldsymbol{X}$.

2) The **Multivariate Central Limit Theorem (MCLT)**: If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $k \times 1$ random vectors with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$, then

$$\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$$

where the sample mean

$$\overline{\boldsymbol{X}}_n = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i.$$

Note: the usual CLT is a special case with $k = 1$.

3) If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid, $E(\|\boldsymbol{X}\|) < \infty$, and $E(\boldsymbol{X}) = \boldsymbol{\mu}$, then

a) WLLN: $\overline{\boldsymbol{X}}_n \xrightarrow{P} \boldsymbol{\mu}$, and

b) SLLN: $\overline{\boldsymbol{X}}_n \xrightarrow{wp1} \boldsymbol{\mu}$.

4) **Continuity Theorem**: Let $\boldsymbol{X}_n$ be a sequence of $k \times 1$ random vectors with characteristic functions $c_{\boldsymbol{X}_n}(\boldsymbol{t})$, and let $\boldsymbol{X}$ be a $k \times 1$ random vector with cf $c_{\boldsymbol{X}}(\boldsymbol{t})$. Then

$$\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X} \text{ iff } \mathrm{c}_{\boldsymbol{X}_\mathrm{n}}(\boldsymbol{t}) \to \mathrm{c}_{\boldsymbol{X}}(\boldsymbol{t})$$

for all $\boldsymbol{t} \in \mathbb{R}^k$.

5) **Theorem: Cramér Wold Device**: Let $\boldsymbol{X}_n$ be a sequence of $k \times 1$ random vectors, and let $\boldsymbol{X}$ be a $k \times 1$ random vector. Then

$$\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X} \text{ iff } \boldsymbol{t}^\mathrm{T}\boldsymbol{X}_\mathrm{n} \xrightarrow{\mathrm{D}} \boldsymbol{t}^\mathrm{T}\boldsymbol{X}$$

for all $\boldsymbol{t} \in \mathbb{R}^k$.

6) **Theorem.** a) If $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, then $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$.

b)

$$\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{c} \text{ iff } \boldsymbol{X}_\mathrm{n} \xrightarrow{\mathrm{D}} \boldsymbol{c}.$$

7) **Continuous Mapping Theorem.** Let $\boldsymbol{X}, \boldsymbol{X}_n \in \mathbb{R}^k$. If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and if the function $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^j$ is continuous, then $\boldsymbol{g}(\boldsymbol{X}_n) \xrightarrow{D} \boldsymbol{g}(\boldsymbol{X})$.

This theorem also holds if $C(\boldsymbol{g})$ is the set of points $\boldsymbol{x}$ for which $\boldsymbol{g}$ is continuous and $P(\boldsymbol{X} \in C(\boldsymbol{g})) = 1$. (Equivalently, $D(\boldsymbol{g})$ is the set of discontinuity points for $\boldsymbol{g}$ and $P(\boldsymbol{X} \in D(\boldsymbol{g})) = 0$.)

8) **Theorem**: Let $\boldsymbol{X}_n = (X_{1n}, ..., X_{kn})^T$ be a sequence of $k \times 1$ random vectors, let $\boldsymbol{Y}_n$ be a sequence of $k \times 1$ random vectors, and let $\boldsymbol{X} = (X_1, ..., X_k)^T$ be a $k \times 1$ random vector. Let $\boldsymbol{W}_n$ be a sequence of $k \times k$ nonsingular random matrices, and let $\boldsymbol{C}$ be a $k \times k$ constant nonsingular matrix.

a) $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$ iff $X_{in} \xrightarrow{P} X_i$ for $i = 1, ..., k$.

b) **Slutsky's Theorem**: If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and $\boldsymbol{Y}_n \xrightarrow{P} \boldsymbol{a}$ for some constant $k \times 1$ vector $\boldsymbol{a}$, then i) $\boldsymbol{X}_n + \boldsymbol{Y}_n \xrightarrow{D} \boldsymbol{X} + \boldsymbol{a}$ and

ii) $\boldsymbol{Y}_n^T \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{a}^T \boldsymbol{X}$.

c) If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and $\boldsymbol{W}_n \xrightarrow{P} \boldsymbol{C}$, then $\boldsymbol{W}_n \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{C}\boldsymbol{X}$, $\boldsymbol{X}_n^T \boldsymbol{W}_n \xrightarrow{D} \boldsymbol{X}^T \boldsymbol{C}$, $\boldsymbol{W}_n^{-1} \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{C}^{-1} \boldsymbol{X}$, and $\boldsymbol{X}_n^T \boldsymbol{W}_n^{-1} \xrightarrow{D} \boldsymbol{X}^T \boldsymbol{C}^{-1}$.

9) If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$, then $X_{in} \xrightarrow{D} X_i$ for $i = 1, ..., k$.

10) In general, $\boldsymbol{X}_{in} \xrightarrow{D} \boldsymbol{X}_i$ for $i = 1, ..., m$ **does not imply that**

$$\begin{bmatrix} \boldsymbol{X}_{1n} \\ \vdots \\ \boldsymbol{X}_{mn} \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_m \end{bmatrix}.$$

That is, marginal convergence in distribution does not imply joint convergence in distribution.

11) Suppose that $\boldsymbol{X}_n \perp\!\!\!\perp \boldsymbol{Y}_n$ for $n = 1, 2, ...$ Suppose $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$, and $\boldsymbol{Y}_n \xrightarrow{D} \boldsymbol{Y}$. Then

$$\begin{bmatrix} \boldsymbol{X}_n \\ \boldsymbol{Y}_n \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix}$$

where $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$.

If the sequence $\{\boldsymbol{X}_n\} \perp\!\!\!\perp \{\boldsymbol{Y}_n\}$ so that $\boldsymbol{X}_i \perp\!\!\!\perp \boldsymbol{Y}_j$ for every i and j, then we should have $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$ even if $\boldsymbol{X} = \boldsymbol{c} = \boldsymbol{Y}$. Roughly, independence is an exception to 10) since independent random vectors have a joint distribution that does not affect the marginal distributions.

3.5 Complements

Theorems 2.5 and 3.4 appears in Olive (2014). Also see Cox (1984) and McCulloch (1988). A similar result to Theorem 3.4 for linear exponential families where $t_i(\boldsymbol{x}) = x_i$, is given by Brown (1986, p. 172).

The multivariate delta method appears, for example, in Ferguson (1996, p. 45), Lehmann (1999, p. 315), Mardia, Kent and Bibby (1979, p. 52), Sen and Singer (1993, p. 136) and Serfling (1980, p. 122).

Suppose $\boldsymbol{\theta} = \boldsymbol{g}^{-1}(\boldsymbol{\eta})$. In analysis, the fact that

$$\boldsymbol{D}^{-1}_{\boldsymbol{g}(\boldsymbol{\theta})} = \boldsymbol{D}_{\boldsymbol{g}^{-1}(\boldsymbol{\eta})}$$

is a corollary of the inverse mapping theorem (or of the inverse function theorem). See Apostol (1957, p. 146), Bickel and Doksum (2007, p. 517), Marsden and Hoffman (1993, p. 393) and Wade (2000, p. 353).

According to Rohatgi (1984, p. 616), if i) $Y_1, ..., Y_n$ are iid with pdf $f(y)$, ii) $Y_{r_n:n}$ is the r_nth order statistic, iii) $r_n/n \to \rho$, iv) $F(\xi_\rho) = \rho$ and v) $f(\xi_\rho) > 0$, then

$$\sqrt{n}(Y_{r_n:n} - \xi_\rho) \xrightarrow{D} N\left(0, \frac{\rho(1-\rho)}{[f(\xi_\rho)]^2}\right).$$

So there are many asymptotically equivalent ways of defining the sample ρ quantile.

3.6 Problems

3.1. Many multiple linear regression estimators $\hat{\boldsymbol{\beta}}$ satisfy

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(0, V(\hat{\boldsymbol{\beta}}, F)\ \boldsymbol{W}) \tag{3.5}$$

when

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \xrightarrow{P} \boldsymbol{W}^{-1}, \tag{3.6}$$

and when the errors e_i are iid with a cdf F and a unimodal pdf f that is symmetric with a unique maximum at 0. When the variance $V(e_i)$ exists,

$$V(OLS, F) = V(e_i) = \sigma^2 \text{ while } \mathrm{V(L_1, F)} = \frac{1}{4[\mathrm{f(0)}]^2}.$$

In the multiple linear regression model,

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{3.7}$$

for $i = 1, \ldots, n$. In matrix notation, these n equations become

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{3.8}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors.

a) What is the ijth element of the matrix

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}?$$

b) Suppose $x_{k,1} = 1$ and that $x_{k,j} \sim X_j$ are iid with $E(X_j) = 0$ and $V(X_j) = 1$ for $k = 1, ..., n$ and $j = 2, ..., p$. Assume that X_i and X_j are independent for $i \neq j$, $i > 1$ and $j > 1$. (Often $x_{k,j} \sim N(0, 1)$ in simulations.) Then what is $\boldsymbol{W}^{-1}$ for model (3.7)?

c) Suppose $p = 2$ and $Y_i = \alpha + \beta X_i + e_i$. Show

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} \frac{\sum X_i^2}{n\sum(X_i - \overline{X})^2} & \frac{-\sum X_i}{n\sum(X_i - \overline{X})^2} \\ \frac{-\sum X_i}{n\sum(X_i - \overline{X})^2} & \frac{n}{n\sum(X_i - \overline{X})^2} \end{bmatrix}.$$

d) Under the conditions of c), let $S_x^2 = \sum(X_i - \overline{X})^2/n$. Show that

$$n(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \left(\frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\right)^{-1} = \begin{bmatrix} \frac{\frac{1}{n}\sum X_i^2}{S_x^2} & \frac{-\overline{X}}{S_x^2} \\ \frac{-\overline{X}}{S_x^2} & \frac{1}{S_x^2} \end{bmatrix}.$$

e) If the X_i are iid with variance $V(X)$ then $n(\boldsymbol{X}^T\boldsymbol{X})^{-1} \xrightarrow{P} \boldsymbol{W}$. What is $\boldsymbol{W}$?

f) Now suppose that n is divisible by 5 and the $n/5$ of X_i are at 0.1, $n/5$ at 0.3, $n/5$ at 0.5, $n/5$ at 0.7 and $n/5$ at 0.9. (Hence if $n = 100$, 20 of the X_i are at 0.1, 0.3, 0.5, 0.7 and 0.9.)

Find $\sum X_i^2/n$, $\overline{X}$ and S_x^2. (Your answers should not depend on n.)

g) Under the conditions of f), estimate $V(\hat{\alpha})$ and $V(\hat{\beta})$ if L_1 is used and if the e_i are iid $N(0, 0.01)$.

Hint: Estimate $\boldsymbol{W}$ with $n(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and $V(\hat{\boldsymbol{\beta}}, F) = V(L_1, F) = \frac{1}{4[f(0)]^2}$. Hence

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \approx N_2 \left[\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \frac{1}{n} \frac{1}{4[f(0)]^2} \begin{pmatrix} \frac{\frac{1}{n}\sum X_i^2}{S_x^2} & \frac{-\overline{X}}{S_x^2} \\ \frac{-\overline{X}}{S_x^2} & \frac{1}{S_x^2} \end{pmatrix} \right].$$

You should get an answer like $0.0648/n$.

3.2. Suppose $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $p \times 1$ random vectors from a multivariate t-distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ with d degrees of freedom. Then $E(\boldsymbol{X}_i) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{X}) = \dfrac{d}{d-2}\boldsymbol{\Sigma}$ for $d > 2$. Assuming $d > 2$, find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{X}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

3.3. Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ be iid $k \times 1$ random vectors where $E(\boldsymbol{X}_i) = (\lambda_1, ..., \lambda_k)^T$ and $Cov(\boldsymbol{X}_i) = diag(\lambda_1^2, ..., \lambda_k^2)$, a diagonal $k \times k$ matrix with jth diagonal entry λ_j^2. The nondiagonal entries are 0. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{X}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

3.4. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $p \times 1$ random vectors where $E(\boldsymbol{x}_i) = e^{0.5}\mathbf{1}$ and $\mathrm{Cov}(\boldsymbol{x}_i) = (e^2 - e)\boldsymbol{I}_p$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

3.5. Assume that

$$\sqrt{n}\left[\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} - \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \right] \xrightarrow{D} N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \right).$$

Find the limiting distribution of

$$\sqrt{n}[(\hat{\beta}_1 - \hat{\beta}_2) - (\beta_1 - \beta_2)] = (1 \quad -1)\sqrt{n}\left[\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} - \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \right].$$

3.6. Suppose $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid 3×1 random vectors from a multinomial distribution with

$$E(\boldsymbol{X}_i) = \begin{bmatrix} m\rho_1 \\ m\rho_2 \\ m\rho_3 \end{bmatrix} \text{ and } \mathrm{Cov}(\boldsymbol{X}_\mathrm{i}) = \begin{bmatrix} m\rho_1(1-\rho_1) & -m\rho_1\rho_2 & -m\rho_1\rho_3 \\ -m\rho_1\rho_2 & m\rho_2(1-\rho_2) & -m\rho_2\rho_3 \\ -m\rho_1\rho_3 & -m\rho_2\rho_3 & m\rho_3(1-\rho_3) \end{bmatrix}$$

where m is a known positive integer and $0 < \rho_i < 1$ with $\rho_1 + \rho_2 + \rho_3 = 1$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{X}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

3.7. Suppose $\boldsymbol{Y}_n \xrightarrow{P} \boldsymbol{Y}$. Then $\boldsymbol{W}_n = \boldsymbol{Y}_n - \boldsymbol{Y} \xrightarrow{P} \boldsymbol{0}$. Define $\boldsymbol{X}_n = \boldsymbol{Y}$ for all n. Then $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{Y}$. Then $\boldsymbol{Y}_n = \boldsymbol{X}_n + \boldsymbol{W}_n \xrightarrow{D} \boldsymbol{Z}$ by Slutsky's Theorem. What is $\boldsymbol{Z}$?

3.8. If $\boldsymbol{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the characteristic function of $\boldsymbol{X}$ is

$$c_{\boldsymbol{X}}(\boldsymbol{t}) = \exp\left(i\boldsymbol{t}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}\right)$$

for $\boldsymbol{t} \in \mathbb{R}^k$. Let $\boldsymbol{a} \in \mathbb{R}^k$ and find the characteristic function of $\boldsymbol{a}^T\boldsymbol{X} = c_{\boldsymbol{a}^T\boldsymbol{X}}(y) = E[\exp(i\ y\ \boldsymbol{a}^T\boldsymbol{X})] = c_{\boldsymbol{X}}(y\boldsymbol{a})$ for any $y \in \mathbb{R}$. Simplify any constants.

3.9. Suppose

$$\sqrt{n}\left(\begin{pmatrix}\hat{\theta}_1\\ \vdots \\ \hat{\theta}_p\end{pmatrix} - \begin{pmatrix}\theta_1\\ \vdots \\ \theta_p\end{pmatrix}\right) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

Let $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^T$ and let $\boldsymbol{g}(\boldsymbol{\theta}) = (e^{\theta_1}, ..., e^{\theta_p})^T$. Find $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}$.

3.10. Let $\boldsymbol{\mu}_i$ be the ith population mean and let $\boldsymbol{\Sigma}_i$ be the nonsingular population covariance matrix of the ith population. Let $\boldsymbol{x}_{i,1}, ...\boldsymbol{x}_{i,n_i}$ be iid from the ith population. Let $\overline{\boldsymbol{x}}_i$ be the $k \times 1$ sample mean from the $\boldsymbol{x}_{i,j}$, $j = 1, ..., n_i$.

a) Find the limiting distribution of $\sqrt{n_i}(\overline{\boldsymbol{x}}_i - \boldsymbol{\mu}_i)$.

b) Assume there are p populations, $n = \sum_{i=1}^p n_i$, and $n_i/n \xrightarrow{P} \pi_i$ where $0 < \pi_i < 1$ and $1 = \sum_{i=1}^p \pi_i$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}}_i - \boldsymbol{\mu}_i)$. Hint: $\sqrt{n} = (\sqrt{n}/\sqrt{n_i})(\sqrt{n_i})$.

3.11. Suppose $\boldsymbol{Z}_n \xrightarrow{D} N_p(\boldsymbol{\mu}, \boldsymbol{I})$. Let $\boldsymbol{a}$ be a $p \times 1$ constant vector. Find the limiting distribution of $\boldsymbol{a}^T(\boldsymbol{Z}_n - \boldsymbol{\mu})$.

3.12. Suppose $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $k \times 1$ random vectors where $E(\boldsymbol{X}_i) = (\mu_1, ..., \mu_k)^T$ and $Cov(\boldsymbol{X}_i) = diag(\sigma_1^2, ..., \sigma_k^2)$, a diagonal $k \times k$ matrix with jth diagonal entry σ_j^2. The nondiagonal entries are 0. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{X}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

3.13. Suppose that $\boldsymbol{\beta}$ is a $p \times 1$ vector and that $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{C})$ where $\boldsymbol{C}$ is a $p \times p$ nonsingular matrix. Let $\boldsymbol{A}$ be a $j \times p$ matrix with full rank j. Suppose that $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}$.

a) What is the limiting distribution of $\sqrt{n}\boldsymbol{A}\hat{\boldsymbol{\beta}}_n$?

b) What is the limiting distribution of $\boldsymbol{Z}_n = \sqrt{n}[\boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^T]^{-1/2}\boldsymbol{A}\hat{\boldsymbol{\beta}}_n$? Hint: for a square symmetric nonsingular matrix $\boldsymbol{D}$, we have $\boldsymbol{D}^{1/2}\boldsymbol{D}^{1/2} = \boldsymbol{D}$, and $\boldsymbol{D}^{-1/2}\boldsymbol{D}^{-1/2} = \boldsymbol{D}^{-1}$, and $\boldsymbol{D}^{-1/2}$ and $\boldsymbol{D}^{1/2}$ are both symmetric.

c) What is the limiting distribution of $\boldsymbol{Z}_n^T\boldsymbol{Z}_n = n\hat{\boldsymbol{\beta}}_n^T\boldsymbol{A}^T[\boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^T]^{-1}\boldsymbol{A}\hat{\boldsymbol{\beta}}_n$? Hint: If $\boldsymbol{Z}_n \xrightarrow{D} \boldsymbol{Z} \sim N_k(\boldsymbol{0}, \boldsymbol{I})$ then $\boldsymbol{Z}_n^T\boldsymbol{Z}_n \xrightarrow{D} \boldsymbol{Z}^T\boldsymbol{Z} \sim \chi_k^2$.

3.14. Suppose

$$\sqrt{n}\left(\begin{pmatrix}\hat{\sigma}_1^2\\ \vdots \\ \hat{\sigma}_p^2\end{pmatrix} - \begin{pmatrix}\sigma_1^2\\ \vdots \\ \sigma_p^2\end{pmatrix}\right) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma}).$$

Let $\boldsymbol{\theta} = (\sigma_1^2, ..., \sigma_p^2)^T$ and let $\boldsymbol{g}(\boldsymbol{\theta}) = (\sqrt{\sigma_1^2}, ..., \sqrt{\sigma_p^2})^T$. Find $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}$.

3.15. Suppose

$$\sqrt{n}\left(\begin{pmatrix}\hat{\sigma}_1\\ \vdots \\ \hat{\sigma}_p\end{pmatrix} - \begin{pmatrix}\sigma_1\\ \vdots \\ \sigma_p\end{pmatrix}\right) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma}).$$

Let $\boldsymbol{\theta} = (\sigma_1, ..., \sigma_p)^T$ and let $\boldsymbol{g}(\boldsymbol{\theta}) = ((\sigma_1)^2, ..., (\sigma_p)^2)^T$. Find $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}$.

3.16. Let $\boldsymbol{w}_B \sim N_p\left(\mathbf{0}, \frac{\boldsymbol{\Sigma}}{B}\right)$. Then $\boldsymbol{w}_B \xrightarrow{D} \boldsymbol{w}$ as $B \to \infty$. Find $\boldsymbol{w}$.

3.17. Suppose $\boldsymbol{Z}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{I})$. Let $\boldsymbol{A}$ be a constant $r \times k$ matrix. Find the limiting distribution of $\boldsymbol{A}(\boldsymbol{Z}_n - \boldsymbol{\mu})$.

3.18. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $p \times 1$ random vectors where

$$\boldsymbol{x}_i \sim (1-\gamma)N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \gamma N_p(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$$

with $0 < \gamma < 1$ and $c > 0$. Then $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{x}_i) = [1 + \gamma(c-1)]\boldsymbol{\Sigma}$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{d})$ for appropriate vector $\boldsymbol{d}$.

3.19. Let $\boldsymbol{\Sigma}_i$ be the nonsingular population covariance matrix of the ith treatment group or population. To simplify the large sample theory, assume $n_i = \pi_i n$ where $0 < \pi_i < 1$ and $\sum_{i=1}^{3} \pi_i = 1$. Let T_i be a multivariate location estimator such that

$\sqrt{n_i}(T_i - \boldsymbol{\mu}_i) \xrightarrow{D} N_m(\mathbf{0}, \boldsymbol{\Sigma}_i)$, and $\sqrt{n}(T_i - \boldsymbol{\mu}_i) \xrightarrow{D} N_m\left(\mathbf{0}, \frac{\boldsymbol{\Sigma}_i}{\pi_i}\right)$ for $i = 1, 2, 3$.

Assume the T_i are independent.

Then

$$\sqrt{n}\begin{bmatrix}T_1 - \boldsymbol{\mu}_1\\ T_2 - \boldsymbol{\mu}_2\\ T_3 - \boldsymbol{\mu}_3\end{bmatrix} \xrightarrow{D} \boldsymbol{u}.$$

a) Find the distribution of $\boldsymbol{u}$.

b) Suggest an estimator $\hat{\pi}_i$ of π_i.

3.20. Suppose $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $k \times 1$ random vectors where $E(\boldsymbol{X}_i) = (\mu_1, ..., \mu_k)^T$ and $Cov(\boldsymbol{X}_i) = (1-\alpha)\boldsymbol{I} + \alpha\mathbf{1}\mathbf{1}^T$, where $\boldsymbol{I}$ is the $k \times k$ identity matrix, $\mathbf{1} = (1, 1, ..., 1)^T$, and $-(k-1)^{-1} < \alpha < 1$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{X}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

3.21. Show the usual Delta Method is a special case of the Multivariate Delta Method if g is a real function ($d = 1$), T_n is a random variable, θ is a scalar and $\boldsymbol{\Sigma} = \sigma^2$ is a scalar ($k = 1$).

3.22. Let $\boldsymbol{X}$ be a $k \times 1$ random vector and $\boldsymbol{X}_n$ be a sequence of $k \times 1$ random vectors and suppose that

$$\boldsymbol{t}^T \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{t}^T \boldsymbol{X}$$

for all $\boldsymbol{t} \in \mathbb{R}^k$. Does $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$? Explain briefly.

3.23. Suppose the $k \times 1$ random vector $\boldsymbol{X}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Hence the asymptotic distribution of $\boldsymbol{X}_n$ is the multivariate normal MVN $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Find the d, $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ for the following problem. Let $\boldsymbol{C}^T$ be the transpose of $\boldsymbol{C}$.

Let $\boldsymbol{C}$ be an $m \times k$ matrix, then $\boldsymbol{C}\boldsymbol{X}_n \xrightarrow{D} N_d(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$.

3.24. Suppose $\boldsymbol{X}_n$ are $k \times 1$ random vectors with characteristic functions $c_{\boldsymbol{X}_n}(\boldsymbol{t})$. Does $c_{\boldsymbol{X}_n}(\boldsymbol{0}) \to a$ for some constant a? Prove or disprove. Here $\boldsymbol{0}$ is a $k \times 1$ vector of zeroes.

3.25. Suppose

$$\sqrt{n}\left(\begin{pmatrix} \hat{\lambda} \\ \hat{\boldsymbol{\eta}} \end{pmatrix} - \begin{pmatrix} \lambda \\ \boldsymbol{\eta} \end{pmatrix} \right) \xrightarrow{D} N_{p+1}\left(\begin{pmatrix} 0 \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \Sigma_\lambda & \boldsymbol{\Sigma}_{\lambda\boldsymbol{\eta}} \\ \boldsymbol{\Sigma}_{\boldsymbol{\eta}\lambda} & \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \end{pmatrix} \right) \sim N_{p+1}(\boldsymbol{0}, \boldsymbol{\Sigma})$$

where λ is a scalar and $\boldsymbol{\eta} = (\eta_1, ..., \eta_p)$. Let

$$\boldsymbol{g}\begin{pmatrix} \lambda \\ \boldsymbol{\eta} \end{pmatrix} = \lambda\boldsymbol{\eta} =$$

$(\lambda\eta_1, ..., \lambda\eta_p)^T$. Then

$$\sqrt{n}(\hat{\lambda}\hat{\boldsymbol{\eta}} - \lambda\boldsymbol{\eta}) \xrightarrow{D} N_p\left(\boldsymbol{0}, \boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} \boldsymbol{\Sigma} \boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\theta})}\right)$$

by the Multivariate Delta Method.

a) Find $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}$.

b) Let $\boldsymbol{A}$ be a $k \times p$ full rank constant matrix with $k \leq p$ and $\boldsymbol{0} = \boldsymbol{A}\boldsymbol{\eta}$. Find $\boldsymbol{A}\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}$.

Note: then $\sqrt{n}(\boldsymbol{A}\hat{\lambda}\hat{\boldsymbol{\eta}} - \boldsymbol{0}) \xrightarrow{D} N_p\left(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} \boldsymbol{\Sigma} \boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\theta})} \boldsymbol{A}^T\right)$.

3.26. Suppose

$$\sqrt{n}\left(\begin{pmatrix} \hat{\sigma}_1^2 \\ \vdots \\ \hat{\sigma}_p^2 \end{pmatrix} - \begin{pmatrix} \sigma_1^2 \\ \vdots \\ \sigma_p^2 \end{pmatrix} \right) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

Let $\boldsymbol{\theta} = (\sigma_1^2, ..., \sigma_p^2)^T$ and let $\boldsymbol{g}(\boldsymbol{\theta}) = (\log(\sigma_1^2), ..., \log(\sigma_p^2))^T$. Find $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}$.

3.27. Let $W \sim N(\mu_W, \sigma_W^2)$ and let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The characteristic function of W is

$$\varphi_W(y) = E(e^{iyW}) = \exp\left(iy\mu_W - \frac{y^2}{2}\sigma_w^2\right).$$

Suppose $W = \boldsymbol{t}^T\boldsymbol{X}$. Then $W \sim N(\mu_W, \sigma_W^2)$. Find μ_W and σ_W^2. Then the characteristic function of $\boldsymbol{X}$ is

$$\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = E(e^{i\boldsymbol{t}^T\boldsymbol{X}}) = \varphi_W(1).$$

Use these results to find $\varphi_{\boldsymbol{X}}(\boldsymbol{t})$.

3.28. Suppose $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $k \times 1$ random vectors where $E(\boldsymbol{X}_i) = \boldsymbol{1} = (1, ..., 1)^T$ and $Cov(\boldsymbol{X}_i) = \boldsymbol{I}_k = diag(1, ..., 1)$, the $k \times k$ identity matrix. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{X}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

3.29. Suppose $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid with $E(\boldsymbol{X}_i) = \boldsymbol{0}$ but $Cov(\boldsymbol{X}_i)$ does not exist. Does $\overline{\boldsymbol{X}}_n \xrightarrow{P} \boldsymbol{c}$ for some constant vector $\boldsymbol{c}$? Explain briefly.

3.30. Suppose $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and $\boldsymbol{Y}_n - \boldsymbol{X}_n \xrightarrow{P} \boldsymbol{0}$. Does $\boldsymbol{Y}_n \xrightarrow{D} \boldsymbol{W}$ for some random vector $\boldsymbol{W}$? [Hint: $\boldsymbol{Y}_n = \boldsymbol{X}_n + (\boldsymbol{Y}_n - \boldsymbol{X}_n)$.]

3.31. a) If $\boldsymbol{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the characteristic function of $\boldsymbol{X}$ is

$$\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp\left(i\boldsymbol{t}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}\right)$$

for $\boldsymbol{t} \in \mathbb{R}^k$. Let $\boldsymbol{a} \in \mathbb{R}^k$ and find the characteristic function of $\boldsymbol{a}^T\boldsymbol{X} = \varphi_{\boldsymbol{a}^T\boldsymbol{X}}(y) = E[\exp(i\ y\ \boldsymbol{a}^T\boldsymbol{X})] = \varphi_{\boldsymbol{X}}(\boldsymbol{t})$ for any $y \in \mathbb{R}$ and some vector $\boldsymbol{t} \in \mathbb{R}^k$ that depends on y. Simplify any constants.

b Suppose $\boldsymbol{X} = \boldsymbol{c}$ for some constant vector $\boldsymbol{c} \in \mathbb{R}^k$. Prove $\boldsymbol{c} \sim N_k(\boldsymbol{c}, \boldsymbol{0})$ where $\boldsymbol{0}$ is the $k \times k$ matrix of zeroes. Hint: find the characteristic function of $\boldsymbol{X}$ where $P(\boldsymbol{X} = \boldsymbol{c}) = 1$, and compare to the characteristic function given in problem 3).

3.32^Q. Suppose that $\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{y}_n$ for $n = 1, 2, ...$ Suppose $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x}$, and $\boldsymbol{y}_n \xrightarrow{D} \boldsymbol{y}$ where $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$. Prove that

$$\begin{bmatrix} \boldsymbol{x}_n \\ \boldsymbol{y}_n \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}.$$

3.33. Suppose we have random variables (x_1, x_2, Y) with $\sigma_i^2 = V(x_i)$ and $\sigma_{iY} = Cov(x_i, Y)$ for $i = 1, 2$. Let $S_i^2 = \hat{\sigma}_i^2$ and let $\hat{\sigma}_{iY}$ estimate σ_{iY}. Suppose

$$\sqrt{n}\left[\begin{pmatrix} s_1^2 \\ s_2^2 \\ \hat{\sigma}_{1Y} \\ \hat{\sigma}_{2Y} \end{pmatrix} - \begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \\ \sigma_{1Y} \\ \sigma_{2Y} \end{pmatrix}\right] = \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N_4(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

Let $\boldsymbol{g}(\boldsymbol{\theta}) = \left(\frac{\sigma_{1Y}}{\sigma_1^2}, \frac{\sigma_{2Y}}{\sigma_2^2}\right)^T$. Find $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}$.

3.34. Find the limiting distribution of

$$\sqrt{n}\left((\hat{\xi}_{n,0.75} - \hat{\xi}_{n,0.25}) \ - \ (\xi_{0.75} - \xi_{0.25})\right)$$

if the data $Y_1, ..., Y_n$ are iid U(0,1). Then $\xi_\alpha = \alpha$ and $f(\xi_\alpha) = 1$ where $0 < \alpha < 1$.

3.35. Find the limiting distribution of

$$\sqrt{n}\left((\hat{\xi}_{n,0.9} - \hat{\xi}_{n,0.1}) \ - \ (\xi_{0.9} - \xi_{0.1})\right).$$

3.36. Let S_M^2 be the method of moments estimator of the variance σ^2. Suppose

$$\sqrt{n}\left(\begin{pmatrix} \overline{X} \\ S_M^2 \end{pmatrix} - \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}\right) \xrightarrow{D} N_2(\mathbf{0}, \boldsymbol{\Sigma}).$$

Let $\boldsymbol{\theta} = (\mu, \sigma^2)^T$ and let $\boldsymbol{g}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) = \mu/\sigma$. Note that if $\tau = \sigma^2$, then $g(\boldsymbol{\theta}) = \mu/\sqrt{\tau}$. Find $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}$.

3.37. Suppose $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x} \sim D(\boldsymbol{\tau})$, a random vector with a distribution that depends on unknown parameters $\boldsymbol{\tau}$. The plug-in principle says approximate $\boldsymbol{x}$ by $\boldsymbol{z}_n \sim D(\hat{\boldsymbol{\tau}}_n)$ where $\hat{\boldsymbol{\tau}}_n$ is a consistent estimator of $\boldsymbol{\tau}$. Interpret $\boldsymbol{z}_n \sim N_p(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n)$ as $\boldsymbol{z}_n = \hat{\boldsymbol{\mu}}_n + \hat{\boldsymbol{\Sigma}}_n^{1/2} N_p(\mathbf{0}, \boldsymbol{I}_p) \xrightarrow{D} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If $\boldsymbol{x}_n = \sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{x} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$, and if $\hat{\boldsymbol{\tau}}_n = \hat{\boldsymbol{\Sigma}}_n$ and $\boldsymbol{\Sigma}$ are invertible, then $n(T_n - \boldsymbol{\theta})^T \boldsymbol{\Sigma}^{-1}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \chi_p^2$ and $d_n = \boldsymbol{x}_n^T \hat{\boldsymbol{\Sigma}}_n^{-1} \boldsymbol{x}_n = n(T_n - \boldsymbol{\theta})^T \hat{\boldsymbol{\Sigma}}_n^{-1}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \chi_p^2$. To help see $d_n \xrightarrow{D} \chi_p^2$, note that $d_n = n(T_n - \boldsymbol{\theta})^T(\hat{\boldsymbol{\Sigma}}_n^{-1} - \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1})(T_n - \boldsymbol{\theta}) = n(T_n - \boldsymbol{\theta})^T \boldsymbol{\Sigma}^{-1}(T_n - \boldsymbol{\theta}) +$ terms like $a_n = \sqrt{n}(T_n - \boldsymbol{\theta})^T(\hat{\boldsymbol{\Sigma}}_n^{-1} - \boldsymbol{\Sigma}^{-1})\sqrt{n}(T_n - \boldsymbol{\theta})$. Note that $\boldsymbol{x}_n = \sqrt{n}(T_n - \boldsymbol{\theta}) = O_P(1)$ since $\boldsymbol{x}_n \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma})$, and $(\hat{\boldsymbol{\Sigma}}_n^{-1} - \boldsymbol{\Sigma}^{-1}) = o_P(1)$ since $\hat{\boldsymbol{\Sigma}}_n^{-1} \xrightarrow{P} \boldsymbol{\Sigma}^{-1}$. Thus $\hat{\boldsymbol{\Sigma}}_n^{-1} - \boldsymbol{\Sigma}^{-1} \xrightarrow{P} \mathbf{0}$. Does a_n converge in probability to c for some constant c? Explain.

3.38. Let $\boldsymbol{w}_n^* \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_n)$. If $\boldsymbol{\Sigma}_n \xrightarrow{P} \boldsymbol{\Sigma}$, then $\boldsymbol{w}_n^* \xrightarrow{D} \boldsymbol{w}$ for large n by the plug-in principle. Find $\boldsymbol{w}$.

3.39. The interquartile range $\text{IQR}(n) = \hat{\xi}_{n,0.75} - \hat{\xi}_{n,0.25}$ and is a popular estimator of scale. Show that

$$\sqrt{n}\frac{1}{2}(IQR(n) - IQR(Y)) \xrightarrow{D} N(0, \sigma_A^2)$$

where

$$\sigma_A^2 = \frac{1}{64}\left[\frac{3}{[f(\xi_{0.75})]^2} - \frac{2}{f(\xi_{0.75})f(\xi_{0.25})} + \frac{3}{[f(\xi_{0.25})]^2}\right].$$

Hint: $\sigma_A^2 = (-1/2 \ \ 1/2)\boldsymbol{\Sigma}(-1/2 \ \ 1/2)^T$ where $\boldsymbol{\Sigma}$ is obtained from Theorem 3.11 for the 0.25 and 0.75 quantiles.

3.40. Suppose the $p \times 1$ random vector $\boldsymbol{u}_n \xrightarrow{D} \boldsymbol{u}$ and the $p \times p$ random matrix $\boldsymbol{C}_n^{-1} \xrightarrow{P} \boldsymbol{C}^{-1}$. Then $D_n^2 = \boldsymbol{u}_n^T \boldsymbol{C}_n^{-1} \boldsymbol{u}_n \xrightarrow{D} D^2$. Find D^2.

3.41. Suppose $\boldsymbol{Z}_n \xrightarrow{D} \boldsymbol{Z} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$. Then $\boldsymbol{Z}_n^T \boldsymbol{Z}_n \xrightarrow{D} W$. What is W? Simplify if possible.

Chapter 4
Prediction Intervals and Prediction Regions

This chapter considers prediction intervals and prediction regions for iid data. In later chapters, prediction intervals for regression and prediction regions for multivariate regression are derived. Inference after variable selection will consider bootstrap hypothesis testing. Applying certain prediction intervals or prediction regions to the bootstrap sample will result in confidence intervals or confidence regions. See Chapter 5.

4.1 Prediction Intervals

Notation: $P(A_n)$ is "eventually bounded below" by $1-\delta$ if $P(A_n)$ gets arbitrarily close to or higher than $1-\delta$ as $n \to \infty$. Hence $P(A_n) > 1-\delta-\epsilon$ for any $\epsilon > 0$ if n is large enough. If $P(A_n) \to 1-\delta$ as $n \to \infty$, then $P(A_n)$ is eventually bounded below by $1-\delta$. The actual coverage is $1-\gamma_n = P(Y_f \in [L_n, U_n])$, the nominal coverage is $1-\delta$ where $0 < \delta < 1$. The 90% and 95% large sample prediction intervals and prediction regions are common.

Definition 4.1. Consider predicting a future test value Y_f given training data $Y_1, ..., Y_n$. A large sample $100(1-\delta)\%$ *prediction interval* (PI) for Y_f has the form $[\hat{L}_n, \hat{U}_n]$ where $P(\hat{L}_n \leq Y_f \leq \hat{U}_n)$ is eventually bounded below by $1-\delta$ as the sample size $n \to \infty$. A large sample $100(1-\delta)\%$ PI is *asymptotically optimal* if it has the shortest asymptotic length: the length of $[\hat{L}_n, \hat{U}_n]$ converges to $U_s - L_s$ as $n \to \infty$ where $[L_s, U_s]$ is the *population shorth*: the shortest interval covering at least $100(1-\delta)\%$ of the mass.

If Y_f has a pdf, we often want $P(\hat{L}_n \leq Y_f \leq \hat{U}_n) \to 1-\delta$ as $n \to \infty$. The interpretation of a 100 $(1-\delta)\%$ PI for a random variable Y_f is similar to that of a confidence interval (CI). Collect data, then form the PI, and repeat for a total of k times where the k trials are independent from the same population. If Y_{fi} is the ith random variable and PI_i is the ith PI, then the probability

that $Y_{fi} \in PI_i$ for j of the PIs approximately follows a binomial$(k, \rho = 1-\delta)$ distribution. Hence if 100 95% PIs are made, $\rho = 0.95$ and $Y_{fi} \in PI_i$ happens about 95 times.

There are two big differences between CIs and PIs. First, the length of the CI goes to 0 as the sample size n goes to ∞ while the length of the PI converges to some nonzero number J, say. Secondly, many confidence intervals work well for large classes of distributions while many prediction intervals assume that the distribution of the data is known up to some unknown parameters. Usually the $N(\mu, \sigma^2)$ distribution is assumed, and the parametric PI may not perform well if the normality assumption is violated. This section will give some PIs that work well for large classes of distributions.

Consider the location model, $Y_i = \mu + e_i$, where $Y_1, ..., Y_n, Y_f$ are iid with the same distribution as Y. Let $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ be the order statistics of the iid training data $Y_1, ..., Y_n$. Then the unknown future value Y_f is the test data. Suppose the sample percentiles $[\hat{L}_n, \hat{U}_n]$ of the training data $Y_1, ..., Y_n$ are consistent estimators of the population percentiles $[L, U]$ of the distribution where $P(Y \in [L, U]) = 1 - \delta$. Then $P(Y_f \in [\hat{L}_n, \hat{U}_n] \rightarrow P(Y_f \in [L, U]) = 1 - \delta$ as $n \rightarrow \infty$. Three common choices are a) $P(Y \leq U) = 1 - \delta/2$ and $P(Y \leq L) = \delta/2$, b) $P(Y^2 \leq U^2) = P(|Y| \leq U) = P(-U \leq Y \leq U) = 1 - \delta$ with $L = -U$, and c) the population shorth is the shortest interval (with length $U - L$) such that $P(Y \in [L, U]) = 1 - \delta$. The PI c) is asymptotically optimal while a) and b) are asymptotically optimal on the class of symmetric zero mean unimodal error distributions.

If the cdf F_Y of Y has jumps, then it may not be possible to find L and U such that $P(Y \in [L, U]) = 1 - \delta$, but it is possible to find L and U such that $P(Y \in [L, U]) \geq 1 - \delta$ for $0 < \delta < 1$. For example, if $P(Y = c) = 1$, then $P(Y \in [c, c]) = 1 \geq 1 - \delta$ for $0 < \delta < 1$. For $Y_1, ..., Y_n$ iid BIN$(n = 1, \rho)$, useful PIs are [0,0], [0,1], and [1,1]. Using open intervals would give 0% coverage.

Let $0 < \alpha < 1$, and let Y_α be a number such that $P(Y \leq Y_\alpha) = \alpha$ if Y_α is a continuity point of the cdf $F_Y(y)$. Let $F(y-) = P(Y < y)$. If Y_α is not a continuity point of $F_Y(y)$, let $F(Y_\alpha -) = \alpha_1 \leq \alpha \leq \alpha_2 = F(Y_\alpha)$ where $0 \leq \alpha_1 < \alpha_2 \leq 1$. Suppose $\alpha_1 < \alpha < \alpha_2$. For example, let $\alpha_1 = 0.89 < \alpha = 0.9 < \alpha_2 = 0.92$. Let $\lceil x \rceil$ be the smallest integer $\geq x$. For example, $\lceil 7.7 \rceil = 8$. Then $\sum_{i=1}^{n} I(Y_i \leq Y_{(\lceil n\alpha \rceil)}) \geq \lceil n\alpha \rceil$ with equality unless there are ties: at least two $Y_i = Y_{(\lceil n\alpha \rceil)}$. Thus if $Y_{(\lceil n\alpha \rceil)} < Y_\alpha$, not enough $Y_i \leq Y_{(\lceil n\alpha \rceil)}$, while if $Y_{(\lceil n\alpha \rceil)} > Y_\alpha$, too many $Y_i \leq Y_{(\lceil n\alpha \rceil)}$. Hence $P(Y_{(\lceil n\alpha \rceil)} = Y_\alpha) \rightarrow 1$, $P(Y_f < Y_{(\lceil n\alpha \rceil)}) \rightarrow \alpha_1 < \alpha$, and $P(Y_f \leq Y_{(\lceil n\alpha \rceil)}) \rightarrow \alpha_2 > \alpha$ as $n \rightarrow \infty$. Similarly, if $\alpha_2 = \alpha$, then $P(Y_{(\lceil n\alpha \rceil)} \geq Y_\alpha) \rightarrow 1$ as $n \rightarrow \infty$. If $\alpha_1 = \alpha$ and $F_Y(y)$ is strictly increasing on the interval $(Y_\alpha - \epsilon, Y_\alpha]$ for some $\epsilon > 0$, then $P(Y \leq Y_{(\lceil n\alpha \rceil)})$ gets arbitrarily close to or higher than α as $n \rightarrow \infty$. If Y_m is the smallest value of y such that $P(Y \leq Y_m) = \alpha$, $\alpha_1 = \alpha$, and $Y_m < Y_\alpha$, then $P(Y_{(\lceil n\alpha \rceil)} \geq Y_m) \rightarrow 1$ as $n \rightarrow \infty$. Hence $P(Y \leq Y_{(\lceil n\alpha \rceil)})$ gets arbitrarily close to or higher than α in all cases. Hence closed intervals have coverage eventually bounded below by $1 - \delta$.

Remark 4.1. Confidence intervals, prediction intervals, confidence regions, and prediction regions should used closed sets not open sets. The closed sets have the same volume as as the open sets, but have coverage at least as high as the open sets with weaker regularity conditions. In particular, confidence and prediction intervals should be closed intervals, not open intervals.

In the following theorem, if the open interval $(Y_{(k_1)}, Y_{(k_2)})$ was used, we would need to add the regularity condition that $Y_{\delta/2}$ and $Y_{1-\delta/2}$ are continuity points of $F_Y(y)$.

Theorem 4.1. Let $Y_1, ..., Y_n, Y_f$ be iid. Let $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ be the order statistics of the training data. Let $k_1 = \lceil n\delta/2 \rceil$ and $k_2 = \lceil n(1-\delta/2) \rceil$ where $0 < \delta < 1$. The large sample $100(1-\delta)\%$ *percentile prediction interval* for Y_f is

$$[Y_{(k_1)}, Y_{(k_2)}]. \tag{4.1}$$

The shorth(c) estimator of the population shorth is useful for making asymptotically optimal prediction intervals. For the uniform distribution, the population shorth is not unique. Of course the length of the population shorth is unique. For a large sample $100(1-\delta)\%$ PI, the nominal coverage is $100(1-\delta)\%$. Undercoverage occurs if the actual coverage is below the nominal coverage. For example, if the actual coverage is 0.93 for a large sample 95% PI, than the undercoverage is 0.02.

Definition 4.2. Let the shortest closed interval containing at least c of the $Y_1, ..., Y_n$ be

$$\text{shorth(c)} = [\text{Y}_{(\text{s})}, \text{Y}_{(\text{s+c}-1)}]. \tag{4.2}$$

Theorem 4.2, Frey (2013). Let $Y_1, ..., Y_n$ be iid. Let

$$k_n = \lceil n(1-\delta) \rceil. \tag{4.3}$$

For large $n\delta$ and iid data, the large sample $100(1-\delta)\%$ shorth(k_n) prediction interval has maximum undercoverage $\approx 1.12\sqrt{\delta/n}$. The maximum undercoverage occurs for the family of uniform $U(\theta_1, \theta_2)$ distributions.

Theorem 4.3, Frey (2013). Let $Y_1, ..., Y_n, Y_f$ be iid. Let $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ be the order statistics of the training data. The large sample $100(1-\delta)\%$ shorth(c) prediction interval for Y_f is

$$[Y_{(s)}, Y_{(s+c-1)}] \text{ where } \text{c} = \min(\text{n}, \lceil \text{n}[1 - \delta + 1.12\sqrt{\delta/\text{n}}\ ]\ \rceil). \tag{4.4}$$

A problem with the prediction intervals that cover $\approx 100(1-\delta)\%$ of the training data cases Y_i (such as (4.2) using $c = k_n$ given by (4.3)), is that they have coverage lower than the nominal coverage of $1-\delta$ for moderate n. This result is not surprising since empirically statistical methods perform worse on test data. For iid data, Frey (2013) used (4.4) to correct for undercoverage.

Theorem 4.4. Let $Y_1, ..., Y_n, Y_f$ be iid. Let $W_{(1)} \leq W_{(2)} \leq \cdots \leq W_{(n)}$ be the order statistics of the squared training data $W_1, ..., W_n$ where $W_i = Y_i^2$ for $i = 1, ..., n$. Let k_n be given by Equation (4.3). Let $L_n = -U_n$ and $U_n = \sqrt{W_{(k_n)}}$. Then $[L_n, U_n]$ is a large sample $100(1-\delta)\%$ PI for Y_f.

Note that $P(0 \leq W_f \leq U_n^2)$ is eventually bounded below by $1-\delta$ as $n \to \infty$.

By Chebyshev's inequality, for $k > 1$,

$$P(\mu - k\sigma \leq Y \leq \mu + k\sigma) \geq P(\mu - k\sigma < Y < \mu + k\sigma) \geq 1 - \frac{1}{k^2}. \quad (4.5)$$

Note that $k = 5$ gives 96% asymptotic coverage. The value $k = 1.96$ gives 95% coverage for the $N(\mu, \sigma^2)$ distribution, but the coverage could be as low as 74%. Use $\hat{\mu} = \overline{Y}$ and $\hat{\sigma} = S$, the square root of the unbiased sample variance estimator.

Theorem 4.5. Let $Y_1, ..., Y_n, Y_f$ be iid. Suppose that $E(Y) = \mu$ and the standard deviation $SD(Y) = \sigma$. Let $\hat{\mu}$ and $\hat{\sigma}$ be consistent estimators of μ and σ. Let $1 - 1/k^2 \geq 1 - \delta$. Let $\mu \pm k\sigma$ be continuity points of $F_Y(y)$. Then

$$[L_n, U_n] = [\hat{\mu} - k\hat{\sigma}, \hat{\mu} + k\hat{\sigma}]$$

is a large sample $100(1-\delta)\%$ *Chebyshev PI* for Y_f.

Remark 4.2. a) The Chebyshev PIs tend to be too long, and need second moments. b) The shorth PI (4.4) often has good coverage for $n \geq 50$ and $0.05 \leq \delta \leq 0.1$, but the convergence of $U_n - L_n$ to the population shorth length $U_s - L_s$ can be quite slow. Under regularity conditions, Grübel (1982) showed that for iid data, the length and center the shorth(k_n) interval are $\sqrt{n}$ consistent and $n^{1/3}$ consistent estimators of the length and center of the population shorth interval, respectively. The correction factor also increases the length. For a unimodal and symmetric error distribution, the percentile PI (4.1), shorth PI (4.4), and Theorem 4.4 PI are asymptotically equivalent, but PI (4.1) can be the shortest PI. c) The percentile PI (4.1) and Theorem 4.4 PI can be much longer than the shorth PI (4.4) if the data distribution is skewed. The Theorem 4.4 PI can very long if Y is a nonnegative random variable.

Example 4.1. Given below were votes for preseason 1A basketball poll from Nov. 22, 2011 WSIL News where the 778 was a typo: the actual value was 78. As shown below, finding shorth(3) from the ordered data is simple. If the outlier was corrected, shorth(3) = [76,78].

```
111   89   778   78   76

order data: 76 78 89 111 778
```

```
13 = 89 - 76

33 = 111 - 78

689 = 778 - 89
shorth(3) = [76,89]
```

Remark 4.3. The large sample $100(1-\delta)\%$ shorth PI (4.4) may or may not be asymptotically optimal if the $100(1-\delta)\%$ population shorth is $[L_s, U_s]$ and $F_Y(y)$ is not strictly increasing in intervals $(L_s - \epsilon, L_s + \epsilon)$ and $(U_s - \epsilon, U_s + \epsilon)$ for some $\epsilon > 0$. To see the issue, suppose Y has probability mass function (pmf) $f(0) = 0.4$, $f(1) = 0.3$, $f(2) = 0.2$, $f(3) = 0.06$, and $f(4) = 0.04$. Then the 90% population shorth is [0,2] and the $100(1-\delta)\%$ population shorth is [0,3] for $(1-\delta) \in (0.9, 0.96]$. Let $W_i = I(Y_i \leq y) = 1$ if $Y_i \leq y$ and 0, otherwise. The empirical cdf

$$\hat{F}_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_{(i)} \leq y)$$

is the sample proportion of $Y_i \leq y$. If $Y_1, ..., Y_n$ are iid, then for fixed y, $n\hat{F}_n(y) \sim binomial(n, F(y))$. Thus $\hat{F}_n(y) \sim AN(F(y), F(y)(1-F(y))/n)$. For the Y with the above pmf, $\hat{F}_n(2) \xrightarrow{P} 0.9$ as $n \to \infty$ with $P(\hat{F}_n(2) < 0.9) \to 0.5$ and $P(\hat{F}_n(2) \geq 0.9) \to 0.5$ as $n \to \infty$. Hence the large sample 90% PI (4.4) will be [0,2] or [0,3] with probabilities $\to 0.5$ as $n \to \infty$ with expected asymptotic length of 2.5 and expected asymptotic coverage converging to 0.93. However, the large sample $100(1-\delta)\%$ PI (4.4) converges to [0,3] and is asymptotically optimal with asymptotic coverage 0.96 for $(1-\delta) \in (0.9, 0.96)$.

For a random variable Y, the $100(1-\delta)\%$ highest density region is a union of $k \geq 1$ disjoint intervals such that the mass within the intervals $\geq 1-\delta$ and the sum of the k interval lengths is as small as possible. Suppose that $f(z)$ is a unimodal pdf that has interval support, and that the pdf $f(z)$ of Y decreases rapidly as z moves away from the mode. Let $[a, b]$ be the shortest interval such that $F_Y(b) - F_Y(a) = 1 - \delta$ where the cdf $F_Y(z) = P(Y \leq z)$. Then the interval $[a, b]$ is the $100(1-\delta)\%$ highest density region. To find the $100(1-\delta)\%$ highest density region of a pdf, move a horizontal line down from the top of the pdf. The line will intersect the pdf or the boundaries of the support of the pdf at $[a_1, b_1], ..., [a_k, b_k]$ for some $k \geq 1$. Stop moving the line when the areas under the pdf corresponding to the intervals is equal to $1-\delta$. As an example, let $f(z) = e^{-z}$ for $z > 0$. See Figure 4.1 where the area under the pdf from 0 to 1 is 0.368. Hence [0,1] is the 36.8% highest density region. The shorth PI estimates the highest density interval which is the highest density region for a distribution with a unimodal pdf. Often the highest density region is an interval $[a, b]$ where $f(a) = f(b)$, especially if the support where $f(z) > 0$ is $(-\infty, \infty)$.

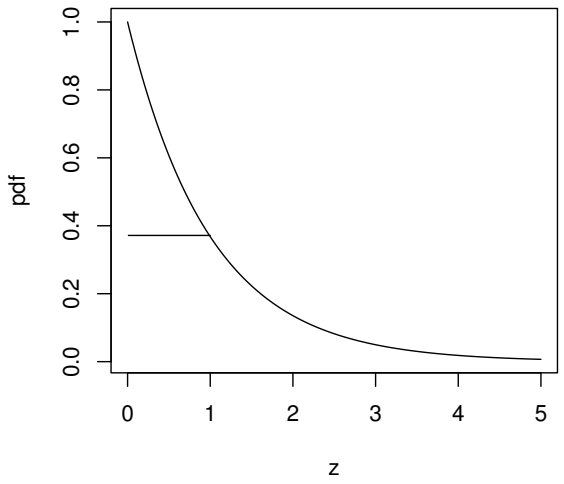

Fig. 4.1 The 36.8% Highest Density Region is [0,1]

Remark 4.4. Note that correction factors $b_n \to 1$ are used in large sample confidence intervals and tests if the limiting distribution is N(0,1) or χ^2_p, but a t_{d_n} or pF_{p,d_n} cutoff is used: $t_{d_n,1-\delta}/z_{1-\delta} \to 1$ and $pF_{p,d_n,1-\delta}/\chi^2_{p,1-\delta} \to 1$ if $d_n \to \infty$ as $n \to \infty$. See Example 2.16 and Theorem 2.34. Using correction factors for large sample confidence intervals, tests, prediction intervals, prediction regions, and confidence regions improves the performance for moderate sample size n.

4.2 Prediction Regions

Consider predicting a $p \times 1$ future test value $\boldsymbol{x}_f$, given past training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ where $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid. Much as confidence regions and intervals give a measure of precision for the point estimator $\hat{\boldsymbol{\theta}}$ of the parameter $\boldsymbol{\theta}$, prediction regions and intervals give a measure of precision of the point estimator $T = \hat{\boldsymbol{x}}_f$ of the future random vector $\boldsymbol{x}_f$.

Definition 4.3. A *large sample* $100(1-\delta)\%$ *prediction region* is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$. A prediction region is *asymptotically optimal* if its volume converges in

probability to the volume of the minimum volume covering region or the highest density region of the distribution of $\boldsymbol{x}_f$.

If $\boldsymbol{x}_f$ has a pdf, we often want $P(\boldsymbol{x}_f \in \mathcal{A}_n) \rightarrow 1 - \delta$ as $n \rightarrow \infty$. A PI is a prediction region where $p = 1$. Highest density regions are usually hard to estimate for p not much larger than four, but many elliptically contoured distributions with a nonsingular population covariance matrix, including the multivariate normal distribution, have highest density regions that can be estimated by the nonparametric prediction region (4.11). For more about highest density regions, see Olive (2017b, pp. 148-155) and Hyndman (1996).

For multivariate data, sample Mahalanobis distances play a role similar to that of residuals in multiple linear regression. Let the observed training data be collected in an $n \times p$ matrix $\boldsymbol{W}$. Let the $p \times 1$ column vector $T = T(\boldsymbol{W})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C} = \boldsymbol{C}(\boldsymbol{W})$ be a dispersion estimator.

Definition 4.4. Let $x_{1j}, ..., x_{nj}$ be measurements on the jth random variable X_j corresponding to the jth column of the data matrix $\boldsymbol{W}$. The jth *sample mean* is $\overline{x}_j = \frac{1}{n}\sum_{k=1}^{n} x_{kj}$. The *sample covariance* S_{ij} estimates $\mathrm{Cov}(X_i, X_j) = \sigma_{ij} = E[(X_i - E(X_i))(X_j - E(X_j))]$, and

$$S_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j).$$

$S_{ii} = S_i^2$ is the *sample variance* that estimates the population variance $\sigma_{ii} = \sigma_i^2$. The *sample correlation* r_{ij} estimates the population correlation $\mathrm{Cor}(X_i, X_j) = \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$, and

$$r_{ij} = \frac{S_{ij}}{S_i S_j} = \frac{S_{ij}}{\sqrt{S_{ii} S_{jj}}} = \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{kj} - \overline{x}_j)^2}}.$$

Definition 4.5. Let $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ be the data where $\boldsymbol{x}_i$ is a $p \times 1$ vector. The **sample mean** or *sample mean vector*

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = (\overline{x}_1, ..., \overline{x}_p)^T = \frac{1}{n}\boldsymbol{W}^T \mathbf{1}$$

where $\mathbf{1}$ is the $n \times 1$ vector of ones. The **sample covariance matrix**

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T = (S_{ij}).$$

That is, the ij entry of $\boldsymbol{S}$ is the sample covariance S_{ij}. The *classical estimator of multivariate location and dispersion* is $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. The **sample correlation matrix**

$$\boldsymbol{R} = (r_{ij}).$$

That is, the ij entry of $\boldsymbol{R}$ is the sample correlation r_{ij}.

It can be shown that $(n-1)\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T - \overline{\boldsymbol{x}}\, \overline{\boldsymbol{x}}^T =$

$$\boldsymbol{W}^T \boldsymbol{W} - \frac{1}{n} \boldsymbol{W}^T \boldsymbol{1}\boldsymbol{1}^T \boldsymbol{W}.$$

Hence if the *centering matrix* $\boldsymbol{G} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T$, then $(n-1)\boldsymbol{S} = \boldsymbol{W}^T \boldsymbol{G} \boldsymbol{W}$.

Definition 4.6. The ith *Mahalanobis distance* $D_i = \sqrt{D_i^2}$ where the ith *squared Mahalanobis distance* is

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W})) \quad (4.6)$$

for each point $\boldsymbol{x}_i$. Notice that D_i^2 is a random variable (scalar valued). Let $(T, \boldsymbol{C}) = (T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}))$. Then

$$D_{\boldsymbol{x}}^2(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T).$$

Hence D_i^2 uses $\boldsymbol{x} = \boldsymbol{x}_i$.

See Definition 1.29 for the population mean and population covariance matrix. The Mahalanobis distance in Definition 4.6 is a random variable that estimates the population Mahalanobis distance of Definition 1.49. Let the $p \times 1$ location vector be $\boldsymbol{\mu}$, often the population mean, and let the $p \times p$ dispersion matrix be $\boldsymbol{\Sigma}$, often the population covariance matrix. Notice that if $\boldsymbol{x}$ is a random vector, then the population squared Mahalanobis distance from Definition 1.49 is

$$D_{\boldsymbol{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \quad (4.7)$$

and that the term $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})$ is the $p-$dimensional analog to the z-score used to transform a univariate $N(\mu, \sigma^2)$ random variable into a $N(0,1)$ random variable. Hence the sample Mahalanobis distance $D_i = \sqrt{D_i^2}$ is an analog of the absolute value $|Z_i|$ of the sample Z-score $Z_i = (X_i - \overline{X})/\hat{\sigma}$. Also notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center $T(\boldsymbol{W})$ is $D_i(T(\boldsymbol{W}), \boldsymbol{I}_p)$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix.

Theorem 4.6. i) Suppose $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}$ be a $q \times p$ constant matrix. Then $\boldsymbol{A}\sqrt{n}(T_n - \boldsymbol{\mu}) = \sqrt{n}(\boldsymbol{A}T_n - \boldsymbol{A}\boldsymbol{\mu}) \xrightarrow{D} N_q(\boldsymbol{A}\boldsymbol{\theta}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

ii) Let $\boldsymbol{\Sigma} > 0$. If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ where $s > 0$ is some constant, then $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + o_P(1)$, so $D^2_{\boldsymbol{x}}(T, \boldsymbol{C})$ is a consistent estimator of $s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

iii) Let $\boldsymbol{\Sigma} > 0$. If $\sqrt{n}(T - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ and if $\boldsymbol{C}$ is a consistent estimator of $\boldsymbol{\Sigma}$, then $n(T - \boldsymbol{\mu})^T \boldsymbol{C}^{-1} (T - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$. In particular, $n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$.

Proof: i) $\boldsymbol{A}\boldsymbol{W}_n \xrightarrow{D} \boldsymbol{A}\boldsymbol{W}$ by Theorem 3.13 iii), and the result follows.

ii) $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) =$
$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T [\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1} + s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)$
$= (\boldsymbol{x} - \boldsymbol{\mu})^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu}) + (\boldsymbol{x} - T)^T [\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - T)$
$+(\boldsymbol{x} - \boldsymbol{\mu})^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{\mu} - T) + (\boldsymbol{\mu} - T)^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu})$
$+(\boldsymbol{\mu} - T)^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{\mu} - T) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(1)$.

(Note that $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(n^{-\delta})$ if $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ with rate n^{δ} where $0 < \delta \leq 0.5$ if $[\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1}] = O_P(n^{-\delta})$.)

Alternatively, $D^2_{\boldsymbol{x}}(T, \boldsymbol{C})$ is a continuous function of $(T, \boldsymbol{C})$ if $\boldsymbol{C} > 0$ for $n > 10p$. Hence $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) \xrightarrow{P} D^2_{\boldsymbol{x}}(\mu, s\boldsymbol{\Sigma})$.

iii) Note that $\boldsymbol{Z}_n = \sqrt{n}\ \boldsymbol{\Sigma}^{-1/2}(T - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{I}_p)$. Thus $\boldsymbol{Z}_n^T \boldsymbol{Z}_n = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (T - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$. Now $n(T - \boldsymbol{\mu})^T \boldsymbol{C}^{-1} (T - \boldsymbol{\mu}) =$
$n(T - \boldsymbol{\mu})^T [\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}](T - \boldsymbol{\mu}) = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (T - \boldsymbol{\mu}) +$
$n(T - \boldsymbol{\mu})^T [\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1}](T - \boldsymbol{\mu}) = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (T - \boldsymbol{\mu}) + o_P(1) \xrightarrow{D} \chi^2_p$ since
$\sqrt{n}(T - \boldsymbol{\mu})^T [\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1}] \sqrt{n}(T - \boldsymbol{\mu}) = O_P(1) o_P(1) O_P(1) = o_P(1)$. □

Next, we derive a prediction region for $\boldsymbol{x}_f$ if $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$, $\boldsymbol{\mu} = E(\boldsymbol{x})$, and $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \mathrm{Cov}(\boldsymbol{x})$ is nonsingular. Let $D = D(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. Then $D_i \xrightarrow{D} D$ and $D_i^2 \xrightarrow{D} D^2$ by Theorem 4.6. Hence the sample percentiles of the D_i are consistent estimators of the population percentiles of D at continuity points of the cdf of D, and the sample percentiles of the D_i^2 are consistent estimators of the population percentiles of D^2 at continuity points of the cdf of D^2. Let $c = k_n = \lceil n(1 - \delta) \rceil$. Then Olive (2013b) showed that the hyperellipsoid

$$\mathcal{A}_n = \{\boldsymbol{x} : D^2_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D^2_{(c)}\} = \{\boldsymbol{x} : D_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D_{(c)}\} \tag{4.8}$$

is a large sample $100(1 - \delta)\%$ prediction region under mild conditions, although regions with smaller volumes may exist.

To improve performance, we will use a correction factor $c = U_n$ where U_n decreases to k_n. U_n is defined under Equation (4.10). A problem with the prediction regions that cover $\approx 100(1 - \delta)\%$ of the training data cases $\boldsymbol{x}_i$ (such as (4.8) for $c = k_n$), is that they have coverage lower than the nominal coverage of $1 - \delta$ for moderate n. This result is not surprising since empirically statistical methods perform worse on test data than on training data. Also see Remark 4.4. Empirically for many distributions, for $n = 20p$, the prediction region (4.8) applied to iid data using $c = k_n = \lceil n(1 - \delta) \rceil$ tended to have

undercoverage as high as $\min(0.05, \delta/2)$. The undercoverage decreases rapidly as n increases. (Referring to the next paragraph, taking $q_n \equiv 1 - \delta$ does not take into account the unknown variability of $(\overline{\boldsymbol{x}}, \boldsymbol{S})$, which is another reason for undercoverage and the need for a correction factor.)

Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + p/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta p/n), \quad \text{otherwise.} \tag{4.9}$$

If $1 - \delta < 0.999$ and $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Using

$$c = \lceil n q_n \rceil \tag{4.10}$$

in (4.8) decreased the undercoverage. Let $D_{(U_n)}$ be the $100q_n$th sample quantile of the D_i.

The nonparametric prediction region is due to Olive (2013b). For the classical prediction region, see Chew (1966) and Johnson and Wichern (1988, pp. 134, 151). A future observation (random vector) $\boldsymbol{x}_f$ is in the region (4.11) if $D_{\boldsymbol{x}_f} \leq D_{(U_n)}^2$. If $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ and $\boldsymbol{x}_f$ are iid, the nonparametric prediction region (4.11) is asymptotically optimal for a large class of elliptically contoured distributions since the volume of (4.11) converges in probability to the volume of the highest density region. (These distributions have a highest density region which is a hyperellipsoid determined by a population Mahalanobis distance. See Section 1.7.) Refer to the above paragraph for $D_{(U_n)}$. Let $P(D^2 \leq D_{1-\delta}^2) = 1 - \delta$ if $D_{1-\delta}^2$ is a continuity point of the cdf $F_{D^2}(y)$ and $D_{\boldsymbol{x}}^2(\overline{\boldsymbol{x}}, \boldsymbol{S}) \xrightarrow{D} D^2 = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$.

Theorem 4.7. Assume that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid from a distribution with mean $E(\boldsymbol{x}) = \boldsymbol{\mu}$ and nonsingular covariance matrix $\mathrm{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$. The large sample $100(1 - \delta)\%$ *nonparametric prediction region* for a future value $\boldsymbol{x}_f$ is

$$\{\boldsymbol{z} : D_{\boldsymbol{z}}^2(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D_{(U_n)}^2\} \tag{4.11}$$

if $D_{1-\delta}^2$ is a continuity point of the cdf $F_{D^2}(y)$.

Theorem 4.8. Assume that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. Then the large sample $100(1 - \delta)\%$ *classical prediction region* is

$$\{\boldsymbol{z} : D_{\boldsymbol{z}}^2(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq \chi_{p,1-\delta}^2\}. \tag{4.12}$$

If p is small, Mahalanobis distances tend to be right skewed with a population shorth that discards the right tail. For $p = 1$ and $n \geq 20$, the finite sample correction factors c/n for c given by (4.4) and (4.10) do not differ by much more than 3% for $0.01 \leq \delta \leq 0.5$. See Figure 4.2 where ol = (Eq. 4.10)$/n$ is plotted versus fr = (Eq. 4.4)$/n$ for $n = 20, 21, ..., 500$. The top plot is for $\delta = 0.01$, while the bottom plot is for $\delta = 0.3$. The identity line is added

to each plot as a visual aid. The value of n increases from 20 to 500 from the right of the plot to the left of the plot. Examining the axes of each plot shows that the correction factors do not differ greatly. R code to create Figure 4.2 is shown below.

```
cmar <- par("mar"); par(mfrow = c(2, 1))
par(mar=c(4.0,4.0,2.0,0.5))
frey(0.01); frey(0.3)
par(mfrow = c(1, 1)); par(mar=cmar)
```

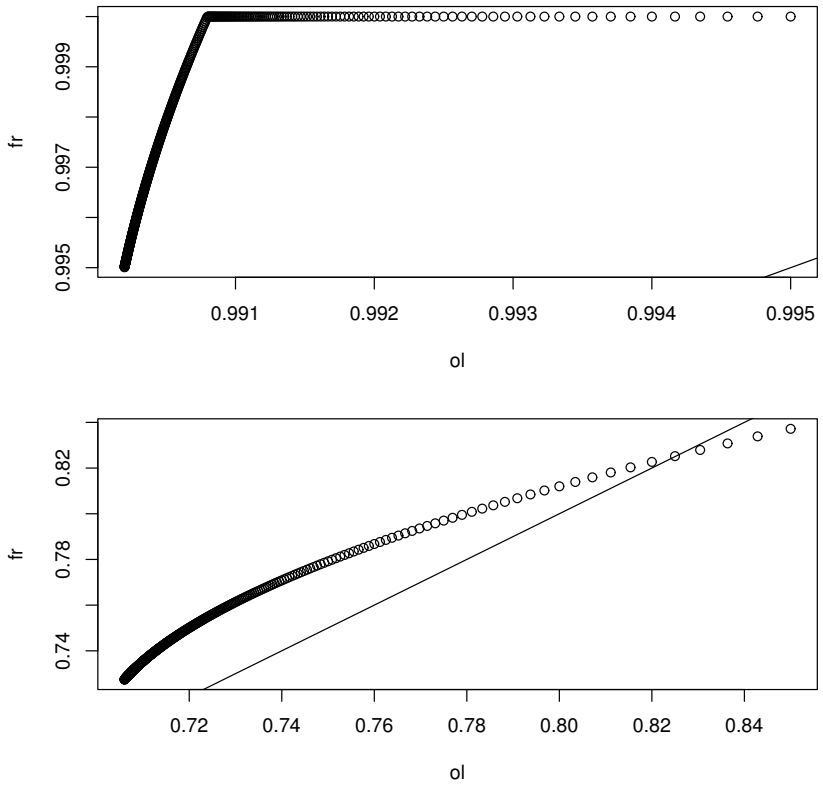

Fig. 4.2 Correction Factor Comparison when $\delta = 0.01$ (Top Plot) and $\delta = 0.3$ (Bottom Plot)

Remark 4.5. The nonparametric prediction region (4.11) is useful if $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid from a distribution with a nonsingular covariance matrix, and the sample size n is large enough. The distribution could be continuous, discrete, or a mixture. The asymptotic coverage is $1 - \delta$ if D has a pdf, although prediction regions with smaller volume may exist. The nonparametric

prediction region (4.11) contains U_n of the training data cases $\boldsymbol{x}_i$ provided that $\boldsymbol{S}$ is nonsingular, even if the model is wrong. For many distributions, the coverage started to be close to $1-\delta$ for $n \geq 10p$ where the coverage is the simulated percentage of times that the prediction region contained $\boldsymbol{x}_f$.

Theorem 4.9, Chen (2011). Multivariate Chebyshev's Inequality: Let $E(\boldsymbol{x}) = \boldsymbol{\mu}$, and let $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \mathrm{Cov}(\boldsymbol{x})$ be nonsingular. Then

$$P(D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \leq \gamma) \geq 1 - p/\gamma > 0$$

for $\gamma > p$.

For more on the above theorem, see Budny (2014) and Navarro (2014, 2016). For $h > 0$, consider the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{z} - \overline{\boldsymbol{x}}) \leq h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}. \quad (4.13)$$

Using $\gamma = h^2 = p/\delta$ in (4.13) usually results in prediction regions with volume and coverage that is too large. Using $\gamma = h^2 = \chi^2_{p,1-\delta}$ in (4.13) gives the classical prediction region (4.12), which usually has volume and coverage that is too low, although bounded above 0 by Theorem 4.9 asymptotically if $0 < \delta < 0.25$. (The median of a chi-square χ^2_p distribution is $\chi^2_{p,0.5} \approx p - 2/3$.) Using $h^2 = D^2(U_n)$ tends to give better volume and coverage.

Remark 4.6. The most used prediction regions assume that the error vectors are iid from a multivariate normal distribution. It can be shown that the ratio of the volumes of regions (4.12) and (4.11) is

$$\left(\frac{\chi^2_{p,1-\delta}}{D^2_{(U_n)}} \right)^{p/2},$$

which can become close to zero rapidly as p gets large if the $\boldsymbol{x}_i$ are not from the light tailed multivariate normal distribution. For example, suppose $\chi^2_{4,0.5} \approx 3.33$ and $D^2_{(U_n)} \approx D^2_{\boldsymbol{x},0.5} = 6$. Then the ratio is $(3.33/6)^2 \approx 0.308$. Hence if the data is not multivariate normal, severe undercoverage can occur if the classical prediction region (4.12) is used, and the undercoverage tends to get worse as the dimension p increases.

Remark 4.7. The nonparametric prediction region (4.11) starts to have good coverage for $n \geq 10p$ for a large class of distributions. Olive (2013b) suggests $n \geq 50p$ may be needed for the prediction region to have a good volume. Of course for any n there are distributions that will have severe undercoverage. Statisticians often say that correction factors are ad hoc, but doing nothing is much more ad hoc than using correction factors. Section 4.3 uses data splitting to derive a prediction region that does not need a correction factor.

For the multivariate lognormal distribution with $n = 20p$, the large sample nonparametric 95% prediction region (4.11) had coverages 0.970, 0.959, and 0.964 for $p = 100, 200$, and 500. Some R code is below.

```
nruns=1000 #lognormal, p = 100, n = 20p = 2000
count<-0
for(i in 1:nruns){
x <- exp(matrix(rnorm(200000),ncol=100,nrow=2000))
xff <- exp(as.vector(rnorm(100)))
count <- count + predrgn(x,xf=xff)$inr}
count #970/1000, may take a few minutes
```

If $\boldsymbol{X}$ and $\boldsymbol{Z}$ have dispersion matrices $\boldsymbol{\Sigma}$ and $c\boldsymbol{\Sigma}$ where $c > 0$, then the dispersion matrices have the same shape. The dispersion matrices determine the shape of the hyperellipsoid $\{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \leq h^2\}$. Figure 4.3 was made with the *Arc* software of Cook and Weisberg (1999). The 10%, 30%, 50%, 70%, 90%, and 98% highest density regions are shown for two multivariate normal (MVN) distributions. Both distributions have $\boldsymbol{\mu} = \mathbf{0}$. In Figure 4.3a),

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 4 \end{pmatrix}.$$

Note that the ellipsoids are narrow with high positive correlation. In Figure 4.3b),

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & -0.4 \\ -0.4 & 1 \end{pmatrix}.$$

Note that the ellipsoids are wide with negative correlation. The highest density ellipsoids are superimposed on a scatterplot of a sample of size 100 from each distribution.

4.3 Prediction Regions If n/p Is Small

Some of the data splitting prediction regions, described in this section, can handle $\boldsymbol{x}_f$ from a distribution where the population mean does not exist. Data splitting divides the training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ into two sets: H and the validation set V where H has n_H of the cases and V has the remaining $n_V = n - n_H$ cases $i_1, ..., i_{n_V}$. A common method of data splitting randomly divides the training data into the two sets H and V. Often $n_H \approx \lceil n/2 \rceil$.

The estimator $(T_H, \boldsymbol{C}_H)$ is computed using the data set H. Then the squared validation distances $D_j^2 = D_{\boldsymbol{x}_{i_j}}^2(T_H, \boldsymbol{C}_H) = (\boldsymbol{x}_{i_j} - T_H)^T \boldsymbol{C}_H^{-1}(\boldsymbol{x}_{i_j} - T_H)$ are computed for the $j = 1, ..., n_V$ cases in the validation set V. Let $D_{(U_V)}^2$ be the U_Vth order statistic of the D_j^2 where

$$U_V = \min(n_V, \lceil (n_V + 1)(1 - \delta) \rceil). \tag{4.14}$$

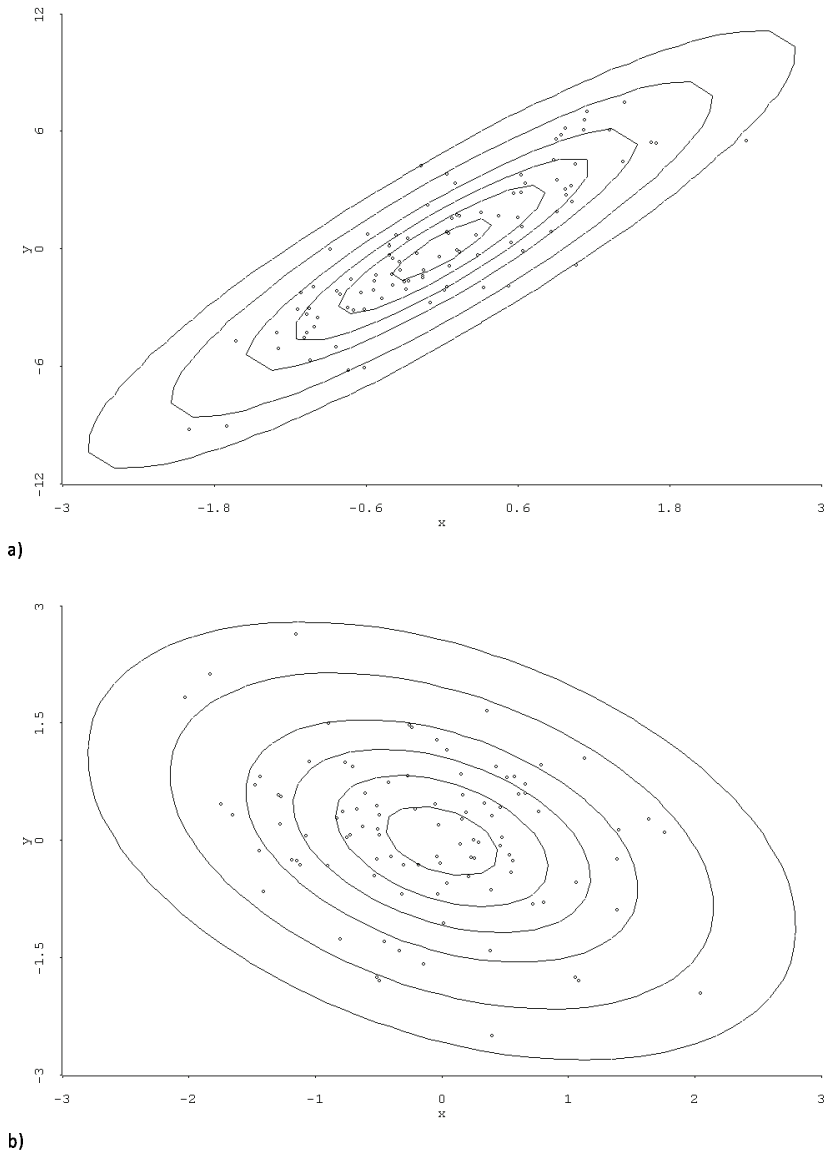

Fig. 4.3 Highest Density Regions for 2 MVN Distributions

Theorem 4.10. Assume that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid and that $\boldsymbol{C}_H^{-1}$ exists. The large sample $100(1-\delta)\%$ *data splitting prediction region* for $\boldsymbol{x}_f$ is

$$\{\boldsymbol{z} : D_{\boldsymbol{z}}^2(T_H, \boldsymbol{C}_H) \leq D_{(U_V)}^2\}. \tag{4.15}$$

Proof. To show that (4.15) is a prediction region, suppose the $\boldsymbol{x}_i$ are iid for $i = 1, ..., n, n+1$ where $\boldsymbol{x}_f = \boldsymbol{x}_{n+1}$. Compute $(T_H, \boldsymbol{C}_H)$ from the cases in H. Consider the squared validation distances D_k^2 for $k = 1, ..., n_V$ and the squared validation distance $D_{n_V+1}^2$ for case $\boldsymbol{x}_f$. Since these $n_V + 1$ cases are iid, the probability that D_t^2 has rank j for $j = 1, ..., n_V + 1$ is $1/(n_V + 1)$ for each t, i.e., the ranks follow the discrete uniform distribution. Let $t = n_V + 1$ and let the $D_{(j)}^2$ be the ordered squared validation distances using $j = 1, ..., n_V$. That is, get the order statistics without using the unknown squared validation distance $D_{n_V+1}^2$. Then $D_{(i)}^2$ has rank i if $D_{(i)}^2 < D_{n_V+1}^2$ but rank $i+1$ if $D_{(i)}^2 > D_{n_V+1}^2$. Thus $D_{(U_V)}^2$ has rank $U_V + 1$ if $D_{\boldsymbol{x}_f}^2 < D_{(U_V)}^2$ and

$$P(\boldsymbol{x}_f \in \{\boldsymbol{z} : D_{\boldsymbol{z}}^2(T_H, \boldsymbol{C}_H) \leq D_{(U_V)}^2\}) = P(D_{\boldsymbol{x}_f}^2 \leq D_{(U_V)}^2) \geq U_V/(1 + n_V) \rightarrow$$

$1 - \delta$ as $n_V \rightarrow \infty$. If there are no tied ranks, then

$$P(D_{\boldsymbol{x}_f}^2 \leq D_{(U_V)}^2) = P(D_{\boldsymbol{x}_f}^2 < D_{(U_V)}^2) = P(\text{rank of } \mathrm{D}_{\boldsymbol{x}_\mathrm{f}}^2 \leq \mathrm{U_V}) = \mathrm{U_V}/(1+\mathrm{n_V}).$$

□

Note that we can get the actual coverage $U_V/(1 + n_V)$ close to $1 - \delta$ for $n_V \geq 20$ for $\delta = 0.05$ even if $(T_H, \boldsymbol{C}_H)$ is a bad estimator. The volume of the prediction region tends to be much larger than that of the highest density region, even if $\boldsymbol{C}_H$ is well conditioned. We likely need $U_V \geq 50$ for $D_{(U_V)}^2$ to approximate the population percentile of $D_j^2 = (\boldsymbol{x}_{i_j} - T_H)^T \boldsymbol{C}_H^{-1} (\boldsymbol{x}_{i_j} - T_H)$.

The above prediction region coverage theory did not depend on the dimension p as long as $\boldsymbol{C}_H = \boldsymbol{C}$ is nonsingular. If $\boldsymbol{C} = \boldsymbol{I}_p$ or $\boldsymbol{C} = diag(S_1^2, ..., S_p^2)$, then prediction region (4.15) can be used for high dimensional data where $p > n$. Regularized covariance matrices or precision matrices could also be used.

Example 4.2. The Wisseman, Hopke, and Schindler-Kaudelka (1987) pottery data consists of a chemical analysis on pottery shards. The data set has 36 cases and 5 groups corresponding to types of pottery shards. The variables $x_1, ..., x_{20}$ correspond to the $p = 20$ chemicals analyzed. Consider the $n = 18$ group 1 cases where the pottery shards were Arretine, a type of Roman pottery. We randomly selected case 4 from group 1 to be $\boldsymbol{x}_f$ and computed the 88.89% data splitting prediction region with the remaining 17 cases, $n_V = 8$, and $(T, \boldsymbol{C}) = (MED(\boldsymbol{W}), \boldsymbol{I}_p)$ where $MED(\boldsymbol{W})$ is the coordinatewise median computed from the 9 cases in H. The cutoff $D_{(U_V)}^2 = 612.2$ and $D^2(\boldsymbol{x}_f) = 353.8$. Hence $\boldsymbol{x}_f$ was in the 88.89% prediction region. Next, we made $\boldsymbol{x}_f$ equal to each of the 36 cases. Then 8 cases $\boldsymbol{x}_f$ were not in the

above prediction region, including 7 of the 18 cases that were not from group 1.

Table 4.1 Data Splitting Nominal 95% Prediction region

n	p	nv	xtype	dtype	cov
50	100	20	1	1	0.9560
50	100	20	2	1	0.9466
50	100	20	3	1	0.9504
50	100	20	1	2	0.9558
50	100	20	2	2	0.9508
50	100	20	3	2	0.9522
100	100	50	1	1	0.9620
100	100	50	2	1	0.9622
100	100	50	3	1	0.9596
100	100	50	1	2	0.9638
100	100	50	2	2	0.9578
100	100	50	3	2	0.9638
100	100	25	1	1	0.9588
100	100	25	2	1	0.9658
100	100	25	3	1	0.9568
100	100	25	1	2	0.9622
100	100	25	2	2	0.9672
100	100	25	3	2	0.9662

The theory for the new prediction regions is simple, so Table 4.1 is more of a check that the programs work than that the theory works. The output gives cov = observed coverage, up $\approx$ actual coverage, and mnhsq = mean cutoff $D^2_{(U_V)}$. With 5000 runs, expect observed coverage $\in [0.94, 0.96]$ if the actual coverage is close to 0.95. The random vector $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{w}$ where $\boldsymbol{x} = \boldsymbol{w} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ for xtype = 3, and $\boldsymbol{x} \sim N_p(\boldsymbol{0}, diag(1, ..., p))$ for xtype = 1. For xtype = 2, $\boldsymbol{w}$ has the w_i iid lognormal(0,1) with $\boldsymbol{A} = diag(1, \sqrt{2}, ..., \sqrt{p})$. The dispersion matrix types are dtype = 1 if $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{I}_p)$ and dtype = 2 if $(T, \boldsymbol{C}) = (\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ where MED$(\boldsymbol{W})$ is the coordinatewise median of the $\boldsymbol{x}_i$.

When xtype=3 and dtype=1, $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{I}_p)$ where $\boldsymbol{x}_i \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$. Then $D^2_{(U_V)}$ should estimate the population percentile $\chi^2_{p,0.95}$ if $n \geq \max(20p, 200)$ and $n_V = 100$. This result did occur in the simulations.

Table 4.1 gives n, p, n_V, a number xtype corresponding to the distribution of $\boldsymbol{x}$, and a number dtype corresponding to $(T, \boldsymbol{C})$ used in prediction region (4.15). High dimensional data was used since $p \geq n$. With $n_V = 20$, the actual coverage is $20/21 = 0.9524$, $n_V = 25$ has actual coverage $25/26 = 0.9615$, and $n_V = 50$ has actual coverage $49/51 = 0.9608$. The observed coverages were close to the actual coverages in Table 4.1.

4.4 Summary

1) Consider predicting a future test value Y_f given training data $Y_1, ..., Y_n$. A large sample $100(1-\delta)\%$ *prediction interval* (PI) for Y_f has the form $[\hat{L}_n, \hat{U}_n]$ where $P(\hat{L}_n \leq Y_f \leq \hat{U}_n)$ is eventually bounded below by $1-\delta$ as the sample size $n \to \infty$. A large sample $100(1-\delta)\%$ PI is *asymptotically optimal* if it has the shortest asymptotic length: the length of $[\hat{L}_n, \hat{U}_n]$ converges to $U_s - L_s$ as $n \to \infty$ where $[L_s, U_s]$ is the *population shorth*: the shortest interval covering at least $100(1-\delta)\%$ of the mass.

2) Let $Y_1, ..., Y_n, Y_f$ be iid. Let $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ be the order statistics of the training data. Let $k_1 = \lceil n\delta/2 \rceil$ and $k_2 = \lceil n(1-\delta/2) \rceil$ where $0 < \delta < 1$. The large sample $100(1-\delta)\%$ *percentile prediction interval* for Y_f is

$$[Y_{(k_1)}, Y_{(k_2)}]. \tag{4.16}$$

3) Let the shortest closed interval containing at least c of the $Y_1, ..., Y_n$ be shorth(c) $= [\mathrm{Y}_{(\mathrm{s})}, \mathrm{Y}_{(\mathrm{s+c-1})}]$.

4) Let $Y_1, ..., Y_n, Y_f$ be iid. Let $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ be the order statistics of the training data. The large sample $100(1-\delta)\%$ shorth(c) prediction interval for Y_f is

$$[Y_{(s)}, Y_{(s+c-1)}] \text{ where } \mathrm{c} = \min(\mathrm{n}, \lceil \mathrm{n}[1 - \delta + 1.12\sqrt{\delta/\mathrm{n}}\ ]\ \rceil).$$

5) Let $Y_1, ..., Y_n, Y_f$ be iid. Let $W_{(1)} \leq W_{(2)} \leq \cdots \leq W_{(n)}$ be the order statistics of the squared training data $W_1, ..., W_n$ where $W_i = Y_i^2$ for $i = 1, ..., n$. Let $k_n = \lceil n(1-\delta) \rceil$. Let $L_n = -U_n$ and $U_n = \sqrt{W_{(k_n)}}$. Then $[L_n, U_n]$ is a large sample $100(1-\delta)\%$ PI for Y_f.

6) Let $Y_1, ..., Y_n, Y_f$ be iid. Suppose that $E(Y) = \mu$ and the standard deviation $SD(Y) = \sigma$. Let $\hat{\mu}$ and $\hat{\sigma}$ be consistent estimators of μ and σ. Let $1 - 1/k^2 \geq 1 - \delta$. Let $\mu \pm k\sigma$ be continuity points of $F_Y(y)$. Then

$$[L_n, U_n] = [\hat{\mu} - k\hat{\sigma}, \hat{\mu} + k\hat{\sigma}]$$

is a large sample $100(1-\delta)\%$ *Chebyshev PI* for Y_f.

Note often $k = 1.96$ is used which is good for a 95% PI for iid normal data, but is usually too short to be a 95% PI for iid data.

7) In a simulation for a PI, prediction region, CI, or confidence region with nominal $100(1-\delta)\%$ coverage, let the actual coverage $1 - \delta_n = P(a_n \in R)$ be $P(Y_f \in PI)$, $P(\boldsymbol{Y}_f \in$ prediction region$)$, $P(\theta \in CI)$, or $P(\boldsymbol{\theta} \in$ confidence region$)$. Then $P(a_n \in R) \sim bin(k, 1-\delta_n) \approx bin(k, 1-\delta)$ where k is the number of runs in the simulation. a) for $k = 5000$, simulated coverage in [0.94,0.95] suggests the actual coverage $1-\delta_n$ is close to the nominal coverage $1-\delta = 0.95$. b) for $k = 100$, simulated coverage in [0.89,1] suggests the actual coverage $1 - \delta_n$ is close to the nominal coverage $1 - \delta = 0.95$.

4.5 Complements

There are many prediction intervals and regions in the literature. For references, see Beran (1990, 1993), Fontana, Zeni, and Vantini (2023), Guan (2023), Olive (2013b, 2018), Steinberger and Leeb (2023), and Tian, Nordman, and Meeker (2022).

See Frey (2013) for references about nonparametric PIs. The shorth PI (4.1) often has good coverage for $n \geq 50$ and $0.05 \leq \delta \leq 0.1$, but the convergence of $U_n - L_n$ to the population shorth length $U_s - L_s$ can be quite slow. Under regularity conditions, Grübel (1982) showed that for iid data, the length and center of the shorth(k_n) interval are $\sqrt{n}$ consistent and $n^{1/3}$ consistent estimators of the length and center of the population shorth interval, respectively. Einmahl and Mason (1992) gave large sample theory for the shorth under slightly milder conditions than Grübel (1982). Chen and Shao (1999) showed that the shorth PI converges to the population shorth under mild conditions for ergodic data.

A method for obtaining an asymptotically optimal PI from a parametric distribution, possibly with right censored data, is given by Olive, Rathnayake, and Haile (2022). The data splitting prediction region of Section 4.3 was based on Haile, Zhang, and Olive (2024).

Prediction intervals and prediction regions can be used to estimate Bayesian credible intervals and Bayesian credible regions. Applying certain prediction regions to bootstrap samples results in confidence regions. See Chapter 5 and Welagedara and Olive (2024).

Software. The simulations were done in *R*. See R Core Team (2020). The function `predrgn` makes the nonparametric prediction region and determines whether $\boldsymbol{x}_f$ is in the region. The function `predreg` also makes the nonparametric prediction region, and determines if $\mathbf{0}$ is in the region. The `shorth3` function computes the shorth(c) intervals with the Frey (2013) correction used when $g = 1$. The function `predsim2` simulates the data splitting prediction region for Table 4.1. The function `predrgn2` computes the prediction region (4.14) using $(T, \boldsymbol{C}) = (\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$.

4.6 Problems

4.1. Consider the Cushny and Peebles data set (see Staudte and Sheather 1990, p. 97) listed below. Find shorth(7). Show work.

```
0.0  0.8  1.0  1.2  1.3  1.3  1.4  1.8  2.4  4.6
```

4.2. Find shorth(5) for the following data set. Show work.

```
6   76   90   90   94   94   95   97   97   1008
```

4.3. Find shorth(5) for the following data set. Show work.

```
66  76   90   90   94   94   95   95   97   98
```

4.4. The data below are a sorted residuals from a least squares regression where $n = 100$ and $p = 4$. Find shorth(97) of the residuals.

```
number      1     2     3     4   ...  97   98   99   100
residual -2.39 -2.34 -2.03 -1.77 ... 1.76 1.81 1.83  2.16
```

R Problems

Use the command *source("G:/lsamppack.txt")* **to download the functions** and the command *source("G:/lsampdata.txt")* **to download the data. See Preface.** Typing the name of the `lsamppack` function, e.g. *predsim*, will display the code for the function. Use the `args` command, e.g. *args(predsim)*, to display the needed arguments for the function. For the following problem, the *R* command can be copied and pasted from (http://parker.ad.siu.edu/Olive/lsamphw.txt) into *R*.

4.5. a) Type the *R* command `predsim()` and paste the output into *Word*.

This program computes $\boldsymbol{x}_i \sim N_4(\mathbf{0}, diag(1, 2, 3, 4))$ for $i = 1, ..., 100$ and $\boldsymbol{x}_f = \boldsymbol{x}_{101}$. One hundred such data sets are made, and ncvr, scvr, and mcvr count the number of times $\boldsymbol{x}_f$ was in the nonparametric, semiparametric, and parametric MVN 90% prediction regions. The volumes of the prediction regions are computed and voln, vols, and volm are the average ratio of the volume of the ith prediction region over that of the semiparametric region. Hence vols is always equal to 1. For multivariate normal data, these ratios should converge to 1 as $n \to \infty$.

b) Were the three coverages near 90%?

More problems:

4.6. For a Poisson regression model, $Y|\boldsymbol{x} \sim$ Poisson[exp($\boldsymbol{x}^T\boldsymbol{\beta}$)]. Suppose $\hat{\boldsymbol{\beta}}_n$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$. Let $W_n \sim$ Poisson[exp($\boldsymbol{x}_f^T\hat{\boldsymbol{\beta}}_n$)]. Treat $\boldsymbol{x}_f$ as a known constant vector. Then W_n approximates W. What is the distribution of W?

(Note: to get a prediction interval for $Y_f|\boldsymbol{x}_f$, generate an iid sample $W_1, ..., W_B$ where $W_i \sim$ Poisson[exp($\boldsymbol{x}_f^T\hat{\boldsymbol{\beta}}_n$)]. Then compute the shorth PI from the W_i. This technique is called the parametric bootstrap. It is not clear whether $W_n \xrightarrow{D} W$.)

Chapter 5
Confidence Regions and the Bootstrap

This chapter follows Olive (2014, ch. 9; 2017b, ∮ 5.3) closely. Also see Olive (2023abcd). Sections 5.1–5.3 consider confidence intervals from asymptotic pivots while Section 5.4 covers bootstrap confidence regions. Closed regions are better than open regions. Again, $0 < \delta < 1$. Applying certain prediction intervals or prediction regions to the bootstrap sample will result in confidence intervals or confidence regions. The prediction intervals and regions are based on samples of size n, while the bootstrap sample size is $B = B_n$.

Notation: As in Chapter 4, $P(A_n)$ is "eventually bounded below" by $1-\delta$ if $P(A_n)$ gets arbitrarily close to or higher than $1-\delta$ as $n \to \infty$. Hence $P(A_n) > 1 - \delta - \epsilon$ for any $\epsilon > 0$ if n is large enough. If $P(A_n) \to 1-\delta$ as $n \to \infty$, then $P(A_n)$ is eventually bounded below by $1-\delta$. The actual coverage is $1 - \gamma_n = P(\theta \in [L_n, U_n])$, the nominal coverage is $1-\delta$ where $0 < \delta < 1$. The 90% and 95% large sample confidence intervals and confidence regions are common.

5.1 Confidence Intervals

Definition 5.1. Let the data $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ have joint pdf or pmf $f(\boldsymbol{y}|\theta)$ with parameter space Θ and support $\mathcal{Y}$. Let $L_n(\boldsymbol{Y})$ and $U_n(\boldsymbol{Y})$ be statistics such that $L_n(\boldsymbol{y}) \leq U_n(\boldsymbol{y})$, $\forall \boldsymbol{y} \in \mathcal{Y}$. Then $[L_n(\boldsymbol{y}), U_n(\boldsymbol{y})]$ is a 100 $(1-\delta)$ % **confidence interval** (CI) for θ if

$$P_\theta(L_n(\boldsymbol{Y}) \leq \theta \leq U_n(\boldsymbol{Y})) = 1 - \delta$$

for all $\theta \in \Theta$. The interval $[L_n(\boldsymbol{y}), U_n(\boldsymbol{y})]$ is a large sample $100(1-\delta)$ % CI for θ if

$$P_\theta(L_n(\boldsymbol{Y}) \leq \theta \leq U_n(\boldsymbol{Y}))$$

is eventually bounded below by $1-\delta$ for all $\theta \in \Theta$ as the sample size $n \to \infty$.

Pivots and asymptotic pivots are used to make CIs. An asymptotic pivot is a random quantity that is not a statistic since the asymptotic pivot depends on the unknown parameters $\boldsymbol{\theta}$.

Definition 5.2. Let the data $Y_1, ..., Y_n$ have joint pdf or pmf $f(\boldsymbol{y}|\boldsymbol{\theta})$ with parameter space Θ and support $\mathcal{Y}$. The quantity $R(\boldsymbol{Y}|\boldsymbol{\theta})$ is a **pivot** or pivotal quantity if the distribution of $R(\boldsymbol{Y}|\boldsymbol{\theta})$ is independent $\boldsymbol{\theta}$. The quantity $R(\boldsymbol{Y}, \boldsymbol{\theta})$ is an **asymptotic pivot** or asymptotic pivotal quantity if the limiting distribution of $R(\boldsymbol{Y}, \boldsymbol{\theta})$ is independent of $\boldsymbol{\theta}$.

The first CI in Definition 5.1 is sometimes called an exact CI. The words "exact" and "large sample" are often omitted. In the following definition, the scaled asymptotic length is closely related to asymptotic relative efficiency of an estimator and high power of a test of hypotheses.

Definition 5.3. Let $[L_n, U_n]$ be a 100 $(1-\delta)$% CI or large sample CI for θ. If

$$n^{\tau}(U_n - L_n) \xrightarrow{P} A_{\delta}$$

where $0 < \tau \leq 1$, then A_{δ} is the *scaled asymptotic length* of the CI. Typically $\tau = 0.5$ but superefficient CIs have $\tau = 1$. For fixed τ and fixed coverage $1-\delta$, a CI with smaller A_{δ} is "better" than a CI with larger A_{δ}. If $A_{1,\delta}$ and $A_{2,\delta}$ are for two competing CIs with the same τ, then $(A_{2,\delta}/A_{1,\delta})^{1/\tau}$ is a measure of "asymptotic relative efficiency."

Definition 5.4. Suppose a nominal $100(1-\delta)$% CI for θ has actual coverage $1-\gamma$, so that $P_{\theta}(L_n(\boldsymbol{Y}) \leq \theta \leq U_n(\boldsymbol{Y})) = 1-\gamma$ for all $\theta \in \Theta$. If $1-\gamma > 1-\delta$, then the CI is *conservative*. If $1-\gamma < 1-\delta$, then the CI is *liberal*. Conservative CIs are generally considered better than liberal CIs. Suppose a nominal $100(1-\delta)$% large sample CI for θ has actual coverage $1-\gamma_n$ where $\gamma_n \to \gamma$ as $n \to \infty$ for all $\theta \in \Theta$. If $1-\gamma > 1-\delta$, then the CI is *asymptotically conservative*. If $1-\gamma < 1-\delta$, then the CI is *asymptotically liberal*. It is possible that $\gamma \equiv \gamma(\theta)$ depends on θ, and that the CI is (asymptotically) conservative or liberal for different values of θ, in that the (asymptotic) coverage is higher or lower than the nominal coverage, depending on θ.

Example 5.1. a) Let $Y_1, ..., Y_n$ be iid $N(\mu, \sigma^2)$ where $\sigma^2 > 0$. Then

$$R(\boldsymbol{Y}|\mu, \sigma^2) = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

is a pivot. A statistic does not depend on any unknown parameters. Hence the above pivot is not a statistic if μ is unknown.

To use this pivot to find a CI for μ, let $t_{p,\delta}$ be the δ percentile of the t_p distribution. Hence $P(T \leq t_{p,\delta}) = \delta$ if $T \sim t_p$. Using $t_{p,\delta} = -t_{p,1-\delta}$ for $0 < \delta < 0.5$, note that

$$1-\delta = P(-t_{n-1,1-\delta/2} \le \frac{\overline{Y}-\mu}{S/\sqrt{n}} \le t_{n-1,1-\delta/2}) =$$

$$P(-t_{n-1,1-\delta/2} \;\; S/\sqrt{n} \le \overline{Y}-\mu \le t_{n-1,1-\delta/2} \;\; S/\sqrt{n}) =$$
$$P(-\overline{Y} - t_{n-1,1-\delta/2} \;\; S/\sqrt{n} \le -\mu \le -\overline{Y} + t_{n-1,1-\delta/2} \;\; S/\sqrt{n}) =$$
$$P(\overline{Y} - t_{n-1,1-\delta/2} \;\; S/\sqrt{n} \le \mu \le \overline{Y} + t_{n-1,1-\delta/2} \;\; S/\sqrt{n}).$$

Thus

$$\overline{Y} \pm t_{n-1,1-\delta/2} \;\; S/\sqrt{n}$$

is a $100(1-\delta)\%$ CI for μ.

b) If $Y_1, ..., Y_n$ are iid with $E(Y) = \mu$ and $\text{VAR}(Y) = \sigma^2 > 0$, then, by the CLT and Slutsky's Theorem,

$$R(\boldsymbol{Y}, \mu, \sigma^2) = \frac{\overline{Y}-\mu}{S/\sqrt{n}} = \frac{\sigma}{S} \;\; \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \xrightarrow{D} N(0,1)$$

is an asymptotic pivot.

To use this asymptotic pivot to find a large sample CI for μ, let z_δ be the δ percentile of the $N(0,1)$ distribution. Hence $P(Z \le z_\delta) = \delta$ if $Z \sim N(0,1)$. Using $z_\delta = -z_{1-\delta}$ for $0 < \delta < 0.5$, note that for large n,

$$1-\delta \approx P(-z_{1-\delta/2} \le \frac{\overline{Y}-\mu}{S/\sqrt{n}} \le z_{1-\delta/2}) =$$

$$P(-z_{1-\delta/2} \;\; S/\sqrt{n} \le \overline{Y}-\mu \le z_{1-\delta/2} \;\; S/\sqrt{n}) =$$
$$P(-\overline{Y} - z_{1-\delta/2} \;\; S/\sqrt{n} \le -\mu \le -\overline{Y} + z_{1-\delta/2} \;\; S/\sqrt{n}) =$$
$$P(\overline{Y} - z_{1-\delta/2} \;\; S/\sqrt{n} \le \mu \le \overline{Y} + z_{1-\delta/2} \;\; S/\sqrt{n}).$$

Thus

$$\overline{Y} \pm z_{1-\delta/2} \;\; S/\sqrt{n} \tag{5.1}$$

is a large sample $100(1-\delta)\%$ CI for μ.

Since $t_{n-1,1-\delta/2} > z_{1-\delta/2}$ but $t_{n-1,1-\delta/2} \to z_{1-\delta/2}$ as $n \to \infty$,

$$\overline{Y} \pm t_{n-1,1-\delta/2} \;\; S/\sqrt{n} \tag{5.2}$$

is also a large sample $100(1-\delta)\%$ CI for μ. This t interval is the same as that in a), and is likely the most widely used confidence interval in statistics. Replacing $z_{1-\delta/2}$ by $t_{n-1,1-\delta/2}$ makes the CI longer and hence less likely to be liberal.

Remark 5.1.

$$\overline{Y} \pm t_{n-1,1-\delta/2} \;\; S/\sqrt{n} = \overline{Y} \pm \frac{t_{n-1,1-\delta/2}}{z_{1-\delta/2}} \;\; z_{1-\delta/2} S/\sqrt{n}$$

where

$$\frac{t_{n-1,1-\delta/2}}{z_{1-\delta/2}} \to 1$$

as $n \to \infty$ is a small sample correction factor. See Example 2.16. The CI (5.2) should be used instead of the CI (5.1). If a large sample $100(1-\delta)\%$ CI for θ is $\hat{\theta} \pm z_{1-\delta/2} SE(\hat{\theta})$, then the large sample $100(1-\delta)\%$ CI $\hat{\theta} \pm t_{d_n,1-\delta/2} SE(\hat{\theta})$ where $d_n \to \infty$ as $n \to \infty$ tends to perform better for small sample sizes. Typically the actual distribution of the asymptotic pivot has heavier tails than the N(0,1) distribution for moderate sample sizes, and using a correction factor improves performance.

5.2 Large Sample CIs and Tests

Large sample theory can be used to construct *confidence intervals* and *hypothesis tests*. Suppose that $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ and that $W_n \equiv W_n(\boldsymbol{Y})$ is an estimator of some parameter μ_W such that

$$\sqrt{n}(W_n - \mu_W) \xrightarrow{D} N(0, \sigma_W^2)$$

where σ_W^2/n is the asymptotic variance of the estimator W_n. The above notation means that if n is large, then for probability calculations

$$W_n - \mu_W \approx N(0, \sigma_W^2/n).$$

Suppose that S_W^2 is a consistent estimator of σ_W^2 so that the (asymptotic) *standard error* of W_n is $\mathrm{SE}(W_n) = S_W/\sqrt{n}$. Using the notation of Example 5.1,

$$P\left(-z_{1-\delta/2} \leq \frac{W_n - \mu_W}{SE(W_n)} \leq z_{1-\delta/2}\right) \to 1 - \delta$$

and a large sample $100(1-\delta)\%$ CI for μ_W is given by

$$[W_n - z_{1-\delta/2} SE(W_n), W_n + z_{1-\delta/2} SE(W_n)]. \tag{5.3}$$

Three common approximate level δ tests of hypotheses all use the *null hypothesis* $H_o : \mu_W = \mu_o$. A right tailed test uses the *alternative hypothesis* $H_A : \mu_W > \mu_o$, a left tailed test uses $H_A : \mu_W < \mu_o$, and a two tail test uses $H_A : \mu_W \neq \mu_o$. The test statistic is

$$t_o = \frac{W_n - \mu_o}{SE(W_n)},$$

and the (approximate) *p-values* are $P(Z > t_o)$ for a right tail test, $P(Z < t_o)$ for a left tail test, and $2P(Z > |t_o|) = 2P(Z < -|t_o|)$ for a two tail test. The null hypothesis H_o is rejected if the p-value $< \delta$.

Remark 5.2. Frequently the large sample CIs and tests can be improved for smaller samples by substituting a t distribution with d_n degrees of freedom for the standard normal distribution Z where d_n is some increasing function of the sample size n. Then the $100(1-\delta)\%$ CI for μ_W is given by

$$[W_n - t_{d_n,1-\delta/2} SE(W_n), W_n + t_{d_n,1-\delta/2} SE(W_n)]. \tag{5.4}$$

The test statistic rarely has an exact t_{d_n} distribution, but CI (5.6) often performs better than the CI (5.5) in small samples. The CI (5.6) is longer than the CI (5.5), and H_0 is less likely to be rejected. Hence the CI (5.6) is more *conservative* than the CI (5.5). This book will typically use very simple rules for d_n and not investigate the exact distribution of the test statistic. Note that the small sample correction factor

$$\frac{t_{d_n,1-\delta/2}}{z_{1-\delta/2}} \to 1$$

if $d_n \equiv p_n \to \infty$ as $n \to \infty$. See Example 2.16.

Paired and two sample procedures can be obtained directly from the one sample procedures. Suppose there are two samples $Y_1, ..., Y_n$ and $X_1, ..., X_m$. If $n = m$ and it is known that (Y_i, X_i) match up in correlated pairs, then *paired* CIs and tests apply the one sample procedures to the differences $D_i = Y_i - X_i$. Otherwise, assume the two samples are independent, that n and m are large, and that

$$\begin{pmatrix} \sqrt{n}(W_n(\boldsymbol{Y}) - \mu_W(Y)) \\ \sqrt{m}(W_m(\boldsymbol{X}) - \mu_W(X)) \end{pmatrix} \xrightarrow{D} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_W(Y) & 0 \\ 0 & \sigma^2_W(X) \end{pmatrix} \right).$$

Then

$$\begin{pmatrix} (W_n(\boldsymbol{Y}) - \mu_W(Y)) \\ (W_m(\boldsymbol{X}) - \mu_W(X)) \end{pmatrix} \approx N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_W(Y)/n & 0 \\ 0 & \sigma^2_W(X)/m \end{pmatrix} \right),$$

and

$$W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X}) - (\mu_W(Y) - \mu_W(X)) \approx N\left(0, \frac{\sigma^2_W(Y)}{n} + \frac{\sigma^2_W(X)}{m}\right).$$

Hence $SE(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})) =$

$$\sqrt{\frac{S^2_W(\boldsymbol{Y})}{n} + \frac{S^2_W(\boldsymbol{X})}{m}} = \sqrt{[SE(W_n(\boldsymbol{Y}))]^2 + [SE(W_m(\boldsymbol{X}))]^2},$$

and the large sample $100(1-\delta)\%$ CI for $\mu_W(Y) - \mu_W(X)$ is given by

$$(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})) \pm z_{1-\delta/2} SE(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})).$$

Often approximate level δ tests of hypotheses use the *null hypothesis* H_o : $\mu_W(Y) = \mu_W(X)$. A right tailed test uses the *alternative hypothesis* H_A : $\mu_W(Y) > \mu_W(X)$, a left tailed test uses $H_A : \mu_W(Y) < \mu_W(X)$, and a two tail test uses $H_A : \mu_W(Y) \neq \mu_W(X)$. The test statistic is

$$t_o = \frac{W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})}{SE(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X}))},$$

and the (approximate) *p-values* are $P(Z > t_o)$ for a right tail test, $P(Z < t_o)$ for a left tail test, and $2P(Z > |t_o|) = 2P(Z < -|t_o|)$ for a two tail test. The null hypothesis H_o is rejected if the p-value $< \delta$.

Remark 5.3. Again a t_{p_n} cutoff will often be used instead of the z cutoff. If d_n is the degrees of freedom used for a single sample procedure when the sample size is n, use $d_{n,m} = \min(d_n, d_m)$ for the two sample procedure if a better formula is not given. Then the large sample $100(1-\delta)\%$ CI for $\mu_W(Y) - \mu_W(X)$ is

$$(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})) \pm t_{d_{n,m},1-\delta/2} SE(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})). \tag{5.5}$$

These CIs are known as *Welch intervals*. See Welch (1937) and Yuen (1974).

Example 5.2. Consider the single sample procedures where $W_n = \overline{Y}_n$. Then $\mu_W = E(Y)$, $\sigma_W^2 = \mathrm{VAR}(Y)$, $S_W = S_n$, and $d_n = n - 1$. Then the classical *t-interval* for $\mu \equiv E(Y)$ is

$$\overline{Y}_n \pm t_{n-1,1-\delta/2} \frac{S_n}{\sqrt{n}}$$

and the *t-test statistic* is

$$t_o = \frac{\overline{Y} - \mu_o}{S_n/\sqrt{n}}.$$

The right tailed p-value is given by $P(t_{n-1} > t_o)$.

Now suppose that there are two samples where $W_n(\boldsymbol{Y}) = \overline{Y}_n$ and $W_m(\boldsymbol{X}) = \overline{X}_m$. Then $\mu_W(Y) = E(Y) \equiv \mu_Y$, $\mu_W(X) = E(X) \equiv \mu_X$, $\sigma_W^2(Y) = \mathrm{VAR}(Y) \equiv \sigma_Y^2$, $\sigma_W^2(X) = \mathrm{VAR}(X) \equiv \sigma_X^2$, and $d_n = n - 1$. Let $d_{n,m} = \min(n-1, m-1)$. Since

$$SE(W_n(\boldsymbol{Y}) - W_m(\boldsymbol{X})) = \sqrt{\frac{S_n^2(\boldsymbol{Y})}{n} + \frac{S_m^2(\boldsymbol{X})}{m}},$$

the *two sample t-interval* for $\mu_Y - \mu_X$ is

$$(\overline{Y}_n - \overline{X}_m) \pm t_{d_{n,m},1-\delta/2}\sqrt{\frac{S_n^2(\boldsymbol{Y})}{n} + \frac{S_m^2(\boldsymbol{X})}{m}}$$

and *two sample t-test statistic* is

$$t_o = \frac{\overline{Y}_n - \overline{X}_m}{\sqrt{\frac{S_n^2(\boldsymbol{Y})}{n} + \frac{S_m^2(\boldsymbol{X})}{m}}}.$$

The right tailed p-value is given by $P(t_{d_n,m} > t_o)$. For sample means, values of the degrees of freedom that are more accurate than $d_{n,m} = \min(n-1, m-1)$ can be computed. See Moore (2007, p. 474) and Example 5.9.

5.3 Some CI Examples

Example 5.3. Suppose that $Y_1, ..., Y_n$ are iid from a one parameter exponential family with parameter τ. Assume that $T_n = \sum_{i=1}^n t(Y_i)$ is a complete sufficient statistic. Then Olive (2014, pp. 92-93), often $T_n \sim G(na, 2b\ \tau)$ where a and b are known positive constants. Then

$$\hat{\tau} = \frac{T_n}{2nab}$$

is the UMVUE and often the MLE of τ. Since $T_n/(b\ \tau) \sim G(na, 2)$, a $100(1-\delta)\%$ confidence interval for τ is

$$\left[\frac{T_n/b}{G(na, 2, 1-\delta/2)}, \frac{T_n/b}{G(na, 2, \delta/2)}\right] \approx \left[\frac{T_n/b}{\chi^2_d(1-\delta/2)}, \frac{T_n/b}{\chi^2_d(\delta/2)}\right] \tag{5.6}$$

where $d = \lfloor 2na \rfloor$, $\lfloor x \rfloor$ is the greatest integer function (e.g. $\lfloor 7.7 \rfloor = \lfloor 7 \rfloor = 7$), $P[G \leq G(\nu, \lambda, \delta)] = \delta$ if $G \sim G(\nu, \lambda)$, and $P[X \leq \chi^2_d(\delta)] = \delta$ if X has a chi-square χ^2_d distribution with d degrees of freedom.

This confidence interval can be inverted to perform two tail tests of hypotheses. By Olive (2014, p. 186: Theorem 7.3), if $w(\theta)$ is increasing, then the uniformly most powerful (UMP) test of $H_o : \tau \leq \tau_o$ versus $H_A : \tau > \tau_o$ rejects H_0 if and only if $T_n > k$ where $P[G > k] = \delta$ when $G \sim G(na, 2b\ \tau_o)$. Hence

$$k = G(na, 2b\ \tau_o, 1-\delta). \tag{5.7}$$

A good approximation to this test rejects H_0 if and only if

$$T_n > b\ \tau_o \chi^2_d(1-\delta)$$

where $d = \lfloor 2na \rfloor$.

Example 5.4. Olive (2014, pp. 264-266): If Y is half normal HN(μ, σ) then the pdf of Y is

$$f(y) = \frac{2}{\sqrt{2\pi}\ \sigma} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)$$

where $\sigma > 0$ and $y > \mu$ and μ is real. Since

$$f(y) = \frac{2}{\sqrt{2\pi}\ \sigma} I[y > \mu] \exp\left[\left(\frac{-1}{2\sigma^2}\right)(y-\mu)^2\right],$$

Y is a 1P–REF if μ is known.

Since $T_n = \sum(Y_i - \mu)^2 \sim G(n/2, 2\sigma^2)$, in Example 5.3 take $a = 1/2$, $b = 1$, $d = n$ and $\tau = \sigma^2$. Then a $100(1-\delta)\%$ confidence interval for σ^2 is

$$\left[\frac{T_n}{\chi^2_n(1-\delta/2)}, \frac{T_n}{\chi^2_n(\delta/2)}\right]. \tag{5.8}$$

The UMP test of $H_0 : \sigma^2 \leq \sigma^2_o$ versus $H_A : \sigma^2 > \sigma^2_o$ rejects H_o if and only if

$$T_n/\sigma^2_o > \chi^2_n(1-\delta).$$

Now consider inference when both μ and σ are unknown. Then the family is no longer an exponential family since the support depends on μ. Let

$$D_n = \sum_{i=1}^{n} (Y_i - Y_{1:n})^2. \tag{5.9}$$

Pewsey (2002) showed that $(\hat{\mu}, \hat{\sigma}^2) = (Y_{1:n}, \frac{1}{n} D_n)$ is the MLE of (μ, σ^2), and that

$$\frac{Y_{1:n} - \mu}{\sigma \Phi^{-1}(\frac{1}{2} + \frac{1}{2n})} \xrightarrow{D} EXP(1)$$

where $Y_{1:n} = Y_{(1)} = \min(Y_1, ..., Y_n)$ is the first order statistic. Since $(\sqrt{\pi/2})/n$ is an approximation to $\Phi^{-1}(\frac{1}{2} + \frac{1}{2n})$ based on a first order Taylor series expansion such that

$$\frac{\Phi^{-1}(\frac{1}{2} + \frac{1}{2n})}{(\sqrt{\pi/2})/n} \to 1,$$

it follows that

$$\frac{n(Y_{1:n} - \mu)}{\sigma\sqrt{\frac{\pi}{2}}} \xrightarrow{D} EXP(1). \tag{5.10}$$

Using this fact, it can be shown that a large sample $100(1-\delta)\%$ CI for μ is

$$[\hat{\mu} + \hat{\sigma} \log(\delta)\ \Phi^{-1}(\frac{1}{2} + \frac{1}{2n})\ (1 + 13/n^2),\quad \hat{\mu}] \tag{5.11}$$

where the term $(1 + 13/n^2)$ is a small sample correction factor.

Note that

$$D_n = \sum_{i=1}^{n}(Y_i - Y_{1:n})^2 = \sum_{i=1}^{n}(Y_i - \mu + \mu - Y_{1:n})^2 =$$

$$\sum_{i=1}^{n}(Y_i - \mu)^2 + n(\mu - Y_{1:n})^2 + 2(\mu - Y_{1:n})\sum_{i=1}^{n}(Y_i - \mu).$$

Hence

$$D_n = T_n + \frac{1}{n}[n(Y_{1:n} - \mu)]^2 - 2[n(Y_{1:n} - \mu)]\frac{\sum_{i=1}^{n}(Y_i - \mu)}{n},$$

or

$$\frac{D_n}{\sigma^2} = \frac{T_n}{\sigma^2} + \frac{1}{n}\frac{1}{\sigma^2}[n(Y_{1:n} - \mu)]^2 - 2[\frac{n(Y_{1:n} - \mu)}{\sigma}]\frac{\sum_{i=1}^{n}(Y_i - \mu)}{n\sigma}. \quad (5.12)$$

Consider the three terms on the right hand side of (5.12). The middle term converges to 0 in distribution while the third term converges in distribution to a $-2EXP(1)$ or $-\chi^2_2$ distribution since $\sum_{i=1}^{n}(Y_i - \mu)/(\sigma n)$ is the sample mean of HN(0,1) random variables and $E(X) = \sqrt{2/\pi}$ when $X \sim HN(0, 1)$.

Let $T_{n-p} = \sum_{i=1}^{n-p}(Y_i - \mu)^2$. Then

$$D_n = T_{n-p} + \sum_{i=n-p+1}^{n}(Y_i - \mu)^2 - V_n \quad (5.13)$$

where

$$\frac{V_n}{\sigma^2} \xrightarrow{D} \chi^2_2.$$

Hence

$$\frac{D_n}{T_{n-p}} \xrightarrow{D} 1$$

and D_n/σ^2 is asymptotically equivalent to a χ^2_{n-p} random variable where p is an arbitrary nonnegative integer. Pewsey (2002) used $p = 1$.

Thus when both μ and σ^2 are unknown, a large sample $100(1 - \delta)\%$ confidence interval for σ^2 is

$$\left[\frac{D_n}{\chi^2_{n-1}(1 - \delta/2)}, \frac{D_n}{\chi^2_{n-1}(\delta/2)}\right]. \quad (5.14)$$

It can be shown that $\sqrt{n}$ CI length converges in probability to $\sigma^2\sqrt{2}(z_{1-\delta/2} - z_{\delta/2})$ for CIs (5.8) and (5.14) while n length CI (5.11) converges in probability to $-\sigma \log(\delta)\sqrt{\pi/2}$.

When μ and σ^2 are unknown, an approximate δ level test of $H_o : \sigma^2 \leq \sigma_o^2$ versus $H_A : \sigma^2 > \sigma_o^2$ that rejects H_o if and only if

$$D_n/\sigma_o^2 > \chi^2_{n-1}(1-\delta) \tag{5.15}$$

has nearly as much power as the δ level UMP test when μ is known if n is large.

Example 5.5. Let $X_1, ..., X_n$ be iid Poisson(θ) random variables. The classical large sample 100 $(1-\delta)$% CI for θ is

$$\overline{X} \pm z_{1-\delta/2}\sqrt{\overline{X}/n}$$

where $P(Z \leq z_{1-\delta/2}) = 1 - \delta/2$ if $Z \sim N(0,1)$.

Following Byrne and Kabaila (2005), a modified large sample 100 $(1-\delta)$% CI for θ is $[L_n, U_n]$ where

$$L_n = \frac{1}{n}\left(\sum_{i=1}^{n} X_i - 0.5 + 0.5z^2_{1-\delta/2} - z_{1-\delta/2}\sqrt{\sum_{i=1}^{n} X_i - 0.5 + 0.25z^2_{1-\delta/2}}\right)$$

and

$$U_n = \frac{1}{n}\left(\sum_{i=1}^{n} X_i + 0.5 + 0.5z^2_{1-\delta/2} + z_{1-\delta/2}\sqrt{\sum_{i=1}^{n} X_i + 0.5 + 0.25z^2_{1-\delta/2}}\right).$$

Following Grosh (1989, p. 59, 197–200), let $W = \sum_{i=1}^{n} X_i$ and suppose that $W = w$ is observed. Let $P(T < \chi^2_d(\delta)) = \delta$ if $T \sim \chi^2_d$. Then an "exact" 100 $(1-\delta)$% CI for θ is

$$\left[\frac{\chi^2_{2w}(\frac{\delta}{2})}{2n}, \frac{\chi^2_{2w+2}(1-\frac{\delta}{2})}{2n}\right]$$

for $w \neq 0$ and

$$\left[0, \frac{\chi^2_2(1-\delta)}{2n}\right]$$

for $w = 0$.

The "exact" CI is conservative: the actual coverage $(1-\delta_n) \geq 1-\delta =$ the nominal coverage. This interval performs well if θ is very close to 0. See Problem 5.2.

Example 5.6. Let $Y_1, ..., Y_n$ be iid bin$(1, \rho)$. Let $\hat{\rho} = \sum_{i=1}^{n} Y_i/n =$ number of "successes"$/n$. The classical large sample 100 $(1-\delta)$% CI for ρ is

$$\hat{\rho} \pm z_{1-\delta/2}\sqrt{\frac{\hat{\rho}(1-\hat{\rho})}{n}}$$

where $P(Z \leq z_{1-\delta/2}) = 1 - \delta/2$ if $Z \sim N(0,1)$.

The Agresti and Coull (1998) CI takes $\tilde{n} = n + z^2_{1-\delta/2}$ and

$$\tilde{\rho} = \frac{n\hat{\rho} + 0.5z^2_{1-\delta/2}}{n + z^2_{1-\delta/2}}.$$

(The method "adds" $0.5z^2_{1-\delta/2}$ "0's" and $0.5z^2_{1-\delta/2}$ "1's" to the sample, so the "sample size" increases by $z^2_{1-\delta/2}$.) Then the large sample 100 $(1-\delta)$% Agresti Coull CI for ρ is

$$\tilde{\rho} \pm z_{1-\delta/2}\sqrt{\frac{\tilde{\rho}(1-\tilde{\rho})}{\tilde{n}}}.$$

Now let $Y_1, ..., Y_n$ be independent bin(m_i, ρ) random variables, let $W = \sum_{i=1}^n Y_i \sim \text{bin}(\sum_{\text{i=1}}^{\text{n}} \text{m}_\text{i}, \rho)$ and let $n_w = \sum_{i=1}^n m_i$. Often $m_i \equiv 1$ and then $n_w = n$. Let $P(F_{d_1,d_2} \leq F_{d_1,d_2}(\delta)) = \delta$ where F_{d_1,d_2} has an F distribution with d_1 and d_2 degrees of freedom. Assume $W = w$ is observed. Then the Clopper Pearson "exact" 100 $(1-\delta)$% CI for ρ is

$$\left[0, \frac{1}{1 \ + \ n_w \ F_{2n_w,2}(\delta)}\right] \ \text{ for } \ \text{w} = 0,$$

$$\left[\frac{n_w}{n_w \ + \ F_{2,2n_w}(1-\delta)}, 1\right] \ \text{ for } \ \text{w} = \text{n}_\text{w},$$

and $[\rho_L, \rho_U]$ for $0 < w < n_w$ with

$$\rho_L = \frac{w}{w + (n_w - w + 1)F_{2(n_w-w+1),2w}(1-\delta/2)}$$

and

$$\rho_U = \frac{w+1}{w + 1 + (n_w - w)F_{2(n_w-w),2(w+1)}(\delta/2)}.$$

The "exact" CI is conservative: the actual coverage $(1-\delta_n) \geq 1-\delta =$ the nominal coverage. This interval performs well if ρ is very close to 0 or 1. The classical interval should only be used if it agrees with the Agresti Coull interval. See Problem 5.3.

Example 5.7. Let $\hat{\rho}$ = number of "successes"$/n$. Consider a taking a simple random sample of size n from a finite population of known size N. Then the classical finite population large sample 100 $(1-\delta)$% CI for ρ is

$$\hat{\rho} \pm z_{1-\delta/2}\sqrt{\frac{\hat{\rho}(1-\hat{\rho})}{n-1}\left(\frac{N-n}{N}\right)} = \hat{\rho} \pm z_{1-\delta/2}SE(\hat{\rho}) \tag{5.16}$$

where $P(Z \leq z_{1-\delta/2}) = 1 - \delta/2$ if $Z \sim N(0,1)$.

Following DasGupta (2008, p. 121), suppose the number of successes Y has a hypergeometric $(C, N-C, n)$ where $p = C/N$. If $n/N \approx \lambda \in (0,1)$ where n and N are both large, then

$$\hat{\rho} \approx N\left(\rho, \frac{\rho(1-\rho)(1-\lambda)}{n}\right).$$

Hence CI (5.16) should be good if the above normal approximation is good.

Let $\tilde{n} = n + z^2_{1-\delta/2}$ and

$$\tilde{\rho} = \frac{n\hat{\rho} + 0.5z^2_{1-\delta/2}}{n + z^2_{1-\delta/2}}.$$

(Heuristically, the method adds $0.5z^2_{1-\delta/2}$ "0's" and $0.5z^2_{1-\delta/2}$ "1's" to the sample, so the "sample size" increases by $z^2_{1-\delta/2}$.) Then a large sample 100 $(1-\delta)$% Agresti Coull type (ACT) finite population CI for ρ is

$$\tilde{\rho} \pm z_{1-\delta/2}\sqrt{\frac{\tilde{\rho}(1-\tilde{\rho})}{\tilde{n}}\left(\frac{N-n}{N}\right)} = \tilde{\rho} \pm z_{1-\delta/2}SE(\tilde{\rho}). \tag{5.17}$$

Notice that a 95% CI uses $z_{1-\delta/2} = 1.96 \approx 2$.

For data from a finite population, large sample theory gives useful approximations as N and $n \to \infty$ and $n/N \to 0$. Hence theory suggests that the ACT CI should have better coverage than the classical CI if the p is near 0 or 1, if the sample size n is moderate, and if n is small compared to the population size N. The coverage of the classical and ACT CIs should be very similar if n is large enough but small compared to N (which may only be possible if N is enormous). As n increases to N, $\hat{\rho}$ goes to p, SE$(\hat{\rho})$ goes to 0, and the classical CI may perform well. SE$(\tilde{\rho})$ also goes to 0, but $\tilde{\rho}$ is a biased estimator of ρ and the ACT CI will not perform well if n/N is too large.

Want an interval that gives good coverage even if ρ is near 0 or 1 or if n/N is large. A simple method is to combine the two intervals. Let $[L_C, U_C]$ and $[L_A, U_A]$ be the classical and ACT $100(1-\delta)$% intervals. Let the modified $100(1-\delta)$% interval be

$$[\max[0, \min(L_C, L_U)], \min[1, \max(U_C, U_A)]]. \tag{5.18}$$

The modified interval seems to perform well. See Problem 5.4.

Example 5.8. Assume $Y_1, ..., Y_n$ are iid with mean μ and variance σ^2. Bickel and Doksum (2007, p. 279) suggest that

$$W_n = n^{-1/2}\left[\frac{(n-1)S^2}{\sigma^2} - n\right]$$

can be used as an asymptotic pivot for σ^2 if $E(Y^4) < \infty$. Notice that $W_n =$

$$n^{-1/2}\left[\frac{\sum(Y_i-\mu)^2}{\sigma^2} - \frac{n(\overline{Y}-\mu)^2}{\sigma^2} - n\right] =$$

$$\sqrt{n}\left[\frac{\sum\left(\frac{Y_i-\mu}{\sigma}\right)^2}{n} - 1\right] - \frac{1}{\sqrt{n}}\, n\, \left(\frac{\overline{Y}-\mu}{\sigma}\right)^2 = X_n - Z_n.$$

Since $\sqrt{n}Z_n \xrightarrow{D} \chi^2_1$, the term $Z_n \xrightarrow{D} 0$. Now $X_n = \sqrt{n}(\overline{U} - 1) \xrightarrow{D} N(0, \tau)$ by the CLT since $U_i = [(Y_i - \mu)/\sigma]^2$ has mean $E(U_i) = 1$ and variance

$$V(U_i) = \tau = E(U_i^2) - (E(U_i))^2 = \frac{E[(Y_i-\mu)^4]}{\sigma^4} - 1 = \kappa + 2$$

where κ is the kurtosis of Y_i. Thus $W_n \xrightarrow{D} N(0, \tau)$.

Hence

$$1 - \alpha \approx P(-z_{1-\alpha/2} < \frac{W_n}{\sqrt{\tau}} < z_{1-\alpha/2}) = P(-z_{1-\alpha/2}\sqrt{\tau} < W_n < z_{1-\alpha/2}\sqrt{\tau})$$

$$= P(-z_{1-\alpha/2}\sqrt{n\tau} < \frac{(n-1)S^2}{\sigma^2} - n < z_{1-\alpha/2}\sqrt{n\tau})$$

$$= P(n - z_{1-\alpha/2}\sqrt{n\tau} < \frac{(n-1)S^2}{\sigma^2} < n + z_{1-\alpha/2}\sqrt{n\tau}).$$

Hence a large sample $100(1-\alpha)\%$ CI for σ^2 is

$$\left[\frac{(n-1)S^2}{n + z_{1-\alpha/2}\sqrt{n\hat{\tau}}}, \frac{(n-1)S^2}{n - z_{1-\alpha/2}\sqrt{n\hat{\tau}}}\right]$$

where

$$\hat{\tau} = \frac{\frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y})^4}{S^4} - 1.$$

Notice that this CI needs $n > z_{1-\alpha/2}\sqrt{n\hat{\tau}}$ for the right endpoint to be positive. It can be shown that $\sqrt{n}$ (length CI) converges to $2\sigma^2 z_{1-\alpha/2}\sqrt{\tau}$ in probability.

Problem 5.7 uses an asymptotically equivalent $100(1-\alpha)\%$ CI of the form

$$\left[\frac{(n-a)S^2}{n+t_{n-1,1-\alpha/2}\sqrt{n\hat{\tau}}},\ \frac{(n+b)S^2}{n-t_{n-1,1-\alpha/2}\sqrt{n\hat{\tau}}}\right]$$

where a and b depend on $\hat{\tau}$. The goal was to make a 95% CI with good coverage for a wide variety of distributions (with 4th moments) for $n \geq 100$. The price is that the CI is too long for some of the distributions with small kurtosis. The $N(\mu, \sigma^2)$ distribution has $\tau = 2$, while the EXP(λ) distribution has $\sigma^2 = \lambda^2$ and $\tau = 8$. The quantity τ is small for the uniform distribution but large for the lognormal LN(0,1) distribution.

By the binomial theorem, if $E(Y^4)$ exists and $E(Y) = \mu$ then

$$E(Y-\mu)^4 = \sum_{j=0}^{4} \binom{4}{j} E[Y^j](-\mu)^{4-j} =$$

$$\mu^4 - 4\mu^3 E(Y) + 6\mu^2(V(Y) + [E(Y)]^2) - 4\mu E(Y^3) + E(Y^4).$$

This fact can be useful for computing

$$\tau = \frac{E[(Y_i-\mu)^4]}{\sigma^4} - 1 = \kappa + 2.$$

Example 5.9. Following DasGupta (2008, p. 402-404), consider the pooled t CI for $\mu_1 - \mu_2$. Let $X_1, ..., X_{n_1}$ be iid with mean μ_1 and variance σ_1^2. Let $Y_1, ..., Y_{n_2}$ be iid with mean μ_2 and variance σ_2^2. Assume that the two samples are independent and that $n_i \to \infty$ for $i = 1, 2$ in such a way that $\hat{\rho} = \hat{\pi}_1 = \frac{n_1}{n_1+n_2} \to \rho = \pi_1 \in (0,1)$. Let $n = n_1 + n_2$ and let $\hat{\pi}_2 = n_2/n = 1 - \hat{\pi}_1$. Let $\theta = \sigma_2^2/\sigma_1^2$, and let the pooled sample variance

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}.$$

Then

$$\begin{pmatrix} \sqrt{n_1}(\overline{X}-\mu_1) \\ \sqrt{n_2}(\overline{Y}-\mu_2) \end{pmatrix} \xrightarrow{D} N_2(\mathbf{0}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \sigma_2^2)$. Hence

$$\sqrt{n}[(\overline{X}-\overline{Y}) - (\mu_1-\mu_2)] \xrightarrow{D} N(0, \frac{\sigma_1^2}{\pi_1} + \frac{\sigma_2^2}{\pi_2}).$$

So

$$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \xrightarrow{D} N(0,1).$$

Thus

$$\frac{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\ \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}\ =\ \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\xrightarrow{D} N(0,\tau^2)$$

where

$$\frac{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}{(\frac{1}{n_1}+\frac{1}{n_2})\frac{n_1\sigma_1^2+n_2\sigma_2^2}{n_1+n_2}}=\frac{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}{\hat{\rho}\sigma_1^2+(1-\hat{\rho})\sigma_2^2}\ \frac{1/\sigma_1^2}{1/\sigma_1^2}\ \frac{n_1n_2}{n_1+n_2}$$

$$=\frac{\frac{1}{n_1}+\frac{\theta}{n_2}}{\hat{\rho}+(1-\hat{\rho})\theta}\ \frac{n_1n_2}{n_1+n_2}\xrightarrow{D}\frac{1-\rho+\rho\theta}{\rho+(1-\rho)\theta}=\tau^2.$$

Now let $\hat{\theta}=S_2^2/S_1^2$ and

$$\hat{\tau}^2=\frac{1-\hat{\rho}+\hat{\rho}\ \hat{\theta}}{\hat{\rho}+(1-\hat{\rho})\ \hat{\theta}}.$$

Notice that $\hat{\tau}=1$ if $\hat{\rho}=1/2$, and $\hat{\tau}=1$ if $\hat{\theta}=1$. Thus the following pooled t CI often performs well if $n_1/n_2\approx 1$.

The usual large sample $(1-\alpha)100\%$ pooled t CI for $(\mu_1-\mu_2)$ is

$$\overline{X}-\overline{Y}\ \pm\ t_{n_1+n_2-2,1-\alpha/2}\ \ S_p\ \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}. \tag{5.19}$$

The large sample theory says that this CI is valid if $\tau=1$, and that

$$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\hat{\tau}\ S_p\ \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\xrightarrow{D} N(0,1).$$

Hence a large sample $(1-\alpha)100\%$ CI for $(\mu_1-\mu_2)$ is

$$\overline{X}-\overline{Y}\ \pm\ z_{1-\alpha/2}\ \hat{\tau}\ S_p\ \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$$

Then the large sample $(1-\alpha)100\%$ modified pooled t CI for $(\mu_1-\mu_2)$ is

$$\overline{X}-\overline{Y}\ \pm\ t_{n_1+n_2-4,1-\alpha/2}\ \hat{\tau}\ S_p\ \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}. \tag{5.20}$$

The large sample $(1-\alpha)100\%$ Welch CI for $(\mu_1-\mu_2)$ is

$$\overline{X}-\overline{Y}\ \pm\ t_{d,1-\alpha/2}\ \sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} \tag{5.21}$$

where $d=\max(1,\lfloor d_0\rfloor)$, and

$$d_0 = \frac{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})^2}{\frac{1}{n_1-1}(\frac{S_1^2}{n_1})^2 + \frac{1}{n_2-1}(\frac{S_2^2}{n_2})^2}.$$

Suppose $n_1/(n_1+n_2) \to \rho$. It can be shown that if the CI length is multiplied by $\sqrt{n_1}$, then the scaled length of the pooled t CI converges in probability to $2z_{1-\alpha/2}\sqrt{\frac{\rho}{1-\rho}\sigma_1^2 + \sigma_2^2}$ while the scaled lengths of the modified pooled t CI and Welch CI both converge in probability to $2z_{1-\alpha/2}\sqrt{\sigma_1^2 + \frac{\rho}{1-\rho}\sigma_2^2}$.

Results from Olive et al. (2024) can also be used to derive a CI for σ^2.

Example 5.10. Hesterberg (2014) gives the following two competitors of the t interval given by Equation (5.2): the skewness adjusted t interval is

$$[\,\overline{Y}+\frac{S}{\sqrt{n}}[\hat{\kappa}(1+2t_{n-1,1-\alpha/2}^2)-t_{n-1,1-\alpha/2}],\ \overline{Y}+\frac{S}{\sqrt{n}}[\hat{\kappa}(1+2t_{n-1,1-\alpha/2}^2)+t_{n-1,1-\alpha/2}]\,], \tag{5.22}$$

and the asymptotic percentile t CI is

$$[\,\overline{Y}+\frac{S}{\sqrt{n}}[\hat{\kappa}(t_{n-1,1-\alpha/2}-1)^2-t_{n-1,1-\alpha/2}],\ \overline{Y}+\frac{S}{\sqrt{n}}[\hat{\kappa}(t_{n-1,1-\alpha/2}-1)^2+t_{n-1,1-\alpha/2}] \tag{5.23}$$

where

$$\hat{\kappa} = \frac{\hat{\gamma}}{6\sqrt{n}} \ \text{ with } \ \hat{\gamma} = \frac{1}{\mathrm{nS^3}}\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{Y_i} - \overline{\mathrm{Y}})^3.$$

Another competitor is the Johnson (1978) CI is

$$[\,\overline{Y} + \frac{\hat{\mu}_3}{6S^2n} - t_{n-1,1-\alpha/2}\ \ S/\sqrt{n},\ \ \overline{Y} + \frac{\hat{\mu}_3}{6S^2n} + t_{n-1,1-\alpha/2}\ \ S/\sqrt{n}\,] \tag{5.24}$$

where $\mu_3 = E[(Y-\mu)^3]$ and

$$\hat{\mu}_3 = S^3\hat{\gamma} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^3.$$

The t–interval (5.2) may perform better if the distribution has second moments but does not have third or fourth moments. McKinney (2021) gave some more competitors. The Johnson (1978) CI (5.24) appeared to be best, but only very slightly better than the usual t– interval (5.2).

5.4 Bootstrap Confidence Regions and Hypothesis Tests

This section shows that, under regularity conditions, applying the nonparametric prediction region of Section 4.2 to a bootstrap sample results in a

confidence region. The volume of a confidence region $\to 0$ as $n \to 0$, while the volume of a prediction region goes to that of a population region that would contain a new $\boldsymbol{x}_f$ with probability $1-\delta$. The nominal coverage is $100(1-\delta)$. If the actual coverage $100(1-\delta_n) > 100(1-\delta)$, then the region is *conservative*. If $100(1-\delta_n) < 100(1-\delta)$, then the region is *liberal*. A region that is 5% conservative is considered "much better" than a region that is 5% liberal.

When teaching confidence intervals, it is often noted that by the central limit theorem, the probability that $\overline{Y}_n$ is within two standard deviations $(2SD(\overline{Y}_n) = 2\sigma/\sqrt{n})$ of $\theta = \mu$ is about 95%. Hence the probability that θ is within two standard deviations of $\overline{Y}_n$ is about 95%. Thus the interval $[\theta - 1.96S/\sqrt{n}, \theta + 1.96S/\sqrt{n}]$ is a large sample 95% prediction interval for a future value of the sample mean $\overline{Y}_{n,f}$ if θ is known, while $[\overline{Y}_n - 1.96S/\sqrt{n}, \overline{Y}_n + 1.96S/\sqrt{n}]$ is a large sample 95% confidence interval for the population mean θ. Note that the lengths of the two intervals are the same. Where the interval is centered, at the parameter θ or the statistic $\overline{Y}_n$, determines whether the interval is a prediction or a confidence interval. See Theorems 5.2 and 5.3 for a similar relationship between confidence regions and prediction regions. Let $\boldsymbol{\theta}$ be a $g \times 1$ vector of parameters.

Definition 5.5. A *large sample* $100(1-\delta)\%$ *confidence region* for a vector of parameters $\boldsymbol{\theta}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$.

If $\mathcal{A}_n$ is based on a squared Mahalanobis distance D^2 with a limiting distribution that has a pdf, we often want $P(\boldsymbol{\theta} \in \mathcal{A}_n) \to 1-\delta$ as $n \to \infty$.

There are several methods for obtaining a bootstrap sample $T_1^*, ..., T_B^*$ where the sample size n is suppressed: $T_i^* = T_{in}^*$. The parametric bootstrap, nonparametric bootstrap, and residual bootstrap will be used. Applying the nonparametric prediction region (4.11) to the bootstrap sample will result in a confidence region for $\boldsymbol{\theta}$. When $g = 1$, applying the percentile PI (4.1) or the shorth PI (4.4) to the bootstrap sample results in a confidence interval for θ. Section 5.4.2 will help clarify ideas.

When $g = 1$, a confidence interval is a special case of a confidence region. One sided confidence intervals give a lower or upper confidence bound for θ. A large sample $100(1-\delta)\%$ lower confidence interval $(-\infty, U_n]$ uses an upper confidence bound U_n and is in the lower tail of the distribution of $\hat{\theta}$. A large sample $100(1-\delta)\%$ upper confidence interval $[L_n, \infty)$ uses a lower confidence bound L_n and is in the upper tail of the distribution of $\hat{\theta}$. These CIs can be useful if $\theta \in [a, b]$ and $\theta = a$ or $\theta = b$ is of interest for a hypothesis test. For example, $[a, b] = [0, 1]$ if $\theta = \rho^2$, the squared population correlation. Then use $[0, U_n]$ and $[L_n, 1]$ as CIs, e.g. if we expect $\theta = 0$ we might test $H_0 : \theta \leq 0.05$ versus $H_0 : \theta > 0.05$, and fail to reject H_0 if $U_n < 0.05$. Again we often want

the probability to converge to $1-\delta$ if the confidence interval is based on a statistic with an asymptotic distribution that has a pdf.

Definition 5.6. The interval $[L_n, U_n]$ is a large sample $100(1-\delta)\%$ *confidence interval* for θ if $P(L_n \leq \theta \leq U_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$. The interval $(-\infty, U_n]$ is a large sample $100(1-\delta)\%$ *lower confidence interval* for θ if $P(\theta \leq U_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$. The interval $[L_n, \infty)$ is large sample $100(1-\delta)\%$ *upper confidence interval* for θ if $P(\theta \geq L_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$.

Next we discuss bootstrap confidence intervals that are obtained by applying prediction intervals (4.1) and (4.4) to the bootstrap sample. Some additional bootstrap CIs are given in Definition 5.16 and are obtained from three bootstrap confidence regions when $g = 1$. See Efron (1982) and Chen (2016) for the percentile method CI. Let T_n be an estimator of a parameter θ such as $T_n = \overline{Z} = \sum_{i=1}^{n} Z_i/n$ with $\theta = E(Z_1)$. Let $T_1^*, ..., T_B^*$ be a bootstrap sample for T_n. Let $T_{(1)}^*, ..., T_{(B)}^*$ be the order statistics of the the bootstrap sample. The percentile CI (5.25) is obtained by applying percentile PI (4.1) to the bootstrap sample with B used instead of n. Hence (5.25) is also a large sample prediction interval for a future value of T_f^* if the T_i^* are iid from the empirical distribution discussed in Section 5.4.1.

Definition 5.7. The bootstrap large sample $100(1-\delta)\%$ *percentile* confidence interval for θ is an interval $[T_{(k_L)}^*, T_{(K_U)}^*]$ containing $\approx \lceil B(1-\delta) \rceil$ of the T_i^*. Let $k_1 = \lceil B\delta/2 \rceil$ and $k_2 = \lceil B(1-\delta/2) \rceil$. A common choice is

$$[T_{(k_1)}^*, T_{(k_2)}^*]. \tag{5.25}$$

The large sample $100(1-\delta)\%$ *lower percentile* CI for θ is $(-\infty, T_{(\lceil B(1-\delta) \rceil)}^*]$. The large sample $100(1-\delta)\%$ *upper percentile* CI for θ is $[T_{(\lceil B\delta \rceil)}^*, \infty)$.

In the next definition, the large sample $100(1-\delta)\%$ *shorth*(c) *CI* uses the interval $[T_{(1)}^*, T_{(c)}^*], [T_{(2)}^*, T_{(c+1)}^*], ..., [T_{(B-c+1)}^*, T_{(B)}^*]$ of shortest length, denoted by $[T_{(s)}^*, T_{(s+c-1)}^*]$. The shorth$(c)$ CI (5.26) is obtained by applying the shorth(c) PI (4.4) on the bootstrap sample.

Definition 5.8. The large sample $100(1-\delta)\%$ *lower shorth* CI for θ is $(-\infty, T_{(c)}^*]$, while the large sample $100(1-\delta)\%$ *upper shorth* CI for θ is $[T_{(B-c+1)}^*, \infty)$. The large sample $100(1-\delta)\%$ *shorth*(c) *CI*

$$[T_{(s)}^*, T_{(s+c-1)}^*] \text{ where } \mathrm{c} = \min(\mathrm{B}, \lceil \mathrm{B}[1-\delta+1.12\sqrt{\delta/\mathrm{B}}\ ]\ \rceil). \tag{5.26}$$

Applied to a bootstrap sample, the shorth CI can be regarded as the shortest percentile method confidence interval, asymptotically. Hence the shorth confidence interval is a practical implementation of the Hall (1988) shortest

bootstrap interval based on all possible bootstrap samples. See Remark 5.8 for some theory for bootstrap CIs such as (5.25) and (5.26).

5.4.1 The Bootstrap

This subsection illustrates the nonparametric bootstrap with some examples. Suppose a statistic T_n is computed from a data set of n cases. The nonparametric bootstrap draws n cases with replacement from that data set. Then T_1^* is the statistic T_n computed from the sample. This process is repeated B times to produce the bootstrap sample $T_1^*, ..., T_B^*$. Sampling cases with replacement uses the empirical distribution.

Definition 5.9. Suppose that data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ has been collected and observed. Often the data is a random sample (iid) from a distribution with cdf F. The *empirical distribution* is a discrete distribution where the $\boldsymbol{x}_i$ are the possible values, and each value is equally likely. If $\boldsymbol{w}$ is a random variable having the empirical distribution, then $p_i = P(\boldsymbol{w} = \boldsymbol{x}_i) = 1/n$ for $i = 1, ..., n$. The *cdf of the empirical distribution* is denoted by F_n.

Example 5.11. Let $\boldsymbol{w}$ be a random variable having the empirical distribution given by Definition 5.9. Show that $E(\boldsymbol{w}) = \overline{\boldsymbol{x}} \equiv \overline{\boldsymbol{x}}_n$ and $\text{Cov}(\boldsymbol{w}) = \frac{n-1}{n}\boldsymbol{S} \equiv \frac{n-1}{n}\boldsymbol{S}_n$.

Solution: Recall that for a discrete random vector, the population expected value $E(\boldsymbol{w}) = \sum \boldsymbol{x}_i p_i$ where $\boldsymbol{x}_i$ are the values that $\boldsymbol{w}$ takes with positive probability p_i. Similarly, the population covariance matrix

$$\text{Cov}(\boldsymbol{w}) = E[(\boldsymbol{w} - E(\boldsymbol{w}))(\boldsymbol{w} - E(\boldsymbol{w}))^T] = \sum (\boldsymbol{x}_i - E(\boldsymbol{w}))(\boldsymbol{x}_i - E(\boldsymbol{w}))^T p_i.$$

Hence

$$E(\boldsymbol{w}) = \sum_{i=1}^{n} \boldsymbol{x}_i \frac{1}{n} = \overline{\boldsymbol{x}},$$

and

$$\text{Cov}(\boldsymbol{w}) = \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \frac{1}{n} = \frac{n-1}{n}\boldsymbol{S}. \quad \square$$

Example 5.12. If $W_1, ..., W_n$ are iid from a distribution with cdf F_W, then the empirical cdf F_n corresponding to F_W is given by

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} I(W_i \leq y)$$

where the indicator $I(W_i \leq y) = 1$ if $W_i \leq y$ and $I(W_i \leq y) = 0$ if $W_i > y$. Fix n and y. Then $nF_n(y) \sim$ binomial $(n, F_W(y))$. Thus $E[F_n(y)] = F_W(y)$ and $V[F_n(y)] = F_W(y)[1 - F_W(y)]/n$. By the central limit theorem,

$$\sqrt{n}(F_n(y) - F_W(y)) \xrightarrow{D} N(0, F_W(y)[1 - F_W(y)]).$$

Thus $F_n(y) - F_W(y) = O_P(n^{-1/2})$, and F_n is a reasonable estimator of F_W if the sample size n is large.

Suppose there is data $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ collected into an $n \times p$ matrix $\boldsymbol{W}$. Let the statistic $T_n = t(\boldsymbol{W}) = T(F_n)$ be computed from the data. Suppose the statistic estimates $\boldsymbol{\mu} = T(F)$, and let $t(\boldsymbol{W}^*) = t(F_n^*) = T_n^*$ indicate that t was computed from an iid sample from the empirical distribution F_n: a sample $\boldsymbol{w}_1^*, ..., \boldsymbol{w}_n^*$ of size n was drawn with replacement from the observed sample $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. This notation is used for von Mises differentiable statistical functions in large sample theory. See Serfling (1980, ch. 6). The empirical distribution is also important for the influence function (widely used in robust statistics). The *nonparametric bootstrap* draws B samples of size n from the rows of $\boldsymbol{W}$, e.g. from the empirical distribution of $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. Then T_{jn}^* is computed from the jth bootstrap sample for $j = 1, ..., B$.

Example 5.13. Suppose the data is 1, 2, 3, 4, 5, 6, 7. Then $n = 7$ and the sample median T_n is 4. Using R, we drew $B = 2$ bootstrap samples (samples of size n drawn with replacement from the original data) and computed the sample median $T_{1,n}^* = 3$ and $T_{2,n}^* = 4$.

```
b1 <- sample(1:7,replace=T)
b1
[1] 3 2 3 2 5 2 6
median(b1)
[1] 3
b2 <- sample(1:7,replace=T)
b2
[1] 3 5 3 4 3 5 7
median(b2)
[1] 4
```

The bootstrap has been widely used to estimate the population covariance matrix of the statistic $\text{Cov}(T_n)$, for testing hypotheses, and for obtaining confidence regions (often confidence intervals). An iid sample $T_{1n}, ..., T_{Bn}$ of size B of the statistic would be very useful for inference, but typically we only have one sample of data and one value $T_n = T_{1n}$ of the statistic. Often $T_n = t(\boldsymbol{w}_1, ..., \boldsymbol{w}_n)$, and the bootstrap sample $T_{1n}^*, ..., T_{Bn}^*$ is formed where $T_{jn}^* = t(\boldsymbol{w}_{j1}^*, ..., \boldsymbol{w}_{jn}^*)$. Theorem 5.1 will show that $\sqrt{B}(T_{1n}^* - T_n), ..., \sqrt{B}(T_{Bn}^* - T_n)$ is pseudodata for $\sqrt{n}(T_{1n} - \boldsymbol{\theta}), ..., \sqrt{n}(T_{Bn} - \boldsymbol{\theta})$ when n and B are large in that $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ and $\sqrt{B}(T^* - T_n) \xrightarrow{D} \boldsymbol{u}$.

Example 5.14. Suppose there is training data $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ for the model $\boldsymbol{y}_i = m(\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$, and it is desired to predict a future test value $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ and the training data. The model can be fit and the residual vectors $\hat{\boldsymbol{\epsilon}}_i$ computed for $i = 1, ..., n$. One method for obtaining a prediction region for $\boldsymbol{y}_f$ is to form the pseudodata $\hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ for $i = 1, ..., n$, and apply the nonparametric prediction region (4.11) to the pseudodata. See Olive (2017b, 2018). The residual bootstrap could also be used to make a bootstrap sample $\hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_1^*, ..., \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_B^*$ where the $\hat{\boldsymbol{\epsilon}}_j^*$ are selected with replacement from the residual vectors $\hat{\boldsymbol{\epsilon}}_i$ for $j = 1, ..., B$. As $B \to \infty$, the bootstrap sample will take on the n values $\hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ (the pseudodata) with probabilities converging to $1/n$ for $i = 1, ..., n$.

Suppose there is a statistic T_n that is a $g \times 1$ vector. Let

$$\overline{T}^* = \frac{1}{B}\sum_{i=1}^{B} T_i^* \quad \text{and} \quad \boldsymbol{S}_T^* = \frac{1}{B-1}\sum_{i=1}^{B}(T_i^* - \overline{T}^*)(T_i^* - \overline{T}^*)^T \tag{5.27}$$

be the sample mean and sample covariance matrix of the bootstrap sample $T_1^*, ..., T_B^*$ where $T_i^* = T_{i,n}^*$. Fix n, and let $E(T_{i,n}^*) = \boldsymbol{\theta}_n$ and $\mathrm{Cov}(T_{i,n}^*) = \boldsymbol{\Sigma}_n$.

We will often assume that $\mathrm{Cov}(T_n) = \boldsymbol{\Sigma}_T$, and $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$ where $\boldsymbol{\Sigma}_A > 0$ is positive definite and nonsingular. Often $n\hat{\boldsymbol{\Sigma}}_T \xrightarrow{P} \boldsymbol{\Sigma}_A$. For example, using least squares and the residual bootstrap for the multiple linear regression model, $\boldsymbol{\Sigma}_n = \dfrac{n-p}{n} MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, $T_n = \boldsymbol{\theta}_n = \hat{\boldsymbol{\beta}}$, $\boldsymbol{\theta} = \boldsymbol{\beta}$, $\hat{\boldsymbol{\Sigma}}_T = MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, and $\boldsymbol{\Sigma}_A = \sigma^2 \lim_{n\to\infty}(\boldsymbol{X}^T\boldsymbol{X}/n)^{-1}$.

Suppose the $T_i^* = T_{i,n}^*$ are iid from some distribution with cdf $\tilde{F}_n$. For example, if $T_{i,n}^* = t(F_n^*)$ where iid samples from F_n are used, then $\tilde{F}_n$ is the cdf of $t(F_n^*)$. With respect to $\tilde{F}_n$, both $\boldsymbol{\theta}_n$ and $\boldsymbol{\Sigma}_n$ are parameters, but with respect to F, $\boldsymbol{\theta}_n$ is a random vector and $\boldsymbol{\Sigma}_n$ is a random matrix. For fixed n, by the multivariate central limit theorem,

$$\sqrt{B}(\overline{T}^* - \boldsymbol{\theta}_n) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_n) \quad \text{and} \quad \mathrm{B}(\overline{\mathrm{T}}^* - \boldsymbol{\theta}_\mathrm{n})^\mathrm{T}[\boldsymbol{S}_\mathrm{T}^*]^{-1}(\overline{\mathrm{T}}^* - \boldsymbol{\theta}_\mathrm{n}) \xrightarrow{\mathrm{D}} \chi^2_\mathrm{r}$$

as $B \to \infty$.

Remark 5.4. For Examples 5.11 and 5.14, the bootstrap works but is expensive compared to alternative methods. For Example 5.11, fix n, then $\overline{T}^* \xrightarrow{P} \boldsymbol{\theta}_n = \overline{\boldsymbol{x}}$ and $\boldsymbol{S}_T^* \xrightarrow{P} (n-1)\boldsymbol{S}/n$ as $B \to \infty$, but using $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ makes more sense. For Example 5.14, use the pseudodata instead of the residual bootstrap. For these two examples, it is known how the bootstrap sample behaves as $B \to \infty$. The bootstrap can be very useful when $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$, but it not known how to estimate $\boldsymbol{\Sigma}_A$ without using a resampling method like the bootstrap. The bootstrap may be useful when $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, but the limiting distribution (the distribution of $\boldsymbol{u}$) is unknown.

The following theorem shows that $\sqrt{m}(T^*_{1,n} - T_n), ..., \sqrt{m}(T^*_{B,n} - T_n)$ are pseudodata for $\sqrt{n}(T_{1,n} - \boldsymbol{\theta}), ..., \sqrt{n}(T_{B,n} - \boldsymbol{\theta})$. Here $T^*_i = T^*_{i,m}$ with n suppressed or $T^*_{i,n} = T^*_{i,n,m}$ where m is the sample size of the bootstrap data set used to compute T^*_i. Often $m = n$ for the nonparametric bootstrap. The first two convergence assumptions are with respect to the data distribution, while the third convergence assumption is with respect to the bootstrap distribution. The technique is similar to using a triangular array, except both $n \to \infty$ and $m \to \infty$. Note that for large n, $N_g(\mathbf{0}, \boldsymbol{\Sigma}_n) \approx N_g(\mathbf{0}, \boldsymbol{\Sigma})$, and often the $N_g(\mathbf{0}, \boldsymbol{\Sigma}_n)$ approximation is used to produce output since $\boldsymbol{\Sigma}$ is unknown. Typically large sample theory is used to prove the three assumptions of the following theorem.

Theorem 5.1, Bootstrap Proof Technique: Suppose $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}_n \xrightarrow{P} \boldsymbol{\Sigma}$ as $n \to \infty$, and for fixed n, $\sqrt{m}(T^*_{n,m} - T_n) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma}_n)$ as $m \to \infty$. Then a) $\sqrt{m}(T^*_{n,m} - T_n) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma})$ as $m, n \to \infty$. Also b) $\sqrt{n}(T^*_n - T_n) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma})$ as $n \to \infty$ where $T^*_n = T^*_{n,n}$ has $m = n$.

Proof: By the three assumptions, $\boldsymbol{u}_n = \sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u} \sim N_g(\mathbf{0}, \boldsymbol{\Sigma})$ as $n \to \infty$, $\boldsymbol{w}^*_{n,m} = \sqrt{m}(T^*_{n,m} - T_n) \xrightarrow{D} \boldsymbol{w}_n \sim N_g(\mathbf{0}, \boldsymbol{\Sigma}_n)$ as $m \to \infty$ for fixed n, and $\boldsymbol{w}_n \xrightarrow{D} \boldsymbol{u}$ as $n \to \infty$. Hence $\boldsymbol{w}^*_{n,m} = \sqrt{m}(T^*_{n,m} - T_n) \xrightarrow{D} \boldsymbol{u} \sim N_g(\mathbf{0}, \boldsymbol{\Sigma})$ as $m, n \to \infty$. Since this result does not depend on m as long as $m \to \infty$, b) follows. (Interpret $\boldsymbol{w}_n \sim N_g(\mathbf{0}, \boldsymbol{\Sigma}_n)$ as $\boldsymbol{w}_n = \boldsymbol{\Sigma}_n^{1/2} N_g(\mathbf{0}, \boldsymbol{I}_g)$.) □

Example 5.15. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $p \times 1$ random vectors with $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{x}_i) = \boldsymbol{\Sigma}$. a) For the parametric bootstrap, let $\boldsymbol{x}^*_1, ..., \boldsymbol{x}^*_m$ be iid $N_p(\overline{\boldsymbol{x}}_n, \boldsymbol{S}_n)$ where $\boldsymbol{S}_n \xrightarrow{P} \boldsymbol{\Sigma}$ as $n \to \infty$. By the multivariate central limit theorem $\sqrt{n}(\overline{\boldsymbol{x}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma})$ and for fixed n, $\sqrt{m}(\overline{\boldsymbol{x}}^*_{n,m} - \overline{\boldsymbol{x}}_n) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{S}_n)$ where $\overline{\boldsymbol{x}}^*_{n,m} = \frac{1}{m}\sum_{i=1}^m \boldsymbol{x}^*_i$ is the sample mean of the bootstrap data set $\boldsymbol{x}^*_1, ..., \boldsymbol{x}^*_m$. Hence $\sqrt{m}(\overline{\boldsymbol{x}}^*_{n,m} - \overline{\boldsymbol{x}}_n) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma})$ as $n, m \to \infty$ by Theorem 5.1. Note that $m = n$ can be used by Theorem 5.1 b).

b) For the nonparametric bootstrap, $E(\overline{\boldsymbol{x}}^*_n) = E(\boldsymbol{w}_n) = \overline{\boldsymbol{x}}_n$, and $\mathrm{Cov}(\overline{\boldsymbol{x}}^*_n) = \mathrm{Cov}(\boldsymbol{w}_n)/n = (n-1)\boldsymbol{S}_n/n^2$ by Example 5.11 where $\boldsymbol{w} = \boldsymbol{w}_n$. The $\boldsymbol{x}^*_i$ are iid with respect to the bootstrap distribution. If the sample mean $\overline{\boldsymbol{x}}^*_{n,m}$ is computed from m $\boldsymbol{x}^*_i$ selected with replacement from the $\boldsymbol{x}_i$, then $\sqrt{m}(\overline{\boldsymbol{x}}^*_{n,m} - \overline{\boldsymbol{x}}_n) \xrightarrow{D} N_p(\mathbf{0}, \frac{n-1}{n}\boldsymbol{S}_n)$ for fixed n by the multivariate CLT. Then by Theorem 5.1 b) with $m = n$, $\sqrt{n}(\overline{\boldsymbol{x}}^*_n - \overline{\boldsymbol{x}}_n) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma})$ as $n \to \infty$.

5.4.2 Bootstrap Confidence Regions for Hypothesis Testing

When the bootstrap is used, a large sample $100(1-\delta)\%$ confidence region for a $g \times 1$ parameter vector $\boldsymbol{\theta}$ is a set $\mathcal{A}_n = \mathcal{A}_{n,B}$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_{n,B})$ is eventually bounded below by $1-\delta$ as $n, B \to \infty$. The B is often suppressed. Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a known $g \times 1$ vector. Then reject H_0 if $\boldsymbol{\theta}_0$ is not in the confidence region $\mathcal{A}_n$. Let the $g \times 1$ vector T_n be an estimator of $\boldsymbol{\theta}$. Let $T_1^*, ..., T_B^*$ be the bootstrap sample for T_n. Let $\boldsymbol{A}$ be a full rank $g \times p$ constant matrix. For variable selection using notation from Chapter 6, consider testing $H_0 : \boldsymbol{A\beta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{A\beta} \neq \boldsymbol{\theta}_0$ with $\boldsymbol{\theta} = \boldsymbol{A\beta}$ where often $\boldsymbol{\theta}_0 = \mathbf{0}$. Then let $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ and let $T_i^* = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{I_{min},0,i}$ for $i = 1, ..., B$. The statistic $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is the variable selection estimator padded with zeroes.

Let $\overline{T}^*$ and $\boldsymbol{S}_T^*$ be the sample mean and sample covariance matrix of the bootstrap sample $T_1^*, ..., T_B^*$. See Equation (5.27). Here $P(X \leq \chi^2_{g,1-\delta}) = 1-\delta$ if $X \sim \chi^2_g$, and $P(X \leq F_{g,d_n,1-\delta}) = 1-\delta$ if $X \sim F_{g,d_n}$. Let $k_B = \lceil B(1-\delta) \rceil$.

Definition 5.10. a) The large sample $100(1-\delta)\%$ *standard bootstrap confidence region* for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{1-\delta}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{1-\delta}\} \tag{5.28}$$

where $D^2_{1-\delta} = \chi^2_{g,1-\delta}$ or $D^2_{1-\delta} = d_n F_{g,d_n,1-\delta}$ where $d_n \to \infty$ as $n \to \infty$.
b) The large sample $100(1-\delta)\%$ *Bickel and Ren confidence region* for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\hat{\boldsymbol{\Sigma}}_A/n]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(k_{BT})}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \hat{\boldsymbol{\Sigma}}_A/n) \leq D^2_{(k_{BT})}\} \tag{5.29}$$

where the cutoff $D^2_{(k_{BT})}$ is the $100k_B$th sample quantile of the
$D_i^2 = (T_i^* - T_n)^T [\hat{\boldsymbol{\Sigma}}_A/n]^{-1} (T_i^* - T_n) = n(T_i^* - T_n)^T [\hat{\boldsymbol{\Sigma}}_A]^{-1} (T_i^* - T_n)$.

Confidence region (5.28) needs $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma}_A)$ and $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{\Sigma}_A > 0$ as $n, B \to \infty$. See Machado and Parente (2005) for regularity conditions for this assumption. Bickel and Ren (2001) have interesting sufficient conditions for (5.29) to be a confidence region when $\hat{\boldsymbol{\Sigma}}_A$ is a consistent estimator of positive definite $\boldsymbol{\Sigma}_A$. Let the vector of parameters $\boldsymbol{\theta} = T(F)$, the statistic $T_n = T(F_n)$, and the bootstrapped statistic $T^* = T(F_n^*)$ where F is the cdf of iid $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, F_n is the empirical cdf, and F_n^* is the empirical cdf of $\boldsymbol{x}_1^*, ..., \boldsymbol{x}_n^*$, a sample from F_n using the nonparametric bootstrap. If $\sqrt{n}(F_n - F) \xrightarrow{D} \boldsymbol{z}_F$, a Gaussian random process, and if T is sufficiently smooth (has a Hadamard derivative $\dot{T}(F)$), then $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ and

$\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$ with $\boldsymbol{u} = \dot{T}(F)\boldsymbol{z}_F$. Note that F_n is a perfectly good cdf "F" and F_n^* is a perfectly good empirical cdf from $F_n =$ "F." Thus if n is fixed, and a sample of size m is drawn with replacement from the empirical distribution, then $\sqrt{m}(T(F_m^*) - T_n) \xrightarrow{D} \dot{T}(F_n)\boldsymbol{z}_{F_n}$. Now let $n \to \infty$ with $m = n$. Then bootstrap theory gives $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \lim_{n\to\infty} \dot{T}(F_n)\boldsymbol{z}_{F_n} = \dot{T}(F)\boldsymbol{z}_F \sim \boldsymbol{u}$.

The following three confidence regions will be used for inference after variable selection. The Olive (2017ab, 2018) prediction region method confidence region applies the nonparametric prediction region (4.11) to the bootstrap sample. Olive (2017ab, 2018) also gave the modified Bickel and Ren confidence region that uses $\hat{\boldsymbol{\Sigma}}_A = n\boldsymbol{S}_T^*$. The hybrid confidence region is due to Pelawa Watagoda and Olive (2021a). Let $q_B = \min(1 - \delta + 0.05, 1 - \delta + g/B)$ for $\delta > 0.1$ and

$$q_B = \min(1 - \delta/2, 1 - \delta + 10\delta g/B), \quad \text{otherwise.} \tag{5.30}$$

If $1 - \delta < 0.999$ and $q_B < 1 - \delta + 0.001$, set $q_B = 1 - \delta$. Let $D_{(U_B)}$ be the $100q_B$th sample quantile of the D_i. Use (5.30) as a correction factor for finite $B \geq 50g$.

Definition 5.11. The large sample $100(1-\delta)\%$ *prediction region method confidence region* for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - \overline{T}^*) \leq D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\} \tag{5.31}$$

where $D^2_{(U_B)}$ is computed from $D_i^2 = (T_i^* - \overline{T}^*)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - \overline{T}^*)$ for $i = 1, ..., B$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(\overline{T}^* - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (\overline{T}^* - \boldsymbol{\theta}_0) > D^2_{(U_B)}$. (This procedure is basically the one sample Hotelling's T^2 test applied to the T_i^* using $\boldsymbol{S}_T^*$ as the estimated covariance matrix and replacing the $\chi^2_{g,1-\delta}$ cutoff by $D^2_{(U_B)}$.)

Definition 5.12. The large sample $100(1-\delta)\%$ *(modified) Bickel and Ren confidence region* is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_{BT})}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_{BT})}\} \tag{5.32}$$

where the cutoff $D^2_{(U_{BT})}$ is the $100q_B$th sample quantile of the $D_i^2 = (T_i^* - T_n)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - T_n)$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(T_n - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (T_n - \boldsymbol{\theta}_0) > D^2_{(U_B,T)}$.

Definition 5.13. Shift region (5.31) to have center T_n, or equivalently, change the cutoff of region (5.32) to $D^2_{(U_B)}$ to get the large sample $100(1-\delta)\%$ *hybrid confidence region*: $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\}. \tag{5.33}$$

Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(T_n - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (T_n - \boldsymbol{\theta}_0) > D^2_{(U_B)}$.

Rajapaksha and Olive (2024) gave the following two confidence regions. The names of these confidence regions were chosen since they are similar to the Bickel and Ren and prediction region method confidence regions.

Definition 5.14. The large sample $100(1-\delta)\%$ *BR confidence region* is

$$\{\boldsymbol{w} : n(\boldsymbol{w} - T_n)^T \boldsymbol{C}_n^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_{BT})}\} =$$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{C}_n/n) \leq D^2_{(U_{BT})}\} \tag{5.34}$$

where the cutoff $D^2_{(U_{BT})}$ is the $100q_B$th sample quantile of the $D_i^2 = n(T_i^* - T_n)^T \boldsymbol{C}_n^{-1} (T_i^* - T_n)$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $n(T_n - \boldsymbol{\theta}_0)^T \boldsymbol{C}_n^{-1} (T_n - \boldsymbol{\theta}_0) > D^2_{(U_{BT})}$.

Definition 5.15. The large sample $100(1-\delta)\%$ *PR confidence region* for $\boldsymbol{\theta}$ is

$$\{\boldsymbol{w} : n(\boldsymbol{w} - \overline{T}^*)^T \boldsymbol{C}_n^{-1} (\boldsymbol{w} - \overline{T}^*) \leq D^2_{(U_B)}\} =$$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{C}_n/n) \leq D^2_{(U_B)}\} \tag{5.35}$$

where $D^2_{(U_B)}$ is computed from $D_i^2 = n(T_i^* - \overline{T}^*)^T \boldsymbol{C}_n^{-1} (T_i^* - \overline{T}^*)$ for $i = 1, ..., B$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $n(\overline{T}^* - \boldsymbol{\theta}_0)^T \boldsymbol{C}_n^{-1} (\overline{T}^* - \boldsymbol{\theta}_0) > D^2_{(U_B)}$.

Hyperellipsoids (5.31) and (5.33) have the same volume since they are the same region shifted to have a different center. The ratio of the volumes of regions (5.31) and (5.32) is

$$\frac{|\boldsymbol{S}_T^*|^{1/2}}{|\boldsymbol{S}_T^*|^{1/2}} \left(\frac{D_{(U_B)}}{D_{(U_B,T)}} \right)^g = \left(\frac{D_{(U_B)}}{D_{(U_B,T)}} \right)^g. \tag{5.36}$$

The volume of confidence region (5.32) tends to be greater than that of (5.31) since the T_i^* are closer to $\overline{T}^*$ than T_n on average.

If $g = 1$, then a hyperellipsoid is an interval, and confidence intervals are special cases of confidence regions. Suppose the parameter of interest is θ, and there is a bootstrap sample $T_1^*, ..., T_B^*$ where the statistic T_n is an estimator of θ based on a sample of size n. The percentile method uses an interval that contains $U_B \approx k_B = \lceil B(1-\delta) \rceil$ of the T_i^*. Let $a_i = |T_i^* - \overline{T}^*|$. Let $\overline{T}^*$ and S_T^{2*} be the sample mean and variance of the T_i^*. Then the squared Mahalanobis distance $D_\theta^2 = (\theta - \overline{T}^*)^2 / S_T^{*2} \leq D^2_{(U_B)}$ is equivalent to $\theta \in [\overline{T}^* - S_T^* D_{(U_B)}, \overline{T}^* + S_T^* D_{(U_B)}] = [\overline{T}^* - a_{(U_B)}, \overline{T}^* + a_{(U_B)}]$, which is an interval centered at $\overline{T}^*$ just long enough to cover U_B of the T_i^*. Hence the prediction region method CI is a special case of the percentile method CI if $g = 1$. See Definition

5.4. Efron (2014) used a similar large sample $100(1-\delta)\%$ confidence interval assuming that $\overline{T}^*$ is asymptotically normal. The CI $[T_n - a_{(U_B,T)}, T_n + a_{(U_B,T)}]$ corresponding to (5.32) is defined similarly, and $[T_n - a_{(U_B)}, T_n + a_{(U_B)}]$ is the CI for (5.33). Note that the three CIs corresponding to (5.31)–(5.33) can be computed without finding S_T^* or $D_{(U_B)}$ even if $S_T^* = 0$. The shorth(c) CI (5.26) computed from the T_i^* can be much shorter than the Efron (2014) or prediction region method confidence intervals. See Remark 5.8 for some theory for bootstrap CIs.

In the following definition, let U_B and U_{BT} be as in Definitions 5.11 to 5.15. Let a_i be as in the above paragraph.

Definition 5.16. a) The large sample $100(1-\delta)\%$ PR CI is $[\overline{T}^* - a_{(U_B)}, \overline{T}^* + a_{(U_B)}]$.

b) The large sample $100(1-\delta)\%$ BR CI is $[T_n - a_{(U_{BT})}, T_n + a_{(U_{BT})}]$.

c) The large sample $100(1-\delta)\%$ hybrid CI is $[T_n - a_{(U_B)}, T_n + a_{(U_B)}]$.

Remark 5.5. From Chapter 6, $\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*) = \dfrac{n-p}{n} MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \dfrac{n-p}{n}\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})$ where $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ starts to give good estimates of $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\Sigma}_T$ for many error distributions if $n \geq 10p$ and $T = \hat{\boldsymbol{\beta}}$. For the residual bootstrap with large B, note that $\boldsymbol{S}_T^* \approx 0.95\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})$ for $n = 20p$ and $\boldsymbol{S}_T^* \approx 0.99\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})$ for $n = 100p$. Hence we may need $n >> p$ before the $\boldsymbol{S}_T^*$ is a good estimator of $\mathrm{Cov}(T) = \boldsymbol{\Sigma}_T$. The distribution of $\sqrt{n}(T_n - \boldsymbol{\theta})$ is approximated by the distribution of $\sqrt{n}(T^* - T_n)$ or by the distribution of $\sqrt{n}(T^* - \overline{T}^*)$, but n may need to be large before the approximation is good.

Suppose the bootstrap sample mean $\overline{T}^*$ estimates $\boldsymbol{\theta}$, and the bootstrap sample covariance matrix $\boldsymbol{S}_T^*$ estimates $c_n\widehat{\mathrm{Cov}}(T_n) \approx c_n\boldsymbol{\Sigma}_T$ where c_n increases to 1 as $n \to \infty$. Then $\boldsymbol{S}_T^*$ is not a good estimator of $\widehat{\mathrm{Cov}}(T_n)$ until $c_n \approx 1$ ($n \geq 100p$ for OLS $\hat{\boldsymbol{\beta}}$), but the squared Mahalanobis distance $D_{\boldsymbol{w}}^{2*}(\overline{T}^*, \boldsymbol{S}_T^*) \approx D_{\boldsymbol{w}}^2(\boldsymbol{\theta}, \boldsymbol{\Sigma}_T)/c_n$ and $D_{(U_B)}^{2*} \approx D_{1-\delta}^2/c_n$. Hence the prediction region method has a cutoff $D_{(U_B)}^{2*}$ that estimates the cutoff $D_{1-\delta}^2/c_n$. Thus the prediction region method may give good results for much smaller n than a bootstrap method that uses a $\chi_{g,1-\delta}^2$ cutoff when a cutoff $\chi_{g,1-\delta}^2/c_n$ should be used for moderate n.

Remark 5.6. For bootstrapping the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I_{min},0}$, we will often want $n \geq 20p$ and $B \geq \max(100, n, 50p)$. If T_n is $g \times 1$, we might replace p by g or replace p by d if d is the model degrees of freedom. Sometimes much larger n is needed to avoid undercoverage. We want $B \geq 50g$ so that $\boldsymbol{S}_T^*$ is a good estimator of $Cov(T_n^*)$. Prediction region theory uses correction factors like (4.10) and (4.4) to compensate for finite n. The bootstrap confidence

regions (5.31)–(5.35) and the shorth CI use the correction factors (5.30) and (5.26) to compensate for finite $B \geq 50g$. Note that the correction factors make the volume of the confidence region larger as B decreases. Hence a test with larger B will have more power.

5.4.3 Theory for Bootstrap Confidence Regions

Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}$ is $g \times 1$. This section gives some theory for bootstrap confidence regions and for the bagging estimator $\overline{T}^*$, also called the smoothed bootstrap estimator. Empirically, bootstrapping with the bagging estimator often outperforms bootstrapping with T_n. See Breiman (1996), Yang (2003), and Efron (2014). See Büchlmann and Yu (2002) and Friedman and Hall (2007) for theory and references for the bagging estimator.

Remark 5.7. Some regularity conditions used for bootstrap confidence regions are i) $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, ii) $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$, iii) $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, iv) $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$, and v) $n\boldsymbol{S}_T^* \xrightarrow{P} \mathrm{Cov}(\boldsymbol{u})$. Regularity condition v) is rather strong by Machado and Parente (2005). Regularity conditions i) and ii) are often shown using large sample theory. Since (5.32) is a large sample confidence region by Bickel and Ren (2001), (5.31) and (5.33) are too, provided $vi) \sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \mathbf{0}$. Also note that (5.32) is a large sample confidence region if the standard confidence region (5.28) is a large sample confidence region.

Olive (2017b: ∮ 5.3.3, 2018) proved that the prediction region method gives a large sample confidence region under v) from Remark 5.7 and $\boldsymbol{u} \sim N_g(\mathbf{0}, \boldsymbol{\Sigma_u})$, but the following Pelawa Watagoda and Olive (2021a) theorem and proof is simpler. Since iii) and iv) hold by Theorem 5.2, the sample percentile will be consistent under much weaker conditions than v) if $\boldsymbol{\Sigma_u}$ is nonsingular.

Theorem 5.2. a) Suppose i) $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, and ii) $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$ with $E(\boldsymbol{u}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma_u}$. Then iii) $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, iv) $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$, and vi) $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \mathbf{0}$.

b) Then the prediction region method gives a large sample confidence region for $\boldsymbol{\theta}$ provided that the sample percentile $\hat{D}^2_{1-\delta}$ of the $D^2_{T_i^*}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - \overline{T}^*)^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(T_i^* - \overline{T}^*)$ is a consistent estimator of the percentile $D^2_{n,1-\delta}$ of the random variable $D^2_{\boldsymbol{\theta}}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(\boldsymbol{\theta} - \overline{T}^*)^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(\boldsymbol{\theta} - \overline{T}^*)$ in that $\hat{D}^2_{1-\delta} - D^2_{n,1-\delta} \xrightarrow{P} 0$.

Proof. With respect to the bootstrap sample, T_n is a constant and the $\sqrt{n}(T_i^* - T_n)$ are iid for $i = 1, ..., B$. Fix B. Then

$$\begin{bmatrix} \sqrt{n}(T_1^* - T_n) \\ \vdots \\ \sqrt{n}(T_B^* - T_n) \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_B \end{bmatrix}$$

where the $\boldsymbol{v}_i$ are iid with the same distribution as $\boldsymbol{u}$. (Use Theorems 3.7 and 3.8, and see Example 3.2.) For fixed B, the average of the $\sqrt{n}(T_i^* - T_n)$ is

$$\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{D} \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim AN_g\left(\mathbf{0}, \frac{\boldsymbol{\Sigma u}}{B}\right)$$

by Theorem 3.12 where $\boldsymbol{z} \sim AN_g(\mathbf{0}, \boldsymbol{\Sigma})$ is an asymptotic multivariate normal approximation. Hence as $B \to \infty$, $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \mathbf{0}$, and iii), iv), and vi) hold. Hence b) follows. □

Remark 5.8. Note that if $\sqrt{n}(T_n - \theta) \xrightarrow{D} U$ and $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} U$ where U has a unimodal probability density function symmetric about zero, then the confidence intervals from the three confidence regions (5.31)–(5.33), the shorth confidence interval (5.26), and the "usual" percentile method confidence interval (5.25) are asymptotically equivalent (use the central proportion of the bootstrap sample, asymptotically).

Assume $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{\Sigma}_A$ as $n, B \to \infty$ where $\boldsymbol{\Sigma}_A$ and $\boldsymbol{S}_T^*$ are nonsingular $g \times g$ matrices, and T_n is an estimator of $\boldsymbol{\theta}$ such that

$$\sqrt{n}\ \ (T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u} \tag{5.37}$$

as $n \to \infty$. Then

$$\sqrt{n}\ \boldsymbol{\Sigma}_A^{-1/2}\ (T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{\Sigma}_A^{-1/2}\boldsymbol{u} = \boldsymbol{z},$$

$$n\ (T_n - \boldsymbol{\theta})^T\ \hat{\boldsymbol{\Sigma}}_A^{-1}\ (T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{z}^T\boldsymbol{z} = D^2$$

as $n \to \infty$ where $\hat{\boldsymbol{\Sigma}}_A$ is a consistent estimator of $\boldsymbol{\Sigma}_A$, and

$$(T_n - \boldsymbol{\theta})^T\ [\boldsymbol{S}_T^*]^{-1}\ (T_n - \boldsymbol{\theta}) \xrightarrow{D} D^2 \tag{5.38}$$

as $n, B \to \infty$. Assume the cumulative distribution function of D^2 is continuous and increasing in a neighborhood of $D^2_{1-\delta}$ where $P(D^2 \le D^2_{1-\delta}) = 1-\delta$. If the distribution of D^2 is known, then we could use the large sample confidence region (5.28) $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \le D^2_{1-\delta}\}$. Often by a central limit theorem or the multivariate delta method, $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma}_A)$, and $D^2 \sim \chi^2_g$. Note that $[\boldsymbol{S}_T^*]^{-1}$ could be replaced by $n\hat{\boldsymbol{\Sigma}}_A^{-1}$. The following remark gives a simple technical explanation for why bootstrap confidence regions and tests work.

Remark 5.9. a) Assume $\boldsymbol{u}_n \xrightarrow{D} \boldsymbol{u}$ where $\boldsymbol{u}_n$ = i) $\sqrt{n}(T_n - \boldsymbol{\theta})$, ii) $\sqrt{n}(T_i^* - T_n)$, iii) $\sqrt{n}(T_i^* - \overline{T}^*)$, or iv) $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta})$, and $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{C}$ where $\boldsymbol{C}$ is nonsingular. Let

$$D_1^2 = D_{T_i^*}^2(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - \overline{T}^*)^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(T_i^* - \overline{T}^*),$$

$$D_2^2 = D_{\boldsymbol{\theta}}^2(T_n, \boldsymbol{S}_T^*) = \sqrt{n}(T_n - \boldsymbol{\theta})^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(T_n - \boldsymbol{\theta}),$$

$$D_3^2 = D_{\boldsymbol{\theta}}^2(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(\overline{T}^* - \boldsymbol{\theta})^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(\overline{T}^* - \boldsymbol{\theta}), \quad \text{and}$$

$$D_4^2 = D_{T_i^*}^2(T_n, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - T_n)^T (n\boldsymbol{S}_T^*)^{-1} \sqrt{n}(T_i^* - T_n).$$

Then $D_j^2 \approx \boldsymbol{u}^T (n\boldsymbol{S}_T^*)^{-1} \boldsymbol{u} \approx \boldsymbol{u}^T \boldsymbol{C}^{-1} \boldsymbol{u}$, and the percentiles of D_1^2 and D_4^2 can be used as cutoffs. If $(n\boldsymbol{S}_T^*)^{-1}$ is "not too ill conditioned" then $D_j^2 \approx \boldsymbol{u}^T (n\boldsymbol{S}_T^*)^{-1} \boldsymbol{u}$ for large n, and the confidence regions (5.31), (5.32), and (5.33) will have coverage near $1 - \delta$. For confidence regions (5.34) and (5.35), want $\boldsymbol{C}_n^{-1} \xrightarrow{P} \boldsymbol{C}^{-1}$ or $\boldsymbol{C}_n^{-1}$ to be "not too ill conditioned." The regularity conditions for (5.31)–(5.35) are weaker when $g = 1$, since $\boldsymbol{S}_T^*$ and $\boldsymbol{C}_n$ do not need to be computed.

b) Both I) $\sqrt{n}(T_{1n}^* - T_n), ..., \sqrt{n}(T_{Bn}^* - T_n)$ and II) $\sqrt{n}(T_{1n}^* - \overline{T}^*), ..., \sqrt{n}(T_{Bn}^* - \overline{T}^*)$ can be used as pseudodata for III) $\sqrt{n}(T_{1n} - \boldsymbol{\theta}), ..., \sqrt{n}(T_{Bn} - \boldsymbol{\theta})$ when n is large since i), ii) and iii) hold. We can't get the random quantities in III) since $\boldsymbol{\theta}$ is unknown, and we only have $B = 1$ value of the statistic T_n. Note that i) would give an asymptotic pivot if the distribution of $\boldsymbol{u}$ was known.

The following Pelawa Watagoda and Olive (2021a) theorem is very useful. The improved proof, due to Rathnayake and Olive (2023), is used. Let $(\overline{T}, \boldsymbol{S}_T)$ be the sample mean and sample covariance matrix computed from $T_1, ..., T_B$ which have the same distribution as T_n where $T_i = T_{in}$. Let $D_{(U_B)}^2$ be the cutoff computed from the $D_i^2(\overline{T}, \boldsymbol{S}_T)$ for $i = 1, ..., B$. The hyperellipsoids corresponding to $D^2(T_n, \boldsymbol{C})$ and $D^2(\overline{T}, \boldsymbol{C})$ are centered at T_n and $\overline{T}$, respectively. Note that $D_{\overline{T}}^2(T_n, \boldsymbol{C}) = D_{T_n}^2(\overline{T}, \boldsymbol{C})$. Thus $D_{\overline{T}}^2(T_n, \boldsymbol{C}) \leq D_{(U_B)}^2$ iff $D_{T_n}^2(\overline{T}, \boldsymbol{C}) \leq D_{(U_B)}^2$. In Theorem 5.3, since R_p contains T_f with probability $1 - \delta_B$, the region R_c contains $\overline{T}$ with probability $1 - \delta_B$. Since T_n depends on the sample size n, we need $(n\boldsymbol{S}_T)^{-1}$ to be fairly well behaved, e.g. $(n\boldsymbol{S}_T)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_A^{-1}$. Note that $T_i = T_{in}$.

Theorem 5.3: Geometric Argument. Suppose $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ with $E(\boldsymbol{u}) = \boldsymbol{0}$ and $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} \neq \boldsymbol{0}$. Assume $T_1, ..., T_B$ are iid with nonsingular covariance matrix $\boldsymbol{\Sigma}_{T_n}$ where $(n\boldsymbol{S}_T)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_A^{-1}$. Then the large sample $100(1 - \delta)\%$ prediction region $R_p = \{\boldsymbol{w} : D_{\boldsymbol{w}}^2(\overline{T}, \boldsymbol{S}_T) \leq D_{(U_B)}^2\}$ centered at $\overline{T}$ contains a future value of the statistic T_f with probability $1 - \delta_B$ which is eventually bounded below by $1 - \delta$ as $B \to \infty$. Hence the region

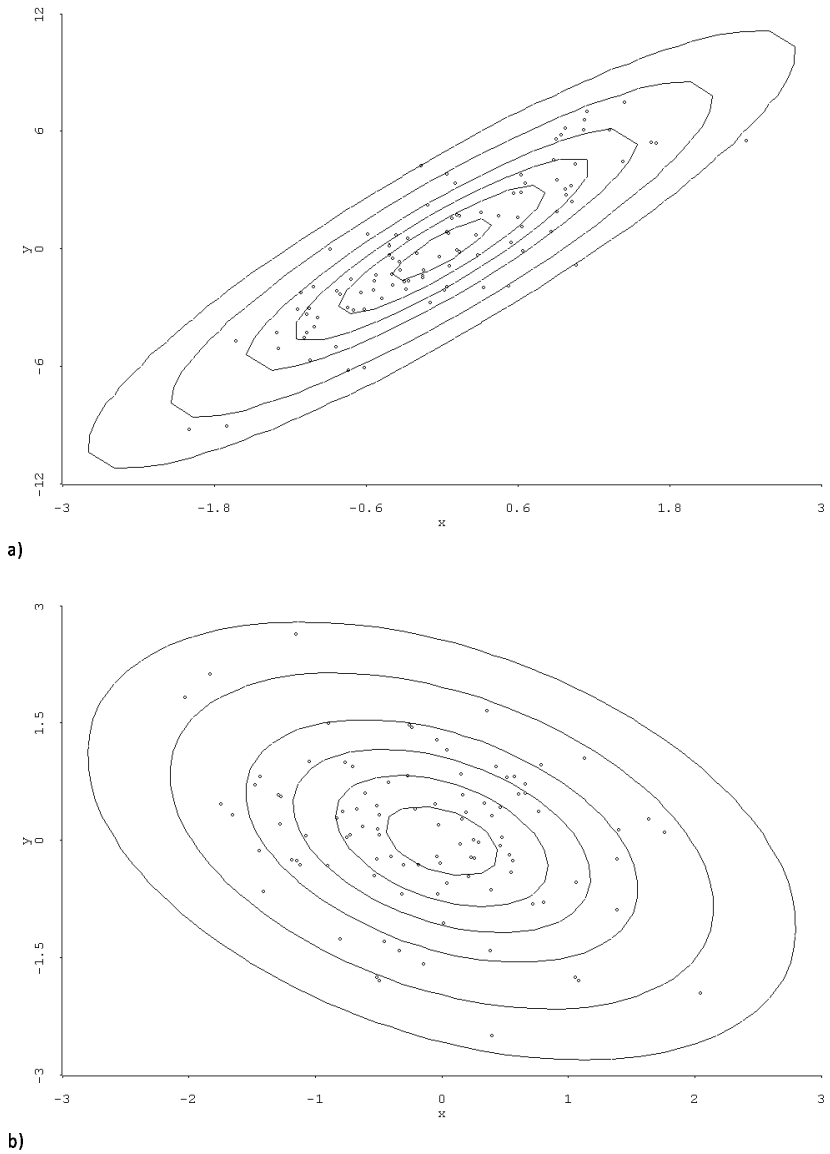

Fig. 5.1 Confidence Regions for 2 Statistics with MVN Distributions

$R_c = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T) \leq D^2_{(U_B)}\}$ is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ where T_n is a randomly selected T_i.

Proof. The region R_c centered at a randomly selected T_n contains $\overline{T}$ with probability $1 - \delta_B$ which is eventually bounded below by $1 - \delta$ as $B \to \infty$. Since the $\sqrt{n}(T_i - \boldsymbol{\theta})$ are iid,

$$\begin{bmatrix} \sqrt{n}(T_1 - \boldsymbol{\theta}) \\ \vdots \\ \sqrt{n}(T_B - \boldsymbol{\theta}) \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_B \end{bmatrix}$$

where the $\boldsymbol{v}_i$ are iid with the same distribution as $\boldsymbol{u}$. (Use Theorems 3.7 and 3.8, and see Example 3.3.) For fixed B, the average of these random vectors is

$$\sqrt{n}(\overline{T} - \boldsymbol{\theta}) \xrightarrow{D} \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim AN_g\left(\boldsymbol{0}, \frac{\boldsymbol{\Sigma_u}}{B}\right)$$

by Theorem 3.12, where AN_g denotes an approximate multivariate normal distribution. Hence $(\overline{T} - \boldsymbol{\theta}) = O_P((nB)^{-1/2})$, and $\overline{T}$ gets arbitrarily close to $\boldsymbol{\theta}$ compared to T_n as $B \to \infty$. Thus R_c is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ as $n, B \to \infty$. $\square$

Examining the iid data cloud $T_1, ..., T_B$ and the bootstrap sample data cloud $T_1^*, ..., T_B^*$ is often useful for understanding the bootstrap. If $\sqrt{n}(T_n - \boldsymbol{\theta})$ and $\sqrt{n}(T_i^* - T_n)$ both converge in distribution to $\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma})$, say, then the bootstrap sample data cloud of $T_1^*, ..., T_B^*$ is like the data cloud of iid $T_1, ..., T_B$ shifted to be centered at T_n. The nonparametric confidence region (5.31) applies the prediction region to the bootstrap. Then the hybrid region (5.33) centers that region at T_n. Hence (5.33) is a confidence region by the geometric argument, and (5.31) is a confidence region if $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \boldsymbol{0}$. Since the T_i^* are closer to $\overline{T}^*$ than T_n on average, $D^2_{(U_{BT})}$ tends to be greater than $D^2_{(U_B)}$. Hence the coverage and volume of (5.32) tend to be at least as large as the coverage and volume of (5.31).

The hyperellipsoid corresponding to the squared Mahalanobis distance $D^2(T_n, \boldsymbol{C})$ is centered at T_n, while the hyperellipsoid corresponding to the squared Mahalanobis distance $D^2(\overline{T}, \boldsymbol{C})$ is centered at $\overline{T}$. Note that $D^2_{\overline{T}}(T_n, \boldsymbol{C}) = (\overline{T} - T_n)^T \boldsymbol{C}^{-1}(\overline{T} - T_n) = (T_n - \overline{T})^T \boldsymbol{C}^{-1}(T_n - \overline{T}) = D^2_{T_n}(\overline{T}, \boldsymbol{C})$. Thus $D^2_{\overline{T}}(T_n, \boldsymbol{C}) \leq D^2_{(U_B)}$ iff $D^2_{T_n}(\overline{T}, \boldsymbol{C}) \leq D^2_{(U_B)}$.

The prediction region method will often simulate well even if B is rather small. If the ellipses are centered at T_n or $\overline{T}^*$, Figure 4.3 shows confidence regions if the plotted points are $T_1^*, ..., T_B^*$ where the T_i^* are approximately multivariate normal. If the ellipses are centered at $\overline{T}$, Figure 5.1 shows 10%, 30%, 50%, 70%, 90%, and 98% prediction regions for a future value of T_f for two multivariate normal statistics. Then the plotted points are iid $T_1, ..., T_B$.

If $nCov(T) \xrightarrow{P} \boldsymbol{\Sigma}_A$, and the T_i^* are iid from the bootstrap distribution, then $Cov(\overline{T}^*) \approx Cov(T)/B \approx \boldsymbol{\Sigma}_A/(nB)$. By Theorem 5.3, if $\overline{T}^*$ is in the 90% prediction region with probability near 90%, then the confidence region should give simulated coverage near 90% and the volume of the confidence region should be near that of the 90% prediction region. If $B = 100$, then $\overline{T}^*$ falls in a covering region of the same shape as the prediction region, but centered near T_n and the lengths of the axes are divided by $\sqrt{B}$. Hence if $B = 100$, then the axes lengths of this covering region are about one tenth of those in Figure 5.1. Hence when T_n falls within the 70% prediction region, the probability that $\overline{T}^*$ falls in the 90% prediction region is near one. If T_n is just within or just without the boundary of the 90% prediction region, $\overline{T}^*$ tends to be just within or just without of the 90% prediction region. Hence the coverage and volume of prediction region confidence region is near that of the nominal coverage 90% and near the volume of the 90% prediction region.

Hence B does not need to be large provided that n and B are large enough so that $S_T^* \approx \text{Cov}(T^*) \approx \boldsymbol{\Sigma}_A/n$. If n is large, the sample covariance matrix starts to be a good estimator of the population covariance matrix when $B \geq Jg$ where $J = 20$ or 50. For small g, using $B = 1000$ often led to good simulations, but $B = \max(50g, 100)$ may work well.

Remark 5.10. Remark 5.5 suggests that even if the statistic T_n is asymptotically normal so the Mahalanobis distances are asymptotically χ^2_g, the prediction region method can give better results for moderate n by using the cutoff $D^2_{(U_B)}$ instead of the cutoff $\chi^2_{g,1-\delta}$. Theorem 5.3 says that the hyperellipsoidal prediction and confidence regions have exactly the same volume. We compensate for the prediction region undercoverage when n is moderate by using $D^2_{(U_n)}$. If n is large, by using $D^2_{(U_B)}$, the prediction region method confidence region compensates for undercoverage when B is moderate, say $B \geq Jg$ where $J = 20$ or 50. See Remark 5.9. This result can be useful if a simulation with $B = 1000$ or $B = 10000$ is much slower than a simulation with $B = Jg$. The price to pay is that the prediction region method confidence region is inflated to have better coverage, so the power of the hypothesis test is decreased if moderate B is used instead of larger B.

5.5 Summary

1) Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a known $g \times 1$ vector. Make a confidence region and reject H_0 if $\boldsymbol{\theta}_0$ is not in the confidence region. Let $\overline{T}^*$ and $\boldsymbol{S}_T^*$ be the sample mean and sample covariance matrix of the bootstrap sample $T_1^*, ..., T_B^*$. a) The prediction region method large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - \overline{T}^*) \leq D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\}$ where $D^2_{(U_B)}$ is computed from $D_i^2 = (T_i^* - \overline{T}^*)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - \overline{T}^*)$ for $i = 1, ..., B$. Note that the corresponding

test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(\overline{T}^* - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (\overline{T}^* - \boldsymbol{\theta}_0) > D^2_{(U_B)}$. This procedure applies the nonparametric prediction region to the bootstrap sample. b) The modified Bickel and Ren (2001) large sample $100(1-\delta)\%$ confidence region is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B,T)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B,T)}\}$ where the cutoff $D^2_{(U_B,T)}$ is the $100q_B$th sample quantile of the $D_i^2 = (T_i^* - T_n)^T [\boldsymbol{S}_T^*]^{-1} (T_i^* - T_n)$. c) The hybrid large sample $100(1-\delta)\%$ confidence region: $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \leq D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\}$.

If $g = 1$, confidence intervals can be computed without $\boldsymbol{S}_T^*$ or D^2 for a), b), and c).

2) **Theorem 5.3: Geometric Argument.** Suppose $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ with $E(\boldsymbol{u}) = \boldsymbol{0}$ and $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$. Assume $T_1, ..., T_B$ are iid with nonsingular covariance matrix $\boldsymbol{\Sigma}_{T_n}$. Then the large sample $100(1-\delta)\%$ prediction region $R_p = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}, \boldsymbol{S}_T) \leq D^2_{(U_B)}\}$ centered at $\overline{T}$ contains a future value of the statistic T_f with probability $1 - \delta_B \to 1 - \delta$ as $B \to \infty$. Hence the region $R_c = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T) \leq D^2_{(U_B)}\}$ is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$.

5.6 Complements

Confidence Intervals

Guenther (1969) is a useful reference for confidence intervals. Agresti and Coull (1998), Brown, Cai and DasGupta (2001, 2002) and Pires and Amado (2008) discuss CIs for a binomial proportion. Agresti and Caffo (2000) discuss CIs for the difference of two binomial proportions $\rho_1 - \rho_2$ obtained from 2 independent samples. Barker (2002), Byrne and Kabaila (2005), Garwood (1936) and Swift (2009) discuss CIs for Poisson (θ) data. Abuhassan and Olive (2008) and Olive (2014) consider CIs for some transformed random variables. Also see Brownstein and Pensky (2008).

Remark 5.11: Correction Factors. Correction factors are used all the time. Let the positive integer $d_n \to \infty$ as $n \to \infty$. In particular, the $z_{1-\delta/2}$ cutoff is replaced by a $t_{d_n,1-\delta/2}$ cutoff for confidence intervals, and the $\chi^2_{k,1-\delta}$ cutoff is replaced by the $kF_{k,d_n,1-\delta}$ cutoff for confidence regions. These cutoffs can be justified by large sample theory. See Example 2.16 and Theorem 2.34. The modified confidence intervals and confidence regions tend to work better in moderate samples because the actual distribution of the statistic tends to have heavier tails than the $N(0,1)$ or χ^2_k distribution. Some statistics need even stronger correction factors. The following correction factors also tend to be asymptotically correct.

A) Confidence intervals: Replace the cutoff $t_{n-1,1-\delta/2}$ by $t_{n-1,up}$ where $up = min(1 - \delta/2 + 0.05, 1 - \delta/2 + 2.5/n)$ if $\delta/2 > 0.1$,

$$up = min(1 - \delta/4, 1 - \delta/2 + 12.5\delta/n)$$

if $\delta/2 \leq 0.1$. If $up < 1 - \delta/2 + 0.001$, then use $up = 1 - \delta/2$. For the nominal 95% CIs, this correction factor uses a cutoff that is between $t_{n-1,0.975}$ and the cutoff $t_{n-1,0.9875}$ that would be used for a 97.5% CI. This technique is like applying a $100[1 - 2(1 - up)]$% CI to the data. See Olive et al. (2024).

B) Confidence regions: Replace the cutoff $\chi^2_{k,1-\delta}$ by $\chi^2_{k,up}$ where $up = \min(1 - \delta + 0.05, 1 - \delta + k/n)$ for $\delta > 0.1$ and

$$up = \min(1 - \delta/2, 1 - \delta + 10\delta k/n), \quad \text{otherwise.} \tag{5.39}$$

If $1 - \delta < 0.999$ and $up < 1 - \delta + 0.001$, set $up = 1 - \delta$. The $kF_{k,d_n,1-\delta}$ cutoff could also be replaced by $kF_{k,d_n,up}$. This technique is like applying a $100up$% confidence region to the data. The "corrected coverage proportion" is increased from the nominal coverage proportion by at most 5% (e.g. 90% to 95%), and by no more than $100\delta/2$% if $\delta \leq 0.05$ (e.g. 95% to 97.5% or 98% to 99%). This correction factor is similar to that used for some of the bootstrap confidence regions and for the nonparametric prediction region. See Equations (5.30) and (4.9).

The Bootstrap

Rajapaksha and Olive (2022) has two more bootstrap confidence regions which have simple large sample theory and which are quick to compute.

Good references for the bootstrap include Efron (1979, 1982), Efron and Hastie (2016, ch. 10–11), and Efron and Tibshirani (1993). Also see Chen (2016) and Hesterberg (2014). One of the sufficient conditions for the bootstrap confidence region is that T has a well behaved Hadamard derivative. Fréchet differentiability implies Hadamard differentiability, and many statistics are shown to be Hadamard differentiable in Bickel and Ren (2001), Clarke (1986, 2000), Fernholtz (1983), Gill (1989), Ren (1991), and Ren and Sen (1995). Bickel and Ren (2001) showed that their method can work when Hadamard differentiability fails.

The double bootstrap technique may be useful. See Hall (1986) and Chang and Hall (2015) for references. The double bootstrap for $\overline{T}^* = \overline{T}^*_B$ says that $T_n = \overline{T}^*$ is a statistic that can be bootstrapped. Let $B_d \geq 50g_{max}$ where $1 \leq g_{max} \leq p$ is the largest dimension of $\boldsymbol{\theta}$ to be tested with the double bootstrap. Draw a bootstrap sample of size B and compute $\overline{T}^* = T^*_1$. Repeat for a total of B_d times. Apply the confidence region (5.31), (5.32), or (5.33) to the double bootstrap sample $T^*_1, ..., T^*_{B_d}$. If $D_{(U_{B_d})} \approx D_{(U_{B_d},T)} \approx \sqrt{\chi^2_{g,1-\delta}}$, then $\overline{T}^*$ may be approximately multivariate normal. The CI (5.31) applied to the double bootstrap sample could be regarded as a modified Frey CI without delta method techniques. Of course the double bootstrap tends to be too computationally expensive to simulate.

Warning: Much of the bootstrap theory in the literature is for when all possible bootstrap samples are taken (the population bootstrap quantities). This theory does not apply when B is fixed, e.g. $B = 1000$, and may not apply if $B = \max(1000, n) \to \infty$ as $n \to \infty$.

Subsampling

The nonparametric bootstrap draws a bootstrap data set $\boldsymbol{x}_1^*, ..., \boldsymbol{x}_n^*$ with replacement from the $\boldsymbol{x}_i$ and computes T_1^* by applying T_n on the bootstrap data set. This process is repeated B times to get a bootstrap sample $T_1^*, ..., T_B^*$. The nonparametric bootstrap has replicates: the proportion of cases in the bootstrap sample that are not replicates is about $1 - e^1 \approx 2/3 \approx 7/11$.

The m out of n bootstrap draws a sample of size m without replacement from the n cases. For $B = 1$, this is a data splitting estimator, and $T_m^* \approx N(0, s_m^2)$ for large enough m and p. Sampling without replacement is also known as subsampling and the delete d jackknife.

Theory for subsampling is given by Politis and Romano (1994) and Wu (1990). Subsampling tends to work well for a large variety of statistics if $m/n \to 0$ with $m \to \infty$. A linear statistic has the form

$$\frac{1}{n}\sum_{i=1}^{n} t(U_i)$$

where $\theta = E[t(U_i)]$ and the U_i are iid. For a linear statistic, subsampling tends to work well if $m/n \to \tau \in [0, 1)$ with $m \to \infty$.

Now let W_i be an indicator random variable with $W_i = 1$ if $\boldsymbol{x}_i^*$ is in the sample and $W_i = 0$, otherwise, for $i = 1, ..., n$. The W_i are binary and identically distributed, but not independent. Hence $P(W_i = 1) = m/n$. Let $W_{ij} = W_i W_j$ with $i \neq j$. Again, the W_{ij} are binary and identically distributed. $P(W_{ij} = 1) = $ P(ordered pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$) was selected in the sample. Hence $P(W_{ij} = 1) = m(m-1)/[n(n-1)]$ since $m(m-1)$ ordered pairs were selected out of $n(n-1)$ possible ordered pairs. Then

$$T_m^* = \frac{1}{m(m-1)} \sum \sum_{k \neq d} \boldsymbol{x}_{i_k}^T \boldsymbol{x}_{i_d} = \frac{1}{m(m-1)} \sum \sum_{i \neq j} W_i W_j \boldsymbol{x}_i^T \boldsymbol{x}_j$$

where the $\boldsymbol{x}_{i_1}, ..., \boldsymbol{x}_{i_m}$ are the m vectors $\boldsymbol{x}_i$ selected in the sample. The first double sum has $m(m-1)$ terms while the second double sum has $n(n-1)$ terms. Hence

$$E(T_m^*) = \frac{1}{m(m-1)} \sum \sum_{i \neq j} E[W_i W_j] \boldsymbol{x}_i^T \boldsymbol{x}_j = T_n.$$

See similar calculations in Buja and Stuetzle (2006). Note that $V(T_m^*) = E([T_m^*]^2) - [T_n]^2 = Cov(T_m^*, T_m^*)$.

Buja and Stuetzle (2006) also show that the nonparametric bootstrap and subsampling with half samples ($m = \lfloor n/2 \rfloor$) often produce similar results.

5.7 Problems

5.1^Q**.** Suppose that $X_1, ..., X_n$ are iid with the Weibull distribution, that is the common pdf is

$$f(x) = \begin{cases} \frac{b}{a} x^{b-1} e^{-\frac{x^b}{a}} & 0 < x \\ 0 & \text{elsewhere} \end{cases}$$

where a is the unknown parameter, but $b(> 0)$ is assumed known.

a) Find a minimal sufficient statistic for a.

b) Assume $n = 10$. Use the Chi-Square Table and the minimal sufficient statistic to find a 95% two sided confidence interval for a.

R Problems

Use a command like *source("G:/lspack.txt")* **to download the functions**. See the Preface. Typing the name of the `lspack` function, e.g. *accisimf*, will display the code for the function. Use the `args` command, e.g. *args(accisimf)*, to display the needed arguments for the function.

5.2. Let $X_1, ..., X_n$ be iid Poisson(θ) random variables.

From the website (http://parker.ad.siu.edu/Olive/lspack.txt), enter the R function `poiscisim` into R. This function simulates the 3 CIs (classical, modified and exact) from Example 5.5. To run the function for $n = 100$ and $\theta = 5$, enter the R command `poiscisim(theta=5)`. Make a table with header "theta ccov clen mcov mlen ecov elen." Fill the table for *theta* = 0.001, 0.1, 1.0, and 5.

The "cov" is the proportion of 500 runs where the CI contained θ and the nominal coverage is 0.95. A coverage between 0.92 and 0.98 gives little evidence that the true coverage differs from the nominal coverage of 0.95. A coverage greater that 0.98 suggests that the CI is conservative while a coverage less than 0.92 suggests that the CI is liberal (too short). Typically want the true coverage $\geq$ the nominal coverage, so conservative intervals are better than liberal CIs. The "len" is the average scaled length of the CI and for large $n\theta$ should be near $2(1.96)\sqrt{\theta}$ for the classical and modified CIs.

From your table, is the classical CI or the modified CI or the "exact" CI better? Explain briefly. (Warning: in a 1999 version of R, there was a bug for the Poisson random number generator for $\theta \geq 10$. The 2011 version of R seems to work.)

5.3. Let $Y_1, ..., Y_n$ be iid binomial($1, \rho$) random variables.

From the website (http://parker.ad.edu/Olive/lspack.txt), enter the R function `bcisim` into R. This function simulates the 3 CIs (classical, Agresti Coull and exact) from Example 5.6, but changes the CI (L,U) to (max(0,L),min(1,U)) to get shorter lengths.

To run the function for $n = 10$ and $\rho \equiv p = 0.001$, enter the R command `bcisim(n=10,p=0.001)`. Make a table with header "n p ccov clen accov aclen ecov elen." Fill the table for $n = 10$ and $p = 0.001, 0.01, 0.5, 0.99, 0.999$ and then repeat for $n = 100$. The "cov" is the proportion of 500 runs where the CI contained p and the nominal coverage is 0.95. A coverage between 0.92 and 0.98 gives little evidence that the true coverage differs from the nominal coverage of 0.95. A coverage greater that 0.98 suggests that the CI is conservative while a coverage less than 0.92 suggests that the CI is liberal. Typically want the true coverage $\geq$ the nominal coverage, so conservative intervals are better than liberal CIs. The "len" is the average scaled length of the CI and for large n should be near $2(1.96)\sqrt{p(1-p)}$.

From your table, is the classical estimator or the Agresti Coull CI better? When is the "exact" interval good? Explain briefly.

5.4. This problem simulates the CIs from Example 5.7.

a) Download the function `accisimf` into R.

b) The function will be used to compare the classical, ACT and modified 95% CIs when the population size $N = 500$ and p is close to 0.01. The function generates such a population, then selects 5000 independent simple random samples from the population. The 5000 CIs are made for both types of intervals, and the number of times the true population p is in the ith CI is counted. The simulated coverage is this count divided by 5000 (the number of CIs). The nominal coverage is 0.95. To run the function for $n = 50$ and $p \approx 0.01$, enter the command `accisimf(n=50,p=0.01)`. Make a table with header "n p ccov clen accov aclen mcov mlen." Fill the table for $n = 50$ and then repeat for $n = 100, 150, 200, 250, 300, 350, 400$ and 450. The "len" is $\sqrt{n}$ times the mean length from the 5000 runs. The "cov" is the proportion of 5000 runs where the CI contained p and the nominal coverage is 0.95. For 5000 runs, an observed coverage between 0.94 and 0.96 gives little evidence that the true coverage differs from the nominal coverage of 0.95. A coverage greater that 0.96 suggests that the CI is conservative while a coverage less than 0.94 suggests that the CI is liberal. Typically want the true coverage $\geq$ the nominal coverage, so conservative intervals are better than liberal CIs. The "ccov" is for the classical CI, "accov" is for the Agresti Coull type (ACT) CI and "mcov" is for the modified interval. Given good coverage > 0.94, want short length.

c) First compare the classical and ACT intervals. From your table, for what values of n is the ACT CI better, for what values of n are the 3 intervals about the same, and for what values of n is the classical CI better?

d) Was the modified CI ever good?

5.5. This problem simulates the CIs from Example 5.1.

a) Download the function `hnsim` into *R*.

The output from this function are the coverages scov, lcov and ccov of the CI for σ^2, μ and of σ^2 if μ is known. The scaled average lengths of the CIs are also given. The lengths of the CIs for σ^2 are multiplied by $\sqrt{n}$ while the length of the CI for μ is multiplied by n.

b) The 5000 CIs are made for 3 intervals, and the number of times the true population parameter $\theta = \mu$ or σ^2 is in the ith CI is counted. The simulated coverage is this count divided by 5000 (the number of CIs). The nominal coverage is 0.95. To run the function for $n = 5$, $\mu = 0$ and $\sigma^2 = 1$ enter the command `hnsim(n=5)`. Make a table with header
"CI for σ^2 CI for μ CI for σ^2, μ known."
Then make a second header "n cov slen cov slen cov slen" where "cov slen" is below each of the three CI headers. Fill the table for $n = 5$ and then repeat for $n = 10, 20, 50, 100$ and 1000. The "cov" is the proportion of 5000 runs where the CI contained θ and the nominal coverage is 0.95. For 5000 runs, an observed coverage between 0.94 and 0.96 gives little evidence that the true coverage differs from the nominal coverage of 0.95. A coverage greater that 0.96 suggests that the CI is conservative while a coverage less than 0.94 suggests that the CI is liberal. As n gets large, the values of slen should get closer to 5.5437, 3.7546 and 5.5437.

5.6. a) Download the function `varcisim` into *R* to simulate a modified version of the CI of Example 5.8.

b) Type the command `varcisim(n = 100, nruns = 1000, type = 1)` to simulate the 95% CI for the variance for iid N(0,1) data. Is the coverage *vcov* close to or higher than 0.95? Is the scaled length *vlen* $= \sqrt{n}$ (CI length) $= 2(1.96)\sigma^2\sqrt{\tau} = 5.554\sigma^2$ close to 5.554?

c) Type the command `varcisim(n = 100, nruns = 1000, type = 2)` to simulate the 95% CI for the variance for iid EXP(1) data. Is the coverage *vcov* close to or higher than 0.95? Is the scaled length *vlen* $= \sqrt{n}$ (CI length) $= 2(1.96)\sigma^2\sqrt{\tau} = 2(1.96)\lambda^2\sqrt{8} = 11.087\lambda^2$ close to 11.087?

d) Type the command `varcisim(n = 100, nruns = 1000, type = 3)` to simulate the 95% CI for the variance for iid LN(0,1) data. Is the coverage *vcov* close to or higher than 0.95? Is the scaled length *vlen* long?

5.7. a) Download the function `pcisim` into *R* to simulate the three CIs of Example 5.9. The modified pooled t CI is almost the same as the Welch CI, but uses degrees of freedom $= n_1 + n_2 - 4$ instead of the more complicated formula for the Welch CI. The pooled t CI should have coverage that is too low if

$$\frac{\rho}{1-\rho}\sigma_1^2 + \sigma_2^2 \ < \ \sigma_1^2 + \frac{\rho}{1-\rho}\sigma_2^2.$$

b) Type the command `pcisim(n1=100,n2=200,var1=10,var2=1)` to simulate the CIs for $N(\mu_i, \sigma_i^2)$ data for $i = 1, 2$. The terms *pcov*, *mpcov*

and *wcov* are the simulated coverages for the pooled, modified pooled and Welch 95% CIs. Record these quantities. Are they near 0.95?

Problems from old qualifying exams are marked with a Q.

5.8^Q**.** Let $X_1, ..., X_n$ be a random sample from a uniform$(0, \theta)$ distribution. Let $Y = \max(X_1, X_2, ..., X_n)$.

a) Find the pdf of Y/θ.

b) To find a confidence interval for θ, can Y/θ be used as a pivot?

c) Find the shortest $(1 - \alpha)\%$ confidence interval for θ.

5.9. Let $Y_1, ..., Y_n$ be iid from a distribution with fourth moments and let S_n^2 be the sample variance. Then

$$\sqrt{n}(S_n^2 - \sigma^2) \xrightarrow{D} N(0, M_4 - \sigma^4)$$

where M_4 is the fourth central moment $E[(Y - \mu)^4]$. Let $\hat{M}_{4,n} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^4$.

a) Use the asymptotic pivot

$$\frac{\sqrt{n}(S_n^2 - \sigma^2)}{\sqrt{\hat{M}_{4,n} - S_n^4}} \xrightarrow{D} N(0, 1)$$

to find a large sample $100(1 - \alpha)\%$ CI for σ^2.

b) Use Equation (5.4) to find a large sample $100(1 - \alpha)\%$ CI for $\sigma_1^2 - \sigma_2^2$.

More problems:

5.10. Suppose $\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma_T^2)$ and that $\hat{\sigma}_T^2$ is a consistent estimator of $\sigma_T^2 > 0$. Then a large sample 95.45% confidence interval for θ is $[T_n - 2SE(T_n), T_n + 2SE(T_n)]$ where $SE(T_n) = \hat{\sigma}_T/\sqrt{n}$. For the test $H_0 : \theta = \theta_0$ versus $H_A : \theta \neq \theta_0$, fail to reject H_0 if θ_0 is in the CI, otherwise reject H_0. The power of the test $= P_\theta$(reject H_0) which goes to 1 as $n \to \infty$ if $\theta \neq \theta_O$ because the length of the CI $\to 0$ as $n \to \infty$. The type I error $=$ P(CI does not contain θ_0) when H_0 is true, and the type I error $\approx 1 - P(-2 < Z < 2) = 1 - 0.9544 = 0.0456$.

Consider the CI

$$[T_n - 2[\log_{10}(n)]^\gamma SE(T_n), T_n + 2[\log_{10}(n)]^\gamma SE(T_n)]$$

where γ is a number like 1/2, 1/3 or 1/4. Take $\gamma = 1/2$.

a) What does the power of the corresponding test converge to as $n \to \infty$?

b) What does the type I error converge to as $n \to \infty$?

c) For what value of n is $2\sqrt{\log_{10}(n)} = 4$?

5.11. Suppose $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$, and $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \mathbf{0}$ where $E(\boldsymbol{u}) = \mathbf{0}$ and $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma u} > 0$.

a) Prove $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$.

b) $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$.

Hint: add $a - a = 0$ to the term in parentheses for a good choice of a, and use Slutsky's Theorem.

5.12. Suppose $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ and $\boldsymbol{C}_n^{-1} \xrightarrow{P} \boldsymbol{C}^{-1}$. Then $n(T_n - \boldsymbol{\theta})^T \boldsymbol{C}_n^{-1}(T_n - \boldsymbol{\theta}) \xrightarrow{D} D^2$.

a) What is D^2 (e.g. is $D^2 = \boldsymbol{u}^T \boldsymbol{C} \boldsymbol{u}$)?

b) If $\boldsymbol{C}_n = \boldsymbol{I}_g$ for all positive integers n, what is D^2?

5.13. Suppose that

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{D} N\left(0, \frac{1}{I_1(\mu)}\right).$$

Find a large sample 95% confidence interval for μ.

5.14. Suppose that $Y_1, ..., Y_n$ are iid from a one parameter exponential family with parameter by τ. Assume that $T_n = \sum_{i=1}^n t(Y_i)$ is a complete sufficient statistics. Suppose, as is often the case, that $T_n \sim G(na, 2b\ \tau)$ where a and b are known positive constants. Then

$$\hat{\tau} = \frac{T_n}{2nab}$$

is the UMVUE and often the MLE of τ. Suggest a $100(1-\alpha)\%$ confidence interval for τ.

Hint: $\dfrac{T_n}{b\tau} \sim G(na, 2)$ and let $P(X \leq G(na, 2, \delta/2)) = \delta/2$ and $P(X \leq G(na, 2, 1 - \delta/2)) = 1 - \delta/2$ if $X \sim G(na, 2)$.

5.15. Suppose that $\boldsymbol{u}_i = (\boldsymbol{x}_i^T, Y_i)^T$ are iid for $i = 1, ..., n$. Let $\boldsymbol{\mu_x} = E(\boldsymbol{x})$ and $\mu_Y = E(Y)$. Let $\tilde{\boldsymbol{\eta}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ and $\boldsymbol{\eta} = \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = Cov(\boldsymbol{x}, Y)$. Then $\sqrt{n}(\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} - \boldsymbol{\Sigma}_{\boldsymbol{x}Y}) = \sqrt{n}(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma_w})$ where $\boldsymbol{\Sigma_w} = Cov(\boldsymbol{w})$ and $\boldsymbol{w}_i = (\boldsymbol{x}_i - \boldsymbol{\mu_x})(Y_i - \boldsymbol{\mu}_Y)$.

The nonparametric bootstrap samples the $\boldsymbol{u}_i = (\boldsymbol{x}_i^T, Y_i)^T$ with replacement. This bootstrap model has the $\boldsymbol{u}_i^* = (\boldsymbol{x}_i^{*T}, Y_i^*)^T$ iid with respect to the bootstrap distribution. Then $E(\boldsymbol{x}_i^*) = \overline{\boldsymbol{x}}$, $E(Y_i^*) = \overline{Y}$, $\boldsymbol{w}_i^* = (\boldsymbol{x}_i^* - \overline{\boldsymbol{x}})(Y_i^* - \overline{Y})$. Fix n. Then $\sqrt{m}(\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}^* - \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}) = \sqrt{m}(\tilde{\boldsymbol{\eta}}^* - \tilde{\boldsymbol{\eta}}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}^*})$. Since the empirical distribution is used,

$$\boldsymbol{\Sigma}_{\boldsymbol{w}^*} = E[(\boldsymbol{w}^* - E(\boldsymbol{w}^*))(\boldsymbol{w}^* - E(\boldsymbol{w}^*))^T] = E[\boldsymbol{w}^* \boldsymbol{w}^{*T}] - E(\boldsymbol{w}^*)[E(\boldsymbol{w}^*)]^T =$$

$$\frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y})[(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y})]^T - [\frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y})][\frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y})]^T$$

$$= \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{w}_i - \overline{\boldsymbol{w}})(\boldsymbol{w}_i - \overline{\boldsymbol{w}})^T = \frac{1}{n}\sum_{i=1}^n \boldsymbol{w}_i \boldsymbol{w}_i^T - \left[\frac{1}{n}\sum_{i=1}^n \boldsymbol{w}_i\right]\left[\frac{1}{n}\sum_{i=1}^n \boldsymbol{w}_i\right]^T$$

$$= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{w}_i \boldsymbol{w}_i^T - \overline{\boldsymbol{w}}[\overline{\boldsymbol{w}}]^T = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i^T - \overline{\boldsymbol{z}}[\overline{\boldsymbol{z}}]^T = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{Z}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{v}_i \boldsymbol{v}_i^T - \overline{\boldsymbol{v}}[\overline{\boldsymbol{v}}]^T = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{V}}$$

where $\boldsymbol{z}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})Y_i$ and $\boldsymbol{v}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y})$.

Use the bootstrap proof technique to find the limiting distribution of $\sqrt{n}(\tilde{\boldsymbol{\Sigma}}^*_{\boldsymbol{x}Y} - \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}) = \sqrt{n}(\tilde{\boldsymbol{\eta}}^* - \tilde{\boldsymbol{\eta}})$.

5.16. The *sample median absolute deviation* is $MAD(Y_i) = MAD(n) = MED(|Y_i - MED(n)|,\ i = 1, \ldots, n)$: find the sample median and go out the distance MAD(n) that covers at least half of the cases. Then $MAD(n)$ estimates the population median absolute deviation MAD(Y): find the population median and go out the distance MAD(Y) that covers at least half of the mass. For $Y_1, ..., Y_n$ iid $N(\mu, \sigma^2)$, $a\ MAD(n) \xrightarrow{P} \sigma$ where $a \approx 1.483$.

a) If X and Y are random variables, show that

$$Cov(X, Y) = [V(X + Y) - V(X - Y)]/4.$$

b) Suppose $(X_i, Y_i)^T$ are iid from a bivariate normal distribution. Suggest a consistent estimator of $Cov(X, Y)$ that is a function of $MAD(X_i + Y_i)$ and $MAD(X_i - Y_i)$.

Hint: $W_i = X_i + Y_i \sim N(E(X) + E(Y), V(X + Y))$ and
$Z_i = X_i - Y_i \sim N(E(X) - E(Y), V(X - Y))$.

5.17. The plug-in principle CI technique when $\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma^2(\theta))$:

$$\left[T_n - z_{1-\delta/2}\frac{\sigma(\hat{\theta})}{\sqrt{n}}, T_n + z_{1-\delta/2}\frac{\sigma(\hat{\theta})}{\sqrt{n}} \right]$$

is a large sample $100(1 - \delta)\%$ CI for θ where $\sigma(\hat{\theta})$ is the estimator of $\sqrt{\sigma^2(\theta)}$.

For the simple linear regression model, $Y_i = \alpha + \beta x_i + e_i$ for $i = 1, ..., n$, it can be shown that

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{D} N[0, \sigma^2/V(x)]$$

where $V(e_i) = \sigma^2$ is estimated by the MSE and $V(X) = V(x_i)$ is estimated by S_x^2. Find a large sample $100(1 - \delta)\%$ CI for β.

5.18. Let $Y_1, ..., Y_n$ be iid $C(\mu, \sigma)$. Then $\sqrt{n}(MED(n) - \mu) \xrightarrow{D} N(0, \pi^2\sigma^2/4)$, and $\hat{\sigma} = MAD(n) \xrightarrow{P} \sigma$ is a consistent estimator of σ. Find a large sample 95% confidence interval for μ. Note: $P(-1.96 < Z < 1.96) = 0.95$ where $Z \sim N(0, 1)$.

Chapter 6
Regression: GLMs, GAMs, Statistical Learning

This chapter considers regression models such as the multiple linear regression model, generalized linear models such as Poisson regression and binomial regression, generalized additive models, and survival regression models such as the Cox proportional hazards regression model. Multivariate linear regression and Statistical Learning methods, such as lasso and ridge regression, are considered. Results for variable selection will be given. See Chapter 10 for some useful plots. **Unless told otherwise, assume the number of predictors p is fixed, while the sample size $n \to \infty$.**

Definition 6.1. For an important class of regression models, **regression** is the study of the conditional distribution $Y|\boldsymbol{A}\boldsymbol{x}$ of the response variable Y given $\boldsymbol{A}\boldsymbol{x}$, where the vector of predictors $\boldsymbol{x} = (x_1, ..., x_p)^T$ and $\boldsymbol{A}$ is a $k \times p$ constant matrix of full rank k with $1 \leq k \leq p$.

Remark 6.1. If $\boldsymbol{A} = \boldsymbol{I}_p$, then $Y|\boldsymbol{A}\boldsymbol{x} = Y|\boldsymbol{x}$. If $\boldsymbol{\beta}$ is a $p \times 1$ coefficient vector and $\boldsymbol{A} = \boldsymbol{\beta}^T$, then $Y|\boldsymbol{A}\boldsymbol{x} = Y|\boldsymbol{\beta}^T\boldsymbol{x} = Y|\boldsymbol{x}^T\boldsymbol{\beta}$.

Definition 6.2. A **quantitative variable** takes on numerical values while a **qualitative variable** takes on categorical values.

Let $\boldsymbol{z} = (z_1, ..., z_k)^T$ where $z_1, ..., z_k$ are k random variables. Often $\boldsymbol{z} = (\boldsymbol{x}^T, Y)^T$ where $\boldsymbol{x}^T = (x_1, ..., x_p)$ is the vector of predictors and Y is the variable of interest, called a response variable. Predictor variables are also called independent variables, covariates, or features. The response variable is also called the dependent variable. Usually context will be used to decide whether $\boldsymbol{z}$ is a random vector or the observed random vector.

Definition 6.3. A **case** or **observation** consists of k random variables measured for one person or thing. The ith case $\boldsymbol{z}_i = (z_{i1}, ..., z_{ik})^T$. The **training data** consists of $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$. A statistical model or method is fit (trained) on the training data. The **test data** consists of $\boldsymbol{z}_{n+1}, ..., \boldsymbol{z}_{n+m}$, and the test data is often used to evaluate the quality of the fitted model.

Definition 6.4. In a **1D regression model**, regression is the study of the conditional distribution of Y given the **sufficient predictor** $SP = h(\boldsymbol{x})$, written

$$Y|SP \text{ or } \mathrm{Y|h}(\boldsymbol{x}), \tag{6.1}$$

where the real valued function $h : \mathbb{R}^p \to \mathbb{R}$. The **estimated sufficient predictor** $\mathrm{ESP} = \hat{h}(\boldsymbol{x})$. An important special case is a model with a linear predictor $h(\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ where $\mathrm{ESP} = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$ and often $\alpha = 0$. This class of models includes the *generalized linear model* (GLM). Another important special case is a *generalized additive model* (GAM), given the *additive predictor* $AP = SP = \alpha + \sum_{j=1}^{p} S_j(x_j)$ for some (usually unknown) functions S_j. The *estimated additive predictor* $\mathrm{EAP} = \mathrm{ESP} = \hat{\alpha} + \sum_{j=1}^{p} \hat{S}_j(x_j)$.

Remark 6.2. The literature often claims that Y is conditionally independent of $\boldsymbol{x}$ given the *sufficient predictor* $SP = h(\boldsymbol{x})$, written

$$Y \perp\!\!\!\perp \boldsymbol{x}|SP \text{ or } \mathrm{Y} \perp\!\!\!\perp \boldsymbol{x}|\mathrm{h}(\boldsymbol{x}). \tag{6.2}$$

The literature also often claims that $Y|\boldsymbol{x} = Y|SP$ or $Y|\boldsymbol{x} = Y|\boldsymbol{\beta}^T\boldsymbol{x}$. This claim is often much too strong.

Notation. Often the conditioning and the index i will be suppressed. For example, the *multiple linear regression model*

$$Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{6.3}$$

for $i = 1, ..., n$ where $\boldsymbol{\beta}$ is a $p \times 1$ unknown vector of parameters, and e_i is a random error. This model could be written $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$. More accurately, $Y|\boldsymbol{\beta}^T\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{\beta} + e$, but the conditioning on $\boldsymbol{\beta}^T\boldsymbol{x}$ will often be suppressed. Often the errors $e_1, ..., e_n$ are **iid** (independent and identically distributed). Often the distribution of the errors is unknown, but often it is assumed that the iid e_i's come from a distribution that is known except for a scale parameter. For example, the e_i's might be iid from a normal (Gaussian) distribution with *mean* 0 and unknown *standard deviation* σ. For this Gaussian model, estimation of $\boldsymbol{\beta}$ and σ is important for inference and for predicting a new future value of the response variable Y_f given a new vector of predictors $\boldsymbol{x}_f$.

Statistical Learning could be defined as the statistical analysis of multivariate data. Machine learning, data mining, big data, analytics, business analytics, data analytics, and predictive analytics are synonymous terms. The techniques are useful for Data Science and Statistics, the science of extracting information from data.

Following James et al. (2013, p. 30), the previously unseen test data is not used to train the Statistical Learning method, but interest is in how well the method performs on the test data. If the training data is $(\boldsymbol{x}_1, Y_1), ..., (\boldsymbol{x}_n, Y_n)$, and the previously unseen test data is $(\boldsymbol{x}_f, Y_f)$, then particular interest is in

the accuracy of the estimator $\hat{Y}_f$ of Y_f obtained when the Statistical Learning method is applied to the predictor $\boldsymbol{x}_f$.

6.1 Multiple Linear Regression

For **multiple linear regression (MLR)**, it is usually useful to have a constant in the model. Sometimes it is convenient to use $Y|\boldsymbol{\beta}^T\boldsymbol{x}$ where $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ and the constant is β_1. Sometimes it is convenient to separate the constant from the nontrivial predictors and use $Y|(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$ where α is the constant. We could also use $\boldsymbol{\beta}^T = (\beta_1, \boldsymbol{\beta}_2^T)$ where β_1 is the intercept and the slopes vector $\boldsymbol{\beta}_2 = (\beta_2, ..., \beta_p)^T$, and $\boldsymbol{x}_i^T = (1, \boldsymbol{u}_i^T)$ where the nontrivial predictors $\boldsymbol{u}_i = (x_{i2}, ..., x_{ip})^T$. Hence we get the following two MLR models. The first model is often used in the theory of linear models, while the second model is often useful for Statistical Learning, MLR with heterogeneity, and high dimensional statistics.

Definition 6.5. Suppose that the response variable Y and at least one predictor variable x_i are quantitative.
a) Let the **MLR model 1** be

$$Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{6.4}$$

for $i = 1, ..., n$. Here n is the sample size and the random variable e_i is the ith error. Assume that the e_i are iid with expected value $E(e_i) = 0$ and variance $V(e_i) = \sigma^2$. In matrix notation, these n equations become $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors.

b) Let the **MLR model 2** be

$$Y_i = \alpha + x_{i,1}\beta_1 + \cdots + x_{i,p}\beta_p + e_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{6.5}$$

for $i = 1, ..., n$. For this model, we may use $\boldsymbol{\phi} = (\alpha, \boldsymbol{\beta}^T)^T$ with $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\phi} + \boldsymbol{e}$.

In matrix notation, suppose the n equations are

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{6.6}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X} = [\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_p]$ is an $n \times p$ matrix of predictors with ith column $\boldsymbol{v}_i$ corresponding to the ith predictor, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. Equivalently,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}. \tag{6.7}$$

For MLR model 1, the first column of $\boldsymbol{X}$ is $\boldsymbol{v}_1 = \mathbf{1}$, the $n \times 1$ vector of ones. The ith **case** $(\boldsymbol{x}_i^T, Y_i)^T = (x_{i1}, x_{i2}, ..., x_{ip}, Y_i)^T$ corresponds to the ith row $\boldsymbol{x}_i^T$ of $\boldsymbol{X}$ and the ith element of $\boldsymbol{Y}$ (if $x_{i1} \equiv 1$, then x_{i1} could be omitted). In the MLR model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$, the Y and e are random variables, but we only have observed values Y_i and $\boldsymbol{x}_i$. MLR is used to estimate the unknown parameters $\boldsymbol{\beta}$ and σ^2.

Definition 6.6. The **constant variance MLR model** uses the assumption that the errors $e_1, ..., e_n$ are iid with mean $E(e_i) = 0$ and variance $\text{VAR}(e_i) = \sigma^2 < \infty$. Also assume that the errors are independent of the predictor variables $\boldsymbol{x}_i$. The predictor variables $\boldsymbol{x}_i$ are assumed to be fixed and measured without error. The cases $(\boldsymbol{x}_i^T, Y_i)^T$ are independent for $i = 1, ..., n$.

If the predictor variables are random variables, then the above MLR model is conditional on the observed values of the $\boldsymbol{x}_i$. That is, observe the $\boldsymbol{x}_i$ and then act as if the observed $\boldsymbol{x}_i$ are fixed.

Definition 6.7. The **unimodal MLR model** has the same assumptions as the constant variance MLR model, as well as the assumption that the zero mean constant variance errors $e_1, ..., e_n$ are iid from a unimodal distribution that is not highly skewed. Note that $E(e_i) = 0$ and $V(e_i) = \sigma^2 < \infty$.

Definition 6.8. The *normal MLR model* or **Gaussian MLR model** has the same assumptions as the unimodal MLR model but adds the assumption that the errors $e_1, ..., e_n$ are iid $N(0, \sigma^2)$ random variables. That is, the e_i are iid normal random variables with zero mean and variance σ^2.

The unknown coefficients for the above 3 models are usually estimated using (ordinary) least squares (OLS).

Notation. The symbol $A \equiv B = f(c)$ means that A and B are equivalent and equal, and that $f(c)$ is the formula used to compute A and B.

Definition 6.9. Given an estimate $\boldsymbol{b}$ of $\boldsymbol{\beta}$, the corresponding vector of *predicted values* or *fitted values* is $\widehat{\boldsymbol{Y}} \equiv \widehat{\boldsymbol{Y}}(\boldsymbol{b}) = \boldsymbol{X}\boldsymbol{b}$. Thus the ith fitted value

$$\hat{Y}_i \equiv \hat{Y}_i(\boldsymbol{b}) = \boldsymbol{x}_i^T\boldsymbol{b} = x_{i,1}b_1 + \cdots + x_{i,p}b_p.$$

The vector of *residuals* is $\boldsymbol{r} \equiv \boldsymbol{r}(\boldsymbol{b}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}(\boldsymbol{b})$. Thus ith residual $r_i \equiv r_i(\boldsymbol{b}) = Y_i - \hat{Y}_i(\boldsymbol{b}) = Y_i - x_{i,1}b_1 - \cdots - x_{i,p}b_p$.

6.1.1 OLS Theory

Ordinary least squares (OLS) large sample theory will be useful. Let $\boldsymbol{X} = (\mathbf{1} \;\; \boldsymbol{X}_1)$. For model (6.4), the ith row of $\boldsymbol{X}$ is $(1, x_{i,2}, ..., x_{i,p})$ while for model (6.5), the ith row of $\boldsymbol{X}$ is $(1, x_{i,1}, ..., x_{i,p})$, and $\boldsymbol{Y} = \alpha\mathbf{1} + \boldsymbol{X}_1\boldsymbol{\beta} + \boldsymbol{e} = \boldsymbol{X}\boldsymbol{\phi} + \boldsymbol{e}$.

Definition 6.10 Using the above notation for MLR model 2 (6.5), let $\boldsymbol{x}_i^T = (x_{i1}, ..., x_{ip})$, let α be the intercept, and let the slopes vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$. Let the population covariance matrices

$$\mathrm{Cov}(\boldsymbol{x}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T] = \boldsymbol{\Sigma}_{\boldsymbol{x}}, \;\; \text{and}$$

$$\mathrm{Cov}(\boldsymbol{x}, Y) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(Y - E(Y))] = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}.$$

If the cases $(\boldsymbol{x}_i, Y_i)$ are iid from some population where $\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ exists and $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ is nonsingular, then the population coefficients from an OLS regression of Y on $\boldsymbol{x}$ (even if a linear model does not hold) are

$$\alpha = \alpha_{OLS} = E(Y) - \boldsymbol{\beta}^T E(\boldsymbol{x}) \;\; \text{and} \;\; \boldsymbol{\beta} = \boldsymbol{\beta}_{\mathrm{OLS}} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}\mathrm{Y}}.$$

Definition 6.11 Let the sample covariance matrices be

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \text{ and } \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y}).$$

Let the method of moments estimators be $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T$ and

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i Y_i - \overline{\boldsymbol{x}}\,\overline{Y}.$$

The method of moment estimators are often called the maximum likelihood estimators, but are the MLE if the $(Y_i, \boldsymbol{x}_i^T)^T$ are iid from a multivariate normal distribution, a very strong assumption. In Theorem 6.1, note that $\boldsymbol{D} = \boldsymbol{X}_1^T\boldsymbol{X}_1 - n\overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^T = (n-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}$.

Theorem 6.1: Seber and Lee (2003, p. 106). Let $\boldsymbol{X} = (\mathbf{1} \;\; \boldsymbol{X}_1)$. Then $\boldsymbol{X}^T\boldsymbol{Y} = \begin{pmatrix} n\overline{Y} \\ \boldsymbol{X}_1^T\boldsymbol{Y} \end{pmatrix} = \begin{pmatrix} n\overline{Y} \\ \sum_{i=1}^{n}\boldsymbol{x}_i Y_i \end{pmatrix}$, $\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} n & n\overline{\boldsymbol{x}}^T \\ n\overline{\boldsymbol{x}} & \boldsymbol{X}_1^T\boldsymbol{X}_1 \end{pmatrix}$,

$$\text{and} \;\; (\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} \frac{1}{n} + \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & -\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & \boldsymbol{D}^{-1} \end{pmatrix}$$

where the $p \times p$ matrix $\boldsymbol{D}^{-1} = [(n-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}]^{-1} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}/(n-1)$.

Under model (6.5), $\hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\phi}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$.

Theorem 6.2: Second way to compute $\hat{\boldsymbol{\phi}}$:
a) If $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}$ exists, then $\hat{\alpha} = \overline{Y} - \hat{\boldsymbol{\beta}}^T\overline{\boldsymbol{x}}$ and

$$\hat{\boldsymbol{\beta}} = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}.$$

b) Suppose that $(Y_i, \boldsymbol{x}_i^T)^T$ are iid random vectors such that σ_Y^2, $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$, and $\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ exist. Then $\hat{\alpha} \xrightarrow{P} \alpha$ and

$$\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta} \text{ as } \mathrm{n} \to \infty$$

where α and $\boldsymbol{\beta}$ are given by Definition 6.10.

Proof. Note that

$$\boldsymbol{Y}^T\boldsymbol{X}_1 = (Y_1 \cdots Y_n)\begin{bmatrix}\boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T\end{bmatrix} = \sum_{i=1}^n Y_i\boldsymbol{x}_i^T$$

and

$$\boldsymbol{X}_1^T\boldsymbol{Y} = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_n]\begin{bmatrix}Y_1 \\ \vdots \\ Y_n\end{bmatrix} = \sum_{i=1}^n \boldsymbol{x}_i Y_i.$$

So

$$\begin{bmatrix}\hat{\alpha} \\ \hat{\boldsymbol{\beta}}\end{bmatrix} = \begin{bmatrix}\frac{1}{n} + \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & -\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & \boldsymbol{D}^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{1}^T \\ \boldsymbol{X}_1^T\end{bmatrix}\boldsymbol{Y} =$$

$$\begin{bmatrix}\frac{1}{n} + \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & -\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & \boldsymbol{D}^{-1}\end{bmatrix}\begin{bmatrix}n\overline{Y} \\ \boldsymbol{X}_1^T\boldsymbol{Y}\end{bmatrix}.$$

Thus $\hat{\boldsymbol{\beta}} = -n\boldsymbol{D}^{-1}\overline{\boldsymbol{x}}\,\overline{Y} + \boldsymbol{D}^{-1}\boldsymbol{X}_1^T\boldsymbol{Y} = \boldsymbol{D}^{-1}(\boldsymbol{X}_1^T\boldsymbol{Y} - n\overline{\boldsymbol{x}}\,\overline{Y}) =$

$$\boldsymbol{D}^{-1}\left[\sum_{i=1}^n \boldsymbol{u}_i Y_i - n\overline{\boldsymbol{x}}\,\overline{Y}\right] = \frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}}{n-1}n\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}. \text{ Then}$$

$\hat{\alpha} = \overline{Y} + n\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}}\,\overline{Y} - \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\boldsymbol{X}_1^T\boldsymbol{Y} = \overline{Y} + [n\overline{Y}\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} - \boldsymbol{Y}^T\boldsymbol{X}_1\boldsymbol{D}^{-1}]\overline{\boldsymbol{x}}$
$= \overline{Y} - \hat{\boldsymbol{\beta}}^T\overline{\boldsymbol{x}}$. The convergence in probability results hold since sample means and sample covariance matrices are consistent estimators of the population means and population covariance matrices. □

Remark 6.3. It is important to note that the convergence in probability results are for iid $(Y_i, \boldsymbol{x}_i^T)^T$ with second moments and nonsingular $\boldsymbol{\Sigma}_{\boldsymbol{x}}$: a linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ does not need to hold. When the linear model does hold, the second method for computing $\hat{\boldsymbol{\beta}}$ is still valid even if $\boldsymbol{X}$ is a

constant matrix, and $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}$ by Theorem 6.3 b). From Theorem 6.3,

$$n(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \hat{\boldsymbol{V}} = \begin{pmatrix} \hat{\boldsymbol{V}}_{11} & \hat{\boldsymbol{V}}_{12} \\ \hat{\boldsymbol{V}}_{21} & \hat{\boldsymbol{V}}_{22} \end{pmatrix} \xrightarrow{P} \boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{pmatrix}.$$

Thus $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \xrightarrow{P} \boldsymbol{V}_{22}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \xrightarrow{P} \boldsymbol{V}_{22}^{-1}$. Note that for Theorem 6.3 b) with iid cases and $\boldsymbol{\mu}_{\boldsymbol{x}} = E(\boldsymbol{x})$,

$$n(\boldsymbol{X}^T\boldsymbol{X})^{-1} \xrightarrow{P} \boldsymbol{V} = \begin{bmatrix} 1 + \boldsymbol{\mu}_{\boldsymbol{x}}^T \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\mu}_{\boldsymbol{x}} & -\boldsymbol{\mu}_{\boldsymbol{x}}^T \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \\ -\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\mu}_{\boldsymbol{x}} & \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \end{bmatrix}.$$

Definition 6.12. For OLS and MLR model 1 from Definition 6.5, $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. Let the *hat matrix* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Then $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Y}}_{OLS} = \boldsymbol{H}\boldsymbol{Y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. The ith leverage $h_i = \boldsymbol{H}_{ii}$ = the ith diagonal element of $\boldsymbol{H}$.

There are many large sample theory results for ordinary least squares. For Theorem 6.3, see, for example, Sen and Singer (1993, p. 280). Theorem 6.3 is analogous to the central limit theorem and the theory for the t–interval for μ based on $\overline{Y}$ and the sample standard deviation (SD) S_Y. If the data $Y_1, ..., Y_n$ are iid with mean 0 and variance σ^2, then $\overline{Y}$ is asymptotically normal and the t–interval will perform well if the sample size is large enough. The results below suggests that the OLS estimators $\hat{Y}_i$ and $\hat{\boldsymbol{\beta}}$ are good if the sample size is large enough. The condition $\max h_i \to 0$ in probability usually holds if the researcher picked the design matrix $\boldsymbol{X}$ or if the $\boldsymbol{x}_i$ are iid random vectors from a well behaved population. Outliers can cause the condition to fail. Theorem 6.3 a) implies that $\hat{\boldsymbol{\beta}} \approx N_p[\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}]$. For Theorem 6.3 a), $\text{rank}(\boldsymbol{X}) = p$ since $\boldsymbol{X}^T\boldsymbol{X}$ is nonsingular. For Theorem 6.3 b), $\text{rank}(\boldsymbol{X}) = p + 1$.

Theorem 6.3, OLS CLTs. Consider the MLR model and assume that the zero mean errors are iid with $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2$. If the $\boldsymbol{x}_i$ are random vectors, assume that the cases $(\boldsymbol{x}_i, Y_i)$ are independent, and that the $\boldsymbol{e}_i$ and $\boldsymbol{x}_i$ are independent. Also assume that $\max_i(h_1, ..., h_n) \to 0$ and

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \to \boldsymbol{V}^{-1}$$

as $n \to \infty$ where the convergence is in probability if the $\boldsymbol{x}_i$ are random vectors (instead of nonstochastic constant vectors).

a) For Equation (6.4), the OLS estimator $\hat{\boldsymbol{\beta}}$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{V}). \tag{6.8}$$

Equivalently,

$$(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{I}_p). \tag{6.9}$$

b) For Equation (6.5), the OLS estimator $\hat{\boldsymbol{\phi}}$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \xrightarrow{D} N_{p+1}(\boldsymbol{0}, \sigma^2\ \boldsymbol{V}). \tag{6.10}$$

c) Suppose the cases $(\boldsymbol{x}_i, Y_i)$ are iid from some population and the Equation (6.5) MLR model $Y_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ holds. Assume that $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$ exist. Then Equation (6.10) holds and

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}) \tag{6.11}$$

where $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$.

Remark 6.4. I) Consider Theorem 6.3. For a) and b), the theory acts as if the $\boldsymbol{x}_i$ are constant even if the $\boldsymbol{x}_i$ are random vectors. The literature says the $\boldsymbol{x}_i$ can be constants, or condition on $\boldsymbol{x}_i$ if the $\boldsymbol{x}_i$ are random vectors. The main assumptions for a) and b) are that the errors are iid with second moments and that $n(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is well behaved. The strong assumptions for c) are much stronger than those for a) and b), but the assumption of iid cases is often reasonable if the cases come from some population.
II) Suppose $Y_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ where the e_i are iid. Then $\hat{\boldsymbol{\beta}}_{OLS} \approx N_p(\boldsymbol{\beta}, MSE\ \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}/n)$ even if the cases are not iid, and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \xrightarrow{P} \boldsymbol{V}_{22}^{-1}$, where $\boldsymbol{V}_{22}^{-1}$ is not necessarily equal to $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, by Remark 6.3. Thus
$(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta})^T\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta})/MSE \xrightarrow{D} \chi_p^2$ as $n \to \infty$. This result is useful since no matrix inversion is required.

Remark 6.5. Consider MLR model (6.5). Let $\boldsymbol{w}_i = \boldsymbol{A}_n\boldsymbol{x}_i$ for $i = 1, ..., n$ where $\boldsymbol{A}_n$ is a full rank $k \times p$ matrix with $1 \le k \le p$.

a) Let $\boldsymbol{\Sigma}^*$ be $\hat{\boldsymbol{\Sigma}}$ or $\tilde{\boldsymbol{\Sigma}}$. Then $\boldsymbol{\Sigma}_{\boldsymbol{w}}^* = \boldsymbol{A}_n\boldsymbol{\Sigma}_{\boldsymbol{x}}^*\boldsymbol{A}_n^T$ and $\boldsymbol{\Sigma}_{\boldsymbol{w}Y}^* = \boldsymbol{A}_n\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^*$.

b) If $\boldsymbol{A}_n$ is a constant matrix, then $\boldsymbol{\Sigma}_{\boldsymbol{w}} = \boldsymbol{A}_n\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}_n^T$ and
$\boldsymbol{\Sigma}_{\boldsymbol{w}Y} = \boldsymbol{A}_n\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$.

c) Let $\hat{\boldsymbol{\beta}}(\boldsymbol{u}, Y)$ and $\boldsymbol{\beta}(\boldsymbol{u}, Y)$ be the estimator and parameter from the OLS regression of Y on $\boldsymbol{u}$. The constant parameter vector should not depend on n. Suppose the cases are iid and $\boldsymbol{A}$ is a constant matrix that does not depend on n. By Theorem 6.2, $\hat{\boldsymbol{\beta}}(\boldsymbol{w}, Y) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Y} = [\boldsymbol{A}_n\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\boldsymbol{A}_n]^{-1}\boldsymbol{A}_n\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = [\boldsymbol{A}_n\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\boldsymbol{A}_n]^{-1}\boldsymbol{A}_n\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\beta}}(\boldsymbol{x}, Y)$. If $\boldsymbol{A}_n \xrightarrow{P} \boldsymbol{A}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \xrightarrow{P} \boldsymbol{\Sigma}_{\boldsymbol{x}}$, and $\hat{\boldsymbol{\beta}}(\boldsymbol{x}, Y) \xrightarrow{P} \boldsymbol{\beta}(\boldsymbol{x}, Y)$, then $\hat{\boldsymbol{\beta}}(\boldsymbol{w}, Y) \xrightarrow{P} \boldsymbol{\beta}(\boldsymbol{w}, Y) = [\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}]^{-1}\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}(\boldsymbol{x}, Y)$.

6.1.2 Ordinary Least Squares

In this subsection, assume MLR model 1 (6.4) from Definition 6.5 holds.

Definition 6.13. The full rank MLR model has $\text{rank}(\boldsymbol{X}) = p$.

Many MLR methods attempt to find an estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ which minimizes some criterion function $Q(\boldsymbol{b})$ of the residuals.

Definition 6.14. The *ordinary least squares (OLS) estimator* $\hat{\boldsymbol{\beta}}_{OLS}$ minimizes

$$Q_{OLS}(\boldsymbol{b}) = \sum_{i=1}^{n} r_i^2(\boldsymbol{b}), \tag{6.12}$$

$$\text{and } \hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}.$$

The vector of *predicted* or *fitted values* $\widehat{\boldsymbol{Y}}_{OLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{H}\boldsymbol{Y}$ where the *hat matrix* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ provided the inverse exists. Typically the subscript OLS is omitted, and the least squares *regression equation* is $\hat{Y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$ where $x_1 \equiv 1$ if the model contains a constant.

Definition 6.15. Let the r_i be the OLS residuals and let

$$\hat{\sigma}^2 = MSE = \frac{1}{n}\sum_{i=1}^{n} r_i^2. \tag{6.13}$$

Theorem 6.4 follows from results in Su and Cook (2012). Also see Freedman (1981). In particular, the iid errors do not need to be from a normal distribution.

Theorem 6.4. Let the MLR model hold and the iid errors e_i satisfy $E(e_i) = 0$ and $V(e_i) = \sigma^2$. Under mild regularity conditions, $\hat{\sigma}^2 = MSE$ is a $\sqrt{n}$ consistent estimator of σ^2.

If $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{V}$, then $\hat{\boldsymbol{\Sigma}}_n = nMSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Hence

$$\hat{\boldsymbol{\beta}} \sim AN_p(\boldsymbol{\beta}, MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}), \text{ and}$$

$$rF_R = \frac{1}{MSE}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c})^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \xrightarrow{D} \chi_r^2 \tag{6.14}$$

as $n \to \infty$ if $H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{c}$ is true so that $\sqrt{n}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \xrightarrow{D} N_r(\boldsymbol{0}, \sigma^2\ \boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T)$.

Remark 6.6. The Cauchy Schwartz inequality says $|\boldsymbol{a}^T\boldsymbol{b}| \le \|\boldsymbol{a}\|\ \ \|\boldsymbol{b}\|$. Suppose $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = O_P(1)$ is bounded in probability. This will occur if $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, e.g. if $\hat{\boldsymbol{\beta}}$ is the OLS estimator. Then

$$|r_i - e_i| = |Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} - (Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})| = |\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})|.$$

Hence

$$\sqrt{n}\max_{i=1,\ldots,n}|r_i - e_i| \le (\max_{i=1,\ldots,n}\|\boldsymbol{x}_i\|)\ \ \|\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\| = O_P(1)$$

since $\max \|\boldsymbol{x}_i\| = O_P(1)$ or there is extrapolation. Hence OLS residuals behave well if the zero mean error distribution of the iid e_i has a finite variance σ^2.

Definition 6.16. A test with test statistic T_n is a *large sample right tail* δ *test* if the test rejects H_0 if $T_n > a_n$ and $P(T_n > a_n) = \delta_n$ where δ_n is eventually bounded above by δ as $n \to \infty$ when H_0 is true.

Often we want $\delta_n \to \delta$ as $n \to \infty$. Typically we want $\delta \leq 0.1$, and the values $\delta = 0.05$ and $\delta = 0.01$ are common. (An analogy is a large sample $100(1-\delta)\%$ confidence interval or prediction interval.)

Remark 6.7. For a test of hypotheses, the p–value $\equiv$ pvalue is the probability of getting a test statistic as extreme as the test statistic actually observed, and H_0 is rejected if the pvalue $\leq \delta$. The pvalue given by output tends to only be correct for the normal MLR model. Hence the output is usually only giving an estimate of the pvalue, which will often be denoted by *pval*. So reject H_0 if pval $\leq \delta$. Often

$$\text{pval} - \text{pvalue} \xrightarrow{\text{P}} 0$$

as the sample size $n \to \infty$. Then the computer output pval is a good estimator of the unknown pvalue. We will use $Fo \equiv F_0$, $Ho \equiv H_0$, and $Ha \equiv H_A \equiv H_1$.

Remark 6.8. Suppose $P(W \leq \chi^2_q(1-\delta)) = 1-\delta$ and $P(W > \chi^2_q(1-\delta)) = \delta$ where $W \sim \chi^2_q$. Suppose $P(W \leq F_{q,d_n}(1-\delta)) = 1-\delta$ when $W \sim F_{q,d_n}$. Also write $\chi^2_q(1-\delta) = \chi^2_{q,1-\delta}$ and $F_{q,d_n}(1-\delta) = F_{q,d_n,1-\delta}$. Suppose $P(W > z_{1-\delta}) = \delta$ when $W \sim N(0,1)$, and $P(W > t_{d_n,1-\delta}) = \delta$ when $W \sim t_{d_n}$.

i) Theorem 6.4 is important because it can often be shown that a statistic $T_n = rW_n \xrightarrow{D} \chi^2_r$ when H_0 is true. Then tests that reject H_0 when $T_n > \chi^2_r(1-\delta)$ or when $T_n/r = W_n > F_{r,d_n}(1-\delta)$ are both large sample right tail δ tests if the positive integer $d_n \to \infty$ as $n \to \infty$. Large sample F tests and intervals are used instead of χ^2 tests and intervals since the F tests and intervals are more accurate for moderate n. See Theorem 2.34.

ii) An analogy is that if test statistic $T_n \xrightarrow{D} N(0,1)$ when H_0 is true, then tests that reject H_0 if $T_n > z_{1-\delta}$ or if $T_n > t_{d_n,1-\delta}$ are both large sample right tail δ tests if the positive integer $d_n \to \infty$ as $n \to \infty$. Large sample t tests and intervals are used instead of Z tests and intervals since the t tests and intervals are more accurate for moderate n.

iii) Often $n \geq 10p$ starts to give good results for the OLS output for error distributions not too far from $N(0,1)$. Larger values of n tend to be needed if the zero mean iid errors have a distribution that is far from a normal distribution.

The following two theorems are useful for proving Theorem 6.7, which shows that the most used F-tests for MLR are large sample tests. The notation $\boldsymbol{\Sigma} > 0$ means the $p \times p$ matrix $\boldsymbol{\Sigma}$ is positive definite and thus nonsingular. Hence $\boldsymbol{x}^T \boldsymbol{\Sigma} \boldsymbol{x} > 0$ unless $\boldsymbol{x} = \boldsymbol{0}$ where $\boldsymbol{x}$ is any $p \times 1$ constant vector. If $>$ is replaced by $\geq$, then $\boldsymbol{\Sigma} \geq 0$ is positive semidefinite. A matrix $\boldsymbol{P}$ is a **projection matrix** if $\boldsymbol{P}$ is symmetric and idempotent: $\boldsymbol{P} = \boldsymbol{P}^T = \boldsymbol{P}\boldsymbol{P}$. Unless told otherwise, assume the matrix $\boldsymbol{A}$ in a quadratic form $\boldsymbol{Z}^T \boldsymbol{A} \boldsymbol{Z}$ is symmetric: $\boldsymbol{A} = \boldsymbol{A}^T$. The trace of a square $p \times p$ matrix $\boldsymbol{A}$ is the sum of the diagonal elements of $\boldsymbol{A}$: if $\boldsymbol{A} = (a_{ij})$ so that the ijth element of $\boldsymbol{A}$ is a_{ij}, then $\text{trace}(\boldsymbol{A}) = tr(\boldsymbol{A}) = \sum_{i=1}^{p} a_{ii}$.

Theorem 6.5: Craig's Theorem. Let $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
a) If $\boldsymbol{\Sigma} > 0$, then $\boldsymbol{Y}^T \boldsymbol{A} \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Y}^T \boldsymbol{B} \boldsymbol{Y}$ iff $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B} = \boldsymbol{0}$ iff $\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{0}$.
b) If $\boldsymbol{\Sigma} \geq 0$, then $\boldsymbol{Y}^T \boldsymbol{A} \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Y}^T \boldsymbol{B} \boldsymbol{Y}$ if $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B} = \boldsymbol{0}$ (or if $\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{A} = \boldsymbol{0}$).
c) If $\boldsymbol{\Sigma} \geq 0$, then $\boldsymbol{Y}^T \boldsymbol{A} \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Y}^T \boldsymbol{B} \boldsymbol{Y}$ iff
$(*)$ $\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}\boldsymbol{\Sigma} = \boldsymbol{0}, \boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}\boldsymbol{\mu} = \boldsymbol{0}, \boldsymbol{\Sigma}\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\mu} = \boldsymbol{0},$ and $\boldsymbol{\mu}^T \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}\boldsymbol{\mu} = 0$.

Theorem 6.6. Let $\boldsymbol{A} = \boldsymbol{A}^T$ be symmetric.
a) If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a projection matrix, then $\boldsymbol{Y}^T \boldsymbol{Y} \sim \chi^2(\text{rank}(\boldsymbol{\Sigma}))$ where $\text{rank}(\boldsymbol{\Sigma}) = tr(\boldsymbol{\Sigma})$.
b) If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{I})$, then $\boldsymbol{Y}^T \boldsymbol{A} \boldsymbol{Y} \sim \chi^2_r$ iff $\boldsymbol{A}$ is idempotent with $\text{rank}(\boldsymbol{A}) = tr(\boldsymbol{A}) = r$.
c) Let $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. Then

$$\frac{\boldsymbol{Y}^T \boldsymbol{A} \boldsymbol{Y}}{\sigma^2} \sim \chi^2_r \text{ or } \boldsymbol{Y}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{Y} \sim \sigma^2 \, \chi^2_{\mathrm{r}}$$

iff $\boldsymbol{A}$ is idempotent of rank r.
d) If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} > 0$, then $\boldsymbol{Y}^T \boldsymbol{A} \boldsymbol{Y} \sim \chi^2_r$ iff $\boldsymbol{A}\boldsymbol{\Sigma}$ is idempotent with $\text{rank}(\boldsymbol{A}) = r = \text{rank}(\boldsymbol{A}\boldsymbol{\Sigma})$.
e) If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ then $\dfrac{\boldsymbol{Y}^T \boldsymbol{Y}}{\sigma^2} \sim \chi^2\left(n, \dfrac{\boldsymbol{\mu}^T \boldsymbol{\mu}}{2\sigma^2}\right)$.
f) If $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{I})$ then $\boldsymbol{Y}^T \boldsymbol{A} \boldsymbol{Y} \sim \chi^2(r, \boldsymbol{\mu}^T \boldsymbol{A}\boldsymbol{\mu}/2)$ iff $\boldsymbol{A}$ is idempotent with $\text{rank}(\boldsymbol{A}) = tr(\boldsymbol{A}) = r$.
g) If $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ then $\dfrac{\boldsymbol{Y}^T \boldsymbol{A} \boldsymbol{Y}}{\sigma^2} \sim \chi^2\left(r, \dfrac{\boldsymbol{\mu}^T \boldsymbol{A}\boldsymbol{\mu}}{2\sigma^2}\right)$ iff $\boldsymbol{A}$ is idempotent with $\text{rank}(\boldsymbol{A}) = tr(\boldsymbol{A}) = r$.

For the following theorem, let $\boldsymbol{P} = \boldsymbol{H}$ be the projection matrix on the column space of $\boldsymbol{X}$. The partial F test is $H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{L}\boldsymbol{\beta} \neq \boldsymbol{0}$ where $\boldsymbol{L}$ is a full rank $r \times p$ matrix with $1 \leq r \leq p$. Let R be the reduced model corresponding to $\boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{0}$, let RSS=SSE(F) be the residual sum of squares of the full model that uses all p predictors, and let RSS(R)=SSE(R) be the residual sum of squares for the reduced model that uses q predictors. This test is for whether the reduced model is good which is equivalent to the test that the $p - q$ predictors not in the reduced model are not needed in the

model given the q predictors in the reduced model are in the model. Note that $\boldsymbol{L} = [\boldsymbol{0} \;\; \boldsymbol{I}_r]$ tests whether the last r coefficients $\beta_i = 0$: hence the reduced model uses the first $p - r$ predictors. Then $r = p - 1$ corresponds to the Anova F test for whether the nontrivial predictors are needed in the model where the first predictor $x_1 = 1$ corresponds to a constant β_1 in the model. Also $\boldsymbol{L} = (0, ..., 1, ..., 0)$ with a 1 in the ith position tests whether $\beta_i = 0$ with a reduced model that omits the ith predictor. This test corresponds to the Wald test for whether the ith predictor is needed in the model given the other predictors are in the model. Let F_R be the test statistic for the partial F test.

Theorem 6.7, Partial F Test Theorem. Suppose $H_0 : \boldsymbol{L\beta} = \boldsymbol{0}$ is true for the partial F test. Under the OLS full rank model, a)

$$F_R = \frac{1}{rMSE}(\boldsymbol{L}\hat{\boldsymbol{\beta}})^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}}).$$

b) If $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, then $F_R \sim F_{r,n-p}$.
c) For a large class of zero mean error distributions $rF_R \xrightarrow{D} \chi^2_r$.
d) The partial F test that rejects $H_0 : \boldsymbol{L\beta} = \boldsymbol{0}$ if $F_R > F_{r,n-p}(1 - \delta)$ is a large sample right tail δ test for the OLS model for a large class of zero mean error distributions.

Proof sketch. a) Seber and Lee (2003, p. 100) show that

$$RSS(R) - RSS = (\boldsymbol{L}\hat{\boldsymbol{\beta}})^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}}).$$

b) Let the full model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ with a constant β_1 in the model: $\boldsymbol{1}$ is the 1st column of $\boldsymbol{X}$. Let the reduced model $\boldsymbol{Y} = \boldsymbol{X}_R\boldsymbol{\beta}_R + \boldsymbol{e}$ also have a constant in the model where the columns of $\boldsymbol{X}_R$ are a subset of k of the columns of $\boldsymbol{X}$. Let $\boldsymbol{P}_R$ be the projection matrix on $C(\boldsymbol{X}_R)$ so $\boldsymbol{P}\boldsymbol{P}_R = \boldsymbol{P}_R$. Then $F_R = \dfrac{SSE(R) - SSE(F)}{rMSE(F)}$ where $r = df_R - df_F = p - k$ = number of predictors in the full model but not in the reduced model. $MSE = MSE(F) = SSE(F)/(n-p)$ where $SSE = SSE(F) = \boldsymbol{Y}(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}$. $SSE(R) - SSE(F) = \boldsymbol{Y}^T(\boldsymbol{P} - \boldsymbol{P}_R)\boldsymbol{Y}$ where $SSE(R) = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P}_R)\boldsymbol{Y}$.

Now assume $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$, and when H_0 is true, $\boldsymbol{Y} \sim N_n(\boldsymbol{X}_R\boldsymbol{\beta}_R, \sigma^2\boldsymbol{I})$. Since $(\boldsymbol{I} - \boldsymbol{P})(\boldsymbol{P} - \boldsymbol{P}_R) = \boldsymbol{0}$, $[SSE(R) - SSE(F)] \perp\!\!\!\perp MSE(F)$ by Craig's Theorem. When H_0 is true, $\boldsymbol{\mu} = \boldsymbol{X}_R\boldsymbol{\beta}_R$ and $\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu} = 0$ where $\boldsymbol{A} = (\boldsymbol{I} - \boldsymbol{P})$ or $\boldsymbol{A} = (\boldsymbol{P} - \boldsymbol{P}_R)$. Hence the noncentrality parameter is 0, and by Theorem 6.6 g), $SSE \sim \sigma^2\chi^2_{n-p}$ and $SSE(R) - SSE(F) \sim \sigma^2\chi^2_{p-k}$ since $rank(\boldsymbol{P} - \boldsymbol{P}_R) = tr(\boldsymbol{P} - \boldsymbol{P}_R) = p - k$. Hence under H_0, $F_R \sim F_{p-k,n-p}$.

Alternatively, let $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}_n)$ where $\boldsymbol{X}$ is an $n \times p$ matrix of rank p. Let $\boldsymbol{X} = [\boldsymbol{X}_1 \;\; \boldsymbol{X}_2]$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T \;\; \boldsymbol{\beta}_2^T)^T$ where $\boldsymbol{X}_1$ is an $n \times k$ matrix and $r = p-k$. Consider testing $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$. (The columns of $\boldsymbol{X}$ can be rearranged so that H_0 corresponds to the partial F test.) Let $\boldsymbol{P}$ be the projection matrix

on $C(\boldsymbol{X})$. Then $\boldsymbol{r}^T\boldsymbol{r} = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y} = \boldsymbol{e}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{e} =$
$(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})$ since $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$ and $\boldsymbol{X}^T\boldsymbol{P} = \boldsymbol{X}^T$ imply that $\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{P}) = \boldsymbol{0}$ and $(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{X} = \boldsymbol{0}$.

Suppose that $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$ is true so that $\boldsymbol{Y} \sim N_n(\boldsymbol{X}_1\boldsymbol{\beta}_1, \sigma^2\boldsymbol{I}_n)$. Let $\boldsymbol{P}_1$ be the projection matrix on $C(\boldsymbol{X}_1)$. By the above argument, $\boldsymbol{r}_R^T\boldsymbol{r}_R = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{Y} = (\boldsymbol{Y}-\boldsymbol{X}_1\boldsymbol{\beta}_1)^T(\boldsymbol{I}-\boldsymbol{P}_1)(\boldsymbol{Y}-\boldsymbol{X}_1\boldsymbol{\beta}_1) = \boldsymbol{e}_R^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{e}_R$ where $\boldsymbol{e}_R \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$ when H_0 is true. Or use RHS $= \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{Y}$

$$-\boldsymbol{\beta}_1^T\boldsymbol{X}_1^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{Y} + \boldsymbol{\beta}_1^T\boldsymbol{X}_1^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_1\boldsymbol{\beta}_1 - \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_1\boldsymbol{\beta}_1,$$

and the last three terms equal 0 since $\boldsymbol{X}_1^T(\boldsymbol{I}-\boldsymbol{P}_1) = \boldsymbol{0}$ and $(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_1 = \boldsymbol{0}$.

Hence

$$\frac{\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}}{\sigma^2} \sim \chi^2_{n-p} \perp\!\!\!\perp \frac{\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}}{\sigma^2} \sim \chi^2_r$$

by Theorem 6.6 c) using $\boldsymbol{e}$ and $\boldsymbol{e}_R$ instead of $\boldsymbol{Y}$, and Craig's Theorem 6.5 b) since $n-p = rank(\boldsymbol{I}-\boldsymbol{P}) = tr(\boldsymbol{I}-\boldsymbol{P})$, $r = rank(\boldsymbol{P}-\boldsymbol{P}_1) = tr(\boldsymbol{P}-\boldsymbol{P}_1) = p-k$, and $(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{P}-\boldsymbol{P}_1) = \boldsymbol{0}$.

If $X_1 \sim \chi^2_{d_1} \perp\!\!\!\perp X_2 \sim \chi^2_{d_2}$, then

$$\frac{X_1/d_1}{X_2/d_2} \sim F_{d_1,d_2}.$$

Hence

$$\frac{\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}/r}{\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}/(n-p)} = \frac{\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}}{rMSE} \sim F_{r,n-p}$$

when H_0 is true. Since $RSS = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}$ and $RSS(R) = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{Y}$, $RSS(R) - RSS = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P}_1-[\boldsymbol{I}-\boldsymbol{P}])\boldsymbol{Y} = \boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}$, and thus

$$F_R = \frac{\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}}{rMSE} \sim F_{r,n-p}.$$

c) Assume H_0 is true. By the OLS CLT, $\sqrt{n}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{L}\boldsymbol{\beta}) = \sqrt{n}\boldsymbol{L}\hat{\boldsymbol{\beta}} \xrightarrow{D} N_r(\boldsymbol{0}, \sigma^2\ \boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T)$. Thus $\sqrt{n}(\boldsymbol{L}\hat{\boldsymbol{\beta}})^T(\sigma^2\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T)^{-1}\sqrt{n}\boldsymbol{L}\hat{\boldsymbol{\beta}} \xrightarrow{D} \chi^2_r$. Let $\hat{\sigma}^2 = MSE$ and $\hat{\boldsymbol{W}} = n(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Then

$$n(\boldsymbol{L}\hat{\boldsymbol{\beta}})^T[MSE\ \boldsymbol{L}n(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{\beta}} = rF_R \xrightarrow{D} \chi^2_r.$$

d) By Theorem 2.34, if $W_n \sim F_{r,d_n}$ then $rW_n \xrightarrow{D} \chi^2_r$ as $n \to \infty$ and $d_n \to \infty$. Hence the result follows by c). □

Remark 6.9, Are Statisticians crazy? Courses on linear models typically assume that the e_i are iid $N(0, \sigma^2)$ and use Theorem 6.7 b). The errors e_i rarely follow a normal distribution. Luckily, the F tests are still large sample theory tests by Theorem 6.7 c) and d). This theory makes OLS a nonparametric method that is widely applicable.

An ANOVA table for the partial F test is shown below, where $k = p_R$ is the number of predictors used by the reduced model, and $r = p - p_R = p - k$ is the number of predictors in the full model that are not in the reduced model.

Source	df	SS	MS	F
Reduced	$n - p_R$	$SSE(R) = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P}_R)\boldsymbol{Y}$	MSE(R)	$F_R = \frac{SSE(R)-SSE}{rMSE} =$
Full	$n - p$	$SSE = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$	MSE	$\frac{\boldsymbol{Y}^T(\boldsymbol{P} - \boldsymbol{P}_R)\boldsymbol{Y}/r}{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}/(n-p)}$

The ANOVA F test is the special case where $k = 1$, $\boldsymbol{X}_R = \boldsymbol{1}$, $\boldsymbol{P}_R = \boldsymbol{P}_1$, and $SSE(R) - SSE(F) = SSTO - SSE = SSR$. This test has the table shown below.

ANOVA table: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ with a constant β_1 in the model: $\boldsymbol{1}$ is the 1st column of $\boldsymbol{X}$. $MS = SS/df$.

$SSTO = \boldsymbol{Y}^T(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T)\boldsymbol{Y} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$, $SSE = \sum_{i=1}^{n} r_i^2$, $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$, $SSTO = SSR + SSE$. SSTO is the SSE (residual sum of squares) for the location model $\boldsymbol{Y} = \boldsymbol{1}\beta_1 + \boldsymbol{e}$ that contains a constant but no nontrivial predictors. The location model has projection matrix $\boldsymbol{P}_1 = \boldsymbol{1}(\boldsymbol{1}^T\boldsymbol{1})^{-1}\boldsymbol{1}^T = \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T$. Hence $\boldsymbol{P}\boldsymbol{P}_1 = \boldsymbol{P}_1$ and $\boldsymbol{P}\boldsymbol{1} = \boldsymbol{P}_1\boldsymbol{1} = \boldsymbol{1}$.

Source	df	SS	MS	F	p-value
Regression	p-1	$SSR = \boldsymbol{Y}^T(\boldsymbol{P} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T)\boldsymbol{Y}$	MSR	$F_0 = \frac{MSR}{MSE}$	for H_0:
Residual	n-p	$SSE = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$	MSE		$\beta_2 = \cdots = \beta_p = 0$

The matrices in the quadratic forms for SSR and SSE are symmetric and idempotent and their product is $\boldsymbol{0}$. Hence if $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ so $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$, then $SSE \perp\!\!\!\perp SSR$ by Craig's Theorem. If H_0 is true under normality, then $\boldsymbol{Y} \sim N_n(\boldsymbol{1}\beta_1, \sigma^2\boldsymbol{I})$, and by Theorem 6.4 g), $SSE \sim \sigma^2\chi^2_{n-p}$ and $SSR \sim \sigma^2\chi^2_{p-1}$ since $rank(\boldsymbol{I} - \boldsymbol{P}) = tr(\boldsymbol{I} - \boldsymbol{P}) = n - p$ and $rank(\boldsymbol{P} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T) = tr(\boldsymbol{P} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T) = p - 1$. Hence under normality, $F_0 \sim F_{p-1,n-p}$.

Let $X \sim t_{n-p}$. Then $X^2 \sim F_{1,n-p}$. The two tail Wald t test for $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$ is equivalent to the corresponding right tailed F test since rejecting H_0 if $|X| > t_{n-p}(1 - \delta)$ is equivalent to rejecting H_0 if $X^2 > F_{1,n-p}(1 - \delta)$.

Theorem 6.8. Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} = \hat{\boldsymbol{Y}} + \boldsymbol{r}$ where $\boldsymbol{X}$ has full rank p, $E(\boldsymbol{e}) = \boldsymbol{0}$, and $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$. i) The least squares estimator $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$: $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$. ii) $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

Proof. i) $E(\hat{\boldsymbol{\beta}}) = E[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E[\boldsymbol{Y}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}$.

ii) $\text{Cov}(\hat{\boldsymbol{\beta}}) = \text{Cov}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}] = \text{Cov}(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}\text{Cov}(\boldsymbol{Y})\boldsymbol{A}^T =$

$$\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}. \ \square$$

6.1.3 L_1

Definition 6.17. Assume the MLR model holds. The L_1 *estimator* or *least absolute deviations estimator* $\hat{\boldsymbol{\beta}}_{L_1}$ minimizes the criterion

$$Q_{L_1}(\boldsymbol{b}) = \sum_{i=1}^{n} |r_i(\boldsymbol{b})| = \sum_{i=1}^{n} |Y_i - \boldsymbol{x}_i^T\boldsymbol{b}|.$$

Theorem 6.9, L_1 **CLT:** Assume the MLR model holds and the errors e_i are iid with a pdf f such that the unique population median is 0 with $f(0) > 0$. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{L_1} - \boldsymbol{\beta}) \xrightarrow{D} N_p\left(\boldsymbol{0}, \frac{1}{4[f(0)]^2}\ \boldsymbol{V}\right) \tag{6.15}$$

when $\boldsymbol{X}^T\boldsymbol{X}/n \to \boldsymbol{V}^{-1}$.

If a constant β_1 is in the model or if the column space of $\boldsymbol{X}$ contains $\boldsymbol{1}$, then the assumption on the pdf is mild, but if the pdf is not symmetric about 0, then the L_1 β_1 tends to differ from the OLS β_1. See Bassett and Koenker (1978) for the theorem. Pollard (1991) discusses some useful extensions. Estimating $f(0)$ can be difficult.

If the pdf is also symmetric about 0 and $V(e_i) = \sigma^2$, then often $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, V(\hat{\boldsymbol{\beta}}, F)\ \boldsymbol{V})$ where F is the cdf of the error distribution. Then $V(\hat{\boldsymbol{\beta}}_{OLS}, F) = V(e_i) = \sigma^2$, and

$$V(\hat{\boldsymbol{\beta}}_{L_1}, F) = \frac{1}{4[f(0)]^2}.$$

6.2 Bootstrapping OLS MLR

Suppose the full model for MLR is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Suppose that there is a minimal subset S such that $\boldsymbol{Y} = \boldsymbol{X}_S\boldsymbol{\beta}_S + \boldsymbol{e}$. Then for any subset I such that $S \subseteq I$, $\boldsymbol{Y} = \boldsymbol{X}_I\boldsymbol{\beta}_I + \boldsymbol{e}$. Assume a constant is in the model and in any submodel I. Then then the OLS residuals sum to 0. Let submodel I contain

a_I predictors, including a constant. If $S \subseteq I$, let

$$\frac{\boldsymbol{X}_I^T \boldsymbol{X}}{n} \to \boldsymbol{V}_I^{-1}.$$

Then by the OLS CLT, $\sqrt{n}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{\Sigma}_I)$ where $\boldsymbol{\Sigma}_I = \sigma^2 \boldsymbol{V}_I$. See Section 6.10 for more on submodel notation.

6.2.1 The Parametric Bootstrap

The parametric bootstrap generates $\boldsymbol{Y}_j^* = (Y_i^*)$ from a parametric distribution. Then regress $\boldsymbol{Y}_j^*$ on $\boldsymbol{X}$ to get $\hat{\boldsymbol{\beta}}_j^*$ for $j = 1, ..., B$. Consider the parametric bootstrap for the MLR model with $\boldsymbol{Y}^* \sim N_n(\boldsymbol{X}\hat{\boldsymbol{\beta}}, \hat{\sigma}_n^2 \boldsymbol{I}) \sim N_n(\boldsymbol{HY}, \hat{\sigma}_n^2 \boldsymbol{I})$ where **we are not assuming** that the $e_i \sim N(0, \sigma^2)$, and

$$\hat{\sigma}_n^2 = MSE = \frac{1}{n-p} \sum_{i=1}^{n} r_i^2$$

where the residuals are from the full OLS model. Then MSE is a $\sqrt{n}$ consistent estimator of σ^2 under mild conditions by Theorem 6.4 and Su and Cook (2012). Hence

$$\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{e}^*$$

where the e_i^* are iid $N(0, MSE)$ and $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS}$.

Thus $\hat{\boldsymbol{\beta}}_I^* = (\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1} \boldsymbol{X}_I^T \boldsymbol{Y}^* \sim N_{a_I}(\hat{\boldsymbol{\beta}}_I, \hat{\sigma}_n^2 (\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1})$ since $E(\hat{\boldsymbol{\beta}}_I^*) = (\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1} \boldsymbol{X}_I^T \boldsymbol{HY} = \hat{\boldsymbol{\beta}}_I$ because $\boldsymbol{HX}_I = \boldsymbol{X}_I$, and $\text{Cov}(\hat{\boldsymbol{\beta}}_I^*) = \hat{\sigma}_n^2 (\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1}$. Hence

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_I^* - \hat{\boldsymbol{\beta}}_I) \sim N_{a_I}(\boldsymbol{0}, n\hat{\sigma}_n^2 (\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1}) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{\Sigma}_I)$$

as $n \to \infty$ if $S \subseteq I$. In particular, for the full model $I = F$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \sim N_p(\boldsymbol{0}, n\hat{\sigma}_n^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$$

as $n \to \infty$, where $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{V}$.

6.2.2 The Residual Bootstrap

The *residual bootstrap* is often useful for additive error regression models of the form $Y_i = m(\boldsymbol{x}_i) + e_i = \hat{m}(\boldsymbol{x}_i) + r_i = \hat{Y}_i + r_i$ for $i = 1, ..., n$ where the ith residual $r_i = Y_i - \hat{Y}_i$. Let $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$, $\boldsymbol{r} = (r_1, ..., r_n)^T$, and let $\boldsymbol{X}$ be an $n \times p$ matrix with ith row $\boldsymbol{x}_i^T$. Then the fitted values $\hat{Y}_i = \hat{m}(\boldsymbol{x}_i)$,

and the residuals are obtained by regressing $\boldsymbol{Y}$ on $\boldsymbol{X}$. Here the errors e_i are iid, and it would be useful to be able to generate B iid samples $e_{1j}, ..., e_{nj}$ from the distribution of e_i where $j = 1, ..., B$. If the $m(\boldsymbol{x}_i)$ were known, then we could form a vector $\boldsymbol{Y}_j$ where the ith element $Y_{ij} = m(\boldsymbol{x}_i) + e_{ij}$ for $i = 1, ..., n$. Then regress $\boldsymbol{Y}_j$ on $\boldsymbol{X}$. Instead, draw samples $r^*_{1j}, ..., r^*_{nj}$ with replacement from the residuals, then form a vector $\boldsymbol{Y}^*_j$ where the ith element $Y^*_{ij} = \hat{m}(\boldsymbol{x}_i) + r^*_{ij}$ for $i = 1, ..., n$. Then regress $\boldsymbol{Y}^*_j$ on $\boldsymbol{X}$. If the residuals do not sum to 0 and $E(e_i) = 0$, then replace r_i by $\epsilon_i = r_i - \overline{r}$, and r^*_{ij} by ϵ^*_{ij}.

For multiple linear regression, $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ is written in matrix form as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Regress $\boldsymbol{Y}$ on $\boldsymbol{X}$ to obtain $\hat{\boldsymbol{\beta}}$, $\boldsymbol{r}$, and $\hat{\boldsymbol{Y}}$ with ith element $\hat{Y}_i = \hat{m}(\boldsymbol{x}_i) = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$. For $j = 1, ..., B$, regress $\boldsymbol{Y}^*_j$ on $\boldsymbol{X}$ to form $\hat{\boldsymbol{\beta}}^*_{1,n}, ..., \hat{\boldsymbol{\beta}}^*_{B,n}$ using the residual bootstrap.

Now examine the OLS model with a constant in the model so the OLS residuals sum to 0. Let $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Y}}_{OLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{H}\boldsymbol{Y}$ be the fitted values from the OLS full model. Let $\boldsymbol{r}^W$ denote an $n \times 1$ random vector of elements selected with replacement from the OLS full model residuals. Following Freedman (1981) and Efron (1982, p. 36),

$$\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{r}^W$$

follows a standard linear model where the elements r^W_i of $\boldsymbol{r}^W$ are iid from the empirical distribution of the OLS full model residuals r_i. Hence

$$E(r^W_i) = \frac{1}{n}\sum_{i=1}^{n} r_i = 0, \quad V(r^W_i) = \sigma^2_n = \frac{1}{n}\sum_{i=1}^{n} r_i^2 = \frac{n-p}{n}MSE,$$

$$E(\boldsymbol{r}^W) = \boldsymbol{0}, \text{ and } \mathrm{Cov}(\boldsymbol{Y}^*) = \mathrm{Cov}(\boldsymbol{r}^{\mathrm{W}}) = \sigma^2_{\mathrm{n}}\boldsymbol{I}_{\mathrm{n}}.$$

Let $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS}$. Then $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}^*$ with $\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*) = \sigma^2_n(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \frac{n-p}{n}MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, and $E(\hat{\boldsymbol{\beta}}^*) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E(\boldsymbol{Y}^*) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}\boldsymbol{Y} = \hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_n$ since $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$. The expectations are with respect to the bootstrap distribution where $\hat{\boldsymbol{Y}}$ acts as a constant. One difference from the usual OLS MLR model is that $\sigma^2_n \xrightarrow{P} \sigma^2$ depends on n. The usual model has $V(e_i) = \sigma^2$ which does not depend on n.

For the OLS estimator $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS}$, the estimated covariance matrix of $\hat{\boldsymbol{\beta}}_{OLS}$ is $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{OLS}) = MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. The sample covariance matrix of the $\hat{\boldsymbol{\beta}}^*$ is estimating $\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*)$ as $B \to \infty$. Hence the residual bootstrap standard error $SE(\hat{\beta}^*_i) \approx \sqrt{\frac{n-p}{n}}\, SE(\hat{\beta}_i)$ for $i = 1, ..., p$ where $\hat{\boldsymbol{\beta}}_{OLS} = \hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$. The OLS CLT Theorem 6.3 says

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \lim_{n\to\infty} n\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{OLS})) \sim N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V})$$

where $n(\boldsymbol{X}^T\boldsymbol{X})^{-1} \to \boldsymbol{V}$. Since $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{r}^W$ follows a standard linear model, it may not be surprising that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}_{OLS}) \xrightarrow{D} N_p(\boldsymbol{0}, \lim_{n\to\infty} n\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}^*)) \sim N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V}). \tag{6.16}$$

Imagine for large fixed $n = N$ we get the OLS residuals. Then we use these residuals for $n > N$ to get $\hat{\boldsymbol{\beta}}^*_{n,N}$. Then by the OLS CLT, $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_{n,N} - \hat{\boldsymbol{\beta}}_{OLS}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2_N\boldsymbol{V})$ as $n \to \infty$, and $N_p(\boldsymbol{0}, \sigma^2_N\boldsymbol{V}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V})$ as $N \to \infty$. Hence Theorem 5.1 is satisfied, and Equation (6.16) holds. See Freedman (1981) for an alternative proof.

Remark 6.10. Both the residual bootstrap and parametric bootstrap for the OLS full model are robust to the unknown error distribution of the iid e_i. For the MLR residual bootstrap with $S \subseteq I$ where I is not the full model, we conjecture that $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_I - \hat{\boldsymbol{\beta}}_I) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{\Sigma}_I)$ as $n \to \infty$ since OLS estimators tend to be asymptotically normal with a distribution that depends on the covariance matrix of the estimator. For the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, the e_i are iid from a distribution that does not depend on n, and $\boldsymbol{\beta}_O = \boldsymbol{0}$ where O denotes the terms in the full model that are not in I. For $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{r}^W$, the distribution of the r^W_i depends on n and $\hat{\boldsymbol{\beta}}_O \neq \boldsymbol{0}$ although $\sqrt{n}\hat{\boldsymbol{\beta}}_O = O_P(1)$.

6.2.3 The Nonparametric Bootstrap

The nonparametric bootstrap (also called the empirical bootstrap, naive bootstrap, the pairwise bootstrap, and the pairs bootstrap) draws a sample of n cases $(Y^*_i, \boldsymbol{x}^*_i)$ with replacement from the n cases $(Y_i, \boldsymbol{x}_i)$, and regresses the Y^*_i on the $\boldsymbol{x}^*_i$ to get $\hat{\boldsymbol{\beta}}^*_{VS,1}$, and then draws another sample to get $\hat{\boldsymbol{\beta}}^*_{MIX,1}$. This process is repeated B times to get the two bootstrap samples for $i = 1, ..., B$.

Then for the full model,

$$\boldsymbol{Y}^* = \boldsymbol{X}^*\hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{r}^W$$

and for a submodel I,

$$\boldsymbol{Y}^* = \boldsymbol{X}^*_I\hat{\boldsymbol{\beta}}_{I,OLS} + \boldsymbol{r}^W_I.$$

Freedman (1981) showed that under regularity conditions for the OLS MLR model, $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V}) \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$. Hence if $S \subseteq I$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^*_I - \hat{\boldsymbol{\beta}}_I) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{\Sigma}_I)$$

as $n \to \infty$. (Treat I as if I is the full model.)

One set of regularity conditions is that the MLR model holds, and if $\boldsymbol{x}_i = (1 \ \boldsymbol{u}_i^T)^T$, then the $\boldsymbol{w}_i = (Y_i \ \boldsymbol{u}_i^T)^T$ are iid from some population with a nonsingular covariance matrix.

The nonparametric bootstrap uses $\boldsymbol{w}_1^*, ..., \boldsymbol{w}_n^*$ where the $\boldsymbol{w}_i^*$ are sampled with replacement from $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. By Example 5.11, $E(\boldsymbol{w}^*) = \overline{\boldsymbol{w}}$, and

$$\text{Cov}(\boldsymbol{w}^*) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_i - \overline{\boldsymbol{w}})(\boldsymbol{w}_i - \overline{\boldsymbol{w}})^T = \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \begin{bmatrix} \tilde{S}_Y^2 & \tilde{\boldsymbol{\Sigma}}_{Y\boldsymbol{u}} \\ \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} & \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}} \end{bmatrix}.$$

Note that $\hat{\boldsymbol{\beta}}$ is a constant with respect to the bootstrap distribution. Assume all inverse matrices exist. Then

$$\hat{\boldsymbol{\beta}}^* = \begin{bmatrix} \hat{\beta}_1^* \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}^* \end{bmatrix} = \begin{bmatrix} \overline{Y}^* - \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}^{*T}\overline{\boldsymbol{u}}^* \\ \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1^*}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y}^* \end{bmatrix} \xrightarrow{P} \begin{bmatrix} \overline{Y} - \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}^T\overline{\boldsymbol{u}} \\ \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{u}} \end{bmatrix} = \hat{\boldsymbol{\beta}}$$

as $B \to \infty$. This result suggests that the nonparametric bootstrap for OLS MLR might work under milder regularity conditions than the $\boldsymbol{w}_i$ being iid from some population with a nonsingular covariance matrix.

6.3 Statistical Learning Methods for MLR

There are many MLR methods, including OLS for the full model, forward selection with OLS, the marginal maximum likelihood estimator (MMLE), elastic net, principal components regression (PCR), partial least squares (PLS), lasso, lasso variable selection, and ridge regression (RR). For the last six methods, it is convenient to use centered or scaled data. Suppose U has observed values $U_1, ..., U_n$. For example, if $U_i = Y_i$ then U corresponds to the response variable Y. The observed values of a random variable V are *centered* if their sample mean is 0. The centered values of U are $V_i = U_i - \overline{U}$ for $i = 1, ..., n$. Let g be an integer near 0. If the sample variance of the U_i is

$$\hat{\sigma}_g^2 = \frac{1}{n-g}\sum_{i=1}^{n}(U_i - \overline{U})^2,$$

then the sample standard deviation of U_i is $\hat{\sigma}_g$. If the values of U_i are not all the same, then $\hat{\sigma}_g > 0$, and the standardized values of the U_i are

$$W_i = \frac{U_i - \overline{U}}{\hat{\sigma}_g}.$$

Typically $g = 1$ or $g = 0$ are used: $g = 1$ gives an unbiased estimator of σ^2 while $g = 0$ gives the method of moments estimator. Note that the

standardized values are centered, $\overline{W} = 0$, and the sample variance of the standardized values

$$\frac{1}{n-g}\sum_{i=1}^{n} W_i^2 = 1. \tag{6.17}$$

Remark 6.11. Let $Y = \alpha + \boldsymbol{x}^T\boldsymbol{\beta} + e$. Let $\boldsymbol{w}_i^T = (w_{i,1}, ..., w_{i,p})$ be the standardized vector of nontrivial predictors for the ith case. Since the standardized predictors are also centered, $\overline{\boldsymbol{w}} = \boldsymbol{0}$. Let the $n \times p$ matrix of standardized nontrivial predictors $\boldsymbol{W}_g = (W_{ij})$ when the predictors are standardized using $\hat{\sigma}_g$. Then the ith row of $\boldsymbol{W}_g$ is $\boldsymbol{w}_i^T$. Thus, $\sum_{i=1}^n W_{ij} = 0$ and $\sum_{i=1}^n W_{ij}^2 = n-g$ for $j = 1, ..., p$. Hence

$$W_{ij} = \frac{x_{i,j} - \overline{x}_j}{\hat{\sigma}_j} \quad \text{where} \quad \hat{\sigma}_{\mathrm{j}}^2 = \frac{1}{\mathrm{n-g}}\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{x_{i,j}} - \overline{\mathrm{x}}_{\mathrm{j}})^2$$

is $\hat{\sigma}_g$ for the jth variable x_j. Then the sample covariance matrix of the $\boldsymbol{w}_i$ is the sample correlation matrix of the $\boldsymbol{x}_i$:

$$\hat{\boldsymbol{\rho}}_{\boldsymbol{x}} = \boldsymbol{R}_{\boldsymbol{x}} = (r_{ij}) = \frac{\boldsymbol{W}_g^T \boldsymbol{W}_g}{n-g}$$

where r_{ij} is the sample correlation of x_i and x_j. Thus the sample correlation matrix $\boldsymbol{R}_{\boldsymbol{x}}$ does not depend on g. Let $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$ where $\overline{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1}$. Since the R software tends to use $g = 0$, let $\boldsymbol{W} = \boldsymbol{W}_0$. Note that $n \times p$ matrix $\boldsymbol{W}$ does not include a vector $\boldsymbol{1}$ of ones. Then regression through the origin is used for the model

$$\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{\epsilon} \tag{6.18}$$

where $\boldsymbol{Z} = (Z_1, ..., Z_n)^T$ and $\boldsymbol{\eta} = (\eta_1, ..., \eta_p)^T$. The vector of fitted values $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$.

Remark 6.12. i) Interest is in model (6.5): estimate $\hat{Y}_f$ and $\hat{\boldsymbol{\beta}}$. For many regression estimators, a method is needed so that everyone who uses the same units of measurements for the predictors and Y gets the same $(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{\beta}})$. Equation (6.18) is a commonly used method for achieving this goal. Suppose $g = 0$. The method of moments estimator of the variance σ_w^2 is

$$\hat{\sigma}_{g=0}^2 = S_M^2 = \frac{1}{n}\sum_{i=1}^{n}(w_i - \overline{w})^2.$$

When data x_i are standardized to have $\overline{w} = 0$ and $S_M^2 = 1$, the standardized data w_i has no units. ii) Hence the estimators $\hat{\boldsymbol{Z}}$ and $\hat{\boldsymbol{\eta}}$ do not depend on the units of measurement of the x_i if standardized data and Equation (6.18) are used. Linear combinations of the $\boldsymbol{w}_i$ are linear combinations of the $\boldsymbol{x}_i$. Thus the estimators $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{\beta}}$ are obtained using $\hat{\boldsymbol{Z}}$, $\hat{\boldsymbol{\eta}}$, and $\overline{\boldsymbol{Y}}$. The linear transformation to obtain $(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{\beta}})$ from $(\hat{\boldsymbol{Z}}, \hat{\boldsymbol{\eta}})$ is unique for a given set of units of measurements for the x_i and Y. Hence everyone using the same units of

measurements gets the same $(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{\beta}})$. iii) Also, since $\overline{W}_j = 0$ and $S_{M,j}^2 = 1$, the standardized predictor variables have similar spread, and the magnitude of $\hat{\eta}_i$ is a measure of the importance of the predictor variable W_j for predicting Y.

Definition 6.18. Consider model (6.4) $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$. If $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$, where the $n \times q$ matrix $\boldsymbol{W}$ has full rank $q = p - 1$, then the *OLS estimator*

$$\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$$

minimizes the OLS criterion $Q_{OLS}(\boldsymbol{\eta}) = \boldsymbol{r}(\boldsymbol{\eta})^T\boldsymbol{r}(\boldsymbol{\eta})$ over all vectors $\boldsymbol{\eta} \in \mathbb{R}^{p-1}$. The vector of *predicted* or *fitted values* $\widehat{\boldsymbol{Z}}_{OLS} = \boldsymbol{W}\hat{\boldsymbol{\eta}}_{OLS} = \boldsymbol{H}\boldsymbol{Z}$ where $\boldsymbol{H} = \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T$. The vector of residuals $\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{Z}, \boldsymbol{W}) = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Z}$.

For model (6.4) $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$, let $\boldsymbol{x} = (1\ \boldsymbol{u})^T$, and let $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{\epsilon}$. Assume that the sample correlation matrix

$$\boldsymbol{R}_{\boldsymbol{u}} = \frac{\boldsymbol{W}^T\boldsymbol{W}}{n} \xrightarrow{P} \boldsymbol{V}^{-1}. \tag{6.19}$$

Note that $\boldsymbol{V}^{-1} = \boldsymbol{\rho}_{\boldsymbol{u}}$, the population correlation matrix of the nontrivial predictors $\boldsymbol{u}_i$, if the $\boldsymbol{u}_i$ are a random sample from a population. Let $\boldsymbol{H} = \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T = (h_{ij})$, and assume that $\max_{i=1,\ldots,n} h_{ii} \xrightarrow{P} 0$ as $n \to \infty$. Section 6.7 examines whether the OLS estimator satisfies

$$\boldsymbol{u}_n = \sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V}). \tag{6.20}$$

Assume that the sample correlation matrix

$$\boldsymbol{R}_{\boldsymbol{u}} = \frac{\boldsymbol{W}^T\boldsymbol{W}}{n} \xrightarrow{P} \boldsymbol{V}^{-1}. \tag{6.21}$$

Note that $\boldsymbol{V}^{-1} = \boldsymbol{\rho}_{\boldsymbol{u}}$, the population correlation matrix of the nontrivial predictors $\boldsymbol{u}_i$, if the $\boldsymbol{u}_i$ are a random sample from a population. Let $\boldsymbol{H} = \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T = (h_{ij})$, and assume that $\max_{i=1,\ldots,n} h_{ii} \xrightarrow{P} 0$ as $n \to \infty$. Then by Theorem 6.1 (the OLS CLT), the OLS estimator satisfies

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V}). \tag{6.22}$$

Definition 6.19. Consider the MLR model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. Let $\boldsymbol{b}$ be a $(p-1) \times 1$ vector. Then the fitted value $\hat{Z}_i(\boldsymbol{b}) = \boldsymbol{w}_i^T\boldsymbol{b}$ and the residual $r_i(\boldsymbol{b}) = Z_i - \hat{Z}_i(\boldsymbol{b})$. The vector of fitted values $\hat{\boldsymbol{Z}}(\boldsymbol{b}) = \boldsymbol{W}\boldsymbol{b}$ and the vector of residuals $\boldsymbol{r}(\boldsymbol{b}) = \boldsymbol{Z} - \hat{\boldsymbol{Z}}(\boldsymbol{b})$.

6.3.1 Ridge Regression

Definition 6.20. a) Consider fitting the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Let $\lambda_{1,n} \geq 0$ be a constant. One *ridge regression estimator* $\hat{\boldsymbol{\beta}}_R$ minimizes the *ridge regression criterion*

$$Q_R(\boldsymbol{\beta}) = \frac{1}{a}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p}\beta_i^2 \tag{6.23}$$

over all vectors $\boldsymbol{\beta} \in \mathbb{R}^p$. Then

$$\hat{\boldsymbol{\beta}}_R = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{Y}. \tag{6.24}$$

The residual sum of squares $RSS(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$. The ridge regression vector of fitted values is $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Y}}_R = \boldsymbol{X}\hat{\boldsymbol{\beta}}_R$, and the ridge regression vector of residuals $\boldsymbol{r}_R = \boldsymbol{r}(\hat{\boldsymbol{\beta}}_R) = \boldsymbol{Y} - \hat{\boldsymbol{Y}}_R$.

b) Another *ridge regression estimator* $\tilde{\boldsymbol{\beta}}_{RR}$ minimizes the *ridge regression criterion*

$$Q_{RR}(\boldsymbol{\beta}) = \frac{1}{a}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{\lambda_{1,n}}{a}\sum_{i=2}^{p}\beta_i^2$$

over all vectors $\boldsymbol{\beta} \in \mathbb{R}^p$.

The following identity from Gunst and Mason (1980, p. 342) is useful for ridge regression inference: $\hat{\boldsymbol{\beta}}_R = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$

$$= (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

$$= (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{A}_n\hat{\boldsymbol{\beta}}_{OLS} =$$

$$[\boldsymbol{I}_p - \lambda_{1,n}(\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}]\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{B}_n\hat{\boldsymbol{\beta}}_{OLS} =$$

$$\hat{\boldsymbol{\beta}}_{OLS} - \frac{\lambda_{1n}}{n}n(\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}\hat{\boldsymbol{\beta}}_{OLS}$$

since $\boldsymbol{A}_n - \boldsymbol{B}_n = \boldsymbol{0}$, where $\boldsymbol{A}_n = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}(\boldsymbol{X}^T\boldsymbol{X}) = \boldsymbol{B}_n$
$= \boldsymbol{I}_p - \lambda_{1,n}(\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}$. See Problem 6.3. Assume

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \to \boldsymbol{V}^{-1}$$

as $n \to \infty$. If $\lambda_{1,n}/n \to 0$ then

$$\frac{\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p}{n} \xrightarrow{P} \boldsymbol{V}^{-1}, \quad \text{and} \quad \mathrm{n}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda_{1,\mathrm{n}}\boldsymbol{I}_{\mathrm{p}})^{-1} \xrightarrow{\mathrm{P}} \boldsymbol{V}.$$

Note that

$$\boldsymbol{A}_n = \boldsymbol{A}_{n,\lambda} = \left(\frac{\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p}{n} \right)^{-1} \frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \xrightarrow{P} \boldsymbol{V}\ \boldsymbol{V}^{-1} = \boldsymbol{I}_p$$

if $\lambda_{1,n}/n \to 0$ since matrix inversion is a continuous function of a positive definite matrix. See, for example, Bhatia et al. (1990), Stewart (1969), and Severini (2005, pp. 348-349).

For model selection, the M values of $\lambda = \lambda_{1,n}$ are denoted by $\lambda_1, \lambda_2, ..., \lambda_M$ where $\lambda_i = \lambda_{1,n,i}$ depends on n for $i = 1, ..., M$. If λ_s corresponds to the model selected, then $\hat{\lambda}_{1,n} = \lambda_s$. The following theorem shows that ridge regression and the OLS full model are asymptotically equivalent if $\hat{\lambda}_{1,n} = o_P(n^{1/2})$ so $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$.

Theorem 6.10, RR CLT (Ridge Regression Central Limit Theorem. Assume p is fixed and that the conditions of the OLS CLT Theorem Equation (6.8) hold for the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$.

a) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V}).$$

b) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$ then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}) \xrightarrow{D} N_p(-\tau\boldsymbol{V}\boldsymbol{\beta}, \sigma^2\boldsymbol{V}).$$

Proof: If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$, then by the above Gunst and Mason (1980) identity,

$$\hat{\boldsymbol{\beta}}_R = [\boldsymbol{I}_p - \hat{\lambda}_{1,n}(\boldsymbol{X}^T\boldsymbol{X} + \hat{\lambda}_{1,n}\boldsymbol{I}_p)^{-1}]\hat{\boldsymbol{\beta}}_{OLS}.$$

Hence

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_{OLS} + \hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) =$$

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) - \sqrt{n}\frac{\hat{\lambda}_{1,n}}{n} n(\boldsymbol{X}^T\boldsymbol{X} + \hat{\lambda}_{1,n}\boldsymbol{I}_p)^{-1}\hat{\boldsymbol{\beta}}_{OLS}$$

$$\xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V}) - \tau\boldsymbol{V}\boldsymbol{\beta} \sim N_p(-\tau\boldsymbol{V}\boldsymbol{\beta}, \sigma^2\boldsymbol{V}). \quad \square$$

For p fixed, Knight and Fu (2000) note i) that $\hat{\boldsymbol{\beta}}_R$ is a consistent estimator of $\boldsymbol{\beta}$ if $\lambda_{1,n} = o(n)$ so $\lambda_{1,n}/n \to 0$ as $n \to \infty$, ii) OLS and ridge regression are asymptotically equivalent if $\lambda_{1,n}/\sqrt{n} \to 0$ as $n \to \infty$, iii) ridge regression is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ if $\lambda_{1,n} = O(\sqrt{n})$ (so $\lambda_{1,n}/\sqrt{n}$ is bounded), and iv) if $\lambda_{1,n}/\sqrt{n} \to \tau \geq 0$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}) \xrightarrow{D} N_p(-\tau\boldsymbol{V}\boldsymbol{\beta}, \sigma^2\boldsymbol{V}).$$

Hence the bias can be considerable if $\tau \neq 0$. If $\tau = 0$, then OLS and ridge regression have the same limiting distribution. The proof of the Theorem 6.10 is due Pelawa Watagoda and Olive (2021b).

Even if p is fixed, there are several problems with ridge regression inference if $\hat{\lambda}_{1,n}$ is selected, e.g. after 10-fold cross validation. For OLS forward selection, the probability that the model I_{min} underfits goes to zero, and each model with $S \subseteq I$ produced a $\sqrt{n}$ consistent estimator $\hat{\boldsymbol{\beta}}_{I,0}$ of $\boldsymbol{\beta}$. Ridge regression with 10-fold CV often shrinks $\hat{\boldsymbol{\beta}}_R$ too much if both i) the number of population active predictors $k_S = a_S - 1$ in Equation (6.41) is greater than about 20, and ii) the predictors are highly correlated. If p is fixed and $\lambda_{1,n} = o_P(\sqrt{n})$, then the OLS full model and ridge regression are asymptotically equivalent, but much larger sample sizes may be needed for the normal approximation to be good for ridge regression since the ridge regression estimator can have large bias for moderate n. Ten fold CV does not appear to guarantee that $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$ or $\hat{\lambda}_{1,n}/n \xrightarrow{P} 0$.

Ridge regression can be a lot better than the OLS full model if i) $\boldsymbol{X}^T\boldsymbol{X}$ is singular or ill conditioned or ii) n/p is small. Ridge regression can be much faster than forward selection if $M = 100$ and n and p are large.

Warning. The R functions `glmnet` and `cv.glmnet` do ridge regression using Definition 6.20 b).

6.3.2 Lasso

Definition 6.21. Consider fitting the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. The *lasso estimator* $\hat{\boldsymbol{\beta}}_L$ minimizes the *lasso criterion*

$$Q_L(\boldsymbol{\beta}) = \frac{1}{a}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{\lambda_{1,n}}{a}\sum_{i=2}^{p}|\beta_i| \tag{6.25}$$

over all vectors $\boldsymbol{\beta} \in \mathbb{R}^p$ where $\lambda_{1,n} \geq 0$ and $a > 0$ are known constants with $a = 1, 2, n$, and $2n$ are common. The residual sum of squares $RSS(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ if $\boldsymbol{X}$ has full rank p. The lasso vector of fitted values is $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Y}}_L = \boldsymbol{X}\hat{\boldsymbol{\beta}}_L$, and the lasso vector of residuals $\boldsymbol{r}(\hat{\boldsymbol{\beta}}_L) = \boldsymbol{Y} - \hat{\boldsymbol{Y}}_L$.

The following identity from Efron and Hastie (2016, p. 308), for example, is useful for inference for the lasso estimator $\hat{\boldsymbol{\eta}}_L$:

$$\frac{-1}{n}\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_L) + \frac{\lambda_{1,n}}{2n}\boldsymbol{s}_n = \boldsymbol{0} \quad \text{or} \quad -\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{L}}) + \frac{\lambda_{1,\mathrm{n}}}{2}\boldsymbol{s}_{\mathrm{n}} = \boldsymbol{0}$$

where $s_{in} \in [-1, 1]$ and $s_{in} = \text{sign}(\hat{\beta}_{i,L})$ if $\hat{\beta}_{i,L} \neq 0$. Here $\text{sign}(\beta_i) = 1$ if $\beta_i > 0$ and $\text{sign}(\beta_i) = -1$ if $\beta_i < 0$. Note that $\boldsymbol{s}_n = \boldsymbol{s}_{n,\hat{\boldsymbol{\beta}}_L}$ depends on $\hat{\boldsymbol{\beta}}_L$.

Thus $\hat{\boldsymbol{\beta}}_L$

$$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} - \frac{\lambda_{1,n}}{2n}\, n(\boldsymbol{X}^T\boldsymbol{X})^{-1}\, \boldsymbol{s}_n = \hat{\boldsymbol{\beta}}_{OLS} - \frac{\lambda_{1,n}}{2n}\, n(\boldsymbol{X}^T\boldsymbol{X})^{-1}\, \boldsymbol{s}_n.$$

If none of the elements of $\boldsymbol{\beta}$ are zero, and if $\hat{\boldsymbol{\beta}}_L$ is a consistent estimator of $\boldsymbol{\beta}$, then $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\beta}}$. If $\lambda_{1,n}/\sqrt{n} \to 0$, then OLS and lasso are asymptotically equivalent even if $\boldsymbol{s}_n$ does not converge to a vector $\boldsymbol{s}$ as $n \to \infty$ since $\boldsymbol{s}_n$ is bounded. For model selection, the M values of λ are denoted by $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_M$ where $\lambda_i = \lambda_{1,n,i}$ depends on n for $i = 1, ..., M$. Also, λ_M is the smallest value of λ such that $\hat{\boldsymbol{\beta}}_{\lambda_M} = \mathbf{0}$. Hence $\hat{\boldsymbol{\beta}}_{\lambda_i} \neq \mathbf{0}$ for $i < M$. If λ_s corresponds to the model selected, then $\hat{\lambda}_{1,n} = \lambda_s$. The following theorem shows that lasso and the OLS full model are asymptotically equivalent if $\hat{\lambda}_{1,n} = o_P(n^{1/2})$ so $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$: thus $\sqrt{n}(\hat{\boldsymbol{\beta}}_L - \hat{\boldsymbol{\beta}}_{OLS}) = o_p(1)$.

Theorem 6.11, Lasso CLT. Assume p is fixed and that the conditions of the OLS CLT Theorem Equation (6.8) hold for the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$.

a) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \sigma^2\boldsymbol{V}).$$

b) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$ and $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\beta}}$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}) \xrightarrow{D} N_p\left(\frac{-\tau}{2}\boldsymbol{V}\boldsymbol{s}, \sigma^2\boldsymbol{V}\right).$$

Proof. If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$ and $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\beta}}$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_L - \hat{\boldsymbol{\beta}}_{OLS} + \hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) =$$

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) - \sqrt{n}\frac{\lambda_{1,n}}{2n} n(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{s}_n \xrightarrow{D} N_p(\mathbf{0}, \sigma^2\boldsymbol{V}) - \frac{\tau}{2}\boldsymbol{V}\boldsymbol{s}$$

$$\sim N_p\left(\frac{-\tau}{2}\boldsymbol{V}\boldsymbol{s}, \sigma^2\boldsymbol{V}\right)$$

since under the OLS CLT, $n(\boldsymbol{X}^T\boldsymbol{X})^{-1} \xrightarrow{P} \boldsymbol{V}$.

Part a) does not need $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s}$ as $n \to \infty$, since $\boldsymbol{s}_n$ is bounded. □

Suppose p is fixed. Knight and Fu (2000) note i) that $\hat{\boldsymbol{\beta}}_L$ is a consistent estimator of $\boldsymbol{\eta}$ if $\lambda_{1,n} = o(n)$ so $\lambda_{1,n}/n \to 0$ as $n \to \infty$, ii) OLS and lasso are asymptotically equivalent if $\lambda_{1,n} \to \infty$ too slowly as $n \to \infty$ (e.g. if $\lambda_{1,n} = \lambda$ is fixed), iii) lasso is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ if $\lambda_{1,n} = O(\sqrt{n})$ (so $\lambda_{1,n}/\sqrt{n}$ is bounded). Note that Theorem 6.11 shows that OLS and lasso are

asymptotically equivalent if $\lambda_{1,n}/\sqrt{n} \to 0$ as $n \to 0$. The proof of Theorem 6.11 is due Pelawa Watagoda and Olive (2021b).

6.3.3 The Elastic Net

Following Hastie et al. (2015, p. 57), let $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_S^T)^T$, let $\lambda_{1,n} \geq 0$, and let $\alpha \in [0, 1]$. Let

$$RSS(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2.$$

For a $k \times 1$ vector $\boldsymbol{\eta}$, the squared (Euclidean) L_2 norm $\|\boldsymbol{\eta}\|_2^2 = \boldsymbol{\eta}^T\boldsymbol{\eta} = \sum_{i=1}^k \eta_i^2$ and the L_1 norm $\|\boldsymbol{\eta}\|_1 = \sum_{i=1}^k |\eta_i|$.

Definition 6.22. The *elastic net* estimator $\hat{\boldsymbol{\beta}}_{EN}$ minimizes the criterion

$$Q_{EN}(\boldsymbol{\beta}) = \frac{1}{2}RSS(\boldsymbol{\beta}) + \lambda_{1,n}\left[\frac{1}{2}(1-\alpha)\|\boldsymbol{\beta}_S\|_2^2 + \alpha\|\boldsymbol{\beta}_S\|_1\right], \text{ or} \tag{6.26}$$

$$Q_2(\boldsymbol{\beta}) = RSS(\boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}_S\|_2^2 + \lambda_2\|\boldsymbol{\beta}_S\|_1 \tag{6.27}$$

where $0 \leq \alpha \leq 1$, $\lambda_1 = (1-\alpha)\lambda_{1,n}$ and $\lambda_2 = 2\alpha\lambda_{1,n}$.

Note that $\alpha = 1$ corresponds to lasso (using $\lambda_{a=0.5}$), and $\alpha = 0$ corresponds to ridge regression. For $\alpha < 1$ and $\lambda_{1,n} > 0$, the optimization problem is *strictly convex* with a unique solution. The elastic net is due to Zou and Hastie (2005). It has been observed that the elastic net can have much better prediction accuracy than lasso when the predictors are highly correlated.

The following theorem is probably for the elastic net estimator that uses the usual ridge regression estimator of Definition 6.20 b), rather that the ridge regression estimator of Definition 6.20 c). Hence Equation (6.27) would need to be modified. Following Jia and Yu (2010), by standard Karush-Kuhn-Tucker (KKT) conditions for convex optimality for the "modified Equation (6.27)," $\hat{\boldsymbol{\beta}}_{EN}$ is optimal if

$$2\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}_{EN} - 2\boldsymbol{X}^T\boldsymbol{Y} + 2\lambda_1\hat{\boldsymbol{\beta}}_{EN} + \lambda_2\boldsymbol{s}_n = \boldsymbol{0}, \text{ or}$$

$$(\boldsymbol{X}^T\boldsymbol{X} + \lambda_1\boldsymbol{I}_p)\hat{\boldsymbol{\beta}}_{EN} = \boldsymbol{X}^T\boldsymbol{Y} - \frac{\lambda_2}{2}\boldsymbol{s}_n, \text{ or}$$

$$\hat{\boldsymbol{\beta}}_{EN} = \hat{\boldsymbol{\beta}}_R - n(\boldsymbol{X}^T\boldsymbol{X} + \lambda_1\boldsymbol{I}_p)^{-1}\frac{\lambda_2}{2n}\boldsymbol{s}_n. \tag{6.28}$$

Hence

$$\hat{\boldsymbol{\beta}}_{EN} = \hat{\boldsymbol{\beta}}_{OLS} - \frac{\lambda_1}{n}\, n(\boldsymbol{X}^T\boldsymbol{X} + \lambda_1\boldsymbol{I}_p)^{-1}\, \hat{\boldsymbol{\beta}}_{OLS} - \frac{\lambda_2}{2n}\, n(\boldsymbol{X}^T\boldsymbol{X} + \lambda_1\boldsymbol{I}_p)^{-1}\, \boldsymbol{s}_n$$

$$= \hat{\boldsymbol{\beta}}_{OLS} - n(\boldsymbol{X}^T\boldsymbol{X} + \lambda_1\boldsymbol{I}_p)^{-1}\,[\frac{\lambda_1}{n}\hat{\boldsymbol{\beta}}_{OLS} + \frac{\lambda_2}{2n}\boldsymbol{s}_n].$$

Note that if $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau$ and $\hat{\alpha} \xrightarrow{P} \psi$, then $\hat{\lambda}_1/\sqrt{n} \xrightarrow{P} (1-\psi)\tau$ and $\hat{\lambda}_2/\sqrt{n} \xrightarrow{P} 2\psi\tau$. The following theorem shows elastic net is asymptotically equivalent to the OLS full model if $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$. Note that we get the RR CLT if $\psi = 0$ and the lasso CLT (using $2\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 2\tau$) if $\psi = 1$. Under these conditions,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{EN} - \boldsymbol{\beta}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) - n(\boldsymbol{X}^T\boldsymbol{X} + \hat{\lambda}_1\boldsymbol{I}_p)^{-1}\,[\frac{\hat{\lambda}_1}{\sqrt{n}}\hat{\boldsymbol{\beta}}_{OLS} + \frac{\hat{\lambda}_2}{2\sqrt{n}}\boldsymbol{s}_n].$$

The following theorem is due to Slawski et al. (2010), and summarized in Pelawa Watagoda and Olive (2021b).

Theorem 6.12, Elastic Net CLT. Assume p is fixed and that the conditions of the OLS CLT Equation (6.8) hold for the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$.

a) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{EN} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V}).$$

b) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$, $\hat{\alpha} \xrightarrow{P} \psi \in [0,1]$, and $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\hat{\boldsymbol{\beta}}}$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{EN} - \boldsymbol{\beta}) \xrightarrow{D} N_p\left(-\boldsymbol{V}[(1-\psi)\tau\boldsymbol{\beta} + \psi\tau\boldsymbol{s}], \sigma^2\boldsymbol{V}\right).$$

Proof. By the above remarks and the RR CLT Theorem 6.10,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{EN} - \boldsymbol{\beta}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_{EN} - \hat{\boldsymbol{\beta}}_R + \hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}) + \sqrt{n}(\hat{\boldsymbol{\beta}}_{EN} - \hat{\boldsymbol{\beta}}_R)$$

$$\xrightarrow{D} N_p\left(-(1-\psi)\tau\boldsymbol{V}\boldsymbol{\beta}, \sigma^2\boldsymbol{V}\right) \quad - \quad \frac{2\psi\tau}{2}\boldsymbol{V}\boldsymbol{s}$$

$$\sim N_p\left(-\boldsymbol{V}[(1-\psi)\tau\boldsymbol{\beta} + \psi\tau\boldsymbol{s}], \sigma^2\boldsymbol{V}\right).$$

The mean of the normal distribution is $\boldsymbol{0}$ under a) since $\hat{\alpha}$ and $\boldsymbol{s}_n$ are bounded. □

Warning. The above theorem uses the ridge regression estimator (6.23). The *R* functions `glmnet` and `cv.glmnet` do ridge regression and elastic net using Definition 6.20 b).

6.3.4 Ridge Type Regression Estimators

See Jin and Olive (2023).

6.4 MLR with Heterogeneity

A multiple linear regression model with heterogeneity is

$$Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i \tag{6.29}$$

for $i = 1, ..., n$ where the e_i are independent with $E(e_i) = 0$ and $V(e_i) = \sigma_i^2$. In matrix form, this model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. Also $E(\boldsymbol{e}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}) = \boldsymbol{\Sigma_e} = diag(\sigma_i^2) = diag(\sigma_1^2, ..., \sigma_n^2)$ is an $n \times n$ positive definite matrix. In Section 2, the constant variance assumption was used: $\sigma_i^2 = \sigma^2$ for all i. Hence heterogeneity means that the constant variance assumption does not hold. A common assumption is that the $e_i = \sigma_i \epsilon_i$ where the ϵ_i are independent and identically distributed (iid) with $V(\epsilon_i) = 1$. See, for example, Zhou, Cook, and Zou (2023).

Weighted least squares (WLS) would be useful if the σ_i^2 were known. Since the σ_i^2 are not known, ordinary least squares (OLS) is often used. The OLS theory for MLR with heterogeneity often assume iid cases. For the following theorem, see Romano and Wolf (2017), Freedman (1981), and White (1980).

Theorem 6.13. Assume $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ for $i = 1, ..., n$ where the cases $(Y_i, \boldsymbol{x}_i^T)^T$ are iid with "fourth moments," $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, the $e_i = e_i(\boldsymbol{x}_i)$ are independent, $E[e_i|\boldsymbol{x}_i] = 0$, $\boldsymbol{V}^{-1} = E[\boldsymbol{x}_i\boldsymbol{x}_i^T]$, $E[e_i^2|\boldsymbol{x}_i] = v(\boldsymbol{x}_i) = \sigma_i^2$, $Cov[\boldsymbol{e}|\boldsymbol{X}] = diag(v(\boldsymbol{x}_1), ..., v(\boldsymbol{x}_n))$ and $\boldsymbol{\Omega} = E[v(\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T] = E[e_i^2\boldsymbol{x}_i\boldsymbol{x}_i^T]$. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}\boldsymbol{\Omega}\boldsymbol{V}). \tag{6.30}$$

Remark 6.13. a) White (1980) showed that the iid cases assumption can be weakened. Assume the cases are independent,

$$\boldsymbol{V}_n = \frac{1}{n}\sum_{i=1}^{n} E[\boldsymbol{x}_i\boldsymbol{x}_i^T] \xrightarrow{P} \boldsymbol{V}^{-1},$$

and

$$\boldsymbol{\Omega}_n = \frac{1}{n}\sum_{i=1}^{n} E[e_i^2\boldsymbol{x}_i\boldsymbol{x}_i^T] \xrightarrow{P} \boldsymbol{\Omega}.$$

Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}\boldsymbol{\Omega}\boldsymbol{V}).$$

b) Under the assumptions of Theorem 6.13,

$$\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T \xrightarrow{P} \boldsymbol{V}^{-1}.$$

Let $\boldsymbol{D} = diag(\sigma_1^2, ..., \sigma_n^2) = \boldsymbol{\Sigma_e}$ and $\hat{\boldsymbol{D}} = diag(r_1^2, ..., r_n^2)$ where r_i^2 is the ith residual from OLS regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$. Then $\hat{\boldsymbol{D}}$ is not a consistent estimator of $\boldsymbol{D}$. The following theorem, due to White (1980), shows that $\hat{\boldsymbol{D}}$ can be used to get a consistent estimator of $\boldsymbol{\Omega}$. This result leads to the sandwich estimators.

Theorem 6.14. Under strong regularity conditions,

$$\frac{1}{n}(\boldsymbol{X}^T\hat{\boldsymbol{D}}\boldsymbol{X}) \xrightarrow{P} \boldsymbol{\Omega} \text{ and } \frac{1}{\mathrm{n}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}) \xrightarrow{\mathrm{P}} \boldsymbol{\Omega}.$$

Hence

$$n(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\hat{\boldsymbol{D}}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1} \xrightarrow{P} \boldsymbol{V}\boldsymbol{\Omega}\boldsymbol{V}.$$

Rajapaksha and Olive (2024) compare several methods for inference for model (6.19). The nonparametric bootstrap worked well. Olive et al. (2024) proved that the OPLS and MMLE estimators, described in the following two sections, are also useful for model (6.29).

6.5 OPLS

Cook, Helland, and Su (2013) showed that the OPLS estimator $\hat{\boldsymbol{\beta}}_{OPLS}$ estimates $\boldsymbol{\beta}_{OPLS}$, and that the OPLS estimator can be computed from the OLS simple linear regression (SLR) of Y on $W = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}^T\boldsymbol{x}$, giving $\hat{Y} = \hat{\alpha}_{OPLS} + \hat{\lambda}W = \hat{\alpha}_{OPLS} + \hat{\boldsymbol{\beta}}_{OPLS}^T\boldsymbol{x}$. Also see Basa et al. (2024) and Wold (1975). Also see Remark 6.5.

Definition 6.23. The *one component partial least squares (OPLS) estimator* $\hat{\boldsymbol{\beta}}_{OPLS} = \hat{\lambda}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ estimates $\lambda\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{\beta}_{OPLS}$ where

$$\lambda = \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T\boldsymbol{\Sigma}_{\boldsymbol{x}Y}}{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x}Y}} \text{ and } \hat{\lambda} = \frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}}{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}} \tag{6.31}$$

for $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} \neq \boldsymbol{0}$. If $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{0}$, then $\boldsymbol{\beta}_{OPLS} = \boldsymbol{0}$.

The following Olive and Zhang (2024) theorem gives some large sample theory for $\hat{\boldsymbol{\eta}} = \widehat{\mathrm{Cov}}(\boldsymbol{x}, Y)$. This theory needs $\boldsymbol{\eta} = \boldsymbol{\eta}_{OPLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ to exist for $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ to be a consistent estimator of $\boldsymbol{\eta}$. Let $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$ and let $\boldsymbol{w}_i$ and $\boldsymbol{z}_i$ be defined below where

$$\mathrm{Cov}(\boldsymbol{w}_i) = \boldsymbol{\Sigma}_{\boldsymbol{w}} = E[(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})^T(Y_i - \mu_Y)^2)] - \boldsymbol{\Sigma}_{\boldsymbol{x}Y}\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T.$$

Then the low order moments are needed for $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}$ to be a consistent estimator of $\boldsymbol{\Sigma}_{\boldsymbol{W}}$. The theory uses milder regularity conditions than the theory in the previous literature. The theory can be used for testing, including some high dimensional tests for low dimensional quantities such as $H_O : \beta_i = 0$ or $H_0 : \beta_i - \beta_j = 0$. These tests depended on iid cases, but not on linearity or the constant variance assumption. Data splitting uses model selection (variable selection is a special case) to reduce the high dimensional problem to a low dimensional problem. Olive et al. (2024) gave alternative proofs, and showed that the results hold for multiple linear regression with heterogeneity.

Theorem 6.15. Assume the cases $(\boldsymbol{x}_i^T, Y_i)^T$ are iid. Assume $E(x_{ij}^k \; Y_i^m)$ exist for $j = 1, ..., p$ and $k, m = 0, 1, 2$. Let $\boldsymbol{\mu}_{\boldsymbol{x}} = E(\boldsymbol{x})$ and $\mu_Y = E(Y)$. Let $\boldsymbol{w}_i = (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(Y_i - \mu_Y)$ with sample mean $\overline{\boldsymbol{w}}_n$. Let $\boldsymbol{\eta} = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. Then a)

$$\sqrt{n}(\overline{\boldsymbol{w}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{W}}), \; \sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{W}}), \tag{6.32}$$

$$\text{and} \;\; \sqrt{\mathrm{n}}(\tilde{\boldsymbol{\eta}}_{\mathrm{n}} - \boldsymbol{\eta}) \xrightarrow{\mathrm{D}} \mathrm{N_p}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{W}}).$$

b) Let $\boldsymbol{z}_i = \boldsymbol{x}_i(Y_i - \overline{Y}_n)$ and $\boldsymbol{v}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n)(Y_i - \overline{Y}_n)$. Then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{W}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}} + O_P(n^{-1/2}) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{V}} + O_P(n^{-1/2})$. Hence $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{W}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{Z}} + O_P(n^{-1/2}) = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{V}} + O_P(n^{-1/2})$.
c) Let $\boldsymbol{A}$ be a $k \times p$ full rank constant matrix with $k \leq p$, assume $H_0 : \boldsymbol{A}\boldsymbol{\beta}_{OPLS} = \boldsymbol{0}$ is true, and assume $\hat{\lambda} \xrightarrow{P} \lambda \neq 0$. Then

$$\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{OPLS} - \boldsymbol{\beta}_{OPLS}) \xrightarrow{D} N_k(\boldsymbol{0}, \lambda^2 \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{W}}\boldsymbol{A}^T). \tag{6.33}$$

Proof. a) Note that $\sqrt{n}(\overline{\boldsymbol{w}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{W}})$ by the multivariate central limit theorem since the $\boldsymbol{w}_i$ are iid with $E(\boldsymbol{w}_i) = \boldsymbol{\eta} = \mathrm{Cov}(\boldsymbol{x}, Y)$ and $\mathrm{Cov}(\boldsymbol{w}) = \boldsymbol{\Sigma}_{\boldsymbol{W}}$. Now $n\tilde{\boldsymbol{\eta}}_n =$

$$\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}} + \boldsymbol{\mu}_{\boldsymbol{x}} - \overline{\boldsymbol{x}})(Y_i - \mu_Y + \mu_Y - \overline{Y}) = \sum_i (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(Y_i - \mu_Y)$$

$$+ \sum_i (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(\mu_Y - \overline{Y}) + (\boldsymbol{\mu}_{\boldsymbol{x}} - \overline{\boldsymbol{x}}) \sum_i (Y_i - \mu_Y) + n(\boldsymbol{\mu}_{\boldsymbol{x}} - \overline{\boldsymbol{x}})(\mu_Y - \overline{Y})$$

$$= \sum_i \boldsymbol{w}_i - n\boldsymbol{a}_n - n\boldsymbol{a}_n + n\boldsymbol{a}_n = \sum_i \boldsymbol{w}_i - n(\boldsymbol{\mu}_{\boldsymbol{x}} - \overline{\boldsymbol{x}})(\mu_Y - \overline{Y}).$$

$$\text{Thus} \;\; \sqrt{\mathrm{n}}\tilde{\boldsymbol{\eta}}_{\mathrm{n}} = \sqrt{\mathrm{n}}\frac{1}{\mathrm{n}}\sum_{\mathrm{i}} \boldsymbol{w}_{\mathrm{i}} - \frac{\sqrt{\mathrm{n}}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_{\boldsymbol{x}})\sqrt{\mathrm{n}}(\overline{\mathrm{Y}} - \mu_{\mathrm{Y}})}{\sqrt{\mathrm{n}}} = \sqrt{\mathrm{n}}\, \overline{\boldsymbol{w}}_{\mathrm{n}} + \mathrm{o_P}(1).$$

$$\text{Hence } \sqrt{\mathrm{n}}(\tilde{\boldsymbol{\eta}}_{\mathrm{n}} - \boldsymbol{\eta}) = \sqrt{\mathrm{n}}(\overline{\boldsymbol{w}}_{\mathrm{n}} - \boldsymbol{\eta}) + \mathrm{o_P}(1).$$

$$\text{Thus } \sqrt{\mathrm{n}}(\tilde{\boldsymbol{\eta}}_{\mathrm{n}} - \boldsymbol{\eta}) \xrightarrow{\mathrm{D}} \mathrm{N_p}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{W}})$$

by Slutsky's theorem. Now

$$\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) = \sqrt{n}\left(\frac{n}{n-1}\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}\right) = \sqrt{n}\left(\frac{n}{n-1}\tilde{\boldsymbol{\eta}} - \frac{n}{n-1}\boldsymbol{\eta} + \frac{n}{n-1}\boldsymbol{\eta} - \boldsymbol{\eta}\right)$$

$$= \sqrt{n}\frac{n}{n-1}(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}) + \sqrt{n}\left(\frac{\boldsymbol{\eta}}{n-1}\right).$$

$$\text{Thus } \sqrt{\mathrm{n}}(\hat{\boldsymbol{\eta}}_{\mathrm{n}} - \boldsymbol{\eta}) \xrightarrow{\mathrm{D}} \mathrm{N_p}(\mathbf{0}, \boldsymbol{\Sigma_w}).$$

b) See Olive et al. (2024).

c) If H_0 is true, then $\boldsymbol{A\eta} = \mathbf{0}$, and

$$\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{OPLS} - \boldsymbol{\beta}_{OPLS}) = \sqrt{n}\boldsymbol{A}(\hat{\lambda}\hat{\boldsymbol{\eta}} - \hat{\lambda}\boldsymbol{\eta} + \hat{\lambda}\boldsymbol{\eta} - \boldsymbol{\beta}_{OPLS}) =$$

$$\hat{\lambda}\boldsymbol{A}\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) + \boldsymbol{A}\sqrt{n}(\hat{\lambda} - \lambda)\boldsymbol{\eta} = \boldsymbol{Z}_n + \boldsymbol{b}_n \xrightarrow{D} N_k(\mathbf{0}, \lambda^2 \boldsymbol{A\Sigma_w A}^T)$$

since $\boldsymbol{b}_n = \mathbf{0}$ when H_0 is true. □

In Theorems 6.15 and 6.16, the scalars λ and $\hat{\lambda}$ are given by Equation (6.31), $\boldsymbol{\eta} = (\eta_1, ..., \eta_p)^T$, and $\boldsymbol{\Sigma_\eta} = \boldsymbol{\Sigma_w}$. Results from Su and Cook (2012) and Olive et al. (2024), for example, show that elements of a sample covariance matrix can be stacked to get large sample theory. Then $\hat{\lambda}$ and $\hat{\boldsymbol{\eta}}$ can be stacked as in Theorem 6.16 by the multivariate delta method. Theorem 6.15 c) and Theorem 6.16 c) are equivalent with different notation. Currently $\boldsymbol{\Sigma}$ from Theorem 6.16 is difficult to estimate.

Theorem 6.16. Assume

$$\sqrt{n}\left(\begin{pmatrix}\hat{\lambda}\\ \hat{\boldsymbol{\eta}}\end{pmatrix} - \begin{pmatrix}\lambda\\ \boldsymbol{\eta}\end{pmatrix}\right) \xrightarrow{D} N_{p+1}\left(\begin{pmatrix}0\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\Sigma_\lambda & \boldsymbol{\Sigma}_{\lambda\boldsymbol{\eta}}\\ \boldsymbol{\Sigma}_{\boldsymbol{\eta}\lambda} & \boldsymbol{\Sigma_\eta}\end{pmatrix}\right) \sim N_{p+1}(\mathbf{0}, \boldsymbol{\Sigma}).$$

a) $\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma_\eta})$.

b) $\sqrt{n}(\hat{\lambda}\hat{\boldsymbol{\eta}} - \lambda\boldsymbol{\eta}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_{OPLS} - \boldsymbol{\beta}_{OPLS}) \xrightarrow{D} N_p\left(\mathbf{0}, \boldsymbol{D\Sigma D}^T\right)$ with $\boldsymbol{D} = [\boldsymbol{\eta} \;\; \lambda \boldsymbol{I}_p]$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix.

c) Let $\boldsymbol{A}$ be a $k \times p$ full rank constant matrix with $k \leq p$ and $\boldsymbol{A\beta}_{OPLS} = \mathbf{0} = \boldsymbol{A\eta}$. Then

$$\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OPLS} - \mathbf{0}) \xrightarrow{D} N_k\left(\mathbf{0}, \lambda^2 \boldsymbol{A\Sigma_\eta A}^T\right).$$

Proof. a) Follows by Equation (6.32) or since joint convergence in distribution implies marginal convergence in distribution.

b) Follows by the Multivariate Delta Method with

$$\boldsymbol{g}\begin{pmatrix}\lambda\\ \boldsymbol{\eta}\end{pmatrix} = \lambda\boldsymbol{\eta} =$$

$(\lambda\eta_1, ..., \lambda\eta_p)^T$, and the Jacobian matrix of partial derivatives $\boldsymbol{D} = \boldsymbol{D_g}$.

c) By b), $\sqrt{\mathrm{n}}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{\mathrm{OPLS}} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{\mathrm{D}} \mathrm{N_k}\left(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\right)$,

but $\boldsymbol{A}\boldsymbol{D} = [\boldsymbol{0} \;\; \lambda\boldsymbol{A}]$. Hence $\boldsymbol{A}\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}^T\boldsymbol{A}^T = \lambda^2\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{A}^T$. □

Some additional useful OPLS and OLS formulas are derived next if the cases are iid. Let $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$. Then $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \mathrm{Cov}(\boldsymbol{x}, Y) = \mathrm{Cov}(\boldsymbol{x})\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}$. Since $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS}$,

$$\boldsymbol{\beta}_{OPLS} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS},\ \boldsymbol{\beta}_{OPLS} = \lambda\mathrm{Cov}(\boldsymbol{x})\boldsymbol{\beta}_{OLS},\ \text{and}$$

$$\boldsymbol{\beta}_{OLS} = \frac{1}{\lambda}[\mathrm{Cov}(\boldsymbol{x})]^{-1}\boldsymbol{\beta}_{OPLS}.$$

6.6 MMLE

The marginal maximum likelihood estimator (MMLE or marginal least squares estimator) is due to Fan and Lv (2008) and Fan and Song (2010). This estimator computes the marginal regression of Y on x_i resulting in the estimator $(\hat{\alpha}_{i,M}, \hat{\beta}_{i,M})$ for $i = 1, ..., p$. Then $\hat{\boldsymbol{\beta}}_{MMLE} = (\hat{\beta}_{1,M}, ..., \hat{\beta}_{p,M})^T$. For multiple linear regression, the marginal estimators are the simple linear regression estimators, and $(\hat{\alpha}_{i,M}, \hat{\beta}_{i,M}) = (\hat{\alpha}_{i,SLR}, \hat{\beta}_{i,SLR})$. Hence

$$\hat{\boldsymbol{\beta}}_{MMLE} = [diag(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}})]^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}. \tag{6.34}$$

If the $\boldsymbol{t}_i$ are the predictors are scaled or standardized to have unit sample variances, then

$$\hat{\boldsymbol{\beta}}_{MMLE} = \hat{\boldsymbol{\beta}}_{MMLE}(\boldsymbol{t}, Y) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{t}Y} = \boldsymbol{I}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t}Y} = \hat{\boldsymbol{\eta}}_{OPLS}(\boldsymbol{t}, Y) \tag{6.35}$$

where $(\boldsymbol{t}, Y)$ denotes that Y was regressed on $\boldsymbol{t}$, and $\boldsymbol{I}$ is the $p \times p$ identity matrix. Olive et al. (2024) gave some large sample theory for the MMLE.

6.7 OLS with Scaled Predictors

See Olive (2024).

6.8 GLMs and Related Regression Models

Definition 6.24. A **parametric 1D regression model** is $Y|h(\boldsymbol{x}) \sim D(h(\boldsymbol{x}), \boldsymbol{\gamma})$ or $Y_i|h(\boldsymbol{x}_i) \sim D(h(\boldsymbol{x}_i), \boldsymbol{\gamma})$, where D is a parametric distribution

that depends on the $p \times 1$ vector of predictors $\boldsymbol{x}$ only through $SP = h(\boldsymbol{x})$, and $\boldsymbol{\gamma}$ is a $q \times 1$ vector of parameters.

An important special case is a *generalized additive model* (GAM) from Definition 6.4. Another large class of parametric 1D regression models uses $SP = h(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$ where $\hat{\boldsymbol{\beta}}$ is the MLE. Generalized linear models are a special case. Some important 1D regression models are defined below. The AER model is a 1D regression model that is not a not a parametric 1D regression model.

Definition 6.25. i) The **additive error regression (AER) model** $Y = SP + e$ has conditional mean function $E(Y|SP) = SP$ and conditional variance function $V(Y|SP) = \sigma^2 = V(e)$. The response plot of ESP versus Y and the residual plot of ESP versus $r = Y - \hat{Y}$ are used just as for multiple linear regression. The estimated model (conditional) mean function is the identity line $Y = ESP$. The *response transformation model* is $Y = t(Z) = SP + e$ where the response transformation $t(Z)$ can be found using a graphical method.

ii) The **binary regression model** is $Y \sim \text{binomial}\left(1, \rho = \frac{\mathrm{e}^{\mathrm{SP}}}{1+\mathrm{e}^{\mathrm{SP}}}\right)$. This model has $E(Y|SP) = \rho = \rho(SP)$ and $V(Y|SP) = \rho(SP)(1 - \rho(SP))$. Then $\hat{\rho} = \dfrac{e^{ESP}}{1 + e^{ESP}}$ is the estimated mean function.

iii) The **binomial regression model** is $Y_i \sim \text{binomial}\left(\mathrm{m_i}, \rho = \frac{\mathrm{e}^{\mathrm{SP}}}{1+\mathrm{e}^{\mathrm{SP}}}\right)$. Then $E(Y_i|SP_i) = m_i\rho(SP_i)$ and $V(Y_i|SP_i) = m_i\rho(SP_i)(1 - \rho(SP_i))$, and $\hat{E}(Y_i|\boldsymbol{x}_i) = m_i\hat{\rho} = \dfrac{m_i e^{ESP}}{1 + e^{ESP}}$ is the estimated mean function.

iv) The **Poisson regression (PR) model** $Y \sim \text{Poisson}\left(\mathrm{e}^{\mathrm{SP}}\right)$ has $E(Y|SP) = V(Y|SP) = \exp(SP)$. The estimated mean and variance functions are $\hat{E}(Y|\boldsymbol{x}) = e^{ESP}$.

v) Suppose Y has a gamma $\mathrm{G}(\nu, \lambda)$ distribution so that $E(Y) = \nu\lambda$ and $V(Y) = \nu\lambda^2$. The **Gamma regression model** $Y \sim G\left(\nu, \lambda = \mu(SP)/\nu\right)$ has $E(Y|SP) = \mu(SP)$ and $V(Y|SP) = [\mu(SP)]^2/\nu$. The estimated mean function is $\hat{E}(Y|\boldsymbol{x}) = \mu(ESP)$. The choices $\mu(SP) = SP$, $\mu(SP) = \exp(SP)$ and $\mu(SP) = 1/SP$ are common. Since $\mu(SP) > 0$, Gamma regression models that use the identity or reciprocal link run into problems if $\mu(ESP)$ is negative for some of the cases.

Alternatives to the binomial and Poisson regression models are needed because often the mean function for the model is good, but the variance function is not: there is overdispersion.

A useful alternative to the binomial regression model is a beta–binomial regression (BBR) model. Following Simonoff (2003, pp. 93-94) and Agresti

(2002, pp. 554-555), let $\delta = \rho/\theta$ and $\nu = (1-\rho)/\theta$, so $\rho = \delta/(\delta+\nu)$ and $\theta = 1/(\delta+\nu)$. Let $B(\delta,\nu) = \dfrac{\Gamma(\delta)\Gamma(\nu)}{\Gamma(\delta+\nu)}$. If Y has a beta–binomial distribution, $Y \sim \text{BB}(\text{m},\rho,\theta)$, then the probability mass function of Y is $P(Y=y) = \binom{m}{y}\dfrac{B(\delta+y,\nu+m-y)}{B(\delta,\nu)}$ for $y = 0,1,2,...,m$ where $0<\rho<1$ and $\theta>0$. Hence $\delta>0$ and $\nu>0$. Then $E(Y) = m\delta/(\delta+\nu) = m\rho$ and $\text{V}(Y) = m\rho(1-\rho)[1+(m-1)\theta/(1+\theta)]$. If $Y|\pi \sim \text{binomial}(m,\pi)$ and $\pi \sim \text{beta}(\delta,\nu)$, then $Y \sim \text{BB}(\text{m},\rho,\theta)$. As $\theta \to 0$, it can be shown that $V(\pi) \to 0$, and the beta–binomial distribution converges to the binomial distribution.

Definition 6.26. The BBR model states that $Y_1,...,Y_n$ are independent random variables where $Y_i|SP_i \sim \text{BB}(\text{m}_\text{i}, \rho(\text{SP}_\text{i}), \theta)$. Hence $E(Y_i|SP_i) = m_i\rho(SP_i)$ and

$$V(Y_i|SP_i) = m_i\rho(SP_i)(1-\rho(SP_i))[1+(m_i-1)\theta/(1+\theta)].$$

The BBR model has the same mean function as the binomial regression model, but allows for overdispersion. As $\theta \to 0$, it can be shown that the BBR model converges to the binomial regression model.

A useful alternative to the PR model is a negative binomial regression (NBR) model. If Y has a (generalized) negative binomial distribution, $Y \sim NB(\mu,\kappa)$, then the probability mass function of Y is

$$P(Y=y) = \frac{\Gamma(y+\kappa)}{\Gamma(\kappa)\Gamma(y+1)}\left(\frac{\kappa}{\mu+\kappa}\right)^\kappa\left(1-\frac{\kappa}{\mu+\kappa}\right)^y$$

for $y = 0,1,2,...$ where $\mu>0$ and $\kappa>0$. Then $E(Y)=\mu$ and $\text{V}(Y) = \mu+\mu^2/\kappa$. (This distribution is a generalization of the negative binomial (κ,ρ) distribution where $\rho = \kappa/(\mu+\kappa)$ and $\kappa>0$ is an unknown real parameter rather than a known integer.)

Definition 6.27. The **negative binomial regression (NBR) model** is $Y|SP \sim \text{NB}(\exp(\text{SP}),\kappa)$. Thus $E(Y|SP) = \exp(SP)$ and

$$V(Y|SP) = \exp(SP)\left(1+\frac{\exp(SP)}{\kappa}\right) = \exp(SP) + \tau\exp(2\ SP).$$

The NBR model has the same mean function as the PR model but allows for overdispersion. Following Agresti (2002, p. 560), as $\tau \equiv 1/\kappa \to 0$, it can be shown that the NBR model converges to the PR model.

For GLMs, $SP = \boldsymbol{x}^T\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$ is the MLE, and the regularity conditions are fairly reasonable because the distributions for the GLMs come from an exponential family. Overdispersion can be a problem. The assumptions on the NBR and BBR models are stronger than those for GLMS.

Remark 6.14. a) For binary logistic regression, the MLE does not exist if the $Y_i = 0$ cases and $Y_i = 1$ cases can be separated in a plot of ESP versus Y (on the vertical axis) by the vertical line at $ESP = 0$. Hence the Y values of 0 and 1 are not nearly perfectly classified by the rule $\hat{Y} = 1$ if $\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} > 0$ and $\hat{Y} = 0$, otherwise.

b) For binomial regression, including binary regression, the MLE tends not to exist if an estimated probability is 0 or one. The MLE tends to converge if $\max(|\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}|) = \max(ESP) \leq 7$.

c) For Poisson regression, the MLE tends to converge if $\max(|\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}|) = \max(ESP) \leq 11$.

For the parametric regression model $Y_i | \boldsymbol{x}_i^T \boldsymbol{\beta} \sim D(\boldsymbol{x}_i^T \boldsymbol{\beta}, \boldsymbol{\gamma})$, assume $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{\beta}))$, and that $\boldsymbol{V}(\hat{\boldsymbol{\beta}}) \xrightarrow{P} \boldsymbol{V}(\boldsymbol{\beta})$ as $n \to \infty$. These assumptions tend to be mild for a parametric regression model where the MLE $\hat{\boldsymbol{\beta}}$ is used. Then $\boldsymbol{V}(\boldsymbol{\beta}) = \boldsymbol{I}^{-1}(\boldsymbol{\beta})$, the inverse Fisher information matrix.

Consider a parametric regression model $Y_i | \boldsymbol{x}_i^T \boldsymbol{\beta} \sim D(\boldsymbol{x}_i^T \boldsymbol{\beta}, \boldsymbol{\gamma})$, Under regularity conditions, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{\beta}))$, and $\boldsymbol{V}(\hat{\boldsymbol{\beta}}) \xrightarrow{P} \boldsymbol{V}(\boldsymbol{\beta})$ as $n \to \infty$. For the parametric regression model, we regress $\boldsymbol{Y}$ on $\boldsymbol{X}$ to obtain $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ where the $n \times 1$ vector $\boldsymbol{Y} = (Y_i)$ and the ith row of the $n \times p$ design matrix $\boldsymbol{X}$ is $\boldsymbol{x}_i^T$. For GLMs, see the following theorem, for example, in Sen and Singer (1993, p. 309). Typically $I(\hat{\boldsymbol{\beta}})$ or $\hat{I}(\hat{\boldsymbol{\beta}})$ is a consistent estimator of $I(\boldsymbol{\beta})$ produced by the MLE method.

Theorem 6.17. For a parametric regression model, let $\hat{\boldsymbol{\beta}}$ be the MLE, and let $\boldsymbol{V}(\boldsymbol{\beta}) = \boldsymbol{I}^{-1}(\boldsymbol{\beta})$, the inverse Fisher information matrix. Then under regularity conditions, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}(\boldsymbol{\beta}))$, and $\boldsymbol{V}(\hat{\boldsymbol{\beta}}) \xrightarrow{P} \boldsymbol{V}(\boldsymbol{\beta})$ as $n \to \infty$.

6.9 Survival Regression

Several important survival regression models are 1D regression models with $SP = \boldsymbol{x}^T \boldsymbol{\beta}$, including the Cox (1972) proportional hazards regression model. The following survival regression models are parametric. The *accelerated failure time model* has $\log(Y) = \alpha + SP_A + \sigma e$ where $SP_A = \boldsymbol{u}^T \boldsymbol{\beta}_A$, $V(e) = 1$, and the e_i are iid from a location scale family. If the Y_i are lognormal, the e_i are normal. If the Y_i are loglogistic, the e_i are logistic. If the Y_i are Weibull, the e_i are from a smallest extreme value distribution. The Weibull regression model is a proportional hazards model using Y_i and an accelerated failure time model using $\log(Y_i)$ with $\boldsymbol{\beta}_P = \boldsymbol{\beta}_A / \sigma$. Let Y hav a Weibull $W(\gamma, \lambda)$ distribution if the pdf of Y is

$$f(y) = \lambda \gamma y^{\gamma - 1} \exp[-\lambda y^{\gamma}]$$

for $y > 0$. Prediction intervals for parametric survival regression models are for survival times Y, not censored survival times. See Section 6.13.

Definition 10.26. The *Weibull proportional hazards regression model* is

$$Y|SP \sim W(\gamma = 1/\sigma, \lambda_0 \exp(SP)),$$

where $\lambda_0 = \exp(-\alpha/\sigma)$.

In the following theorem, right censoring is allowed by the regularity conditions. The Cox PH estimator is computed by maximizing a partial likelihood and is known as a PMLE. If the Weibull regression estimator is the MLE, Theorem 6.17 applies.

Theorem 6.18. For the Cox PH estimator $\hat{\boldsymbol{\beta}}$, under regularity conditions, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\beta}))$, and $\boldsymbol{V}(\hat{\boldsymbol{\beta}}) \xrightarrow{P} \boldsymbol{V}(\boldsymbol{\beta})$ as $n \to \infty$.

6.10 Bootstrapping Some Regression Models

6.10.1 Parametric Bootstrap

For the parametric regression model $Y_i|\boldsymbol{x}_i^T\boldsymbol{\beta} \sim D(\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{\gamma})$ of Definition 6.24, assume $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\beta}))$, and that $\boldsymbol{V}(\hat{\boldsymbol{\beta}}) \xrightarrow{P} \boldsymbol{V}(\boldsymbol{\beta})$ as $n \to \infty$. These assumptions tend to be mild for a parametric regression model where the MLE $\hat{\boldsymbol{\beta}}$ is used. Then $\boldsymbol{V}(\boldsymbol{\beta}) = \boldsymbol{I}^{-1}(\boldsymbol{\beta})$, the inverse Fisher information matrix. For GLMs, see, for example, Sen and Singer (1993, p. 309). For the parametric regression model, we regress $\boldsymbol{Y}$ on $\boldsymbol{X}$ to obtain $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ where the $n \times 1$ vector $\boldsymbol{Y} = (Y_i)$ and the ith row of the $n \times p$ design matrix $\boldsymbol{X}$ is $\boldsymbol{x}_i^T$. See Section 6.2 for the parametric bootstrap for the OLS MLR model.

The parametric bootstrap uses $\boldsymbol{Y}_j^* = (Y_i^*)$ where $Y_i^*|\boldsymbol{x}_i \sim D(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ for $i = 1, ..., n$. Regress $\boldsymbol{Y}_j^*$ on $\boldsymbol{X}$ to get $\hat{\boldsymbol{\beta}}_j^*$ for $j = 1, ..., B$. The large sample theory for $\hat{\boldsymbol{\beta}}^*$ is simple. Note that if $Y_i^*|\boldsymbol{x}_i \sim D(\boldsymbol{x}_i^T\boldsymbol{b}, \hat{\boldsymbol{\gamma}})$ where $\boldsymbol{b}$ does not depend on n, then $(\boldsymbol{Y}^*, \boldsymbol{X})$ follows the parametric regression model with parameters $(\boldsymbol{b}, \hat{\boldsymbol{\gamma}})$. Hence $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \boldsymbol{b}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}(\boldsymbol{b}))$. Now fix large integer n_0, and let $\boldsymbol{b} = \hat{\boldsymbol{\beta}}_{n_o}$. Then $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}_{n_o}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}(\hat{\boldsymbol{\beta}}_{n_o}))$. Since $N_p(\mathbf{0}, \boldsymbol{V}(\hat{\boldsymbol{\beta}})) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\beta}))$, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\beta})) \tag{6.36}$$

as $n \to \infty$. See Theorem 5.1.

Now suppose $S \subseteq I$. Without loss of generality, let $\boldsymbol{\beta} = (\boldsymbol{\beta}_I^T, \boldsymbol{\beta}_O^T)^T$ and $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}(I)^T, \hat{\boldsymbol{\beta}}(O)^T)^T$. Then $(\boldsymbol{Y}, \boldsymbol{X}_I)$ follows the parametric regression model with

parameters $(\boldsymbol{\beta}_I, \boldsymbol{\gamma})$. Hence $\sqrt{n}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I) \xrightarrow{D} N_{a_I}(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\beta}_I))$. Now $(\boldsymbol{Y}^*, \boldsymbol{X}_I)$ only follows the parametric regression model asymptotically, since $\hat{\boldsymbol{\beta}}(O) \neq \mathbf{0}$. Then showing $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_{I_j} - \hat{\boldsymbol{\beta}}_{I_j}) \xrightarrow{D} N_{a_j}(\mathbf{0}, \boldsymbol{V}_j)$ is often difficult.

6.10.2 Nonparametric Bootstrap

The nonparametric bootstrap (also called the empirical bootstrap, naive bootstrap, and the pairs bootstrap) draws a sample of n cases $(Y_i^*, \boldsymbol{x}_i^*)$ with replacement from the n cases $(Y_i, \boldsymbol{x}_i)$, and regresses the Y_i^* on the $\boldsymbol{x}_i^*$ to get $\hat{\boldsymbol{\beta}}^*_{VS,1}$, and then draws another sample to get $\hat{\boldsymbol{\beta}}^*_{MIX,1}$. This process is repeated B times to get the two bootstrap samples for $i = 1, ..., B$. If $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V})$ for the full model, then $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_{I_j} - \hat{\boldsymbol{\beta}}_{I_j}) \xrightarrow{D} N_{a_j}(\mathbf{0}, \boldsymbol{V}_j)$ when $S \subseteq I_j$: just use I_j as the new full model. The method is used for multiple linear regression, Cox proportional hazards regression with right censored Y_i, and GLMs. See, for example, Burr (1994), Efron and Tibshirani (1986), Freedman (1981), and Shao and Tu (1995, pp. 335-349).

6.11 Variable Selection

Variable selection, also called subset or model selection, is the search for a subset of predictor variables that can be deleted with little loss of information if n/p is large.

Consider 1D regression models that study the conditional distribution $Y|\boldsymbol{x}^T\boldsymbol{\beta}$ of the response variable Y given a single linear combination of the predictors $\boldsymbol{x}^T\boldsymbol{\beta}$. Many important regression models satisfy this condition, including multiple linear regression, the Nelder and Wedderburn (1972) generalized linear models (GLMs), and the Cox (1972) proportional hazards regression model. Forward selection or backward elimination with the Akaike (1973) AIC criterion or Schwarz (1978) BIC criterion are often used for variable selection.

Sparse regression methods can also be used for variable selection even if n/p is not large: the regression submodel, such as a Nelder and Wedderburn (1972) generalized linear model (GLM), uses the predictors that had nonzero sparse regression estimated coefficients. These methods include least angle regression, lasso, relaxed lasso, elastic net, and sparse regression by projection. Least angle regression variable selection is the LARS-OLS hybrid estimator of Efron et al. (2004, p. 421). Lasso variable selection is called relaxed lasso by Hastie, Tibshirani, and Wainwright (2015, p. 12), and the relaxed lasso estimator with $\phi = 0$ by Meinshausen (2007, p. 376). Also see Fan and Li (2001), Friedman, Hastie, and Tibshirani (2010), Qi et al. (2015), Simon et

al. (2011), Tay, Narasimhan, and Hastie (2023), Tibshirani (1996), and Zou and Hastie (2005). The Meinshausen (2007) relaxed lasso estimator fits lasso with penalty λ_n to get a subset of variables with nonzero coefficients, and then fits lasso with a smaller penalty ϕ_n to this subset of variables where n is the sample size.

Following Olive and Hawkins (2005), a *model for variable selection* can be described by

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S \tag{6.37}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated from the model given that the subset S is in the model.

Since S is unknown, candidate subsets will be examined. Let $\boldsymbol{x}_I$ be the vector of a terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I + \boldsymbol{x}_O^T\boldsymbol{\beta}_O.$$

Suppose that S is a subset of I and that model (6.37) holds. Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \boldsymbol{0}$ and the sample correlation $\text{corr}(\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{x}_{I,i}^T\boldsymbol{\beta}_I) = 1.0$ for the population model if $S \subseteq I$. The estimated sufficient predictor (ESP) is $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$, and *a submodel I is worth considering if the correlation* $\text{corr}(ESP, ESP(I)) \geq 0.95$.

To clarify notation, suppose $p = 4$, a constant $x_1 = 1$ corresponding to β_1 is always in the model, and $\boldsymbol{\beta} = (\beta_1, \beta_2, 0, 0)^T$. Then there are $J = 2^{p-1} = 8$ possible subsets of $\{1, 2, ..., p\}$ that contain 1, including $I_1 = \{1\}$ and $S = I_2 = \{1, 2\}$. There are $2^{p-a_S} = 4$ subsets such that $S \subseteq I_j$. Let $\hat{\boldsymbol{\beta}}_{I_2} = (\hat{\beta}_1, \hat{\beta}_2)^T$ and $\boldsymbol{x}_{I_2} = (x_1, x_2)^T$. We may use the notation $I = F$ for the full model in the following definition.

Definition 6.28. The model $Y|\boldsymbol{x}^T\boldsymbol{\beta}$ that uses all of the predictors is called the *full model.* A model $Y|\boldsymbol{x}_I^T\boldsymbol{\beta}_I$ that uses a subset $\boldsymbol{x}_I$ of the predictors is called a *submodel.* The **full model is always a submodel**. The full model has *sufficient predictor* $SP = \boldsymbol{x}^T\boldsymbol{\beta}$ and the submodel has $SP = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$.

Let I_{min} correspond to the set of predictors selected by a variable selection method such as forward selection or lasso variable selection. If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, use zero padding to form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$, then the observed variable selection estimator $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. As

a statistic, $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{k,0}}$ with probabilities $\pi_{kn} = P(I_{min} = I_k)$ for $k = 1, ..., J$ where there are J subsets, e.g. $J = 2^p - 1$.

The large sample theory for $\hat{\boldsymbol{\beta}}_{MIX}$, defined below, is useful for explaining the large sample theory of $\hat{\boldsymbol{\beta}}_{VS}$. Review Section 1.8 for mixture distributions.

Definition 6.29. The *variable selection estimator* $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0}$, and $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{k,0}}$ with probabilities $\pi_{kn} = P(I_{min} = I_k)$ for $k = 1, ..., J$ where there are J subsets.

Definition 6.30. Let $\hat{\boldsymbol{\beta}}_{MIX}$ be a random vector with a mixture distribution of the $\hat{\boldsymbol{\beta}}_{I_{k,0}}$ with probabilities equal to π_{kn}. Hence $\hat{\boldsymbol{\beta}}_{MIX} = \hat{\boldsymbol{\beta}}_{I_{k,0}}$ with same probabilities π_{kn} of the variable selection estimator $\hat{\boldsymbol{\beta}}_{VS}$, but the I_k are randomly selected.

Inference will consider bootstrap hypothesis testing with confidence intervals (CIs) and regions. Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a known $g \times 1$ vector. A large sample $100(1 - \delta)\%$ confidence region for $\boldsymbol{\theta}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_n)$ is eventually bounded below by $1 - \delta$ as the sample size $n \to \infty$. Then reject H_0 if $\boldsymbol{\theta}_0$ is not in the confidence region. Let the $g \times 1$ vector T_n be an estimator of $\boldsymbol{\theta}$. Let $T_1^*, ..., T_B^*$ be the bootstrap sample for T_n. Let $\boldsymbol{A}$ be a full rank $g \times p$ constant matrix. For variable selection, test $H_0 : \boldsymbol{A\beta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{A\beta} \neq \boldsymbol{\theta}_0$ with $\boldsymbol{\theta} = \boldsymbol{A\beta}$. Then let $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{SEL}$ and let $T_i^* = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{SEL}^*$ for $i = 1, ..., B$ and SEL is VS or MIX. See Section 5.4 for the bootstrap confidence regions that will be used for variable selection inference.

6.11.1 Large Sample Theory for Variable Selection Estimators

The Theorems 6.19 and 6.20 in this subsection are due to Rathnayake and Olive (2023), and generalize the Pelawa Watagoda and Olive (2021ab) theory for multiple linear regression to many other models. The theory assumes that there is a "true model" S and that at least one subset I is considered such that $S \subseteq I$. For example, with forward selection and backward elimination, the theory assumes that the full model contains S. The theory does not hold if the true model S is not a subset of any of the considered models. For example, S could contain some interactions that were not included in the "full" model. Checking that the full model is good is important.

Assume p is fixed. Suppose model (6.37) holds, and that if $S \subseteq I_j$ where the dimension of I_j is a_j, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\mathbf{0}, \boldsymbol{V}_j)$ where $\boldsymbol{V}_j$ is the covariance matrix of the asymptotic multivariate normal distribution. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}_{j,0}) \tag{6.38}$$

where $\boldsymbol{V}_{j,0}$ adds columns and rows of zeros corresponding to the x_i not in I_j, and $\boldsymbol{V}_{j,0}$ is singular unless I_j corresponds to the full model. This large sample theory holds for many models, including multiple linear regression fit by least squares (OLS), GLMs fit by maximum likelihood, and Cox regression fit by maximum partial likelihood. See, for example, Sen and Singer (1993, pp. 280, 309).

The first assumption in Theorem 6.19 is $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Then the variable selection estimator corresponding to I_{min} underfits with probability going to zero, and the assumption holds under regularity conditions if BIC or AIC is used for many parametric regression models such as GLMs. See Charkhi and Claeskens (2018) and Claeskens and Hjort (2008, pp. 70, 101, 102, 114, 232). This assumption is a necessary condition for a variable selection estimator to be a consistent estimator. See Zhao and Yu (2006). Thus if a sparse estimator that does variable selection is a consistent estimator of $\boldsymbol{\beta}$, then $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Hence Theorem 6.19c) proves that the lasso variable selection and elastic net variable selection estimators are $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$ if lasso and elastic net are consistent. Also see Theorem 6.20. The assumption on $\boldsymbol{u}_{jn}$ in Theorem 6.19 is reasonable by (6.38) since $S \subseteq I_j$ for each π_j, and since $\hat{\boldsymbol{\beta}}_{MIX}$ uses random selection.

Consider the assumption $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$ for multiple linear regression. Charkhi and Claeskens (2018) proved the assumption holds for AIC for a wide variety of error distributions. Shao (1993) gave similar results for AIC, BIC, and C_p. The assumption holds for lasso variable selection and elastic net variable selection provided that $\hat{\lambda}_n/n \to 0$ as $n \to \infty$ so lasso and elastic net are consistent estimators. Here $\hat{\lambda}_n$ is the shrinkage penalty parameter selected after k-fold cross validation. See Theorems 6.11, 6.12, Pelawa Watogoda and Olive (2021b) and Knight and Fu (2000).

Next we will consider $P(S \subseteq I_{min}) \to 1$ for multiple linear regression with the Mallows (1973) C_p criterion. For MLR, recall that if the candidate model of $\boldsymbol{x}_I$ has k terms (including the constant), then the partial F statistic for testing whether the $p-k$ predictor variables in $\boldsymbol{x}_O$ can be deleted is

$$F_I = \frac{SSE(I) - SSE}{(n-k)-(n-p)} / \frac{SSE}{n-p} = \frac{n-p}{p-k}\left[\frac{SSE(I)}{SSE} - 1\right]$$

where SSE is the error sum of squares from the full model, and SSE(I) is the error sum of squares from the candidate submodel. An important criterion for variable selection is the C_p criterion.

Definition 6.31.

$$C_p(I) = \frac{SSE(I)}{MSE} + 2k - n = (p-k)(F_I - 1) + k$$

where MSE is the error mean square for the full model.

Note that when $H_0 : \boldsymbol{\beta}_O = \mathbf{0}$ is true, $(p-k)(F_I - 1) + k \xrightarrow{D} \chi^2_{p-k} + 2k - p$ for a large class of iid error distributions. Minimizing $C_p(I)$ is equivalent to minimizing $MSE\ [C_p(I)] = SSE(I) + (2k-n)MSE = \boldsymbol{r}^T(I)\boldsymbol{r}(I) + (2k - n)MSE$.

Assume each submodel contains a constant. Let submodel I have $k \leq p$ predictors including a constant. Then $C_p(I) \geq -p$. Assume the full model F is one of the submodels considered with $C_p(F) = p$, e.g. forward selection, backward elimination, stepwise selection, and all subsets selection. Then $-p \leq C_p(I_{min}) \leq p$. Let $\boldsymbol{r}$ be the residual vector for the full model and $\boldsymbol{r}_I$ that for the submodel. Then the correlation

$$corr(r, r_I) = \sqrt{\frac{n-p}{C_p(I) + n - 2k}}$$

by Theorem 10.3 and Olive and Hawkins (2005). Thus $corr(r, r_{I_{min}}) \to 1$ as $n \to \infty$. Referring to Equation (6.37), if $P(S \subseteq I_{min})$ does not go to 1 as $n \to \infty$, then the above correlation would not go to one. Hence $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. This result is due to Rathnayake and Olive (2023).

Theorem 6.19 a) proves that $\boldsymbol{u}$ is a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j, $E(\boldsymbol{u}) = \mathbf{0}$, and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \sum_j \pi_j \boldsymbol{V}_{j,0}$. Some of the submodels I_k will have $\pi_k = 0$. For example, since the probability of underfitting goes to zero, every submodel I_k that underfits has $\pi_k = 0$. Hence $S \subseteq I_j$ corresponding to the $\pi_j > 0$. If $\pi_d = 1$, then submodel I_d is picked with probability going to 1 as $n \to \infty$, and I_d is the only submodel with a positive π_k. Often $\pi_d = \pi_S$ in the literature. For $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX}$ with $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta}$, we have $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{v}$ by (6.40) where $E(\boldsymbol{v}) = \mathbf{0}$, and $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \sum_j \pi_j \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T$.

Theorem 6.19. Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{MIX} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive π_k by π_j. Assume $\boldsymbol{u}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_p(\mathbf{0}, \boldsymbol{V}_{j,0})$. a) Then

$$\boldsymbol{u}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \tag{6.39}$$

where the cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$. Thus $\boldsymbol{u}$ has a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j, $E(\boldsymbol{u}) = \mathbf{0}$, and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \sum_j \pi_j \boldsymbol{V}_{j,0}$.

b) Let $\boldsymbol{A}$ be a $g \times p$ full rank matrix with $1 \leq g \leq p$. Then

$$\boldsymbol{v}_n = \boldsymbol{A}\boldsymbol{u}_n = \sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} = \boldsymbol{v} \tag{6.40}$$

where $\boldsymbol{v}$ has a mixture distribution of the $\boldsymbol{v}_j = \boldsymbol{A}\boldsymbol{u}_j \sim N_g(\mathbf{0}, \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T)$ with probabilities π_j.

c) The estimator $\hat{\boldsymbol{\beta}}_{VS}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$: $\sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) = O_P(1)$.

d) If $\pi_d = 1$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{SEL} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \sim N_p(\mathbf{0}, \boldsymbol{V}_{d,0})$ where SEL is VS or MIX.

Proof. a) Since $\boldsymbol{u}_n$ has a mixture distribution of the $\boldsymbol{u}_{kn}$ with probabilities π_{kn}, the cdf of $\boldsymbol{u}_n$ is $F_{\boldsymbol{u}_n}(\boldsymbol{t}) = \sum_k \pi_{kn} F_{\boldsymbol{u}_{kn}}(\boldsymbol{t}) \to F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$ at continuity points of the $F_{\boldsymbol{u}_j}(\boldsymbol{t})$ as $n \to \infty$.

b) Since $\boldsymbol{u}_n \xrightarrow{D} \boldsymbol{u}$, then $\boldsymbol{A}\boldsymbol{u}_n \xrightarrow{D} \boldsymbol{A}\boldsymbol{u}$.

c) The result follows since selecting from a finite number J of $\sqrt{n}$ consistent estimators (even on a set that goes to one in probability) results in a $\sqrt{n}$ consistent estimator by Pratt (1959).

d) If $\pi_d = 1$, there is no selection bias, asymptotically. The result also follows by Pötscher (1991, Lemma 1). □

The following subscript notation is useful. Subscripts before the MIX are used for subsets of $\hat{\boldsymbol{\beta}}_{MIX} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$. Let $\hat{\boldsymbol{\beta}}_{i,MIX} = \hat{\beta}_i$. Similarly, if $I = \{i_1, ..., i_a\}$, then $\hat{\boldsymbol{\beta}}_{I,MIX} = (\hat{\beta}_{i_1}, ..., \hat{\beta}_{i_a})^T$. Subscripts after MIX denote the ith vector from a sample $\hat{\boldsymbol{\beta}}_{MIX,1}, ..., \hat{\boldsymbol{\beta}}_{MIX,B}$. Similar notation is used for other estimators such as $\hat{\boldsymbol{\beta}}_{VS}$. The subscript 0 is still used for zero padding. We may use $FULL$ to denote the full model $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{FULL}$.

Typically the mixture distribution is not asymptotically normal unless a $\pi_d = 1$ (e.g. if S is the full model), or if for each π_j, $\boldsymbol{A}\boldsymbol{u}_j \sim N_g(\mathbf{0}, \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T) = N_g(\mathbf{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. Then $\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} \sim N_g(\mathbf{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. This special case occurs for $\hat{\boldsymbol{\beta}}_{S,MIX}$ if $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V})$ where the asymptotic covariance matrix $\boldsymbol{V}$ is diagonal and nonsingular. Then $\hat{\boldsymbol{\beta}}_{S,MIX}$ and $\hat{\boldsymbol{\beta}}_{S,FULL}$ have the same multivariate normal limiting distribution. For several criteria, this result should hold for $\hat{\boldsymbol{\beta}}_{VS}$ since asymptotically, $\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{A}\boldsymbol{\beta})$ is selecting from the $\boldsymbol{A}\boldsymbol{u}_j$ which have the same distribution. In the simulations when $\boldsymbol{V}$ is diagonal, the confidence regions applied to $\boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{SEL} = \boldsymbol{B}\hat{\boldsymbol{\beta}}^*_{S,SEL}$ had similar volume and cutoffs where SEL is MIX, VS, or $FULL$.

Theorem 6.19 can be used to justify prediction intervals after variable selection. See Olive, Rathnayake, and Haile (2022). Theorem 6.19d) is useful for *variable selection consistency* and the *oracle property* where $\pi_d = \pi_S = 1$ if $P(I_{min} = S) \to 1$ as $n \to \infty$. See Claeskens and Hjort (2008, pp. 101-114) and Fan and Li (2001) for references. A necessary condition for $P(I_{min} = S) \to 1$ is that S is one of the models considered with probability going to one. This condition holds under very strong regularity conditions for fast methods. See Wieczorek and Lei (2022) for forward selection and Hastie, Tibshirani, and Wainwright (2015, pp. 295-302) for lasso, where the predictors need a "near orthogonality" condition.

Remark 6.15. If $A_1, A_2, ..., A_k$ are pairwise disjoint and if $\cup_{i=1}^k A_i = S$, then the collection of sets $A_1, A_2, ..., A_k$ is a *partition* of S. Then the *Law of Total Probability* states that if $A_1, A_2, ..., A_k$ form a partition of S such that

$P(A_i) > 0$ for $i = 1, ..., k$, then

$$P(B) = \sum_{j=1}^{k} P(B \cap A_j) = \sum_{j=1}^{k} P(B|A_j)P(A_j).$$

Let sets $A_{k+1}, ..., A_m$ satisfy $P(A_i) = 0$ for $i = k+1, ..., m$. Define $P(B|A_j) = 0$ if $P(A_j) = 0$. Then a Generalized Law of Total Probability is

$$P(B) = \sum_{j=1}^{m} P(B \cap A_j) = \sum_{j=1}^{m} P(B|A_j)P(A_j),$$

and will be used in the proof of the result in the following paragraph.

Pötscher (1991) used the conditional distribution of $\hat{\boldsymbol{\beta}}_{VS}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})$ to find the distribution of $\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta})$. Let $\hat{\boldsymbol{\beta}}^C_{I_k,0}$ be a random vector from the conditional distribution $\hat{\boldsymbol{\beta}}_{I_k,0}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})$. Let $\boldsymbol{w}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta})|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}) \sim \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_k,0} - \boldsymbol{\beta})$. Denote $F_{\boldsymbol{z}}(\boldsymbol{t}) = P(z_1 \le t_1, ..., z_p \le t_p)$ by $P(\boldsymbol{z} \le \boldsymbol{t})$. Then Pötscher (1991) and Pelawa Watagoda and Olive (2021b) show

$$F_{\boldsymbol{w}_n}(\boldsymbol{t}) = P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \le \boldsymbol{t}] = \sum_{k=1}^{J} F_{\boldsymbol{w}_{kn}}(\boldsymbol{t})\pi_{kn}.$$

Hence $\hat{\boldsymbol{\beta}}_{VS}$ has a mixture distribution of the $\hat{\boldsymbol{\beta}}^C_{I_k,0}$ with probabilities π_{kn}, and $\boldsymbol{w}_n$ has a mixture distribution of the $\boldsymbol{w}_{kn}$ with probabilities π_{kn}.

Proof: Let $W = W_{VS} = k$ if $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ where $P(W_{VS} = k) = \pi_{kn}$ for $k = 1, ..., J$. Then $(\hat{\boldsymbol{\beta}}_{VS:n}, W_{VS:n}) = (\hat{\boldsymbol{\beta}}_{VS}, W_{VS})$ has a joint distribution where the sample size n is usually suppressed. Note that $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_W,0}$. Then by Remark 6.15,

$$F_{\boldsymbol{w}_n}(\boldsymbol{t}) = P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \le \boldsymbol{t}] =$$

$$\sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \le \boldsymbol{t}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})]P(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}) =$$

$$\sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta}) \le \boldsymbol{t}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})]\pi_{kn}$$

$$= \sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}^C_{I_k,0} - \boldsymbol{\beta}) \le \boldsymbol{t}]\pi_{kn} = \sum_{k=1}^{J} F_{\boldsymbol{w}_{kn}}(\boldsymbol{t})\pi_{kn}. \quad \square$$

Charkhi and Claeskens (2018) showed that $\boldsymbol{w}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w}_j$ if $S \subseteq I_j$ for the maximum likelihood estimator (MLE) with AIC, and gave a forward selection example. They claim that $\boldsymbol{w}_j$ is a multivariate truncated

normal distribution (where no truncation is possible) that is symmetric about $\mathbf{0}$. Hence $E(\boldsymbol{w}_j) = 0$, and $\text{Cov}(\boldsymbol{w}_j) = \boldsymbol{\Sigma}_j$ exits. Note that both $\sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta})$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta})$ are selecting from the $\boldsymbol{u}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta})$ and asymptotically from the $\boldsymbol{u}_j$. The random selection for $\hat{\boldsymbol{\beta}}_{MIX}$ does not change the distribution of $\boldsymbol{u}_{jn}$, but selection bias does change the distribution of the selected $\boldsymbol{u}_{jn}$ and $\boldsymbol{u}_j$ to that of $\boldsymbol{w}_{jn}$ and $\boldsymbol{w}_j$. The assumption that $\boldsymbol{w}_{jn} \xrightarrow{D} \boldsymbol{w}_j$ may not be mild. The proof for Equation (6.41) is the same as that for (6.39). Theorem 6.20 proves that $\boldsymbol{w}$ is a mixture distribution of the $\boldsymbol{w}_j$ with probabilities π_j.

Theorem 6.20. Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive π_k by π_j. Assume $\boldsymbol{w}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w}_j$. Then

$$\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w} \tag{6.41}$$

where the cdf of $\boldsymbol{w}$ is $F_{\boldsymbol{w}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{w}_j}(\boldsymbol{t})$.

Proof. Since $\boldsymbol{w}_n$ has a mixture distribution of the $\boldsymbol{w}_{kn}$ with probabilities π_{kn}, the cdf of $\boldsymbol{w}_n$ is $F_{\boldsymbol{w}_n}(\boldsymbol{t}) = \sum_k \pi_{kn} F_{\boldsymbol{w}_{kn}}(\boldsymbol{t}) \to F_{\boldsymbol{w}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{w}_j}(\boldsymbol{t})$ at continuity points of the $F_{\boldsymbol{w}_j}(\boldsymbol{t})$ as $n \to \infty$. □

Remark 6.16. If $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, then $\hat{\boldsymbol{\beta}}_{VS}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ since selecting from a finite number J of $\sqrt{n}$ consistent estimators (even on a set that goes to one in probability) results in a $\sqrt{n}$ consistent estimator by Pratt (1959). By both this result and Theorems 6.19 and 6.20, the lasso variable selection and elastic net variable selection estimators are $\sqrt{n}$ consistent if lasso and elastic net are consistent (which has been shown for MLR).

Remark 6.17. It could be argued that $\boldsymbol{\beta}_E = \mathbf{0}$ in Equation (6.37) and $\boldsymbol{\beta}_O = \mathbf{0}$ if $S \subseteq I$ is a very strong regularity condition that is easy to simulate but rarely holds for real data sets. Empirically, when n/p is large, good variable selection methods select a subset I such that $cor(\boldsymbol{x}^T\hat{\boldsymbol{\beta}}_I, \boldsymbol{x}^T\hat{\boldsymbol{\beta}})$ is quite high. Data splitting can also be used for inference after variable selection. See Section 6.12.

Example 6.1. This is an example where the $\pi_{kn} \to \pi_k$ as $n \to \infty$. Assume $S \subseteq I$ where I has a predictors, including a constant. Then for a wide variety of iid error distributions, $F_I \xrightarrow{D} X/(p-a)$ where $X \sim \chi^2_{p-a}$. Let F denote the full model, and let $S = I = I_i$ be the model that deletes predictor x_i with $a = p-1$. Then from Definition 6.28, $C_p(I) \xrightarrow{D} X + p - 2$ where $X \sim \chi^2_1$. Let F denote the full model and consider all subsets variable selection with C_p. Since only S and F do not underfit, only π_S and π_F are positive. Since $C_p(F) = p$, $I = S$ is selected if $C_p(I) < p$. Hence $\pi_S = P(\chi^2_1 + p - 2 < p) = P(\chi^2_1 < 2) = 0.8427$, and $\pi_F = 1 - \pi_S = 0.1573$. This result also holds for backward

elimination since the probability that x_i will be the first predictor deleted goes to 1 as $n \to \infty$ because $C_p(I_i) = C_p(S)$ is bounded in probability while $C_p(I_j)$ diverges as $n \to \infty$ for $j \neq i$. For forward selection with correlated predictors, expect that $\pi_S < P(\chi^2_1 < 2)$, and hence $\pi_F > 1 - P(\chi^2_1 < 2)$.

For the R code below, $\boldsymbol{\beta} = (1, ..., 1, 0, ..., 0)^T$ is a $p \times 1$ vector with $k+1$ ones and $p - k + 1$ zeroes. Hence $k = p - 2$ deletes the predictor x_p. The function `belimsim` generates 1000 data sets, performs backward elimination, and finds the proportion of time the full model was selected, which was $0.158 \approx 0.1573$.

```
belimsim(n=100,p=5,k=3,nruns=1000)
$fullprop
[1] 0.158
```

6.12 Bootstrapping Variable Selection Estimators

Obtaining the bootstrap samples for $\hat{\boldsymbol{\beta}}_{VS}$ and $\hat{\boldsymbol{\beta}}_{MIX}$ is simple. Generate $\boldsymbol{Y}^*$ and $\boldsymbol{X}^*$ that would be used to produce $\hat{\boldsymbol{\beta}}^*$ if the full model estimator $\hat{\boldsymbol{\beta}}$ was being bootstrapped. Instead of computing $\hat{\boldsymbol{\beta}}^*$, compute the variable selection estimator $\hat{\boldsymbol{\beta}}^*_{VS,1} = \hat{\boldsymbol{\beta}}^{*C}_{I_{k_1},0}$. Then generate another $\boldsymbol{Y}^*$ and $\boldsymbol{X}^*$ and compute $\hat{\boldsymbol{\beta}}^*_{MIX,1} = \hat{\boldsymbol{\beta}}^*_{I_{k_1},0}$ (using the same subset I_{k_1}). This process is repeated B times to get the two bootstrap samples for $i = 1, ..., B$. Let the selection probabilities for the bootstrap variable selection estimator be ρ_{kn}. Then this bootstrap procedure bootstraps both $\hat{\boldsymbol{\beta}}_{VS}$ and $\hat{\boldsymbol{\beta}}_{MIX}$ with $\pi_{kn} = \rho_{kn}$. Then apply the confidence regions (5.31), (5.32), and (5.33) on the bootstrap sample $T^*_1, ..., T^*_B$ where $T^*_i = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{SEL,i}$ where SEL is VS or MIX.

By Subsection 6.11.1, we expect the confidence regions to simulate well (have coverage close to or higher than the nominal level so that the type I error is close to or less than the nominal level) if $\pi_d = 1$ or if the asymptotic covariance matrix for the full model is nonsingular and diagonal, but these conditions are very strong. In simulations for $\hat{\boldsymbol{\beta}}_{VS}$ with $n \geq 20p$, if the confidence regions (5.31) and (5.32) simulated well for the full model bootstrap, then (5.31) and (5.32) also simulated well for $\hat{\boldsymbol{\beta}}_{VS}$. The hybrid confidence region (5.33) had poorer performance, and confidence regions for $\hat{\boldsymbol{\beta}}_{VS}$ tended to have less undercoverage than confidence regions for $\hat{\boldsymbol{\beta}}^*_{MIX}$.

Undercoverage can occur if the bootstrap data cloud is less variable than the iid data cloud, e.g., if $n < 20p$. Heuristically, if $n \geq 20p$, then coverage can be higher than the nominal coverage for two reasons: i) the bootstrap data cloud $T^*_1, ..., T^*_B$ is more variable than the iid data cloud of $T_1, ..., T_B$, and ii) zero padding. In the simulations for $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{B}\boldsymbol{\beta}_S = \boldsymbol{\theta}$, the simulated coverage for confidence intervals and confidence regions (5.31) and (5.32) was roughly 2% less than to 2% higher than the nominal 95% coverage due to

i). In the simulations for $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{B}\boldsymbol{\beta}_E = \boldsymbol{0}$, the simulated coverage for confidence intervals and confidence regions (5.31) and (5.32) tended to be close to 99% when the nominal coverage was 95%, but the nominal 95% confidence intervals tended to be shorter than those for the full model, and the confidence region volumes were often much smaller than those for the full model. See Pelawa Watagoda and Olive (2021a) for more on why zero padding tends to increase the coverage while decreasing the volume of the confidence regions and confidence intervals. The simulations also used $B \geq \max(200, 50p)$ so that $\boldsymbol{S}_T^*$ is a good estimator of $\mathrm{Cov}(T^*)$.

For $H_0 : \boldsymbol{\beta}_S = \boldsymbol{\theta}$, we expect $[\boldsymbol{S}_T^*]^{-1}$ and $\widehat{\mathrm{Cov}}(\boldsymbol{x}_S) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}_S}$ to be close by Remark 6.4 II).

The matrix $\boldsymbol{S}_T^*$ can be singular due to one or more columns of zeros in the bootstrap sample for $\beta_1, ..., \beta_p$. The variables corresponding to these columns are likely not needed in the model given that the other predictors are in the model. A simple remedy is to add d bootstrap samples of the full model estimator $\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}}^*_{FULL}$ to the bootstrap sample. For example, take $d = \lceil cB \rceil$ with $c = 0.01$. A confidence interval $[L_n, U_n]$ can be computed without $\boldsymbol{S}_T^*$ for (5.31), (5.32), and (5.33). Using the confidence interval $[\max(L_n, T^*_{(1)}), \min(U_n, T^*_{(B)})]$ can give a shorter covering region.

Next we examine why the bootstrap data cloud tends to be more variable than the iid data cloud. Let B_{jn} count the number of times $T_i^* = T_{ij}^*$ in the bootstrap sample. Then the bootstrap sample $T_1^*, ..., T_B^*$ can be written as

$$T^*_{1,1}, ..., T^*_{B_{1n},1}, ..., T^*_{1,J}, ..., T^*_{B_{Jn},J}.$$

Denote $T^*_{1j}, ..., T^*_{B_{jn},j}$ as the jth bootstrap component of the bootstrap sample with sample mean $\overline{T}_j^*$ and sample covariance matrix $\boldsymbol{S}^*_{T,j}$. Similarly, we can define the jth component of the iid sample $T_1, ..., T_B$ to have sample mean $\overline{T}_j$ and sample covariance matrix $\boldsymbol{S}_{T,j}$.

Let $T_n = \hat{\boldsymbol{\beta}}_{MIX}$. If $S \subseteq I_j$, assume $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{V}_j)$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_{I_j} - \hat{\boldsymbol{\beta}}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{V}_j)$. Then by Equation (6.38),

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}_{j,0}) \text{ and } \sqrt{\mathrm{n}}(\hat{\boldsymbol{\beta}}^*_{\mathrm{I_j},0} - \hat{\boldsymbol{\beta}}_{\mathrm{I_j},0}) \xrightarrow{\mathrm{D}} \mathrm{N_p}(\boldsymbol{0}, \boldsymbol{V}_{\mathrm{j},0}). \quad (6.42)$$

If Equation (6.42) holds, then the component clouds have the same variability asymptotically, and the confidence regions will shrink to a point at $\boldsymbol{\beta}$ as $n \to \infty$, giving good test power, asymptotically. The iid data component clouds are all centered at $\boldsymbol{\beta}$. If the bootstrap data component clouds were all centered at the same value $\tilde{\boldsymbol{\beta}}$, then the bootstrap cloud would be like an iid data cloud shifted to be centered at $\tilde{\boldsymbol{\beta}}$, and (5.32) and (5.33) would be confidence regions for $\boldsymbol{\theta} = \boldsymbol{\beta}$ by Theorem 5.3. Instead, the bootstrap data component clouds are shifted slightly from a common center, and are each centered at a $\hat{\boldsymbol{\beta}}_{I_j,0}$. Geometrically, the shifting of the bootstrap component data clouds makes the bootstrap data cloud more variable than the iid data cloud, asymptotically

(we want $n \geq 20p$). The shifting also makes the T_i^* further from $\overline{T}^*$ than if there is no shifting. A similar argument can be given for $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX}$ and $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta}$. Region (5.31) has the same volume as region (5.33), but tends to have higher coverage since empirically, the bagging estimator $\overline{T}^*$ tends to estimate $\boldsymbol{\theta}$ at least as well as T_n for a mixture distribution.

The above argument is heuristic since we have not been able to prove that the coverage is $\geq 1 - \delta$, asymptotically, except under strong regularity conditions. Then the type I error $\leq \delta$, asymptotically. Confidence region (5.32) rejects H_0 if $(T_n - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (T_n - \boldsymbol{\theta}_0) > D^2_{(U_{BT})}$. If an iid data cloud was available, the cutoff $D^2_{(U_B)}(T_n, \boldsymbol{S}_T^*)$ could be computed from $D_i^2 = (T_i - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (T_i - \boldsymbol{\theta}_0)$ for $i = 1, ..., B$. Hence the type I error is controlled if $D^2_{(U_{BT})}$ tends to be larger than $D^2_{(U_B)}(T_n, \boldsymbol{S}_T^*)$.

The bootstrap component clouds for $\hat{\boldsymbol{\beta}}^*_{VS}$ are again separated compared to the iid clouds for $\hat{\boldsymbol{\beta}}_{VS}$, which are centered about $\boldsymbol{\beta}$. Heuristically, most of the selection bias is due to predictors in E, not to the predictors in S. Hence $\hat{\boldsymbol{\beta}}^*_{S,VS}$ is roughly similar to $\hat{\boldsymbol{\beta}}^*_{S,MIX}$. Typically the distributions of $\hat{\boldsymbol{\beta}}^*_{E,VS}$ and $\hat{\boldsymbol{\beta}}^*_{E,MIX}$ are not similar, but use the same zero padding.

Next we will examine when Equation (6.42) holds. If $S \subseteq I_j$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{V}_j)$ by the large sample theory (6.38) for the estimator. Bootstrap theory should show that $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V})$, but showing $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_{I_j} - \hat{\boldsymbol{\beta}}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{V}_j)$ is often difficult.

6.12.1 The Parametric Bootstrap

Section 6.10.1 shows that showing $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_{I_j} - \hat{\boldsymbol{\beta}}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{V}_j)$ is often difficult for the parametric bootstrap. Next, we will show that an exception is multiple linear regression.

For the multiple linear regression model, $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, assume a constant x_1 is in the model, and the zero mean e_i are iid with variance $V(e_i) = \sigma^2$. Let $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. For each I with $S \subseteq I$, assume the maximum leverage $\max_{i=1,...,n} \boldsymbol{x}_{iI}^T (\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1} \boldsymbol{x}_{iI} \to 0$ in probability as $n \to \infty$. For least squares with $S \subseteq I$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}_I)$ where $(\boldsymbol{X}_I^T \boldsymbol{X}_I)/(n\sigma^2) \xrightarrow{P} \boldsymbol{V}_I^{-1}$. See Theorem 6.3.

Consider the parametric bootstrap for the above model with $\boldsymbol{Y}^* \sim N_n(\boldsymbol{X}\hat{\boldsymbol{\beta}}, \hat{\sigma}_n^2 \boldsymbol{I}) \sim N_n(\boldsymbol{H}\boldsymbol{Y}, \hat{\sigma}_n^2 \boldsymbol{I})$ where **we are not assuming** that the $e_i \sim N(0, \sigma^2)$, and

$$\hat{\sigma}_n^2 = MSE = \frac{1}{n-p} \sum_{i=1}^{n} r_i^2$$

where the residuals are from the full OLS model. Then MSE is a $\sqrt{n}$ consistent estimator of σ^2 under mild conditions by Su and Cook (2012). Thus $\hat{\boldsymbol{\beta}}_I^* = (\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y}^* \sim N_{a_I}(\hat{\boldsymbol{\beta}}_I, \hat{\sigma}_n^2(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1})$ since $E(\hat{\boldsymbol{\beta}}_I^*) = (\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{H}\boldsymbol{Y} = \hat{\boldsymbol{\beta}}_I$ because $\boldsymbol{H}\boldsymbol{X}_I = \boldsymbol{X}_I$, and $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_I^*) = \hat{\sigma}_n^2(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}$. Hence

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_I^* - \hat{\boldsymbol{\beta}}_I) \sim N_{a_I}(\boldsymbol{0}, n\hat{\sigma}_n^2(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}_I)$$

as $n, B \to \infty$ if $S \subseteq I$. Hence Equation (6.42) holds under mild conditions.

When $\boldsymbol{V}$ is diagonal, $\sqrt{n}(\hat{\boldsymbol{\beta}}_{S,full} - \boldsymbol{\beta}_S) \xrightarrow{D} N_{a_S}(\boldsymbol{0}, \boldsymbol{V}_S)$ where $\boldsymbol{V}_S$ is a diagonal matrix using the relevant diagonal elements of $\boldsymbol{V}$. For multiple linear regression with the parametric bootstrap, the full model $\hat{\boldsymbol{\beta}}^* \sim N_p(\hat{\boldsymbol{\beta}}, \hat{\sigma}_n^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \approx N_p(\hat{\boldsymbol{\beta}}, \boldsymbol{V}/n)$. If the columns of $\boldsymbol{X}$ are orthogonal and $S \subseteq I$, then $\hat{\boldsymbol{\beta}}_{S,I}^* = \hat{\boldsymbol{\beta}}_{S,full}^*$ and $\hat{\boldsymbol{\beta}}_{S,I} = \hat{\boldsymbol{\beta}}_{S,full}$. Hence $\sqrt{n}(\hat{\boldsymbol{\beta}}_{S,MIX}^* - \hat{\boldsymbol{\beta}}_{S,full}) \xrightarrow{D} N_{a_S}(\boldsymbol{0}, \boldsymbol{V}_S)$. When $\boldsymbol{V}$ is diagonal, the columns of $\boldsymbol{X}$ are asymptotically orthogonal. Hence if $S \subseteq I$, $\hat{\boldsymbol{\beta}}_{S,I} \approx \hat{\boldsymbol{\beta}}_{S,full} \approx \overline{T}^*$, and the bootstrap component clouds have the same asymptotic variability as the iid data clouds. Hence we expect the bootstrap cutoffs for $\boldsymbol{A}\hat{\boldsymbol{\beta}}_{S,MIX}^*$ to be near $\chi^2_{g,1-\delta}$. Results in Section 6.2 suggest that the residual bootstrap behaves similarly to the parametric bootstrap, with $\hat{\sigma}_n^2 = MSE$ replaced by $\tilde{\sigma}_n^2 = (n-p)MSE/n$.

The weighted least squares formulation of the GLM maximum likelihood estimator, given for example by Hillis and Davis (1994) and Sen and Singer (1993, p. 307), suggests that similar results hold for the GLM when $\boldsymbol{V}$ is diagonal.

6.12.2 The Residual Bootstrap

The *residual bootstrap* was described in Subsection 6.2.2. Review this subsection for MLR with OLS. For this residual bootstrap, $\hat{\boldsymbol{\beta}}_{I_j}^* = (\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}_{I_j}^T\boldsymbol{Y}^* = \boldsymbol{D}_j\boldsymbol{Y}^*$ with $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{I_j}^*) = \sigma_n^2(\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}$ and $E(\hat{\boldsymbol{\beta}}_{I_j}^*) = (\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}_{I_j}^T E(\boldsymbol{Y}^*) = (\boldsymbol{X}_{I_j}^T\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}_{I_j}^T\boldsymbol{H}\boldsymbol{Y} = \hat{\boldsymbol{\beta}}_{I_j}$ since $\boldsymbol{H}\boldsymbol{X}_{I_j} = \boldsymbol{X}_{I_j}$. The expectations are with respect to the bootstrap distribution where $\hat{\boldsymbol{Y}}$ acts as a constant.

Thus for $S \subseteq I$ and the residual bootstrap using residuals from the full OLS model, $E(\hat{\boldsymbol{\beta}}_I^*) = \hat{\boldsymbol{\beta}}_I$ and $n\mathrm{Cov}(\hat{\boldsymbol{\beta}}_I^*) = n[(n-p)/n]\hat{\sigma}_n^2(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1} \xrightarrow{P} \boldsymbol{V}_I$ as $n \to \infty$ with $\hat{\sigma}_n^2 = MSE$. Hence $\hat{\boldsymbol{\beta}}_I^* - \hat{\boldsymbol{\beta}}_I \xrightarrow{P} \boldsymbol{0}$ as $n \to \infty$ by Lai et al. (1979). Note that $\hat{\boldsymbol{\beta}}_I^* = \hat{\boldsymbol{\beta}}_{I,n}^*$ and $\hat{\boldsymbol{\beta}}_I = \hat{\boldsymbol{\beta}}_{I,n}$ depend on n.

6.12.3 The Nonparametric Bootstrap

From results from Subsection 6.10.2, Equation (6.42) should hold for the nonparametric bootstrap.

For the full MLR model with the nonparametric bootstrap,

$$\boldsymbol{Y}^* = \boldsymbol{X}^* \hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{r}^W$$

and for a submodel I,

$$\boldsymbol{Y}^* = \boldsymbol{X}_I^* \hat{\boldsymbol{\beta}}_{I,OLS} + \boldsymbol{r}_I^W.$$

Freedman (1981) showed that under regularity conditions for the OLS MLR model, $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2 \boldsymbol{W}) \sim N_p(\boldsymbol{0}, \boldsymbol{V})$. Hence if $S \subseteq I$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_I^* - \hat{\boldsymbol{\beta}}_I) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}_I)$$

as $n \to \infty$. (Treat I as if I is the full model.)

One set of regularity conditions is that the MLR model holds, and if $\boldsymbol{x}_i = (1 \ \boldsymbol{u}_i^T)^T$, then the $\boldsymbol{w}_i = (Y_i \ \boldsymbol{u}_i^T)^T$ are iid from some population with a nonsingular covariance matrix.

The nonparametric bootstrap uses $\boldsymbol{w}_1^*, ..., \boldsymbol{w}_n^*$ where the $\boldsymbol{w}_i^*$ are sampled with replacement from $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. By Example 5.11, $E(\boldsymbol{w}^*) = \overline{\boldsymbol{w}}$, and

$$\text{Cov}(\boldsymbol{w}^*) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_i - \overline{\boldsymbol{w}})(\boldsymbol{w}_i - \overline{\boldsymbol{w}})^T = \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \begin{bmatrix} \tilde{S}_Y^2 & \tilde{\boldsymbol{\Sigma}}_{Y\boldsymbol{u}} \\ \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} & \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}} \end{bmatrix}.$$

Note that $\hat{\boldsymbol{\beta}}$ is a constant with respect to the bootstrap distribution. Assume all inverse matrices exist. Then it can be shown that

$$\hat{\boldsymbol{\beta}}^* = \begin{bmatrix} \hat{\beta}_1^* \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}^* \end{bmatrix} = \begin{bmatrix} \overline{Y}^* - \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}^{*T} \overline{\boldsymbol{u}}^* \\ \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1^*} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y}^* \end{bmatrix} \xrightarrow{P} \begin{bmatrix} \overline{Y} - \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}^T \overline{\boldsymbol{u}} \\ \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{u}} \end{bmatrix} = \hat{\boldsymbol{\beta}}$$

as $B \to \infty$. This result suggests that the nonparametric bootstrap for OLS MLR might work under milder regularity conditions than the $\boldsymbol{w}_i$ being iid from some population with a nonsingular covariance matrix.

6.13 Model Selection PLS and Model Selection PCR

In the fixed p setting, model selection PLS and model selection PCR can be shown to give predictions similar to that of the OLS full model. To see this, variable selection with the Mallows (1973) $C_p(I)$ criterion will be useful. Consider the OLS regression of Y on a constant and $\boldsymbol{w} = (W_1, ..., W_p)^T$

where, for example, $W_j = x_j$ or $W_j = \hat{\boldsymbol{\eta}}_j^T \boldsymbol{x}$. Let I index the variables in the model so $I = \{1, 2, 4\}$ means that W_1, W_2, and W_4 were selected. The full model $I = F$ uses all p predictors and the constant with $\boldsymbol{\beta}_I = \boldsymbol{\beta}_F = \boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$. Then by Theorem 10.3 (with $p+1$ parameters), suppose model I uses k predictors including a constant with $2 \leq k \leq p+1$. Then the model I with k predictors that minimizes $C_p(I)$ maximizes $\text{corr}(r, r_I)$, that

$$\text{corr}(r, r_I) = \sqrt{\frac{n-(p+1)}{C_p(I)+n-2k}},$$

and under linearity, $\text{corr}(r, r_I) \to 1$ forces

$$\text{corr}(\hat{\alpha} + \boldsymbol{w}^{\mathrm{T}}\hat{\boldsymbol{\beta}}, \hat{\alpha}_{\mathrm{I}} + \boldsymbol{w}_{\mathrm{I}}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{I}}) = \text{corr}(\mathrm{ESP}, \mathrm{ESP(I)}) = \text{corr}(\hat{\mathrm{Y}}, \hat{\mathrm{Y}}_{\mathrm{I}}) \to 1.$$

Thus $C_p(I) \leq 2k$ implies that $\text{corr}(\mathrm{r}, \mathrm{r_I}) \geq \sqrt{1 - \dfrac{\mathrm{p}+1}{\mathrm{n}}}$. Let the model I_{min} minimize the C_p criterion among the models considered with $C_p(I) \leq 2k_I$. Then $C_p(I_{min}) \leq C_p(F) = p + 1$, and if PLS or PCR is selected using model selection (on models $I_1, ..., I_p$ with $I_j = \{1, 2, ..., j\}$ corresponding to the j-component regression) with the $C_p(I)$ criterion, and $n \geq 20(p+1)$, then $\text{corr}(r, r_I) \geq \sqrt{19/20} = 0.974$. Hence the correlation of ESP(I) and ESP(F) will typically also be high. (For PCR, the following variant should work better: take $U_j = \hat{\boldsymbol{\eta}}_j(PCR)^T \boldsymbol{x}$ and W_1 the U_j with the highest absolute correlation with Y, W_2 the U_j with the second highest absolute correlation, etc.)

Good model selection criterion (such as k-fold cross validation) tend to be similar to $C_p(I)$, and also select model I such that $\text{corr}(r, r_I)$ and $\text{corr}(ESP, ESP(I))$ are high. Hence if the full model is good and $n >> p$ is large, predictions from the model selection PLS and model selection PCR will be similar to that of the full OLS model. Since PLS chooses components that are correlated with Y, typically fewer PLS components should be needed than PCR components, and model selection PLS will often outperform model selection PCR.

For example, let $\boldsymbol{\Sigma}_{\boldsymbol{x}} = diag(1, 2, ..., p)$ and $\boldsymbol{\beta} = \mathbf{1} = (1, ..., 1)^T$. Let the sample size $n = 2000$ and $p = 100$. Then $\boldsymbol{\beta} = \sum_{i=1}^{100} \boldsymbol{\eta}_i(PCR)$, and model selection PCR chose the $k = 100 = p$ OLS estimator while model selection PLS chose $k = 6$. Using $\boldsymbol{\beta} = (0, ..., 0, 1) = \boldsymbol{d}_{100}$ corresponds to H_1. Then model selection PLS chose $k = 2$ components while model selection PCR again chose $k = 100$ OLS. PCR and PLS were done using scaled predictors. If unscaled predictors were used, then model selection PCR chose $k = 89$ components while model selection PLS chose $k = 5$. In all cases, the correlations of the model selection residuals and OLS residuals were greater than 0.99. Computations were done in R with the Mevik, Wehrens, and Liland (2015) `pls` package.

```
library(pls)
set.seed(974)
n<-2000
p<- 100
A <- diag(sqrt(1:p))
beta <- 0*1:p + 1
x <- matrix(rnorm(n * p), nrow = n, ncol = p)
x <- x %*% A
SP <-  x%*%beta
y <- SP + rnorm(n)
#MLRplot(x,y)
#OPLSplot(x,y)
#OPLSEEplot(x,y)
#plot(cor(x,y))

z <- as.data.frame(cbind(y,x))

out<-pcr(V1~.,data=z,scale=T,validation="CV")
tem<-MSEP(out)
cvmse<-tem$val[,,1:(out$ncomp+1)][1,]
npcr <-max(which.min(cvmse)-1,1)  #100
respcr <- out$residuals[,,npcr]
resols <- out$residuals[,,p]

out<-plsr(V1~.,data=z,scale=T,validation="CV")
tem<-MSEP(out)
cvmse<-tem$val[,,1:(out$ncomp+1)][1,]
npls <-max(which.min(cvmse)-1,1)  #6
res <- out$residuals[,,npls]
resols <- out$residuals[,,p]
cor(res,resols)
#[1] 0.9999812
plot(cvmse[2:101])
plot(cvmse[3:101])
plot(cvmse[4:101])
plot(cvmse[5:101])
plot(cvmse[6:101])
plot(cvmse[7:101])

beta <- 0*1:p
beta[p] <- 1
SP <-  x%*%beta
y <- SP + rnorm(n)
z <- as.data.frame(cbind(y,x))
out<-pcr(V1~.,data=z,scale=F,validation="CV")
```

```
tem<-MSEP(out)
cvmse<-tem$val[,,1:(out$ncomp+1)][1,]
npcr <-max(which.min(cvmse)-1,1)
respcr <- out$residuals[,,npcr]
resols <- out$residuals[,,p]
#npcr=89

out<-plsr(V1~.,data=z,scale=F,validation="CV")
tem<-MSEP(out)
cvmse<-tem$val[,,1:(out$ncomp+1)][1,]
npls <-max(which.min(cvmse)-1,1)
res <- out$residuals[,,npls]
resols <- out$residuals[,,p]
cor(res,resols)
#[1] 0.9974041
npls
#[1] 5
```

6.14 Prediction Intervals

This section follows Olive, Rathnayake, and Haile (2022) closely. For a parametric 1D regression model $Y|h(\boldsymbol{x}_f) \sim D(h(\boldsymbol{x}_f), \boldsymbol{\gamma})$, we need $(\hat{h}(\boldsymbol{x}_f), \hat{\boldsymbol{\gamma}})$ to be a consistent estimator of $(h(\boldsymbol{x}_f), \boldsymbol{\gamma})$. Then draw a parametric bootstrap sample $Y_1^*, ..., Y_B^*$ where the Y_i^* are iid from the distribution $\hat{D}(\hat{h}(\boldsymbol{x}_f), \hat{\boldsymbol{\gamma}})$. Then apply the shorth$(c)$ prediction interval (4.3) to the bootstrap sample to get a large sample $100(1-\delta)\%$ prediction interval for Y_f:

$$[Y_{(s)}^*, Y_{(s+c-1)}^*] \text{ with } \text{c} = \min(\text{B}, \lceil \text{B}[1 - \delta + 1.12\sqrt{\delta/\text{B}}\ ]\ \rceil). \tag{6.43}$$

The next PI is for use after variable selection. The prediction interval (6.43) can have undercoverage if n is small compared to the number of estimated parameters. The modified shorth PI (6.44) inflates PI (6.43) to compensate for parameter estimation and model selection. Let d be the number of variables $x_1^*, ..., x_d^*$ used by the full model, forward selection, backward elimination, lasso, or lasso variable selection. (We could let $d = j$ if j is the degrees of freedom of the selected model if that model was chosen in advance without model or variable selection. Hence $d = j$ is not the model degrees of freedom if model selection was used. For a GAM full model, suppose the "degrees of freedom" d_i for $S(x_i)$ is bounded by k. We could let $d = 1 + \sum_{i=2}^{p} d_i$ with $p \leq d \leq pk$.) We want $n \geq 10d$, and the prediction interval length will be increased (penalized) if n/d is not large. For the second new prediction interval, let $q_n = \min(1 - \delta + 0.05, 1 - \delta + d/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta d/n), \quad \text{otherwise.}$$

If $1 - \delta < 0.999$ and $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Then compute the shorth(c_{mod}) PI

$$[Y^*_{(s)}, Y^*_{(s+c_{mod}-1)}] \text{ with } \mathrm{c_{mod}} = \min(\mathrm{B}, \lceil \mathrm{B}[\mathrm{q_n} + 1.12\sqrt{\delta/\mathrm{B}}\ ]\ \rceil). \quad (6.44)$$

Olive (2007, 2013a, 2018) and Pelawa Watagoda and Olive (2021b) used similar correction factors for additive error regression models $Y = h(\boldsymbol{x}) + e$ since the maximum simulated undercoverage was about 0.05 when $n = 20d$. If a $q \times 1$ vector of parameters $\boldsymbol{\gamma}$ is also estimated, we may need to replace d by $d_q = d + q$.

Remark 6.18. a) To show that (6.43) and (6.44) are large sample prediction intervals for a parametric 1D regression model with $SP = \boldsymbol{x}^T\boldsymbol{\beta}$, we need to show that $(\hat{\boldsymbol{\beta}}_{I_{min},0}, \hat{\gamma}_{I_{min}})$ is a consistent estimator of $(\boldsymbol{\beta}, \gamma)$ where the full model $\boldsymbol{\beta} = \boldsymbol{\beta}_F$. Hence we need $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$ as in Section 6.11.

b) Prediction intervals (6.43) and (6.44) often have higher than the nominal coverage if n is large and Y_f can only take on a few values. Consider binary regression where $Y_f \in \{0, 1\}$ and the PIs (6.433) and (6.44) are [0,1] with 100% coverage, [0,0], or [1,1]. If [0,0] or [1,1] is the PI, coverage tends to be higher than nominal coverage unless $P(Y_f = 1|\boldsymbol{x}_f)$ is near δ or $1 - \delta$, e.g., if $P(Y_f = 1|\boldsymbol{x}_f) = 0.01$, then [0,0] has coverage near 99% even if $1 - \delta < 0.99$.

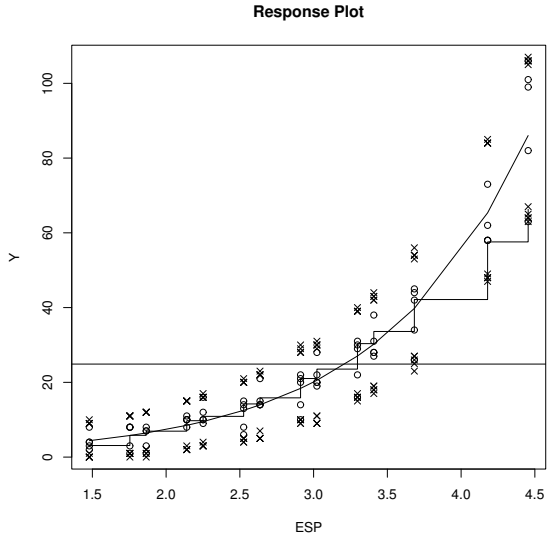

Fig. 6.1 Ceriodaphnia Data Response Plot.

Example 6.2. For the Ceriodaphnia data of Myers et al. (2002, pp. 136-139), the response variable Y is the number of Ceriodaphnia organisms counted in a container. The sample size was $n = 70$, and the predictors were a constant (x_1), seven concentrations of jet fuel (x_2), and an indicator for two strains of organism (x_3). The jet fuel was believed to impair reproduction so high concentrations should have smaller counts. Figure 6.1 shows the response plot of ESP versus Y for this data. In this plot, the lowess curve is represented as a jagged curve to distinguish it from the estimated Poisson regression mean function (the exponential curve). The horizontal line corresponds to the sample mean $\overline{Y}$. We also computed PI (6.44) using $\boldsymbol{x}_f = \boldsymbol{x}_i$ for $i = 1, ..., n$ corresponding to the observed training data $(\boldsymbol{x}_i, Y_i)$. The circles correspond to the Y_i and the $\times$'s to the PIs (6.44) with $d = 3$. The $n = 70$ large sample 95% PIs contained 97% of the Y_i. There was no evidence of overdispersion for this example. There were 5 replications for each of the 14 strain–species combinations, which helps show the bootstrap PI variability tracks the data variability when $B = 1000$. Increasing B from 1000 decreases the average PI length slightly, but using $B = 1000000$ gave a plot very similar to Figure 1 with similar coverage. Using $B = 50$ had longer PIs and sometimes had undercoverage. Using $B = 1000$ several times gave coverage between 97% and 100%.

This example illustrates a useful goodness of fit diagnostic: if the model D is a useful approximation for the data and n is large enough, we expect the coverage on the training data to be close to or higher than the nominal coverage $1 - \delta$. For example, there may be undercoverage if a Poisson regression model is used when a negative binomial regression model is needed, as illustrated in the following example.

Example 6.3. For the species data of Johnson and Raven (1973), the response variable is the total *number of species* recorded on each of $n = 29$ islands in the Galápagos Archipelago. We used a constant and the logarithm of four predictors *endem* = the number of endemic species (those that were not introduced from elsewhere), the *area* of the island, the *distance* to the closest island, the *areanear* = the area of the closest island. The Poisson regression response plot looks good, but Olive (2017b, pp. 438-440) showed that there is overdispersion and that a negative binomial regression model fits the data well. When the incorrect Poisson regression model was used, the n large sample 95% PIs (6.44) contained 89.7% of the Y_i.

Example 6.4. The Flury and Riedwyl (1988, pp. 5-6) banknote data consists of 100 counterfeit and 100 genuine Swiss banknotes. The response variable is an indicator for whether the banknote is counterfeit. The six predictors are measurements on the banknote: *bottom, diagonal, left, length, right,* and *top*. We used a constant, right, and bottom as predictors to get a model that did not have perfect classification. The response plot for this model is shown in the left plot of Figure 6.2 with $Z = Z_i = Y_i/m_i = Y_i$ and the large sample 95% PIs for $Z_i = Y_i$. The circles correspond to the Y_i and the $\times$'s to the PIs (6.44) with $d = 3$, and 199 of the 200 PIs contain Y_i. The PI [0,0] that

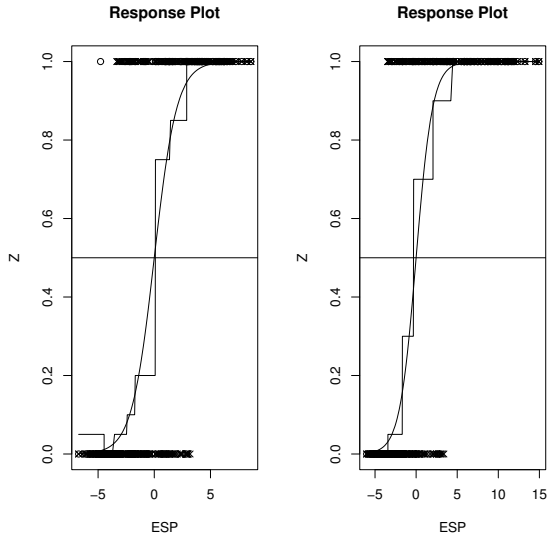

Fig. 6.2 Banknote Data GLM and GAM Response Plots.

did not contain Y_i corresponds to the circle in the upper left corner. The PIs were [0,0], [0,1], or [1,1] since the data is binary. The mean function is the smooth curve and the step function gives the sample proportion of ones in the interval. The step function approximates the smooth curve closely, hence the binary logistic regression model seems reasonable. The right plot of Figure 6.2 shows the GAM using right and bottom with $d = 3$. The coverage was 100% for the training data and the GAM had many [1,1] intervals.

6.15 Multivariate Linear Regression

Multivariate linear regression with $m \geq 2$ response variables is nearly as easy to use, at least if m is small, as multiple linear regression which has 1 response variable. *For multivariate linear regression, at least one predictor variable is quantitative.* We will assume that a constant is in the model unless told otherwise.

Definition 6.32. The **response variables** are the variables that you want to predict. The **predictor variables** are the variables used to predict the response variables.

Definition 6.33. The **multivariate linear regression model**

$$\boldsymbol{y}_i = \boldsymbol{B}^T \boldsymbol{x}_i + \boldsymbol{\epsilon}_i$$

for $i = 1, ..., n$ has $m \geq 2$ response variables $Y_1, ..., Y_m$ and p predictor variables $x_1, x_2, ..., x_p$ where $x_1 \equiv 1$ is the trivial predictor. The ith case is $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T)^T = (1, x_{i2}, ..., x_{ip}, Y_{i1}, ..., Y_{im})^T$ where the 1 could be omitted. The model is written in matrix form as $\boldsymbol{Z} = \boldsymbol{XB} + \boldsymbol{E}$ where the matrices are defined below. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = (\sigma_{ij})$ for $k = 1, ..., n$. Then the $p \times m$ coefficient matrix $\boldsymbol{B} = \left[\boldsymbol{\beta}_1 \ \boldsymbol{\beta}_2 \ \ldots \ \boldsymbol{\beta}_m \right]$ and the $m \times m$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{XB}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T \boldsymbol{\beta}_j$. The $\boldsymbol{\epsilon}_i$ are assumed to be iid. Multiple linear regression corresponds to $m = 1$ response variable, and is written in matrix form as $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$. Subscripts are needed for the m multiple linear regression models $\boldsymbol{Y}_j = \boldsymbol{X\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where $E(\boldsymbol{e}_j) = \boldsymbol{0}$. For the multivariate linear regression model, $\text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij} \ \boldsymbol{I}_n$ for $i, j = 1, ..., m$ where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

Notation. The **multiple linear regression model** uses $m = 1$. See Definition 6.5. The **multivariate linear model** $\boldsymbol{y}_i = \boldsymbol{B}^T \boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$ has $m \geq 2$, and multivariate linear regression and MANOVA models are special cases. This chapter will use $x_1 \equiv 1$ for the multivariate linear regression model. The **multivariate location and dispersion model** is the special case where $\boldsymbol{X} = \boldsymbol{1}$ and $p = 1$.

The data matrix $\boldsymbol{W} = [\boldsymbol{X} \ \ \boldsymbol{Z}]$ except usually the first column $\boldsymbol{1}$ of $\boldsymbol{X}$ is omitted for software. The $n \times m$ matrix

$$\boldsymbol{Z} = \begin{bmatrix} Y_{1,1} & Y_{1,2} & \ldots & Y_{1,m} \\ Y_{2,1} & Y_{2,2} & \ldots & Y_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n,1} & Y_{n,2} & \ldots & Y_{n,m} \end{bmatrix} = \left[\boldsymbol{Y}_1 \ \boldsymbol{Y}_2 \ \ldots \ \boldsymbol{Y}_m \right] = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_n^T \end{bmatrix}.$$

The $n \times p$ design matrix of predictor variables is

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{bmatrix} = \left[\boldsymbol{v}_1 \ \boldsymbol{v}_2 \ \ldots \ \boldsymbol{v}_p \right] = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}$$

where $\boldsymbol{v}_1 = \boldsymbol{1}$.

The $p \times m$ matrix

$$\boldsymbol{B} = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \ldots & \beta_{1,m} \\ \beta_{2,1} & \beta_{2,2} & \ldots & \beta_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p,1} & \beta_{p,2} & \ldots & \beta_{p,m} \end{bmatrix} = \left[\boldsymbol{\beta}_1 \ \boldsymbol{\beta}_2 \ \ldots \ \boldsymbol{\beta}_m \right].$$

The $n \times m$ matrix

$$\boldsymbol{E} = \begin{bmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \cdots & \epsilon_{1,m} \\ \epsilon_{2,1} & \epsilon_{2,2} & \cdots & \epsilon_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n,1} & \epsilon_{n,2} & \cdots & \epsilon_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \ldots & \boldsymbol{e}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{\epsilon}_1^T \\ \vdots \\ \boldsymbol{\epsilon}_n^T \end{bmatrix}.$$

Considering the ith row of $\boldsymbol{Z}, \boldsymbol{X}$, and $\boldsymbol{E}$ shows that $\boldsymbol{y}_i^T = \boldsymbol{x}_i^T \boldsymbol{B} + \boldsymbol{\epsilon}_i^T$.

Each response variable in a multivariate linear regression model follows a multiple linear regression model $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where it is assumed that $E(\boldsymbol{e}_j) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$. Hence the errors corresponding to the jth response are uncorrelated with variance $\sigma_j^2 = \sigma_{jj}$. Notice that the **same design matrix** $\boldsymbol{X}$ of predictors is used for each of the m models, but the jth response variable vector $\boldsymbol{Y}_j$, coefficient vector $\boldsymbol{\beta}_j$, and error vector $\boldsymbol{e}_j$ change and thus depend on j.

Now consider the ith case $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T)^T$ which corresponds to the ith row of $\boldsymbol{Z}$ and the ith row of $\boldsymbol{X}$. Then

$$\begin{bmatrix} Y_{i1} = \beta_{11}x_{i1} + \cdots + \beta_{p1}x_{ip} + \epsilon_{i1} = \boldsymbol{x}_i^T\boldsymbol{\beta}_1 + \epsilon_{i1} \\ Y_{i2} = \beta_{12}x_{i1} + \cdots + \beta_{p2}x_{ip} + \epsilon_{i2} = \boldsymbol{x}_i^T\boldsymbol{\beta}_2 + \epsilon_{i2} \\ \vdots \\ Y_{im} = \beta_{1m}x_{i1} + \cdots + \beta_{pm}x_{ip} + \epsilon_{im} = \boldsymbol{x}_i^T\boldsymbol{\beta}_m + \epsilon_{im} \end{bmatrix}$$

or $\boldsymbol{y}_i = \boldsymbol{\mu}_{\boldsymbol{x}_i} + \boldsymbol{\epsilon}_i = E(\boldsymbol{y}_i) + \boldsymbol{\epsilon}_i$ where

$$E(\boldsymbol{y}_i) = \boldsymbol{\mu}_{\boldsymbol{x}_i} = \boldsymbol{B}^T\boldsymbol{x}_i = \begin{bmatrix} \boldsymbol{x}_i^T\boldsymbol{\beta}_1 \\ \boldsymbol{x}_i^T\boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{x}_i^T\boldsymbol{\beta}_m \end{bmatrix}.$$

The notation $\boldsymbol{y}_i|\boldsymbol{x}_i$ and $E(\boldsymbol{y}_i|\boldsymbol{x}_i)$ is more accurate, but usually the conditioning is suppressed. Taking $\boldsymbol{\mu}_{\boldsymbol{x}_i}$ to be a constant (or condition on $\boldsymbol{x}_i$ if the predictor variables are random variables), $\boldsymbol{y}_i$ and $\boldsymbol{\epsilon}_i$ have the same covariance matrix. In the multivariate regression model, this covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ does not depend on i. Observations from different cases are uncorrelated (often independent), but the m errors for the m different response variables for the *same case* are correlated. If $\boldsymbol{X}$ is a random matrix, then assume $\boldsymbol{X}$ and $\boldsymbol{E}$ are independent and that expectations are conditional on $\boldsymbol{X}$.

Example 6.5. Suppose it is desired to predict the response variables $Y_1 =$ *height* and $Y_2 =$ *height at shoulder* of a person from partial skeletal remains. A model for prediction can be built from nearly complete skeletons or from living humans, depending on the population of interest (e.g. ancient Egyptians or modern US citizens). The predictor variables might be $x_1 \equiv 1$, $x_2 =$

femur length, and $x_3 =$ *ulna length*. The two heights of individuals with $x_2 = 200mm$ and $x_3 = 140mm$ should be shorter on average than the two heights of individuals with $x_2 = 500mm$ and $x_3 = 350mm$. In this example Y_1, Y_2, x_2, and x_3 are quantitative variables. If $x_4 =$ *gender* is a predictor variable, then gender (coded as male = 1 and female = 0) is qualitative.

Definition 6.34. Least squares is the classical method for fitting multivariate linear regression. The **least squares estimators** are

$$\hat{\boldsymbol{B}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z} = \left[\, \hat{\boldsymbol{\beta}}_1 \, \hat{\boldsymbol{\beta}}_2 \, \ldots \, \hat{\boldsymbol{\beta}}_m \,\right].$$

The *predicted values* or *fitted values*

$$\hat{\boldsymbol{Z}} = \boldsymbol{X}\hat{\boldsymbol{B}} = \left[\, \hat{\boldsymbol{Y}}_1 \, \hat{\boldsymbol{Y}}_2 \, \ldots \, \hat{\boldsymbol{Y}}_m \,\right] = \begin{bmatrix} \hat{Y}_{1,1} & \hat{Y}_{1,2} & \ldots & \hat{Y}_{1,m} \\ \hat{Y}_{2,1} & \hat{Y}_{2,2} & \ldots & \hat{Y}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{Y}_{n,1} & \hat{Y}_{n,2} & \ldots & \hat{Y}_{n,m} \end{bmatrix}.$$

The *residuals* $\hat{\boldsymbol{E}} = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = \boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}} =$

$$\begin{bmatrix} \hat{\boldsymbol{\epsilon}}_1^T \\ \hat{\boldsymbol{\epsilon}}_2^T \\ \vdots \\ \hat{\boldsymbol{\epsilon}}_n^T \end{bmatrix} = \left[\, \boldsymbol{r}_1 \, \boldsymbol{r}_2 \, \ldots \, \boldsymbol{r}_m \,\right] = \begin{bmatrix} \hat{\epsilon}_{1,1} & \hat{\epsilon}_{1,2} & \ldots & \hat{\epsilon}_{1,m} \\ \hat{\epsilon}_{2,1} & \hat{\epsilon}_{2,2} & \ldots & \hat{\epsilon}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\epsilon}_{n,1} & \hat{\epsilon}_{n,2} & \ldots & \hat{\epsilon}_{n,m} \end{bmatrix}.$$

These quantities can be found from the m multiple linear regressions of $\boldsymbol{Y}_j$ on the predictors: $\hat{\boldsymbol{\beta}}_j = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}_j$, $\hat{\boldsymbol{Y}}_j = \boldsymbol{X}\hat{\boldsymbol{\beta}}_j$, and $\boldsymbol{r}_j = \boldsymbol{Y}_j - \hat{\boldsymbol{Y}}_j$ for $j = 1, ..., m$. Hence $\hat{\epsilon}_{i,j} = Y_{i,j} - \hat{Y}_{i,j}$ where $\hat{\boldsymbol{Y}}_j = (\hat{Y}_{1,j}, ..., \hat{Y}_{n,j})^T$. Finally, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d} =$

$$\frac{(\boldsymbol{Z} - \hat{\boldsymbol{Z}})^T(\boldsymbol{Z} - \hat{\boldsymbol{Z}})}{n-d} = \frac{(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}})^T(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}})}{n-d} = \frac{\hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}}{n-d} = \frac{1}{n-d}\sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T.$$

The choices $d = 0$ and $d = p$ are common. If $d = 1$, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d=1} = \boldsymbol{S}_r$, the sample covariance matrix of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$, since the sample mean of the $\hat{\boldsymbol{\epsilon}}_i$ is $\boldsymbol{0}$. Let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},p}$ be the unbiased estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. Also,

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d} = (n-d)^{-1}\boldsymbol{Z}^T[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}]\boldsymbol{Z},$$

and

$$\hat{\boldsymbol{E}} = [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}]\boldsymbol{Z}.$$

6.15.1 Testing Hypotheses

This section considers testing a linear hypothesis $H_0 : \boldsymbol{LB} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{LB} \neq \boldsymbol{0}$ where $\boldsymbol{L}$ is a full rank $r \times p$ matrix.

Definition 6.35. Assume $\text{rank}(\boldsymbol{X}) = p$. The *total corrected (for the mean) sum of squares and cross products matrix* is

$$\boldsymbol{T} = \boldsymbol{R} + \boldsymbol{W}_e = \boldsymbol{Z}^T \left(\boldsymbol{I}_n - \frac{1}{n} \boldsymbol{1}\boldsymbol{1}^T \right) \boldsymbol{Z}.$$

Note that $\boldsymbol{T}/(n-1)$ is the usual sample covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{y}}$ if all n of the $\boldsymbol{y}_i$ are iid, e.g. if $\boldsymbol{B} = \boldsymbol{0}$. The *regression sum of squares and cross products matrix* is

$$\boldsymbol{R} = \boldsymbol{Z}^T \left[\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T \right] \boldsymbol{Z} = \boldsymbol{Z}^T \boldsymbol{X} \hat{\boldsymbol{B}} - \frac{1}{n} \boldsymbol{Z}^T \boldsymbol{1}\boldsymbol{1}^T \boldsymbol{Z}.$$

Let $\boldsymbol{H} = \hat{\boldsymbol{B}}^T \boldsymbol{L}^T [\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1} \boldsymbol{L}\hat{\boldsymbol{B}}$. The *error or residual sum of squares and cross products matrix* is

$$\boldsymbol{W}_e = (\boldsymbol{Z} - \hat{\boldsymbol{Z}})^T (\boldsymbol{Z} - \hat{\boldsymbol{Z}}) = \boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{X}\hat{\boldsymbol{B}} = \boldsymbol{Z}^T [\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{Z}.$$

Note that $\boldsymbol{W}_e = \hat{\boldsymbol{E}}^T \hat{\boldsymbol{E}}$ and $\boldsymbol{W}_e/(n-p) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$.

Warning: *SAS* output uses $\boldsymbol{E}$ instead of $\boldsymbol{W}_e$.

The MANOVA table is shown below.

Summary MANOVA Table

Source	matrix	df
Regression or Treatment	$\boldsymbol{R}$	$p-1$
Error or Residual	$\boldsymbol{W}_e$	$n-p$
Total (corrected)	$\boldsymbol{T}$	$n-1$

Definition 6.36. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ be the ordered eigenvalues of $\boldsymbol{W}_e^{-1}\boldsymbol{H}$. Then there are four commonly used test statistics.

The *Roy's maximum root statistic* is $\lambda_{max}(\boldsymbol{L}) = \lambda_1$.

The *Wilks' Λ statistic* is $\Lambda(\boldsymbol{L}) = |(\boldsymbol{H} + \boldsymbol{W}_e)^{-1}\boldsymbol{W}_e| = |\boldsymbol{W}_e^{-1}\boldsymbol{H} + \boldsymbol{I}|^{-1} = \prod_{i=1}^{m} (1+\lambda_i)^{-1}$.

The *Pillai's trace statistic* is $V(\boldsymbol{L}) = tr[(\boldsymbol{H} + \boldsymbol{W}_e)^{-1}\boldsymbol{H}] = \sum_{i=1}^{m} \frac{\lambda_i}{1+\lambda_i}$.

The *Hotelling-Lawley trace statistic* is $U(\boldsymbol{L}) = tr[\boldsymbol{W}_e^{-1}\boldsymbol{H}] = \sum_{i=1}^{m} \lambda_i$.

Typically some function of one of the four above statistics is used to get pval, the estimated pvalue. Output often gives the pvals for all four test statistics. Be cautious about inference if the last three test statistics do not lead to the same conclusions (Roy's test may not be trustworthy for $r > 1$). Theory and simulations developed below for the four statistics will provide more information about the sample sizes needed to use the four test statistics. See the paragraphs after the following theorem for the notation used in that theorem.

Theorem 6.21. *The Hotelling-Lawley trace statistic*

$$U(\boldsymbol{L}) = \frac{1}{n-p}[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]. \quad (6.45)$$

Proof. Using the Searle (1982, p. 333) identity
$tr(\boldsymbol{A}\boldsymbol{G}^T\boldsymbol{D}\boldsymbol{G}\boldsymbol{C}) = [vec(\boldsymbol{G})]^T[\boldsymbol{C}\boldsymbol{A} \otimes \boldsymbol{D}^T][vec(\boldsymbol{G})]$, it follows that
$(n-p)U(\boldsymbol{L}) = tr[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\hat{\boldsymbol{B}}^T\boldsymbol{L}^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}]$
$= [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] = T$ where $\boldsymbol{A} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}$, $\boldsymbol{G} = \boldsymbol{L}\hat{\boldsymbol{B}}, \boldsymbol{D} = [\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}$, and $\boldsymbol{C} = \boldsymbol{I}$. Hence (6.45) holds. □

Some notation is useful to show (6.45) and to show that $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$ under mild conditions if H_0 is true. Following Henderson and Searle (1979), let matrix $\boldsymbol{A} = [\boldsymbol{a}_1 \ \ \boldsymbol{a}_2 \ \ \ldots \ \ \boldsymbol{a}_p]$. Then the vec operator stacks the columns of $\boldsymbol{A}$ on top of one another so

$$vec(\boldsymbol{A}) = \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_p \end{pmatrix}.$$

Let $\boldsymbol{A} = (a_{ij})$ be an $m \times n$ matrix and $\boldsymbol{B}$ a $p \times q$ matrix. Then the Kronecker product of $\boldsymbol{A}$ and $\boldsymbol{B}$ is the $mp \times nq$ matrix

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1n}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2n}\boldsymbol{B} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1}\boldsymbol{B} & a_{m2}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B} \end{bmatrix}.$$

An important fact is that if $\boldsymbol{A}$ and $\boldsymbol{B}$ are nonsingular square matrices, then $[\boldsymbol{A} \otimes \boldsymbol{B}]^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$. The following assumption is important.

Assumption D1: Let h_i be the ith diagonal element of $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Assume $\max_{1\leq i\leq n} h_i \xrightarrow{P} 0$ as $n\to\infty$, assume that the zero mean iid error vectors have finite fourth moments, and assume that $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} \xrightarrow{P} \boldsymbol{W}^{-1}$.

Su and Cook (2012) proved a central limit type theorem for $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ and $\hat{\boldsymbol{B}}$ for the partial envelopes estimator, and the least squares estimator is a special case. These results prove the following theorem. Their theorem also shows that for multiple linear regression ($m=1$), $\hat{\sigma}^2 = MSE$ is a $\sqrt{n}$ consistent estimator of σ^2.

Theorem 6.22: Multivariate Least Squares Central Limit Theorem (MLS CLT). For the least squares estimator, if assumption D1 holds, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ and

$$\sqrt{n}\ \ vec(\hat{\boldsymbol{B}}-\boldsymbol{B}) \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\otimes\boldsymbol{W}).$$

Theorem 6.23. If assumption D1 holds and if H_0 is true, then $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$.

Proof. By Theorem 6.22, $\sqrt{n}\ \ vec(\hat{\boldsymbol{B}}-\boldsymbol{B}) \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\otimes\boldsymbol{W})$. Then under H_0, $\sqrt{n}\ \ vec(\boldsymbol{L}\hat{\boldsymbol{B}}) \xrightarrow{D} N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\otimes\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T)$, and $n\ \ [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\otimes(\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}$. This result also holds if $\boldsymbol{W}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are replaced by $\hat{\boldsymbol{W}} = n(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$. Hence under H_0 and using the proof of Theorem 6.21,

$$T = (n-p)U(\boldsymbol{L}) = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\otimes(\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}.$$

□

Some more details on the above results may be useful. Consider testing a linear hypothesis $H_0 : \boldsymbol{L}\boldsymbol{B} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{L}\boldsymbol{B} \neq \boldsymbol{0}$ where $\boldsymbol{L}$ is a full rank $r\times p$ matrix. For now assume the error distribution is multivariate normal $N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$. Then

$$vec(\hat{\boldsymbol{B}}-\boldsymbol{B}) = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \\ \vdots \\ \hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m \end{pmatrix} \sim N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\otimes(\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

where

$$\boldsymbol{C} = \boldsymbol{\Sigma_\epsilon} \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} \sigma_{11}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{12}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \cdots & \sigma_{1m}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ \sigma_{21}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{22}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \cdots & \sigma_{2m}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{m1}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \sigma_{m2}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \cdots & \sigma_{mm}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{bmatrix}.$$

Now let $\boldsymbol{A}$ be an $rm \times pm$ block diagonal matrix: $\boldsymbol{A} = diag(\boldsymbol{L}, ..., \boldsymbol{L})$. Then $\boldsymbol{A}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) = vec(\boldsymbol{L}(\hat{\boldsymbol{B}} - \boldsymbol{B})) =$

$$\begin{pmatrix} \boldsymbol{L}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \\ \boldsymbol{L}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2) \\ \vdots \\ \boldsymbol{L}(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m) \end{pmatrix} \sim N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)$$

where $\boldsymbol{D} = \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T = \boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^T =$

$$\begin{bmatrix} \sigma_{11}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \sigma_{12}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \cdots & \sigma_{1m}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T \\ \sigma_{21}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \sigma_{22}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \cdots & \sigma_{2m}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{m1}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \sigma_{m2}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \cdots & \sigma_{mm}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T \end{bmatrix}.$$

Under H_0, $vec(\boldsymbol{L}\boldsymbol{B}) = \boldsymbol{A}\ \ vec(\boldsymbol{B}) = \boldsymbol{0}$, and

$$vec(\boldsymbol{L}\hat{\boldsymbol{B}}) = \begin{pmatrix} \boldsymbol{L}\hat{\boldsymbol{\beta}}_1 \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_2 \\ \vdots \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_m \end{pmatrix} \sim N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T).$$

Hence under H_0,

$$[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \sim \chi^2_{rm},$$

and

$$T = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}. \tag{6.46}$$

A large sample level δ test will reject H_0 if $pval \leq \delta$ where

$$pval = P\left(\frac{T}{rm} < F_{rm,n-mp}\right). \tag{6.47}$$

Since least squares estimators are asymptotically normal, if the $\boldsymbol{\epsilon}_i$ are iid for a large class of distributions,

$$\sqrt{n}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) = \sqrt{n}\ \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \\ \vdots \\ \hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m \end{pmatrix} \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{W})$$

where

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \xrightarrow{P} \boldsymbol{W}^{-1}.$$

Then under H_0,

$$\sqrt{n}\ \ vec(\boldsymbol{L}\hat{\boldsymbol{B}}) = \sqrt{n}\ \begin{pmatrix} \boldsymbol{L}\hat{\boldsymbol{\beta}}_1 \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_2 \\ \vdots \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_m \end{pmatrix} \xrightarrow{D} N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T),$$

and

$$n\ \ [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}.$$

Hence (6.46) holds, and (6.47) gives a large sample level δ test if the least squares estimators are asymptotically normal.

Kakizawa (2009) showed, under stronger assumptions than Theorem 6.23, that for a large class of iid error distributions, the following test statistics have the same χ^2_{rm} limiting distribution when H_0 is true, and the same noncentral $\chi^2_{rm}(\omega^2)$ limiting distribution with noncentrality parameter ω^2 when H_0 is false under a local alternative. Hence the three tests are robust to the assumption of normality. The limiting null distribution is well known when the zero mean errors are iid from a multivariate normal distribution. See Khattree and Naik (1999, p. 68): $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$, $(n-p)V(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$, and $-[n-p-0.5(m-r+3)]\log(\Lambda(\boldsymbol{L})) \xrightarrow{D} \chi^2_{rm}$. Results from Kshirsagar (1972, p. 301) suggest that the third chi-square approximation is very good if $n \geq 3(m+p)^2$ for multivariate normal error vectors.

Theorems 6.21 and 6.23 are useful for relating multivariate tests with the partial F test for multiple linear regression that tests whether a reduced model that omits some of the predictors can be used instead of the full model that uses all p predictors. The partial F test statistic is

$$F_R = \left[\frac{SSE(R) - SSE(F)}{df_R - df_F}\right] / MSE(F)$$

where the residual sums of squares $SSE(F)$ and $SSE(R)$ and degrees of freedom df_F and df_r are for the full and reduced model while the mean square error $MSE(F)$ is for the full model. Let the null hypothesis for the partial F test be $H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{0}$ where $\boldsymbol{L}$ sets the coefficients of the predictors in the full model but not in the reduced model to 0. Seber and Lee (2003, p. 100) shows that

$$F_R = \frac{[\boldsymbol{L}\hat{\boldsymbol{\beta}}]^T(\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}[\boldsymbol{L}\hat{\boldsymbol{\beta}}]}{r\hat{\sigma}^2}$$

is distributed as $F_{r,n-p}$ if H_0 is true and the errors are iid $N(0,\sigma^2)$. Note that for multiple linear regression with $m = 1$, $F_R = (n-p)U(\boldsymbol{L})/r$ since $\hat{\boldsymbol{\Sigma}}_\epsilon^{-1} = 1/\hat{\sigma}^2$. Hence the scaled Hotelling Lawley test statistic is the partial F test statistic extended to $m > 1$ predictor variables by Theorem 6.21.

By Theorem 6.23, for example, $rF_R \xrightarrow{D} \chi^2_r$ for a large class of nonnormal error distributions. If $Z_n \sim F_{k,d_n}$, then $Z_n \xrightarrow{D} \chi^2_k/k$ as $d_n \to \infty$. Hence using the $F_{r,n-p}$ approximation gives a large sample test with correct asymptotic level, and the partial F test is robust to nonnormality.

Similarly, using an $F_{rm,n-pm}$ approximation for the following test statistics gives large sample tests with correct asymptotic level by Kakizawa (2009) and similar power for large n. The large sample test will have correct asymptotic level as long as the denominator degrees of freedom $d_n \to \infty$ as $n \to \infty$, and $d_n = n - pm$ reduces to the partial F test if $m = 1$ and $U(\boldsymbol{L})$ is used. Then the three test statistics are

$$\frac{-[n-p-0.5(m-r+3)]}{rm}\ \log(\Lambda(\boldsymbol{L})),\ \ \frac{n-p}{rm}\ V(\boldsymbol{L}),\ \text{and}\ \ \frac{\text{n}-\text{p}}{\text{rm}}\ \text{U}(\boldsymbol{L}).$$

By Berndt and Savin (1977) and Anderson (1984, pp. 333, 371),

$$V(\boldsymbol{L}) \le -\log(\Lambda(\boldsymbol{L})) \le U(\boldsymbol{L}).$$

Hence the Hotelling Lawley test will have the most power and Pillai's test will have the least power.

Following Khattree and Naik (1999, pp. 67-68), there are several approximations used by the SAS software. For the Roy's largest root test, if $h = \max(r, m)$, use

$$\frac{n-p-h+r}{h}\lambda_{max}(\boldsymbol{L}) \approx F(h, n-p-h+r).$$

The simulations in Olive (2017b) suggest that this approximation is good for $r = 1$ but poor for $r > 1$. Anderson (1984, p. 333) stated that Roy's largest root test has the greatest power if $r = 1$ but is an inferior test for $r > 1$. Let $g = n-p-(m-r+1)/2$, $u = (rm-2)/4$ and $t = \sqrt{r^2m^2-4}/\sqrt{m^2+r^2-5}$ for $m^2+r^2-5 > 0$ and $t = 1$, otherwise. Assume H_0 is true. Thus $U \xrightarrow{P} 0, V \xrightarrow{P} 0$, and $\Lambda \xrightarrow{P} 1$ as $n \to \infty$. Then

$$\frac{gt-2u}{rm}\ \ \frac{1-\Lambda^{1/t}}{\Lambda^{1/t}} \approx F(rm, gt-2u)\ \ \text{or}\ \ \frac{(\text{n}-\text{p})\text{t}}{\text{rm}}\ \ \frac{1-\Lambda^{1/\text{t}}}{\Lambda^{1/\text{t}}} \approx \chi^2_{\text{rm}}.$$

For large n and $t > 0$, $-\log(\Lambda) = -t\log(\Lambda^{1/t}) = -t\log(1+\Lambda^{1/t}-1) \approx t(1-\Lambda^{1/t}) \approx t(1-\Lambda^{1/t})/\Lambda^{1/t}$. If it can not be shown that

$$(n-p)[-\log(\Lambda) - t(1-\Lambda^{1/t})/\Lambda^{1/t}] \xrightarrow{P} 0 \text{ as } \mathrm{n} \to \infty,$$

then it is possible that the approximate χ^2_{rm} distribution may be the limiting distribution for only a small class of iid error distributions. When the $\boldsymbol{\epsilon}_i$ are iid $N_m(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, there are some exact results. For $r = 1$,

$$\frac{n-p-m+1}{m} \frac{1-\Lambda}{\Lambda} \sim F(m, n-p-m+1).$$

For $r = 2$,

$$\frac{2(n-p-m+1)}{2m} \frac{1-\Lambda^{1/2}}{\Lambda^{1/2}} \sim F(2m, 2(n-p-m+1)).$$

For $m = 2$,

$$\frac{2(n-p)}{2r} \frac{1-\Lambda^{1/2}}{\Lambda^{1/2}} \sim F(2r, 2(n-p)).$$

Let $s = \min(r, m)$, $m_1 = (|r-m|-1)/2$ and $m_2 = (n-p-m-1)/2$. Note that $s(|r-m|+s) = \min(r,m)\max(r,m) = rm$. Then

$$\frac{n-p}{rm} \frac{V}{1-V/s} = \frac{n-p}{s(|r-m|+s)} \frac{V}{1-V/s} \approx \frac{2m_2+s+1}{2m_1+s+1} \frac{V}{s-V} \approx$$

$$F(s(2m_1+s+1), s(2m_2+s+1)) \approx F(s(|r-m|+s), s(n-p)) = F(rm, s(n-p)).$$

This approximation is asymptotically correct by Slutsky's theorem since $1 - V/s \xrightarrow{P} 1$. Finally, $\frac{n-p}{rm}U =$

$$\frac{n-p}{s(|r-m|+s)}U \approx \frac{2(sm_2+1)}{s^2(2m_1+s+1)}U \approx F(s(2m_1+s+1), 2(sm_2+1))$$

$$\approx F(s(|r-m|+s), s(n-p)) = F(rm, s(n-p)).$$

This approximation is asymptotically correct for a wide range of iid error distributions.

Multivariate analogs of tests for multiple linear regression can be derived with appropriate choice of $\boldsymbol{L}$. Assume a constant $x_1 = 1$ is in the model. As a textbook convention, use $\delta = 0.05$ if δ is not given.

The four step MANOVA test of linear hypotheses is useful.
i) State the hypotheses $H_0 : \boldsymbol{LB} = \mathbf{0}$ and $H_1 : \boldsymbol{LB} \neq \mathbf{0}$.
ii) Get test statistic from output.
iii) Get pval from output.
iv) State whether you reject H_0 or fail to reject H_0. If pval $\leq \delta$, reject H_0 and conclude that $\boldsymbol{LB} \neq \mathbf{0}$. If pval $> \delta$, fail to reject H_0 and conclude that $\boldsymbol{LB} = \mathbf{0}$ or that there is not enough evidence to conclude that $\boldsymbol{LB} \neq \mathbf{0}$.

The MANOVA test of $H_0 : \boldsymbol{B} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{B} \neq \boldsymbol{0}$ is the special case corresponding to $\boldsymbol{L} = \boldsymbol{I}$ and $\boldsymbol{H} = \hat{\boldsymbol{B}}^T \boldsymbol{X}^T \boldsymbol{X} \hat{\boldsymbol{B}} = \hat{\boldsymbol{Z}}^T \hat{\boldsymbol{Z}}$, but is usually not a test of interest.

The analog of the ANOVA F test for multiple linear regression is the MANOVA F test that uses $\boldsymbol{L} = [\boldsymbol{0} \;\; \boldsymbol{I}_{p-1}]$ to test whether the nontrivial predictors are needed in the model. This test should reject H_0 if the response and residual plots look good, n is large enough, and at least one response plot does not look like the corresponding residual plot. A response plot for Y_j will look like a residual plot if the identity line appears almost horizontal, hence the range of $\hat{Y}_j$ is small. Response and residual plots are often useful for $n \geq 10p$.

The 4 step **MANOVA** F **test** of hypotheses uses $\boldsymbol{L} = [\boldsymbol{0} \;\; \boldsymbol{I}_{p-1}]$.
i) State the hypotheses H_0: the nontrivial predictors are not needed in the mreg model H_1: at least one of the nontrivial predictors is needed.
ii) Find the test statistic F_0 from output.
iii) Find the pval from output.
iv) If pval $\leq \delta$, reject H_0. If pval $> \delta$, fail to reject H_0. If H_0 is rejected, conclude that there is a mreg relationship between the response variables $Y_1, ..., Y_m$ and the predictors x_2, ..., x_p. If you fail to reject H_0, conclude that there is a not a mreg relationship between $Y_1, ..., Y_m$ and the predictors x_2, ..., x_p. (Or there is not enough evidence to conclude that there is a mreg relationship between the response variables and the predictors. Get the variable names from the story problem.)

The F_j test of hypotheses uses $\boldsymbol{L}_j = [0, ..., 0, 1, 0, ..., 0]$, where the 1 is in the jth position, to test whether the jth predictor x_j is needed in the model given that the other $p-1$ predictors are in the model. This test is an analog of the t tests for multiple linear regression. Note that x_j is not needed in the model corresponds to $H_0 : \boldsymbol{B}_j = \boldsymbol{0}$ while x_j needed in the model corresponds to $H_1 : \boldsymbol{B}_j \neq \boldsymbol{0}$ where $\boldsymbol{B}_j^T$ is the jth row of $\boldsymbol{B}$.

The 4 step F_j **test** of hypotheses uses $\boldsymbol{L}_j = [0, ..., 0, 1, 0, ..., 0]$ where the 1 is in the jth position.
i) State the hypotheses H_0 : x_j is not needed in the model
H_1 : x_j is needed.
ii) Find the test statistic F_j from output.
iii) Find pval from output.
iv) If pval $\leq \delta$, reject H_0. If pval $> \delta$, fail to reject H_0. Give a nontechnical sentence restating your conclusion in terms of the story problem. If H_0 is rejected, then conclude that x_j is needed in the mreg model for $Y_1, ..., Y_m$ given that the other predictors are in the model. If you fail to reject H_0, then conclude that x_j is not needed in the mreg model for $Y_1, ..., Y_m$ given that the other predictors are in the model. (Or there is not enough evidence to conclude that x_j is needed in the model. Get the variable names from the story problem.)

The Hotelling Lawley statistic

$$F_j = \frac{1}{d_j}\hat{\boldsymbol{B}}_j^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \hat{\boldsymbol{B}}_j = \frac{1}{d_j}(\hat{\beta}_{j1}, \hat{\beta}_{j2}, ..., \hat{\beta}_{jm}) \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \begin{pmatrix} \hat{\beta}_{j1} \\ \hat{\beta}_{j2} \\ \vdots \\ \hat{\beta}_{jm} \end{pmatrix}$$

where $\hat{\boldsymbol{B}}_j^T$ is the jth row of $\hat{\boldsymbol{B}}$ and $d_j = (\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}$, the jth diagonal entry of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. The statistic F_j could be used for forward selection and backward elimination in variable selection.

The 4 step **MANOVA partial F test** of hypotheses has a full model using all of the variables and a reduced model where r of the variables are deleted. The ith row of $\boldsymbol{L}$ has a 1 in the position corresponding to the ith variable to be deleted. Omitting the jth variable corresponds to the F_j test while omitting variables $x_2, ..., x_p$ corresponds to the MANOVA F test. Using $\boldsymbol{L} = [\boldsymbol{0} \ \ \boldsymbol{I}_k]$ tests whether the last k predictors are needed in the multivariate linear regression model given that the remaining predictors are in the model.
i) State the hypotheses H_0: the reduced model is good H_1: use the full model.
ii) Find the test statistic F_R from output.
iii) Find the pval from output.
iv) If pval $\leq \delta$, reject H_0 and conclude that the full model should be used. If pval $> \delta$, fail to reject H_0 and conclude that the reduced model is good.

The *lspack* function `mltreg` produces the m response and residual plots, gives $\hat{\boldsymbol{B}}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$, the MANOVA partial F test statistic and pval corresponding to the reduced model that leaves out the variables given by indices (so x_2 and x_4 in the output below with $F = 0.77$ and pval $= 0.614$), F_j and the pval for the F_j test for variables 1, 2, ..., p (where $p = 4$ in the output below so $F_2 = 1.51$ with pval $= 0.284$), and F_0 and pval for the MANOVA F test (in the output below $F_0 = 3.15$ and pval$= 0.06$). Right click `Stop` on the plots m times to advance the plots and to get the cursor back on the command line in *R*.

The command `out <- mltreg(x,y,indices=c(2))` would produce a MANOVA partial F test corresponding to the F_2 test while the command `out <- mltreg(x,y,indices=c(2,3,4))` would produce a MANOVA partial F test corresponding to the MANOVA F test for a data set with $p = 4$ predictor variables. The Hotelling Lawley trace statistic is used in the tests.

```
out <- mltreg(x,y,indices=c(2,4))
$Bhat
             [,1]         [,2]         [,3]
[1,] 47.96841291 623.2817463 179.8867890
```

```
[2,]  0.07884384   0.7276600  -0.5378649
[3,] -1.45584256 -17.3872206   0.2337900
[4,] -0.01895002   0.1393189  -0.3885967
$Covhat
          [,1]      [,2]     [,3]
[1,]  21.91591  123.2557  132.339
[2,] 123.25566 2619.4996 2145.780
[3,] 132.33902 2145.7797 2954.082
$partial
      partialF      Pval
[1,] 0.7703294 0.6141573

$Ftable
             Fj      pvals
[1,] 6.30355375 0.01677169
[2,] 1.51013090 0.28449166
[3,] 5.61329324 0.02279833
[4,] 0.06482555 0.97701447

$MANOVA
      MANOVAF       pval
[1,] 3.150118 0.06038742
```

```
#Output for Example 6.6
y<-marry[,c(2,3)]; x<-marry[,-c(2,3)];
mltreg(x,y,indices=c(3,4))
$partial

      partialF      Pval
[1,] 0.2001622 0.9349877
$Ftable
               Fj      pvals
[1,]   4.35326807 0.02870083
[2,] 600.57002201 0.00000000
[3,]   0.08819810 0.91597268
[4,]   0.06531531 0.93699302
$MANOVA
     MANOVAF         pval
[1,] 295.071 1.110223e-16
```

Example 6.6. The above output is for the Hebbler (1847) data from the 1843 Prussia census. Sometimes if the wife or husband was not at the household, then s/he would not be counted. Y_1 = number of married civilian men in the district, Y_2 = number of women married to civilians in the district, x_2 = population of the district in 1843, x_3 = number of married military men

in the district, and x_4 = number of women married to military men in the district. The reduced model deletes x_3 and x_4. The constant uses $x_1 = 1$.

a) Do the MANOVA F test.

b) Do the F_2 test.

c) Do the F_4 test.

d) Do an appropriate 4 step test for the reduced model that deletes x_3 and x_4.

e) The output for the reduced model that deletes x_1 and x_2 is shown below. Do an appropriate 4 step test.

```
$partial
     partialF Pval
[1,] 569.6429    0
```

Solution:

a) i) H_0: the nontrivial predictors are not needed in the mreg model
H_1: at least one of the nontrivial predictors is needed

ii) $F_0 = 295.071$

iii) pval = 0

iv) Reject H_0, the nontrivial predictors are needed in the mreg model.

b) i) H_0: x_2 is not needed in the model H_1: x_2 is needed

ii) $F_2 = 600.57$

iii) pval = 0

iv) Reject H_0, *population of the district* is needed in the model.

c) i) H_0: x_4 is not needed in the model H_1: x_4 is needed

ii) $F_4 = 0.065$

iii) pval = 0.937

iv) Fail to reject H_0, *number of women married to military men* is not needed in the model given that the other predictors are in the model.

d) i) H_0: the reduced model is good H_1: use the full model.

ii) $F_R = 0.200$

iii) pval = 0.935

iv) Fail to reject H_0, so the reduced model is good.

e) i) H_0: the reduced model is good H_1: use the full model.

ii) $F_R = 569.6$

iii) pval = 0.00

iv) Reject H_0, so use the full model.

6.15.2 Asymptotically Optimal Prediction Regions

In this section, we will consider a more general multivariate regression model, and then consider the multivariate linear model as a special case. Given n

cases of training or past data $(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_n, \boldsymbol{y}_n)$ and a vector of predictors $\boldsymbol{x}_f$, suppose it is desired to predict a future test vector $\boldsymbol{y}_f$.

Definition 6.37. A *large sample* $100(1-\delta)\%$ *prediction region* is a set $\mathcal{A}_n$ such that $P(\boldsymbol{y}_f \in \mathcal{A}_n) \rightarrow 1-\delta$ as $n \rightarrow \infty$, and is *asymptotically optimal* if the volume of the region converges in probability to the volume of the population minimum volume covering region.

The classical large sample $100(1-\delta)\%$ prediction region for a future value $\boldsymbol{x}_f$ given iid data $\boldsymbol{x}_1, ..., , \boldsymbol{x}_n$ is $\{\boldsymbol{x} : D^2_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq \chi^2_{p,1-\delta}\}$, while for multivariate linear regression, the classical large sample $100(1-\delta)\%$ prediction region for a future value $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ and past data $(\boldsymbol{x}_1, \boldsymbol{y}_i), ..., (\boldsymbol{x}_n, \boldsymbol{y}_n)$ is $\{\boldsymbol{y} : D^2_{\boldsymbol{y}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq \chi^2_{m,1-\delta}\}$. See Johnson and Wichern (1988, pp. 134, 151, 312). This region may work for multivariate normal $\boldsymbol{x}_i$ or $\boldsymbol{\epsilon}_i$, but otherwise tends to have undercoverage. Section 4.2 and Olive (2013a) replaced $\chi^2_{p,1-\delta}$ by the order statistic $D^2_{(U_n)}$ where U_n decreases to $\lceil n(1-\delta) \rceil$. This section will use a similar technique from Olive (2018) to develop possibly the first practical large sample prediction region for the multivariate linear model with unknown error distribution. The following technical theorem will be needed to prove Theorem 6.25.

Theorem 6.24. Let $a > 0$ and assume that $(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n)$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$.

a) $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n) - \frac{1}{a} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = o_P(1)$.

b) Let $0 < \delta \leq 0.5$. If $(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n) - (\boldsymbol{\mu}, a\boldsymbol{\Sigma}) = O_p(n^{-\delta})$ and $a\hat{\boldsymbol{\Sigma}}_n^{-1} - \boldsymbol{\Sigma}^{-1} = O_P(n^{-\delta})$, then

$$D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n) - \frac{1}{a} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = O_P(n^{-\delta}).$$

Proof. Let B_n denote the subset of the sample space on which $\hat{\boldsymbol{\Sigma}}_n$ has an inverse. Then $P(B_n) \rightarrow 1$ as $n \rightarrow \infty$. Now

$$D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\Sigma}}_n) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n)^T \hat{\boldsymbol{\Sigma}}_n^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n) =$$

$$(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a} - \frac{\boldsymbol{\Sigma}^{-1}}{a} + \hat{\boldsymbol{\Sigma}}_n^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n) =$$

$$(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n)^T \left(\frac{-\boldsymbol{\Sigma}^{-1}}{a} + \hat{\boldsymbol{\Sigma}}_n^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n) + (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n) =$$

$$\frac{1}{a} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n)^T (-\boldsymbol{\Sigma}^{-1} + a\, \hat{\boldsymbol{\Sigma}}_n^{-1}) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_n) \; +$$

$$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_n)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a} \right) (\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_n)$$

$$= \frac{1}{a}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + \frac{2}{a}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_n) +$$

$$\frac{1}{a}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_n)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_n) + \frac{1}{a}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_n)^T[a\hat{\boldsymbol{\Sigma}}_n^{-1} - \boldsymbol{\Sigma}^{-1}](\boldsymbol{x}-\hat{\boldsymbol{\mu}}_n)$$

on B_n, and the last three terms are $o_P(1)$ under a) and $O_P(n^{-\delta})$ under b). □

Now suppose a prediction region for an $m \times 1$ random vector $\boldsymbol{y}_f$ given a vector of predictors $\boldsymbol{x}_f$ is desired for the multivariate linear model. If we had many cases $\boldsymbol{z}_i = \boldsymbol{B}^T\boldsymbol{x}_f + \boldsymbol{\epsilon}_i$, then we could use the multivariate prediction region for m variables from Section 4.2. Instead, Theorem 6.25 will use the nonparametric prediction region from Section 4.2 on the pseudodata $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{B}}^T\boldsymbol{x}_f + \hat{\boldsymbol{\epsilon}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ for $i = 1, ..., n$. This takes the data cloud of the n residual vectors $\hat{\boldsymbol{\epsilon}}_i$ and centers the cloud at $\hat{\boldsymbol{y}}_f$. Note that $\hat{\boldsymbol{z}}_i = (\boldsymbol{B} - \boldsymbol{B} + \hat{\boldsymbol{B}})^T\boldsymbol{x}_f + (\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_i + \hat{\boldsymbol{\epsilon}}_i) = \boldsymbol{z}_i + (\hat{\boldsymbol{B}} - \boldsymbol{B})^T\boldsymbol{x}_f + \hat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i = \boldsymbol{z}_i + (\hat{\boldsymbol{B}} - \boldsymbol{B})^T\boldsymbol{x}_f - (\hat{\boldsymbol{B}} - \boldsymbol{B})^T\boldsymbol{x}_i = \boldsymbol{z}_i + O_P(n^{-1/2})$. Hence the distances based on the $\boldsymbol{z}_i$ and the distances based on the $\hat{\boldsymbol{z}}_i$ have the same quantiles, asymptotically (for quantiles that are continuity points of the distribution of $\boldsymbol{z}_i$).

If the $\boldsymbol{\epsilon}_i$ are iid from an $EC_m(\boldsymbol{0}, \boldsymbol{\Sigma}, g)$ distribution with continuous decreasing g and nonsingular covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = c\boldsymbol{\Sigma}$ for some constant $c > 0$, then the population asymptotically optimal prediction region is $\{\boldsymbol{y} : D_{\boldsymbol{y}}(\boldsymbol{B}^T\boldsymbol{x}_f, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}\}$ where $P(D_{\boldsymbol{y}}(\boldsymbol{B}^T\boldsymbol{x}_f, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}) = 1 - \delta$. For example, if the iid $\boldsymbol{\epsilon}_i \sim N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, then $D_{1-\delta} = \sqrt{\chi^2_{m,1-\delta}}$. If the error distribution is not elliptically contoured, then the above region still has $100(1-\delta)\%$ coverage, but prediction regions with smaller volume may exist.

A natural way to make a large sample prediction region is to estimate the target population minimum volume covering region, but for moderate samples and many error distributions, the natural estimator that covers $\lceil n(1-\delta) \rceil$ of the cases tends to have undercoverage as high as $min(0.05, \delta/2)$. This empirical result is not too surprising since it is well known that the performance of a prediction region on the training data is superior to the performance on future test data, due in part to the unknown variability of the estimator. To compensate for the undercoverage, let q_n be as in Theorem 6.25.

Theorem 6.25. Suppose $\boldsymbol{y}_i = E(\boldsymbol{y}_i|\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i = \hat{\boldsymbol{y}}_i + \hat{\boldsymbol{\epsilon}}_i$ where $\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} > 0$, and where the zero mean $\boldsymbol{\epsilon}_f$ and the $\boldsymbol{\epsilon}_i$ are iid for $i = 1, ..., n$. Given $\boldsymbol{x}_f$, suppose the fitted model produces $\hat{\boldsymbol{y}}_f$ and nonsingular $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$. Let $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ and

$$D_i^2 \equiv D_i^2(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) = (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)^T\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}(\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)$$

for $i = 1, ..., n$. Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + m/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta m/n), \quad \text{otherwise.}$$

If $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Let $0 < \delta < 1$ and $h = D_{(U_n)}$ where $D_{(U_n)}$ is the 100 q_nth sample quantile of the Mahalanobis distances D_i. Let the nominal $100(1-\delta)\%$ prediction region for $\boldsymbol{y}_f$ be given by

$$\{\boldsymbol{z} : (\boldsymbol{z} - \hat{\boldsymbol{y}}_f)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\boldsymbol{z} - \hat{\boldsymbol{y}}_f) \leq D^2_{(U_n)}\} =$$

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D^2_{(U_n)}\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{(U_n)}\}. \tag{6.48}$$

a) Consider the n prediction regions for the data where $(\boldsymbol{y}_{f,i}, \boldsymbol{x}_{f,i}) = (\boldsymbol{y}_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. If the order statistic $D_{(U_n)}$ is unique, then U_n of the n prediction regions contain $\boldsymbol{y}_i$ where $U_n/n \to 1 - \delta$ as $n \to \infty$.

b) If $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is a consistent estimator of $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, then (6.48) is a large sample $100(1-\delta)\%$ prediction region for $\boldsymbol{y}_f$.

c) If $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is a consistent estimator of $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, and the $\boldsymbol{\epsilon}_i$ come from an elliptically contoured distribution such that the unique highest density region is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}\}$, then the prediction region (6.48) is asymptotically optimal.

Proof. a) Suppose $(\boldsymbol{x}_f, \boldsymbol{y}_f) = (\boldsymbol{x}_i, \boldsymbol{y}_i)$. Then

$$D^2_{\boldsymbol{y}_i}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) = (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i) = \hat{\boldsymbol{\epsilon}}_i^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \hat{\boldsymbol{\epsilon}}_i = D^2_{\hat{\boldsymbol{\epsilon}}_i}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}).$$

Hence $\boldsymbol{y}_i$ is in the ith prediction region $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{(U_n)}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$ iff $\hat{\boldsymbol{\epsilon}}_i$ is in prediction region $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{(U_n)}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$, but exactly U_n of the $\hat{\boldsymbol{\epsilon}}_i$ are in the latter region by construction, if $D_{(U_n)}$ is unique. Since $D_{(U_n)}$ is the $100(1-\delta)$th percentile of the D_i asymptotically, $U_n/n \to 1 - \delta$.

b) Let $P[D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})] = 1 - \delta$. Since $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} > 0$, Theorem 6.24 shows that if $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \xrightarrow{P} (E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ then $D(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \xrightarrow{D} D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$. Hence the percentiles of the distances converge in distribution, and the probability that $\boldsymbol{y}_f$ is in $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$ converges to $1 - \delta =$ the probability that $\boldsymbol{y}_f$ is in $\{\boldsymbol{z} : D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$ at continuity points $D_{1-\delta}$ of the distribution of $D(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$.

c) The asymptotically optimal prediction region is the region with the smallest volume (hence highest density) such that the coverage is $1 - \delta$, as $n \to \infty$. This region is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$ if the asymptotically optimal region for the $\boldsymbol{\epsilon}_i$ is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$. Hence the result follows by b). □

Notice that if $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}$ exists, then $100q_n\%$ of the n training data $\boldsymbol{y}_i$ are in their corresponding prediction region with $\boldsymbol{x}_f = \boldsymbol{x}_i$, and $q_n \to 1-\delta$ even if $(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is not a good estimator or if the regression model is misspecified. Hence the coverage q_n of the training data is robust to model assumptions. Of course the volume of the prediction region could be large if a poor estimator $(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is

used or if the $\boldsymbol{\epsilon}_i$ do not come from an elliptically contoured distribution. The response, residual, and DD plots can be used to check model assumptions. If the plotted points in the RMVN DD plot cluster tightly about some line through the origin and if $n \geq \max[3(m+p)^2, mp+30]$, we expect the volume of the prediction region may be fairly low for the least squares estimators.

If n is too small, then multivariate data is sparse and the covering ellipsoid for the training data may be far too small for future data, resulting in severe undercoverage. Also notice that $q_n = 1-\delta/2$ or $q_n = 1-\delta+0.05$ for $n \leq 20p$. At the training data, the coverage $q_n \geq 1-\delta$, and q_n converges to the nominal coverage $1-\delta$ as $n \to \infty$. Suppose $n \leq 20p$. Then the nominal 95% prediction region uses $q_n = 0.975$ while the nominal 50% prediction region uses $q_n = 0.55$. Prediction distributions depend both on the error distribution and on the variability of the estimator $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$. This variability is typically unknown but converges to 0 as $n \to \infty$. Also, residuals tend to underestimate errors for small n. For moderate n, ignoring estimator variability and using $q_n = 1-\delta$ resulted in undercoverage as high as $\min(0.05, \delta/2)$. Letting the "coverage" q_n decrease to the nominal coverage $1-\delta$ inflates the volume of the prediction region for small n, compensating for the unknown variability of $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$.

Consider the multivariate linear regression model. Let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d=p}$, $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$, and $D_i^2(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) = (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)^T \boldsymbol{S}_r^{-1} (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)$ for $i = 1, ..., n$. Then the large sample nonparametric $100(1-\delta)\%$ prediction region is

$$\{\boldsymbol{z} : D_{\boldsymbol{z}}^2(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \leq D_{(U_n)}^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \leq D_{(U_n)}\}. \tag{6.49}$$

Theorem 6.26 will show that this prediction region (6.49) can also be found by applying the nonparametric prediction region (4.11) on the $\hat{\boldsymbol{z}}_i$. Recall that $\boldsymbol{S}_r$ defined in Definition 6.34 is the sample covariance matrix of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$. For the multivariate linear regression model, if $D_{1-\delta}$ is a continuity point of the distribution of D, Assumption D1 above Theorem 6.22 holds, and the $\boldsymbol{\epsilon}_i$ have a nonsingular covariance matrix, then (6.49) is a large sample $100(1-\delta)\%$ prediction region for $\boldsymbol{y}_f$.

Theorem 6.26. For multivariate linear regression, when least squares is used to compute $\hat{\boldsymbol{y}}_f$, $\boldsymbol{S}_r$, and the pseudodata $\hat{\boldsymbol{z}}_i$, prediction region (6.49) is the nonparametric prediction region (4.11) applied to the $\hat{\boldsymbol{z}}_i$.

Proof. Multivariate linear regression with least squares satisfies Theorem 6.25 by Su and Cook (2012). (See Theorem 6.22.) Let $(T, \boldsymbol{C})$ be the sample mean and sample covariance matrix applied to the $\hat{\boldsymbol{z}}_i$. The sample mean and sample covariance matrix of the residual vectors is $(\boldsymbol{0}, \boldsymbol{S}_r)$ since least squares was used. Hence the $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ have sample covariance matrix $\boldsymbol{S}_r$, and sample mean $\hat{\boldsymbol{y}}_f$. Hence $(T, \boldsymbol{C}) = (\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$, and the $D_i(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ are used to compute $D_{(U_n)}$. □

The nonparametric prediction region for multivariate linear regression of Theorem 6.26 uses $(T, \boldsymbol{C}) = (\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ in (6.48), and has simple geometry. Let

R_r be the nonparametric prediction region (6.49) applied to the residuals $\hat{\boldsymbol{\epsilon}}_i$ with $\hat{\boldsymbol{y}}_f = \mathbf{0}$. Then R_r is a hyperellipsoid with center $\mathbf{0}$, and the nonparametric prediction region is the hyperellipsoid R_r translated to have center $\hat{\boldsymbol{y}}_f$. Hence in a DD plot, all points to the left of the line $MD = D_{(U_n)}$ correspond to $\boldsymbol{y}_i$ that are in their prediction region, while points to the right of the line are not in their prediction region.

The nonparametric prediction region has some interesting properties. This prediction region is asymptotically optimal if the $\boldsymbol{\epsilon}_i$ are iid for a large class of elliptically contoured $EC_m(\mathbf{0}, \boldsymbol{\Sigma}, g)$ distributions. Also, if there are 100 different values $(\boldsymbol{x}_{jf}, \boldsymbol{y}_{jf})$ to be predicted, we only need to update $\hat{\boldsymbol{y}}_{jf}$ for $j = 1, ..., 100$, we do not need to update the covariance matrix $\boldsymbol{S}_r$.

It is common practice to examine how well the prediction regions work on the training data. That is, for $i = 1, ..., n$, set $\boldsymbol{x}_f = \boldsymbol{x}_i$ and see if $\boldsymbol{y}_i$ is in the region with probability near to $1 - \delta$ with a simulation study. Note that $\hat{\boldsymbol{y}}_f = \hat{\boldsymbol{y}}_i$ if $\boldsymbol{x}_f = \boldsymbol{x}_i$. Simulation is not needed for the nonparametric prediction region (6.49) for the data since the prediction region (6.49) centered at $\hat{\boldsymbol{y}}_i$ contains $\boldsymbol{y}_i$ iff R_r, the prediction region centered at $\mathbf{0}$, contains $\hat{\boldsymbol{\epsilon}}_i$ since $\hat{\boldsymbol{\epsilon}}_i = \boldsymbol{y}_i - \hat{\boldsymbol{y}}_i$. Thus $100q_n\%$ of prediction regions corresponding to the data $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ contain $\boldsymbol{y}_i$, and $100q_n\% \to 100(1-\delta)\%$. Hence the prediction regions work well on the training data and should work well on $(\boldsymbol{x}_f, \boldsymbol{y}_f)$ similar to the training data. Of course simulation should be done for test data $(\boldsymbol{x}_f, \boldsymbol{y}_f)$ that are not equal to training data cases.

This training data result holds provided that the multivariate linear regression using least squares is such that the sample covariance matrix $\boldsymbol{S}_r$ of the residual vectors is nonsingular, **the multivariate regression model need not be correct**. Hence the coverage at the n training data cases $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is robust to model misspecification. Of course, the prediction regions may be very large if the model is severely misspecified, but severity of misspecification can be checked with the response and residual plots. Coverage for a future value $\boldsymbol{y}_f$ can also be arbitrarily bad if there is extrapolation or if $(\boldsymbol{x}_f, \boldsymbol{y}_f)$ comes from a different population than that of the data.

6.16 Data Splitting

Data splitting divides the training data set of n cases into two sets: H and the validation set V where H has n_H of the cases and V has the remaining $n_V = n - n_H$ cases $i_1, ..., i_{n_V}$. An application of data splitting is to use a variable selection method, such as forward selection or lasso, on H to get submodel I_{min} with a predictors, then fit the selected model to the cases in the validation set V using standard inference. See, for example, Rinaldo et al. (2019).

To help understand data splitting when the cases in H are randomly selected, let I denote the predictors selected using H, possibly after variable

selection or after looking at the data and building the model. Let $\hat{\boldsymbol{\beta}}_E(\boldsymbol{x}_I, Y)$ be the estimator obtained by regressing Y on $\boldsymbol{x}_I$ using the cases in V. Then $\hat{\boldsymbol{\beta}}_E(\boldsymbol{x}_I, Y)$ estimates $\boldsymbol{\beta}_I = \boldsymbol{\beta}_I(\boldsymbol{x}_I, Y)$. For example, if the cases are iid with enough low order moments, then $\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{x}_I, Y)$ estimates $\boldsymbol{\beta}_I = \boldsymbol{\Sigma}_{\boldsymbol{x}_I}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}_I, Y}$ while $\hat{\boldsymbol{\beta}}_{OPLS}(\boldsymbol{x}_I, Y)$ estimates $\boldsymbol{\beta}_I = \lambda_I \boldsymbol{\Sigma}_{\boldsymbol{x}_I, Y}$. If the model is sparse, check the fitted model with the same checks used for low dimensional data. For data splitting in low dimensions, if the full model is good, then often model (6.37) works well in that we can eliminate predictors and often do nearly as well or better than the full model. In high dimensions, we often do not know if the full model, that regresses Y on $\boldsymbol{x}$, is good. The data splitting and high dimensional regression literature often claims that $\boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y) = \boldsymbol{\beta}_E(\boldsymbol{x}, Y)$. For example, $\boldsymbol{\beta}_{OPLS} = \boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{OLS}(\boldsymbol{x}, Y)$, or model (6.37) holds with $S \subseteq I_{min}$ and $\boldsymbol{\beta}_{I_{min}}$ a $k \times 1$ vector with $a_S \leq k \leq n/10$. While these claims can be true, the regularity conditions often become too strong as $n/p \to 0$.

Table 6.1 Regression Summary

low dimensions	data splitting with sparse I	high dim. regularity conditions are too strong
general: $\boldsymbol{\beta}(\boldsymbol{x}, Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y)$	$\boldsymbol{\beta}_I(\boldsymbol{x}_I, Y)$	$\boldsymbol{\beta}(\boldsymbol{x}, Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y)$
data splitting: $\boldsymbol{\beta}(\boldsymbol{x}, Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y)$	$\boldsymbol{\beta}_I(\boldsymbol{x}_I, Y)$	$\boldsymbol{\beta}(\boldsymbol{x}, Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y)$
lasso: $\boldsymbol{\beta}_{lasso}$	$\boldsymbol{\beta}_I(\boldsymbol{x}_I, Y)$	$\boldsymbol{\beta}(\boldsymbol{x}, Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y)$
OPLS: $\boldsymbol{\beta}_{OPLS} = \lambda \boldsymbol{\Sigma}_{\boldsymbol{x}, Y}$	$\boldsymbol{\beta}_{I,OPLS} = \lambda_I \boldsymbol{\Sigma}_{\boldsymbol{x}_I, Y}$	$\boldsymbol{\beta}_{OPLS} = \boldsymbol{\beta}_{OLS}$
MMLE: $\boldsymbol{\beta}_{MMLE} = \boldsymbol{\Sigma}_{\boldsymbol{u}, Y}$	$\boldsymbol{\beta}_{I,MMLE} = \boldsymbol{\Sigma}_{\boldsymbol{u}_I, Y}$	$\boldsymbol{\beta}_{MMLE} = \boldsymbol{\beta}_{OLS}$

Table 6.1 summarizes what the regression estimators tend to estimate in low dimensions or after data splitting with a sparse fitted model I. The third column of Table 6 gives some results in the high dimensional literature where the regularity conditions are often too strong. In particular, often the regularity conditions are too strong for low dimensional results to hold in high dimensions.

6.17 Summary

1) a) **MLR model 1** is

$$Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i$$

for $i = 1, ..., n$. Here n is the sample size and the random variable e_i is the ith error. Assume that the e_i are iid with expected value $E(e_i) = 0$ and variance $V(e_i) = \sigma^2$. In matrix notation, these n equations become $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$.

b) **MLR model 2** is

$$Y_i = \alpha + x_{i,1}\beta_1 + \cdots + x_{i,p}\beta_p + e_i = \alpha + \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i$$

for $i = 1, ..., n$. For this model, we may use $\boldsymbol{\phi} = (\alpha, \boldsymbol{\beta}^T)^T$ with $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\phi} + \boldsymbol{e}$.

2) For MLR model 1, the ordinary least squares OLS full model estimator $\hat{\boldsymbol{\beta}}_{OLS}$ minimizes $Q_{OLS}(\boldsymbol{\beta}) = \sum_{i=1}^{n} r_i^2(\boldsymbol{\beta}) = RSS(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$. In the estimating equations $Q_{OLS}(\boldsymbol{\beta})$, the vector $\boldsymbol{\beta}$ is a dummy variable. The minimizer $\hat{\boldsymbol{\beta}}_{OLS}$ estimates the parameter vector $\boldsymbol{\beta}$ for the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Note that $\hat{\boldsymbol{\beta}}_{OLS} \sim AN_p(\boldsymbol{\beta}, MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1})$.

3) Given an estimate $\boldsymbol{b}$ of $\boldsymbol{\beta}$, the corresponding vector of *predicted values* or *fitted values* is $\widehat{\boldsymbol{Y}} \equiv \widehat{\boldsymbol{Y}}(\boldsymbol{b}) = \boldsymbol{X}\boldsymbol{b}$. Thus the ith fitted value

$$\hat{Y}_i \equiv \hat{Y}_i(\boldsymbol{b}) = \boldsymbol{x}_i^T\boldsymbol{b} = x_{i,1}b_1 + \cdots + x_{i,p}b_p.$$

The vector of *residuals* is $\boldsymbol{r} \equiv \boldsymbol{r}(\boldsymbol{b}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}(\boldsymbol{b})$. Thus ith residual $r_i \equiv r_i(\boldsymbol{b}) = Y_i - \hat{Y}_i(\boldsymbol{b}) = Y_i - x_{i,1}b_1 - \cdots - x_{i,p}b_p$. A *response plot* for MLR is a plot of $\hat{Y}_i$ versus Y_i. A *residual plot* is a plot of $\hat{Y}_i$ versus r_i. If the e_i are iid from a unimodal distribution that is not highly skewed, the plotted points should scatter about the identity line and the $r = 0$ line.

4) **OLS CLTs.** Consider the MLR model and assume that the zero mean errors are iid with $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2$. If the $\boldsymbol{x}_i$ are random vectors, assume that the cases $(\boldsymbol{x}_i, Y_i)$ are independent, and that the $\boldsymbol{e}_i$ and $\boldsymbol{x}_i$ are independent. Also assume that $\max_i(h_1, ..., h_n) \to 0$ and

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \to \boldsymbol{V}^{-1}$$

as $n \to \infty$ where the convergence is in probability if the $\boldsymbol{x}_i$ are random vectors (instead of nonstochastic constant vectors).

a) For MLR model 1, the OLS estimator $\hat{\boldsymbol{\beta}}$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{V}).$$

Equivalently,

$$(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{I}_p).$$

b) For MLR model 2, the OLS estimator $\hat{\boldsymbol{\phi}}$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \xrightarrow{D} N_{p+1}(\boldsymbol{0}, \sigma^2\ \boldsymbol{V}).$$

c) Suppose the cases $(\boldsymbol{x}_i, Y_i)$ are iid from some population and the MLR model 2 $Y_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ holds. Assume that $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$ exist. Then 4b) holds and

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1})$$

where $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$.

5) A *model for variable selection* is $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$ where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ is a $p \times 1$ vector of predictors, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$

is a $(p-a_S)\times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$. Let $\boldsymbol{x}_I$ be the vector of a terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). If $S \subseteq I$, then $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$ where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \boldsymbol{0}$ if $S \subseteq I$. Note that $\boldsymbol{\beta}_E = \boldsymbol{0}$. Let $k_S = a_S - 1 =$ the number of population active nontrivial predictors. Then $k = a - 1$ is the number of active predictors in the candidate submodel I.

6) If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$, then $\hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. For the OLS model with $S \subseteq I$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \sigma^2\boldsymbol{V}_I)$ where $(\boldsymbol{X}_I^T\boldsymbol{X}_I)/n \xrightarrow{P} \boldsymbol{V}_I^{-1}$.

7) **Theorem 6.19.** Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{MIX} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive π_k by π_j. Assume $\boldsymbol{u}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_p(\boldsymbol{0}, \boldsymbol{V}_{j,0})$. a) Then

$$\boldsymbol{u}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \tag{6.50}$$

where the cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$. Thus $\boldsymbol{u}$ has a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j, $E(\boldsymbol{u}) = \boldsymbol{0}$, and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \sum_j \pi_j \boldsymbol{V}_{j,0}$.

b) Let $\boldsymbol{A}$ be a $g \times p$ full rank matrix with $1 \le g \le p$. Then

$$\boldsymbol{v}_n = \boldsymbol{A}\boldsymbol{u}_n = \sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} = \boldsymbol{v} \tag{6.51}$$

where $\boldsymbol{v}$ has a mixture distribution of the $\boldsymbol{v}_j = \boldsymbol{A}\boldsymbol{u}_j \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T)$ with probabilities π_j.

c) The estimator $\hat{\boldsymbol{\beta}}_{VS}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$: $\sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) = O_P(1)$.

d) If $\pi_d = 1$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{SEL} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \sim N_p(\boldsymbol{0}, \boldsymbol{V}_{d,0})$ where SEL is VS or MIX.

8) **Theorem 6.20.** Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive π_k by π_j. Assume $\boldsymbol{w}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0}^C - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w}_j$. Then

$$\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w} \tag{6.52}$$

where the cdf of $\boldsymbol{w}$ is $F_{\boldsymbol{w}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{w}_j}(\boldsymbol{t})$.

6.18 Complements

Multiple Linear Regression

For linear model theory based on large sample theory, see Olive (2023b). White (1984) also has important theory. Pelawa Watagoda and Olive (2021b) simplified the theory for ridge regression, lasso, and the elastic net.

Some OLS consistency results are given by Lai, Robbins, and Wei (1979). For example, a sufficient condition for $\hat{\boldsymbol{\beta}}_{OLS}$ to be a consistent estimator of $\boldsymbol{\beta}$ is $\text{Cov}(\hat{\boldsymbol{\beta}}_{OLS}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \to \mathbf{0}$ as $n \to \infty$.

Principal components regression (PCR) and partial least squares are MLR estimators. PCR tends to be an inconsistent estimator of $\boldsymbol{\beta}$ unless the probability that the PCR estimator is equal to the OLS estimator goes to 1. PLS may or may not give a consistent estimator of $\boldsymbol{\beta}$ if p/n does not go to zero: rather strong regularity conditions have been used to prove consistency or inconsistency if p/n does not go to zero. See Chun and Keleş (2010), Cook (2018), Cook et al. (2013), and Cook and Forzani (2018, 2019).

Liu (1993, 2003) has some ridge type regression estimators. See Jin and Olive (2023) for large sample theory.

Multivariate Regression

For multivariate regression with more than one response variable, envelope methods are important. See Cook (2018) for references. The theory in Section 6.10 followed Olive (2017b) and Olive, Pelawa Watagoda, and Rupasinghe Arachchige Don (2015) closely.

Variable Selection: An early reference for forward selection is Efroymson (1960). The variable selection theory in this chapter followed Rathnayake and Olive (2023), and Pelawa Watagoda and Olive (2021ab) closely.

Ridge Regression: An important ridge regression paper is Hoerl and Kennard (1970). Also see Gruber (1998). Ridge regression is known as Tikhonov regularization in the numerical analysis literature.

KKT conditions: For MLR, the large sample theory was often simplified using the KKT conditions. Some papers giving KKT conditions include Sun and Zhang (2012), Tibshirani (2013), Zhang and Cheng (2017).

Other Regression Methods

Olive (2004b) and Olive and Hawkins (2005) used 1D regression models with $h(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$, as did Olive (2013a). Olive (2017ab) may be the first publications using general $h(\boldsymbol{x}) = SP$ in the definition of a 1D regression model.

Yee (2015) considers the MLE for many regression models. There are many Econometrics regression methods. See White (1984) and Koenker (2015).

Tay, Narasimhan, and Hastie (2021) describe methods for computing lasso, elastic net, elastic net variable selection, and lasso variable selection for many regression models. Hastie, Tibshirani, and Tibshirani (2020) suggest that lasso variable selection performs well.

Data Splitting

The Olive and Zhang (2023) sequential data splitting algorithm is simple. Let $\lfloor x \rfloor$ be the integer part of x, e.g. $\lfloor 7.7 \rfloor = 7$. Denote the ceiling function by $\lceil x \rceil$, e.g. $\lceil 7.7 \rceil = 8$. Initially, randomly divide the data set into two sets: H_1 with $n_1 \leq n/2$ cases and V_1 with $n - n_1$ cases. Apply lasso on H_1 to get a set of

a_1 predictors, including a constant if a constant is in the model. If $n_1 \geq 10a_1$, set $H = H_1$ and $V = V_1$. Otherwise, randomly select n_1 cases from V_1 to add to H_1 to form H_2. Let V_2 have the remaining cases from V_1. Apply lasso on H_2 to get a set of a_2 predictors. If $n_2 \geq 10a_2$, set $H = H_2$ and $V = V_2$. Continue in this manner, forming sets $(H_1, V_1), (H_2, V_2), ..., (H_d, V_d)$ where H_i has $n_i = in_1$. Stop when $n_d \geq 10a_d$ or $n_{d+1} > \lfloor (n - J)/2 \rfloor$ where $J = 5$ was often used in the simulations. For the second case, use $n_d = \lfloor (n-J)/2 \rfloor$. Then $H = H_d$ and $V = V_d$. Use the model I_d with a_d predictors as the full model for inference with the data in $V = V_d$.

Lasso uses up to n_d active predictors and a constant. If J is an integer between 0 and 5, set $n_1 = \max(1, \lfloor (n-J)/2 \rfloor)$ if $n < 40$. Otherwise, we often used $n_1 = 30$, but changed n_1 to $\lfloor n/2000 \rfloor$ if initially $\lfloor n/(2n_1) \rfloor > 1000$. If $n >> p$, let $n_1 = Kp$ with K a positive integer, such as $K = 10$ or $K = 20$, or use $n_1 \approx Kp \approx n/(2M)$ with $M = \lceil n/(2Kp) \rceil$. If n/p is not large, options include $M = 10$ or $n_1 = Ka_0$ where a_0 is, for example, a guess of a lower bound for the number of active predictors.

6.19 Problems

Also see Problems 3.1 and 3.13.

6.1. For ridge regression, suppose $\boldsymbol{V} = \boldsymbol{\rho}_{\boldsymbol{u}}^{-1}$. Show that if p/n and $\lambda/n = \lambda_{1,n}/n$ are both small, then

$$\hat{\boldsymbol{\eta}}_R \approx \hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda}{n} \boldsymbol{V} \hat{\boldsymbol{\eta}}_{OLS}.$$

6.2. Consider choosing $\hat{\boldsymbol{\eta}}$ to minimize the criterion

$$Q(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) + \frac{\lambda_{1,n}}{a} \sum_{i=1}^{p-1} |\eta_i|^j$$

where $\lambda_{1,n} \geq 0$, $a > 0$, and $j > 0$ are known constants. Consider the regression methods OLS, forward selection, lasso, PLS, PCR, ridge regression, and relaxed lasso.
a) Which method corresponds to $j = 1$?
b) Which method corresponds to $j = 2$?
c) Which method corresponds to $\lambda_{1,n} = 0$?

6.3. For ridge regression, let $\boldsymbol{A}_n = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{W}$ and $\boldsymbol{B}_n = [\boldsymbol{I}_{p-1} - \lambda_{1,n}(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}]$. Show $\boldsymbol{A}_n - \boldsymbol{B}_n = \boldsymbol{0}$.

6.4. Suppose $\hat{\boldsymbol{Y}} = \boldsymbol{HY}$ where $\boldsymbol{H}$ is an $n \times n$ hat matrix. Then the degrees of freedom $df(\hat{\boldsymbol{Y}}) = tr(\boldsymbol{H})$ = sum of the diagonal elements of $\boldsymbol{H}$. An estimator with low degrees of freedom is inflexible while an estimator with high degrees of freedom is flexible. If the degrees of freedom is too low, the estimator tends to underfit while if the degrees of freedom is to high, the estimator tends to overfit.

a) Find $df(\hat{\boldsymbol{Y}})$ if $\hat{\boldsymbol{Y}} = \overline{Y}\mathbf{1}$ which uses $\boldsymbol{H} = (h_{ij})$ where $h_{ij} \equiv 1/n$ for all i and j. This inflexible estimator uses the sample mean $\overline{Y}$ of the response variable as $\hat{Y}_i$ for $i = 1, ..., n$.

b) Find $df(\hat{\boldsymbol{Y}})$ if $\hat{\boldsymbol{Y}} = \boldsymbol{Y} = \boldsymbol{I}_n\boldsymbol{Y}$ which uses $\boldsymbol{H} = \boldsymbol{I}_n$ where $h_{ii} = 1$. This bad flexible estimator interpolates the response variable.

6.5. Suppose $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$, $\boldsymbol{Z} = \boldsymbol{W\eta} + \boldsymbol{e}$, $\hat{\boldsymbol{Z}} = \boldsymbol{W}\hat{\boldsymbol{\eta}}$, $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$, and $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Z}} + \overline{\boldsymbol{Y}}$. Let the $n \times p$ matrix $\boldsymbol{W}_1 = [\mathbf{1} \quad \boldsymbol{W}]$ and the $p \times 1$ vector $\hat{\boldsymbol{\eta}}_1 = (\overline{Y} \quad \hat{\boldsymbol{\eta}}^T)^T$ where the scalar $\overline{Y}$ is the sample mean of the response variable. Show $\hat{\boldsymbol{Y}} = \boldsymbol{W}_1\hat{\boldsymbol{\eta}}_1$.

6.6. Let $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_S^T)^T$. Consider choosing $\hat{\boldsymbol{\beta}}$ to minimize the criterion

$$Q(\boldsymbol{\beta}) = RSS(\boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}_S\|_2^2 + \lambda_2\|\boldsymbol{\beta}_S\|_1$$

where $\lambda_i \geq 0$ for $i = 1, 2$.

a) Which values of λ_1 and λ_2 correspond to ridge regression?

b) Which values of λ_1 and λ_2 correspond to lasso?

c) Which values of λ_1 and λ_2 correspond to elastic net?

d) Which values of λ_1 and λ_2 correspond to the OLS full model?

6.7. Suppose that $Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i,5} + e_i$.

a) Testing $H_0 : \beta_2 = \beta_4 = \beta_5 = 0$ is equivalent to testing $H_0 : \boldsymbol{A\beta} = \mathbf{0}$. What is $\boldsymbol{A}$? Hint: want $\boldsymbol{A\beta} = (\beta_2, \beta_4, \beta_5)^T$.

b) Testing $H_0 : \beta_2 = \beta_4 = \beta_5$ is equivalent to testing $H_0 : \boldsymbol{A\beta} = \mathbf{0}$. What is $\boldsymbol{A}$? Hint: want, for example, $\boldsymbol{A\beta} = (\beta_2 - \beta_4, \beta_2 - \beta_5)^T$.

6.8. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid 7×1 random vectors where $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{x}_i) = \sum_j \pi_j \boldsymbol{\Sigma}_j$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{d})$ for appropriate vector $\boldsymbol{d}$.

6.9. Let $\boldsymbol{\Sigma}_i$ be the nonsingular population covariance matrix of the ith treatment group or population. To simplify the large sample theory, assume $n_i = \pi_i n$ where $0 < \pi_i < 1$ and $\sum_{i=1}^2 \pi_i = 1$. Let T_i be a multivariate location estimator such that $\sqrt{n_i}(T_i - \boldsymbol{\mu}_i) \xrightarrow{D} N_m(\mathbf{0}, \boldsymbol{\Sigma}_i)$, and $\sqrt{n}(T_i - \boldsymbol{\mu}_i) \xrightarrow{D} N_m\left(\mathbf{0}, \frac{\boldsymbol{\Sigma}_i}{\pi_i}\right)$ for $i = 1, 2$. Assume $T_1 \perp\!\!\!\perp T_2$.

Then

$$\sqrt{n}\begin{bmatrix} T_1 - \boldsymbol{\mu}_1 \\ T_2 - \boldsymbol{\mu}_2 \end{bmatrix} \xrightarrow{D} \boldsymbol{u}.$$

Find the distribution of $\boldsymbol{u}$.

6.10. When the errors e_i are iid, a common assumption for OLS theory is

$$n(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \hat{\boldsymbol{V}} = \begin{pmatrix} \hat{\boldsymbol{V}}_{11} & \hat{\boldsymbol{V}}_{12} \\ \hat{\boldsymbol{V}}_{21} & \hat{\boldsymbol{V}}_{22} = n\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}/(n-1) \end{pmatrix} \xrightarrow{P} \boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{pmatrix}.$$

Then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \xrightarrow{P} \boldsymbol{A}$. What is $\boldsymbol{A}$? Hint: $\boldsymbol{A}$ is not $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$, in general.

6.11. Let $\boldsymbol{\Sigma}_i$ be the nonsingular population covariance matrix of the ith treatment group or population. To simplify the large sample theory, assume $n_i = \pi_i n$ where $0 < \pi_i < 1$ and $\sum_{i=1}^{2} \pi_i = 1$. Let T_i be a multivariate location estimator such that $\sqrt{n_i}(T_i - \boldsymbol{\mu}_i) \xrightarrow{D} N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$, and $\sqrt{n}(T_i - \boldsymbol{\mu}_i) \xrightarrow{D} N_m\left(\boldsymbol{0}, \frac{\boldsymbol{\Sigma}_i}{\pi_i}\right)$ for $i = 1, 2$. Assume $T_1 \perp\!\!\!\perp T_2$.

Then

$$\sqrt{n}\begin{bmatrix} T_1 - \boldsymbol{\mu}_1 \\ T_2 - \boldsymbol{\mu}_2 \end{bmatrix} \xrightarrow{D} \boldsymbol{u}. \tag{6.53}$$

You found the distribution of $\boldsymbol{u}$ in Problem 6.9.

Now

$$\sqrt{n}[T_1 - T_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)] \xrightarrow{D} \boldsymbol{w}.$$

Find the distribution of $\boldsymbol{w}$. Hint: multiply both sides of (1) by $\boldsymbol{A} = [\boldsymbol{I}_m \ -\boldsymbol{I}_m]$ and find the distribution of $\boldsymbol{w} = \boldsymbol{A}\boldsymbol{u}$.

6.12. GLMs are fit by maximum likelihood. Thus $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$. What is $\boldsymbol{\Sigma}$?

6.13. Suppose $Y_i \sim D(\boldsymbol{x}_i^T\boldsymbol{\beta})$ where the Y_i are independent, D is a parametric distribution that depends on $\boldsymbol{x}_i$ only though $\boldsymbol{x}_i^T\boldsymbol{\beta}$ for $i = 1, ..., n$, and $\boldsymbol{\beta}$ contains the unknown parameters. Several GLMs have this form, e.g. $Y_i \sim Poisson[\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})]$. Then the MLE $\hat{\boldsymbol{\beta}}_n$ satisfies $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{I}_1^{-1}(\boldsymbol{\beta}))$ where $\boldsymbol{I}_1^{-1}(\boldsymbol{\beta})$ is the inverse Fisher information matrix and the $\boldsymbol{x}_i$ are treated as constants. The MLE is obtained by regression the Y_i on the $\boldsymbol{x}_i$.

The parametric bootstrap generates independent $Y_i^* \sim D(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_n)$. Fix n and generate Y_i^* for $i = 1, ..., m$. Obtain the bootstrap statistic $\hat{\boldsymbol{\beta}}^*$ by regressing the Y_i^* on the $\boldsymbol{x}_i$. Then the Y_i^* with $\hat{\boldsymbol{\beta}}_n$ in place of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}^*$ in place of $\hat{\boldsymbol{\beta}}_n$ satisfy the above theory.

For fixed n, find the limiting distribution of

$$\sqrt{m}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}_n).$$

6.14. For the parametric bootstrap, $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{e}^*$ where $\boldsymbol{e}^* = (e_1^*, ..., e_m^*)^T$ and the e_i^* are iid $N(0, \sigma_n^2)$ random variables. Assume this model satisfies the OLS CLT as $m \to \infty$ where the OLS estimator is $\hat{\boldsymbol{\beta}}^*$ and $\boldsymbol{\beta}$ is replaced by $\hat{\boldsymbol{\beta}}$. Note that σ^2 is replaced by σ_n^2. Assume $\hat{\boldsymbol{\beta}}$ is a $p \times 1$ vector and $\boldsymbol{Y}^*$ is an $m \times 1$ vector. a) Then use the OLS CLT to find the limiting distribution of $\sqrt{m}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}})$ as $m \to \infty$.

b) In a) you should get $\sqrt{m}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma}_n)$ where $\boldsymbol{\Sigma}_n \xrightarrow{P} \boldsymbol{\Sigma}$. The bootstrap proof technique says that suppose $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma}_n \xrightarrow{P} \boldsymbol{\Sigma}$ as $n \to \infty$, and for fixed n, $\sqrt{m}(T^*_{n,m} - T_n) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma}_n)$ as $m \to \infty$. Then $\sqrt{n}(T^*_n - T_n) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma})$ as $n \to \infty$. Find the limiting distribution $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}})$ as $n \to \infty$ where you may plug in $\boldsymbol{\Sigma}$ in to the result (you do not need to compute $\boldsymbol{\Sigma}$).

6.15. Use the following results. A) Suppose $\boldsymbol{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then
i) $\boldsymbol{AX} \sim N_q(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^T)$.
ii) $\boldsymbol{a} + b\boldsymbol{X} \sim N_k(\boldsymbol{a} + b\boldsymbol{\mu}, b^2\boldsymbol{\Sigma})$.
iii) $\boldsymbol{AX} + \boldsymbol{d} \sim N_q(\boldsymbol{A\mu} + \boldsymbol{d}, \boldsymbol{A\Sigma A}^T)$.
(Find the mean and covariance matrix of the left hand side and plug in those values for the right hand side. **Be careful with the dimension** k or q.)

B) Suppose $\boldsymbol{X}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then
i) $\boldsymbol{AX}_n \xrightarrow{D} N_q(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^T)$.
ii) $\boldsymbol{a} + b\boldsymbol{X}_n \xrightarrow{D} N_k(\boldsymbol{a} + b\boldsymbol{\mu}, b^2\boldsymbol{\Sigma})$.
iii) $\boldsymbol{AX}_n + \boldsymbol{d} \xrightarrow{D} N_q(\boldsymbol{A\mu} + \boldsymbol{d}, \boldsymbol{A\Sigma A}^T)$.
(The behavior of convergence in distribution to a MVN distribution is much like the behavior of the MVN distributions in A).)

By the OLS CLT, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \sigma^2\ \boldsymbol{W})$. Hence the limiting distribution of of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is the $N_p(\mathbf{0}, \sigma^2\ \boldsymbol{W})$ distribution. Let $\boldsymbol{A}$ be a constant $r \times p$ matrix. Find the limiting distribution of $\boldsymbol{A}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.

6.16. Use the following results. Suppose $\boldsymbol{X}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}$ be a $q \times k$ constant matrix. Let $\boldsymbol{a}$ be a $k \times 1$ constant vector and let b be a constant. Let $\boldsymbol{d}$ be a $q \times 1$ constant vector. Then
i) $\boldsymbol{AX}_n \xrightarrow{D} N_q(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^T)$.
ii) $\boldsymbol{a} + b\boldsymbol{X}_n \xrightarrow{D} N_k(\boldsymbol{a} + b\boldsymbol{\mu}, b^2\boldsymbol{\Sigma})$.
iii) $\boldsymbol{AX}_n + \boldsymbol{d} \xrightarrow{D} N_q(\boldsymbol{A\mu} + \boldsymbol{d}, \boldsymbol{A\Sigma A}^T)$.

Suppose $\boldsymbol{X}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

a) Let $\boldsymbol{C}$ be a $k \times k$ constant matrix. Then find the limiting distribution of $\boldsymbol{CX}_n$: that is, $\boldsymbol{CX}_n \xrightarrow{D} \boldsymbol{Z}$. Find $\boldsymbol{Z}$.

b) Suppose $\boldsymbol{C}_n \xrightarrow{P} \boldsymbol{C}$. Thus $\boldsymbol{C}_n - \boldsymbol{C} = o_P(1)$. Find the limiting distribution of $\boldsymbol{C}_n\boldsymbol{X}_n = (\boldsymbol{C}_n - \boldsymbol{C} + \boldsymbol{C})\boldsymbol{X}_n = (\boldsymbol{C}_n - \boldsymbol{C})\boldsymbol{X}_n + \boldsymbol{CX}_n$.

6.17. Let the linear model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ has full rank p, $E(\boldsymbol{e}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$. By the OLS CLT, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \sigma^2\ \boldsymbol{W})$. Let $\boldsymbol{a}$ be a $p \times 1$ constant vector. Then for a large class of iid error distributions, what is the limiting distribution of $\boldsymbol{a}^T\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \sqrt{n}(\boldsymbol{a}^T\hat{\boldsymbol{\beta}} - \boldsymbol{a}^T\boldsymbol{\beta})$?

6.18. Suppose that $Y_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ where the $e_i = \sigma_i\epsilon_i$ where the ϵ_i iid with $E(\epsilon_i) = 0$ and $V(\epsilon_i) = 1$. Then the e_i are independent with $E(e_i) = 0$ and $V(e_i) = \sigma_i^2$. This MLR model can be written as $\boldsymbol{Y} = \alpha\mathbf{1} + \boldsymbol{X\beta} + \boldsymbol{e}$. We will

assume that the cases $(\boldsymbol{x}_i^T, Y_i)^T$ are iid. Fit the model with OLS to get $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ and the residuals r_i. The nonparametric bootstrap samples the $(\boldsymbol{x}_i, Y_i, r_i)$ with replacement to form the MLR model $\boldsymbol{Y}^* = \hat{\alpha}\mathbf{1} + \boldsymbol{X}^*\hat{\boldsymbol{\beta}} + \boldsymbol{r}^*$ where with respect to the bootstrap distribution, the r_i^* are iid with $E(r_i^*) = 0$. This bootstrap model has the $(\boldsymbol{x}_i^{*T}, Y_i^*)^T$ iid with respect to the bootstrap distribution.

The MLR model $\boldsymbol{Y}^* = \hat{\alpha}\mathbf{1} + \boldsymbol{X}^*\hat{\boldsymbol{\beta}} + \boldsymbol{r}^*$ is the bootstrap data set, and OLS is fit to the model to obtain the bootstrapped statistic $(\hat{\alpha}^* = \overline{Y^*} - \hat{\boldsymbol{\beta}}^{*T}\overline{\boldsymbol{x}^*}, \hat{\boldsymbol{\beta}}^* = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}^*}^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}^*Y^*})$.

a) By the second method to compute OLS, $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$. Since the bootstrap distribution for the nonparametric bootstrap is the empirical distribution, it can be shown that $[\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^*]^{-1} \xrightarrow{P} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}^* \xrightarrow{P} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$. Prove that $\hat{\boldsymbol{\beta}}^* = [\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^*]^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}^* \xrightarrow{P} \hat{\boldsymbol{\beta}}$.

b) By the second method to compute OLS, $\hat{\alpha} = \overline{Y} - \hat{\boldsymbol{\beta}}^T\overline{\boldsymbol{x}}$. It can be shown that $\overline{Y^*} \xrightarrow{P} \overline{Y}$ and $\overline{\boldsymbol{x}^*} \xrightarrow{P} \overline{\boldsymbol{x}}$. Prove that $\hat{\alpha}^* = \overline{Y^*} - \hat{\boldsymbol{\beta}}^{*T}\overline{\boldsymbol{x}^*} \xrightarrow{P} \hat{\alpha}$.

6.19. Suppose $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$. Find the distribution of $\boldsymbol{HY}$ if $\boldsymbol{H}$ is an $n \times n$ constant matrix such that $\boldsymbol{HX} = \boldsymbol{X}$ and $\boldsymbol{H} = \boldsymbol{H}^T = \boldsymbol{HH} = \boldsymbol{H}^2$. Simplify.

6.20. Parametric bootstrap: With respect to the bootstrap distribution, quantities with an asterisk are random, while quantities from the sample act as constant scalars, vectors or matrices. Thus $\boldsymbol{X}$, $\sigma_n^2 = MSE$, $\boldsymbol{X}_I$, $\boldsymbol{P}$, $\hat{\boldsymbol{\beta}}_I$, and $\hat{\boldsymbol{\beta}}$ are not random with respect to the bootstrap distribution. An exception is that the ϵ_i are random variables with respect to the bootstrap distribution.

Suppose $\boldsymbol{Y}^* \sim N_n(\boldsymbol{X}\hat{\boldsymbol{\beta}}, \sigma_n^2\boldsymbol{I}_n)$. Hence $Y_i^* = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} + \epsilon_i$ where $E(\epsilon_i) = 0$ and $V(\epsilon_i) = \sigma_n^2$. Hence $\boldsymbol{AY}^* \sim N_g(\boldsymbol{AX}\hat{\boldsymbol{\beta}}, \sigma_n^2\boldsymbol{AA}^T)$ if $\boldsymbol{A}$ is a $g \times n$ constant matrix. Recall that $\boldsymbol{X}$ is an $n \times p$ constant matrix. **Simplify quantities when possible**.

a) What is the distribution of $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}^*$?

b) Using a), what is $E(\hat{\boldsymbol{\beta}}^*)$?

c) Recall that $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{PY}$ where $\boldsymbol{P} = \boldsymbol{H}$, $\boldsymbol{PX} = \boldsymbol{X}$, $\boldsymbol{PX}_I = \boldsymbol{X}_I$, and $\boldsymbol{X}_I^T\boldsymbol{P} = \boldsymbol{X}_I^T$. What is the distribution of $\hat{\boldsymbol{\beta}}_I^* = (\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y}^*$ if $\hat{\boldsymbol{\beta}}_I^*$ is $k \times 1$? Hint: Note that $\hat{\boldsymbol{\beta}}_I = (\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y}$. The mean of the distribution is $(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T E(\boldsymbol{Y}^*) = (\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{PY}$.

6.21. Estimating the $\boldsymbol{\eta}_i$ and performing the OLS regression of Y on $(\hat{\boldsymbol{\eta}}_1^T\boldsymbol{x}, \hat{\boldsymbol{\eta}}_2^T\boldsymbol{x}, ..., \hat{\boldsymbol{\eta}}_k^T\boldsymbol{x})$ and a constant gives the k-component estimator, e.g. the k-component PLS estimator $\hat{\boldsymbol{\beta}}_{kPLS}$ or the k-component PCR estimator, for $k = 1, ..., J$ where $J \leq p$ and the p-component estimator is the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$.

In the fixed p setting, model selection PLS and model selection PCR can be shown to give predictions similar to that of the OLS full model. To see this, variable selection with the $C_p(I)$ criterion will be useful. Consider the OLS regression of Y on a constant and $\boldsymbol{w} = (W_1, ..., W_p)^T$ where, for example,

$W_j = x_j$ or $W_j = \hat{\boldsymbol{\eta}}_j^T \boldsymbol{x}$. Let I index the variables in the model so $I = \{1, 2, 4\}$ means that W_1, W_2, and W_4 were selected. The full model $I = F$ uses all p predictors and the constant with $\boldsymbol{\beta}_I = \boldsymbol{\beta}_F = \boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$. Let r be the residuals from the full OLS model and let r_I be the residuals from model I that uses $\hat{\boldsymbol{\beta}}_I$. Suppose model I uses $k = k_I$ predictors including a constant with $2 \leq k \leq p+1$. It can be shown that $C_p(I) \leq 2k$ implies that

$$\mathrm{corr(r, r_I)} \geq \sqrt{1 - \frac{\mathrm{p+1}}{\mathrm{n}}}. \tag{6.54}$$

Let the model I_{min} minimize the C_p criterion among the models considered with $C_p(I) \leq 2k_I$. Then $C_p(I_{min}) \leq C_p(F) = p + 1$.

If PLS or PCR is selected using model selection (on models $I_1, ..., I_p$ with $I_j = \{1, 2, ..., j\}$ corresponding to the j-component regression) with the $C_p(I)$ criterion, and $n \geq 20(p+1)$, then $\mathrm{corr}(r, r_{I_{min}}) \geq \sqrt{a} = b$. Find a and b.

Hint: in Equation (6.54), replace n by $20(p+1)$.

6.22. Consider the MLR model $Y_i = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i + e_i$ for $i = 1, ..., n$ where the e_i are iid with $E(e_i) = 0$ and $V(e_i) = \sigma^2$. Conditional on the $\boldsymbol{x}_i$, under mild regularity conditions

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) = \sqrt{n}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}).$$

Find the limiting distribution of

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \sqrt{n}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} - \boldsymbol{\beta}).$$

6.23. Consider the MLR model $Y_i = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i + e_i$ for $i = 1, ..., n$ where the $(\boldsymbol{x}_i^T, Y_i)^T$ are iid. Then under mild regularity conditions,

$$\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) = \sqrt{n}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} - \boldsymbol{\Sigma}_{\boldsymbol{x}Y}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}).$$

Let $\boldsymbol{A}$ be an $r \times p$ constant matrix of full rank r. Find the limiting distribution of

$$\boldsymbol{A}\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) = \boldsymbol{A}\sqrt{n}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} - \boldsymbol{\Sigma}_{\boldsymbol{x}Y}).$$

(Note: the model in Problem 6.22 is not the model in problem 6.23, so the large sample theory differs.)

6.24. Let the linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ has full rank p, $E(\boldsymbol{e}) = \boldsymbol{0}$ and $Cov(\boldsymbol{e}) = \sigma^2 \boldsymbol{I}$. Assume $\boldsymbol{X}^T\boldsymbol{X}/n \to \boldsymbol{W}^{-1}$ as $n \to \infty$. Then for a large class of iid error distributions, what is the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$? Hint: use the least squares central limit theorem.

6.25. Let the full model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \boldsymbol{e}$. Suppose that a submodel I uses the constant and $k-1$ nontrivial predictors $X_{i1}, ..., X_{i,k-1}$. Let $X_{i,k}, X_{i,k+1}, ..., X_{i,p-1}$ denote the predictors left out of the model. Then the partial F test statistic F_I tests whether submodel I is good or whether at least one of the predictors left out of the model is needed. Let $\boldsymbol{r}$ denote the

residuals from the full model, let $\boldsymbol{r}_I$ denote the residuals from the submodel, let $C_p(I)$ denote the C_p criterion for the submodel I, and let n be the sample size. Then it can be shown that

$$\mathrm{corr}(\boldsymbol{r}, \boldsymbol{r}_\mathrm{I}) = \sqrt{\frac{\mathrm{n-p}}{\mathrm{C_p(I)+n-2k}}} = \sqrt{\frac{\mathrm{n-p}}{\mathrm{(p-k)F_I+n-p}}}.$$

Assume that $-p \leq C_p(I) \leq k$ and $0 \leq F_I \leq 1$. Then what happens to $\mathrm{corr}(\boldsymbol{r}, \boldsymbol{r}_\mathrm{I})$ as $n \to \infty$?

6.26. Suppose $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{r}^W$ where $E(\boldsymbol{r}^W) = \boldsymbol{0}$ and $Cov(\boldsymbol{r}^W) = Cov(\boldsymbol{Y}^*) = diag(r_i^2) = diag(r_1^2, ..., r_n^2)$. Then $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}^*$ is the least squares estimator from regressing $\boldsymbol{Y}^*$ on $\boldsymbol{X}$, an $n \times p$ constant matrix. This model is used for the wild bootstrap. Simplify quantities when possible. (Can simplify a), but can't simplify b) much. With respect to the bootstrap distribution, $\boldsymbol{Y}^*$ and $\boldsymbol{r}^W$ are random vectors, but $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is a constant vector with respect to the bootstrap distribution.)

a) What is $E(\hat{\boldsymbol{\beta}}^*)$?

b) What is $Cov(\hat{\boldsymbol{\beta}}^*)$?

6.27. Assume that

$$\sqrt{n}\left[\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix} - \begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}\right] \xrightarrow{D} N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \theta\\ \theta & \sigma_2^2\end{pmatrix}\right) \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Find the limiting distribution of

$$\sqrt{n}[(\hat{\beta}_1 - \hat{\beta}_2) - (\beta_1 - \beta_2)] = (1 \quad -1)\sqrt{n}\left[\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix} - \begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}\right].$$

Hint: $\boldsymbol{A} = (1 \ -1)$. Find $\boldsymbol{A}\boldsymbol{\mu}$ and $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T$.

6.28. Suppose $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ satisfies the OLS CLT where $\boldsymbol{X}$ is $n \times p$. Let $r_i = Y_i - \hat{Y}_i$ be the ith OLS residual for $i = 1, ..., n$ where a constant is in the model. The residual bootstrap draws the residuals with replacement to form the model $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{r}^*$.

Suppose that for some large fixed m, OLS is fit to find $\hat{\boldsymbol{\beta}}_m$ and the m OLS residuals. Then as $n \to \infty$, the m residuals are drawn with replacement to form $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_m + \boldsymbol{r}^*$ where $Y_i^* = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_m + r_i^*$ for $i = 1, ..., n$. This model satisfies the OLS CLT with the r_i^* iid with respect to the bootstrap distribution, $E(r_i^*) = 0$, and $V(r_i^*) = \sigma_m^2 = \frac{1}{m}\sum_{i=1}^{m} r_i^2 = \frac{m-1}{m}MSE(m)$ where $MSE(m) \xrightarrow{P} \sigma^2 = V(e_i)$ as $m \to \infty$. The bootstrap estimator $\hat{\boldsymbol{\beta}}^*$ is found by fitting OLS to the model $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_m + \boldsymbol{r}^*$. By the OLS CLT,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}_m) \xrightarrow{D} \boldsymbol{u}.$$

What is the distribution of $\boldsymbol{u}$?

Chapter 7
Experimental Design and One Way MANOVA

7.1 Introduction

Definition 7.1. The **response variables** are the variables that you want to predict. The **predictor variables** are the variables used to predict the response variables.

Notation. A multivariate linear model has $m \geq 2$ response variables. A *multiple linear model = univariate linear model* has $m = 1$ response variable, but at least two nontrivial predictors, and usually a constant (so $p \geq 3$). A simple linear model has $m = 1$, one nontrivial predictor, and usually a constant (so $p = 2$). Multiple linear regression models and ANOVA models are special cases of multiple linear models.

Definition 7.2. The **multivariate linear model**

$$\boldsymbol{y}_i = \boldsymbol{B}^T \boldsymbol{x}_i + \boldsymbol{\epsilon}_i$$

for $i = 1, ..., n$ has $m \geq 2$ response variables $Y_1, ..., Y_m$ and p predictor variables $x_1, x_2, ..., x_p$. The ith case is $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T) = (x_{i1}, x_{i2}, ..., x_{ip}, Y_{i1}, ..., Y_{im})$. If a constant $x_{i1} = 1$ is in the model, then x_{i1} could be omitted from the case. The model is written in matrix form as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$ where the matrices are the same as those in Section 6.14. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = (\sigma_{ij})$ for $k = 1, ..., n$. Then the $p \times m$ coefficient matrix $\boldsymbol{B} = \left[\boldsymbol{\beta}_1 \ \boldsymbol{\beta}_2 \ \ldots \ \boldsymbol{\beta}_m \right]$ and the $m \times m$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{X}\boldsymbol{B}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T \boldsymbol{\beta}_j$. The $\boldsymbol{\epsilon}_i$ are assumed to be iid. The univariate linear model corresponds to $m = 1$ response variable, and is written in matrix form as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Subscripts are needed for the m univariate linear models $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where $E(\boldsymbol{e}_j) = \boldsymbol{0}$. For the multivariate linear model, $\text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij} \ \ \boldsymbol{I}_n$ for $i, j = 1, ..., m$ where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

Definition 7.3. The multivariate analysis of variance (**MANOVA model**) $\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$ has $m \geq 2$ response variables $Y_1, ..., Y_m$ and p predictor variables $X_1, X_2, ..., X_p$. The MANOVA model is a special case of the multivariate linear model. For the MANOVA model, the predictors are not quantitative variables, so the predictors are indicator variables. Sometimes the trivial predictor $\mathbf{1}$ is also in the model. In matrix form, the MANOVA model is $\boldsymbol{Z} = \boldsymbol{XB} + \boldsymbol{E}$. The model has $E(\boldsymbol{\epsilon}_k) = \mathbf{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma_\epsilon} = (\sigma_{ij})$ for $k = 1, ..., n$. Also $E(\boldsymbol{e}_i) = \mathbf{0}$ while $\text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij}\boldsymbol{I}_n$ for $i, j = 1, ..., m$. Then $\boldsymbol{B}$ and $\boldsymbol{\Sigma_\epsilon}$ are unknown matrices of parameters to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{XB}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T\boldsymbol{\beta}_j$.

The data matrix $\boldsymbol{W}_d = [\boldsymbol{X} \quad \boldsymbol{Z}]$. If the model contains a constant, then usually the first column of ones $\mathbf{1}$ of $\boldsymbol{X}$ is omitted from the data matrix for software such as *R* and *SAS*.

Each response variable in a MANOVA model follows an ANOVA model $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where it is assumed that $E(\boldsymbol{e}_j) = \mathbf{0}$ and $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$. Hence the errors corresponding to the jth response are uncorrelated with variance $\sigma_j^2 = \sigma_{jj}$. Notice that the **same design matrix** $\boldsymbol{X}$ of predictors is used for each of the m models, but the jth response variable vector $\boldsymbol{Y}_j$, coefficient vector $\boldsymbol{\beta}_j$, and error vector $\boldsymbol{e}_j$ change and thus depend on j. Hence for a one way MANOVA model, each response variable follows a one way ANOVA model, while for a two way MANOVA model, each response variable follows a two way ANOVA model for $j = 1, ..., m$.

Once the ANOVA model is fixed, e.g. a one way ANOVA model, the design matrix $\boldsymbol{X}$ depends on the parameterization of the ANOVA model. The fitted values and residuals are the same for each parameterization, but the interpretation of the parameters depends on the parameterization.

Now consider the ith case $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T)$ which corresponds to the ith row of $\boldsymbol{X}$ and the ith row of $\boldsymbol{Z}$. Then $\boldsymbol{y}_i = E(\boldsymbol{y}_i) + \boldsymbol{\epsilon}_i$ where

$$E(\boldsymbol{y}_i) = \boldsymbol{B}^T\boldsymbol{x}_i = \begin{bmatrix} \boldsymbol{x}_i^T\boldsymbol{\beta}_1 \\ \boldsymbol{x}_i^T\boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{x}_i^T\boldsymbol{\beta}_m \end{bmatrix}.$$

The notation $\boldsymbol{y}_i|\boldsymbol{B}^T\boldsymbol{x}_i$ and $E(\boldsymbol{y}_i|\boldsymbol{B}^T\boldsymbol{x}_i)$ is more accurate, but usually the conditioning is suppressed. Taking $E(\boldsymbol{y}_i|\boldsymbol{B}^T\boldsymbol{x}_i)$ to be a constant, $\boldsymbol{y}_i$ and $\boldsymbol{\epsilon}_i$ have the same covariance matrix. In the MANOVA model, this covariance matrix $\boldsymbol{\Sigma_\epsilon}$ does not depend on i. Observations from different cases are uncorrelated (often independent), but the m errors for the m different response variables for the *same case* are correlated.

Let $\hat{\boldsymbol{B}}$ be the MANOVA estimator of $\boldsymbol{B}$. MANOVA models are often fit by least squares. Then the **least squares estimators** are

$$\hat{\boldsymbol{B}} = \hat{\boldsymbol{B}}_g = (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{Z} = \left[\hat{\boldsymbol{\beta}}_1\ \hat{\boldsymbol{\beta}}_2\ \ldots\ \hat{\boldsymbol{\beta}}_m\right]$$

where $(\boldsymbol{X}^T\boldsymbol{X})^-$ is a generalized inverse of $\boldsymbol{X}^T\boldsymbol{X}$. Here $\hat{\boldsymbol{B}}_g$ depends on the generalized inverse. If $\boldsymbol{X}$ has full rank p then $(\boldsymbol{X}^T\boldsymbol{X})^- = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and $\hat{\boldsymbol{B}}$ is unique.

Definition 7.4. The *predicted values* or *fitted values*

$$\hat{\boldsymbol{Z}} = \boldsymbol{X}\hat{\boldsymbol{B}} = \left[\hat{\boldsymbol{Y}}_1\ \hat{\boldsymbol{Y}}_2\ \ldots\ \hat{\boldsymbol{Y}}_m\right] = \begin{bmatrix} \hat{Y}_{1,1} & \hat{Y}_{1,2} & \ldots & \hat{Y}_{1,m} \\ \hat{Y}_{2,1} & \hat{Y}_{2,2} & \ldots & \hat{Y}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{Y}_{n,1} & \hat{Y}_{n,2} & \ldots & \hat{Y}_{n,m} \end{bmatrix}.$$

The *residuals* $\hat{\boldsymbol{E}} = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = \boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}} =$

$$\begin{bmatrix} \hat{\boldsymbol{\epsilon}}_1^T \\ \hat{\boldsymbol{\epsilon}}_2^T \\ \vdots \\ \hat{\boldsymbol{\epsilon}}_n^T \end{bmatrix} = \left[\hat{\boldsymbol{r}}_1\ \hat{\boldsymbol{r}}_2\ \ldots\ \hat{\boldsymbol{r}}_m\right] = \begin{bmatrix} \hat{\epsilon}_{1,1} & \hat{\epsilon}_{1,2} & \ldots & \hat{\epsilon}_{1,m} \\ \hat{\epsilon}_{2,1} & \hat{\epsilon}_{2,2} & \ldots & \hat{\epsilon}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\epsilon}_{n,1} & \hat{\epsilon}_{n,2} & \ldots & \hat{\epsilon}_{n,m} \end{bmatrix}.$$

These quantities can be found by fitting m ANOVA models $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ to get $\widehat{\boldsymbol{\beta}}_j$, $\widehat{\boldsymbol{Y}}_j = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j$, and $\hat{\boldsymbol{r}}_j = \boldsymbol{Y}_j - \hat{\boldsymbol{Y}}_j$ for $j = 1, ..., m$. Hence $\hat{\epsilon}_{i,j} = Y_{i,j} - \hat{Y}_{i,j}$ where $\hat{\boldsymbol{Y}}_j = (\hat{Y}_{1,j}, ..., \hat{Y}_{n,j})^T$. Finally, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d} =$

$$\frac{(\boldsymbol{Z} - \hat{\boldsymbol{Z}})^T(\boldsymbol{Z} - \hat{\boldsymbol{Z}})}{n-d} = \frac{(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}})^T(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}})}{n-d} = \frac{\hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}}{n-d} = \frac{1}{n-d}\sum_{i=1}^{n}\hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T.$$

The choices $d = 0$ and $d = p$ are common. Let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ be the usual estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ for the MANOVA model. If least squares is used with a full rank $\boldsymbol{X}$, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d=p}$.

7.2 One Way MANOVA

Using double subscripts will be useful for describing the one way MANOVA model. Suppose there are independent random samples of size n_i from p different populations (treatments), or n_i cases are randomly assigned to p treatment groups. Then $n = \sum_{i=1}^{p} n_i$ and the group sample sizes are n_i for $i = 1, ..., p$. Assume that m response variables $\boldsymbol{y}_{ij} = (Y_{ij1}, ..., Y_{ijm})^T$ are measured for the ith treatment group and the jth case (often an individual or thing) in the group. Hence $i = 1, ..., p$ and $j = 1, ..., n_i$. The Y_{ijk} follow different one way ANOVA models for $k = 1, ..., m$. Assume $E(\boldsymbol{y}_{ij}) = \boldsymbol{\mu}_i$ and

$\text{Cov}(\boldsymbol{y}_{ij}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. Hence the p treatments have different mean vectors $\boldsymbol{\mu}_i$, but common covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$.

The one way MANOVA is used to test $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_p$. Often $\boldsymbol{\mu}_i = \boldsymbol{\mu} + \boldsymbol{\tau}_i$, so H_0 becomes $H_0 : \boldsymbol{\tau}_1 = \cdots = \boldsymbol{\tau}_p$. If $m = 1$, the one way MANOVA model is the one way ANOVA model. MANOVA is useful since it takes into account the correlations between the m response variables. Performing m ANOVA tests fails to account for these correlations, but can be a useful diagnostic. The Hotelling's T^2 test that uses a common covariance matrix is a special case of the one way MANOVA model with $p = 2$.

Let $\boldsymbol{\mu}_i = \boldsymbol{\mu} + \boldsymbol{\tau}_i$ where $\sum_{i=1}^{p} n_i \boldsymbol{\tau}_i = 0$. The jth case from the ith population or treatment group is $\boldsymbol{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\tau}_i + \boldsymbol{\epsilon}_{ij}$ where $\boldsymbol{\epsilon}_{ij}$ is an error vector, $i = 1, ..., p$ and $j = 1, ..., n_i$. Let $\overline{\boldsymbol{y}} = \hat{\boldsymbol{\mu}} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} \boldsymbol{y}_{ij}/n$ be the overall mean. Let $\overline{\boldsymbol{y}}_i = \sum_{j=1}^{n_i} \boldsymbol{y}_{ij}/n_i$ so $\hat{\boldsymbol{\tau}}_i = \overline{\boldsymbol{y}}_i - \overline{\boldsymbol{y}}$. Let the residual vector $\hat{\boldsymbol{\epsilon}}_{ij} = \boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_i = \boldsymbol{y}_{ij} - \hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\tau}}_i$. Then $\boldsymbol{y}_{ij} = \overline{\boldsymbol{y}} + (\overline{\boldsymbol{y}}_i - \overline{\boldsymbol{y}}) + (\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_i) = \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\tau}}_i + \hat{\boldsymbol{\epsilon}}_{ij}$.

Several $m \times m$ matrices will be useful. Let $\boldsymbol{S}_i$ be the sample covariance matrix corresponding to the ith treatment group. Then the within sum of squares and cross products matrix is $\boldsymbol{W} = \boldsymbol{W}_e = (n_1 - 1)\boldsymbol{S}_1 + \cdots + (n_p - 1)\boldsymbol{S}_p = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_i)(\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_i)^T$. Then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \boldsymbol{W}/(n - p)$. The treatment or between sum of squares and cross products matrix is

$$\boldsymbol{B}_T = \sum_{i=1}^{p} n_i (\overline{\boldsymbol{y}}_i - \overline{\boldsymbol{y}})(\overline{\boldsymbol{y}}_i - \overline{\boldsymbol{y}})^T.$$

The total corrected (for the mean) sum of squares and cross products matrix is $\boldsymbol{T} = \boldsymbol{B}_T + \boldsymbol{W} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}})(\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}})^T$. Note that $\boldsymbol{S} = \boldsymbol{T}/(n-1)$ is the usual sample covariance matrix of the $\boldsymbol{y}_{ij}$ if it is assumed that all n of the $\boldsymbol{y}_{ij}$ are iid so that the $\boldsymbol{\mu}_i \equiv \boldsymbol{\mu}$ for $i = 1, ..., p$.

The one way MANOVA model is $\boldsymbol{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\tau}_i + \boldsymbol{\epsilon}_{ij}$ where the $\boldsymbol{\epsilon}_{ij}$ are iid with $E(\boldsymbol{\epsilon}_{ij}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_{ij}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. The MANOVA table is shown below.

Summary One Way MANOVA Table

Source	matrix	df
Treatment or Between	$\boldsymbol{B}_T$	$p-1$
Residual or Error or Within	$\boldsymbol{W}$	$n-p$
Total (corrected)	$\boldsymbol{T}$	$n-1$

If all n of the $\boldsymbol{y}_{ij}$ are iid with $E(\boldsymbol{y}_{ij}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{y}_{ij}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, it can be shown that $\boldsymbol{A}/df \xrightarrow{P} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ where $\boldsymbol{A} = \boldsymbol{W}, \boldsymbol{B}_T$, or $\boldsymbol{T}$, and df is the corresponding degrees of freedom. Let t_0 be the test statistic. Often Pillai's trace statistic, the Hotelling Lawley trace statistic, or Wilks' lambda are used. Wilks' lambda

$$\Lambda = \frac{|\boldsymbol{W}|}{|\boldsymbol{B}_T + \boldsymbol{W}|} = \frac{|\boldsymbol{W}|}{|\boldsymbol{T}|} = \frac{|\sum_{i=1}^{p}(n_i - 1)\boldsymbol{S}_i|}{|(n-1)\boldsymbol{S}|} =$$

$$\frac{|\sum_{i=1}^{p}\sum_{j=1}^{n_i}(\boldsymbol{y}_{ij}-\overline{\boldsymbol{y}}_i)(\boldsymbol{y}_{ij}-\overline{\boldsymbol{y}}_i)^T|}{|\sum_{i=1}^{p}\sum_{j=1}^{n_i}(\boldsymbol{y}_{ij}-\overline{\boldsymbol{y}})(\boldsymbol{y}_{ij}-\overline{\boldsymbol{y}})^T|}.$$

Then $t_o = -[n - 0.5(m+p-2)]\log(\Lambda)$ and pval $= P(\chi^2_{m(p-1)} > t_0)$. Hence reject H_0 if $t_0 > \chi^2_{m(p-1)}(1-\alpha)$. See Johnson and Wichern (1988, p. 238).

The four steps of the one way MANOVA test follow.
i) State the hypotheses $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_p$ and H_1 : not H_0.
ii) Get t_0 from output.
iii) Get pval from output.
iv) State whether you reject H_0 or fail to reject H_0. If pval $\leq \alpha$, reject H_0 and conclude that not all of the p treatment means are equal. If pval $> \alpha$, fail to reject H_0 and conclude that all p treatment means are equal or that there is not enough evidence to conclude that not all of the p treatment means are equal. As a textbook convention, use $\alpha = 0.05$ if α is not given.

Another way to perform the one way MANOVA test is to get R output. The default test is Pillai's test, but other tests can be obtained with the R output shown below.

```
summary(out$out) #default is Pillai's test
summary(out$out, test = "Wilks")
summary(out$out, test = "Hotelling-Lawley")
summary(out$out, test = "Roy")
```

Following Mardia et al. (1979, p. 335), let $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_m$ be the eigenvalues of $\boldsymbol{W}^{-1}\boldsymbol{B}_T$. Then $1+\lambda_i$ for $i = 1, ..., m$ are the eigenvalues of $\boldsymbol{W}^{-1}\boldsymbol{T}$ and $\Lambda = \prod_{i=1}^{m}(1+\lambda_i)^{-1}$.

Following Fujikoshi (2002), let the Hotelling Lawley trace statistic $U = tr(\boldsymbol{B}_T\boldsymbol{W}^{-1}) = tr(\boldsymbol{W}^{-1}\boldsymbol{B}_T) = \sum_{i=1}^{m}\lambda_i$, and let Pillai's trace statistic $V = tr(\boldsymbol{B}_T\boldsymbol{T}^{-1}) = tr(\boldsymbol{T}^{-1}\boldsymbol{B}_T) = \sum_{i=1}^{m}\frac{\lambda_i}{1+\lambda_i}$. If the $\boldsymbol{y}_{ij} - \boldsymbol{\mu}_j$ are iid with common covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, and if H_0 is true, then under regularity conditions $-[n - 0.5(m+p-2)]\log(\Lambda) \xrightarrow{D} \chi^2_{m(p-1)}$, $(n-m-p-1)U \xrightarrow{D} \chi^2_{m(p-1)}$, and $(n-1)V \xrightarrow{D} \chi^2_{m(p-1)}$.

Remark 7.1, Are Statisticians crazy? Note that the common covariance matrix assumption implies that each of the p treatment groups or populations has the same covariance matrix $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ for $i = 1, ..., p$, an extremely strong assumption. There are several possible remedies.
i) If the n_i for each group are large, use a large sample theory test. This test may start to outperform the one way MANOVA test if $n \geq (m+p)^2$ and $n_i \geq 40m$ for $i = 1, ..., p$. See Section 7.2.
ii) Use the bootstrap to get better cutoffs. See Rajapaksha and Olive (2024).
iii) Adapt high dimensional analogs of the ANOVA tests to low dimensions.

7.3 An Alternative Test Based on Large Sample Theory

Large sample theory can be also be used to derive a competing test. Let $\boldsymbol{\Sigma}_i$ be the nonsingular population covariance matrix of the ith treatment group or population. To simplify the large sample theory, assume $n_i = \pi_i n$ where $0 < \pi_i < 1$ and $\sum_{i=1}^{p} \pi_i = 1$. Let T_i be a multivariate location estimator such that $\sqrt{n_i}(T_i - \boldsymbol{\mu}_i) \xrightarrow{D} N_m(\mathbf{0}, \boldsymbol{\Sigma}_i)$, and $\sqrt{n}(T_i - \boldsymbol{\mu}_i) \xrightarrow{D} N_m\left(\mathbf{0}, \frac{\boldsymbol{\Sigma}_i}{\pi_i}\right)$. Let $\boldsymbol{T} = (T_1^T, T_2^T, ..., T_p^T)^T$, $\boldsymbol{\nu} = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T, ..., \boldsymbol{\mu}_p^T)^T$, and $\boldsymbol{A}$ be a full rank $r \times mp$ matrix with rank r, then a large sample test of the form $H_0 : \boldsymbol{A\nu} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{A\nu} \neq \boldsymbol{\theta}_0$ uses

$$\boldsymbol{A}\sqrt{n}(\boldsymbol{T} - \boldsymbol{\nu}) \xrightarrow{D} \boldsymbol{u} \sim N_r\left(\mathbf{0}, \boldsymbol{A}\ diag\left(\frac{\boldsymbol{\Sigma}_1}{\pi_1}, \frac{\boldsymbol{\Sigma}_2}{\pi_2}, ..., \frac{\boldsymbol{\Sigma}_p}{\pi_p}\right)\ \boldsymbol{A}^T\right). \tag{7.1}$$

Let the Wald-type statistic

$$t_0 = [\boldsymbol{AT} - \boldsymbol{\theta}_0]^T \left[\boldsymbol{A}\ diag\left(\frac{\hat{\boldsymbol{\Sigma}}_1}{n_1}, \frac{\hat{\boldsymbol{\Sigma}}_2}{n_2}, ..., \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p}\right)\ \boldsymbol{A}^T\right]^{-1} [\boldsymbol{AT} - \boldsymbol{\theta}_0]. \tag{7.2}$$

These results prove the following theorem.

Theorem 7.1. Under the above conditions, $t_0 \xrightarrow{D} \chi^2_r$ if H_0 is true.

This test is due to Rupasinghe Arachchige Don and Olive (2019), and a special case was used by Zhang and Liu (2013) and Konietschke et al. (2015) with $T_i = \overline{\boldsymbol{y}}_i$ and $\hat{\boldsymbol{\Sigma}}_i = \boldsymbol{S}_i$. The $p = 2$ case gives analogs to the two sample Hotelling's T^2 test. See Rupasinghe Arachchige Don and Pelawa Watagoda (2018). The $m = 1$ case gives analogs of the one way ANOVA test. If $m = 1$, see competing tests in Brown and Forsythe (1974a,b), Olive (2017a, pp. 200-202), and Welch (1947, 1951).

For the one way MANOVA type test, let $\boldsymbol{A}$ be the block matrix

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I} & \mathbf{0} & \mathbf{0} & \dots & -\boldsymbol{I} \\ \mathbf{0} & \boldsymbol{I} & \mathbf{0} & \dots & -\boldsymbol{I} \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{I} & -\boldsymbol{I} \end{bmatrix}.$$

Let $\boldsymbol{\mu}_i \equiv \boldsymbol{\mu}$, let $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_p$ or, equivalently, $H_0 : \boldsymbol{A\nu} = \mathbf{0}$, and let

$$\boldsymbol{w} = \boldsymbol{A}\boldsymbol{T} = \begin{bmatrix} T_1 - T_p \\ T_2 - T_p \\ \vdots \\ T_{p-2} - T_p \\ T_{p-1} - T_p \end{bmatrix}. \tag{7.3}$$

Then $\sqrt{n}\boldsymbol{w} \xrightarrow{D} N_{m(p-1)}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}})$ if H_0 is true with $\boldsymbol{\Sigma}_{\boldsymbol{w}} = (\boldsymbol{\Sigma}_{ij})$ where $\boldsymbol{\Sigma}_{ij} = \frac{\boldsymbol{\Sigma}_p}{\pi_p}$ for $i \neq j$, and $\boldsymbol{\Sigma}_{ii} = \frac{\boldsymbol{\Sigma}_i}{\pi_i} + \frac{\boldsymbol{\Sigma}_p}{\pi_p}$ for $i = j$. Hence

$$t_0 = n\boldsymbol{w}^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1} \boldsymbol{w} = \boldsymbol{w}^T \left(\frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}}{n} \right)^{-1} \boldsymbol{w} \xrightarrow{D} \chi^2_{m(p-1)}$$

as the $n_i \to \infty$ if H_0 is true. Here

$$\frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}}{n} = \begin{bmatrix} \frac{\hat{\boldsymbol{\Sigma}}_1}{n_1} + \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} & \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} & \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} & \cdots & \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} \\ \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} & \frac{\hat{\boldsymbol{\Sigma}}_2}{n_2} + \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} & \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} & \cdots & \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} \\ \vdots & \vdots & \vdots & & \vdots \\ \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} & \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} & \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} & \cdots & \frac{\hat{\boldsymbol{\Sigma}}_{p-1}}{n_{p-1}} + \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p} \end{bmatrix} \tag{7.4}$$

is a block matrix where the off diagonal block entries equal $\hat{\boldsymbol{\Sigma}}_p/n_p$ and the ith diagonal block entry is $\frac{\hat{\boldsymbol{\Sigma}}_i}{n_i} + \frac{\hat{\boldsymbol{\Sigma}}_p}{n_p}$ for $i = 1, ..., (p-1)$.

Reject H_0 if

$$t_0 > m(p-1)F_{m(p-1),d_n}(1-\delta) \tag{7.5}$$

where $d_n = \min(n_1, ..., n_p)$. See Theorem 2.34. It may make sense to relabel the groups so that n_p is the largest n_i or $\hat{\boldsymbol{\Sigma}}_p/n_p$ has the smallest generalized variance of the $\hat{\boldsymbol{\Sigma}}_i/n_i$. This test may start to outperform the one way MANOVA test if $n \geq (m+p)^2$ and $n_i \geq 40m$ for $i = 1, ..., p$.

If $\boldsymbol{\Sigma}_i \equiv \boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}_i$ is replaced by $\hat{\boldsymbol{\Sigma}}$, we will show that for the one way MANOVA test that $t_0 = (n-p)U$ where U is the Hotelling Lawley statistic. See Theorem 7.2. For the proof, some results on the vec and Kronecker product will be useful. Following Henderson and Searle (1979), $vec(\boldsymbol{G})$ and $vec(\boldsymbol{G}^T)$ contain the same elements in different sequences. Define the permutation matrix $\boldsymbol{P}_{r,m}$ such that

$$vec(\boldsymbol{G}) = \boldsymbol{P}_{r,m} vec(\boldsymbol{G}^T) \tag{7.6}$$

where $\boldsymbol{G}$ is $r \times m$. Then $\boldsymbol{P}_{r,m}^T = \boldsymbol{P}_{m,r}$, and $\boldsymbol{P}_{r,m}\boldsymbol{P}_{m,r} = \boldsymbol{P}_{m,r}\boldsymbol{P}_{r,m} = \boldsymbol{I}_{rm}$. If $\boldsymbol{C}$ is $s \times m$ and $\boldsymbol{D}$ is $p \times r$, then

$$\boldsymbol{C} \otimes \boldsymbol{D} = \boldsymbol{P}_{p,s}(\boldsymbol{D} \otimes \boldsymbol{C})\boldsymbol{P}_{m,q}. \tag{7.7}$$

Also

$$(\boldsymbol{C} \otimes \boldsymbol{D})vec(\boldsymbol{G}) = vec(\boldsymbol{D}\boldsymbol{G}\boldsymbol{C}^T) = \boldsymbol{P}_{p,s}(\boldsymbol{D} \otimes \boldsymbol{C})vec(\boldsymbol{G}^T). \tag{7.8}$$

If $\boldsymbol{C}$ is $m \times m$and $\boldsymbol{D}$ is $r \times r$, then $\boldsymbol{C} \otimes \boldsymbol{D} = \boldsymbol{P}_{r,m}(\boldsymbol{D} \otimes \boldsymbol{C})\boldsymbol{P}_{m,r}$, and

$$[vec(\boldsymbol{G})]^T(\boldsymbol{C} \otimes \boldsymbol{D})vec(\boldsymbol{G}) = [vec(\boldsymbol{G}^T)]^T(\boldsymbol{D} \otimes \boldsymbol{C})vec(\boldsymbol{G}^T). \tag{7.9}$$

Remark 7.2. Another method for one way MANOVA is to use the model $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$ or

$$\begin{bmatrix} Y_{111} & Y_{112} & \cdots & Y_{11m} \\ \vdots & \vdots & \cdots & \vdots \\ Y_{1,n_1,1} & Y_{1,n_1,2} & \cdots & Y_{1,n_1,m} \\ Y_{211} & Y_{211} & \cdots & Y_{21m} \\ \vdots & \vdots & \cdots & \vdots \\ Y_{2,n_2,1} & Y_{2,n_2,2} & \cdots & Y_{2,n_2,m} \\ \vdots & \vdots & \cdots & \vdots \\ Y_{p,11} & Y_{p,1m} & \cdots & Y_{p,1m} \\ \vdots & \vdots & \cdots & \vdots \\ Y_{p,n_p,1} & Y_{p,n_p,2} & \cdots & Y_{p,n_p,m} \end{bmatrix} = \begin{bmatrix} 1\,1\,0\ldots 0 \\ \vdots\,\vdots\,\vdots\quad\vdots \\ 1\,1\,0\ldots 0 \\ 1\,0\,1\ldots 0 \\ \vdots\,\vdots\,\vdots\quad\vdots \\ 1\,0\,1\ldots 0 \\ \vdots\,\vdots\,\vdots\quad\vdots \\ 1\,0\,0\ldots 1 \\ \vdots\,\vdots\,\vdots\quad\vdots \\ 1\,0\,0\ldots 1 \\ 1\,0\,0\ldots 0 \\ \vdots\,\vdots\,\vdots\quad\vdots \\ 1\,0\,0\ldots 0 \end{bmatrix} \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \ldots & \beta_{1,m} \\ \beta_{2,1} & \beta_{2,2} & \ldots & \beta_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p,1} & \beta_{p,2} & \ldots & \beta_{p,m} \end{bmatrix} + \boldsymbol{E}.$$

Then $\boldsymbol{X}$ is full rank where the ith column of $\boldsymbol{X}$ is an indicator for group $i-1$ for $i = 2, ..., p$, $\hat{\beta}_{1k} = \overline{Y}_{pok} = \hat{\mu}_{pk}$ for $k = 1, ..., m$, and

$$\hat{\beta}_{ik} = \overline{Y}_{i-1,ok} - \overline{Y}_{pok} = \hat{\mu}_{i-1,k} - \hat{\mu}_{pk}$$

for $k = 1, ..., m$ and $i = 2, ..., p$. Thus testing $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_p$ is equivalent to testing $H_0 : \boldsymbol{L}\boldsymbol{B} = \boldsymbol{0}$ where $\boldsymbol{L} = [\boldsymbol{0}\ \boldsymbol{I}_{p-1}]$. Then $\boldsymbol{y}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_{ij}$ and

$$\boldsymbol{B}_T = \boldsymbol{B} = \begin{bmatrix} \boldsymbol{\mu}_p^T \\ \boldsymbol{\mu}_1^T - \boldsymbol{\mu}_p^T \\ \boldsymbol{\mu}_2^T - \boldsymbol{\mu}_p^T \\ \vdots \\ \boldsymbol{\mu}_{p-2}^T - \boldsymbol{\mu}_p^T \\ \boldsymbol{\mu}_{p-1}^T - \boldsymbol{\mu}_p^T \end{bmatrix}. \tag{7.10}$$

Consider testing a linear hypothesis $H_0 : \boldsymbol{L}\boldsymbol{B} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{L}\boldsymbol{B} \neq \boldsymbol{0}$ where $\boldsymbol{L}$ is a full rank $r \times p$ matrix. Let

$$\boldsymbol{H} = \hat{\boldsymbol{B}}^T \boldsymbol{L}^T [\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}.$$

If $\boldsymbol{L} = (\boldsymbol{0}\ \boldsymbol{I}_{p-1})$ then the multivariate linear regression Hotelling Lawley test statistic for testing $H_0 : \boldsymbol{LB} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{LB} \neq \boldsymbol{0}$ is $U = tr(\boldsymbol{W}^{-1}\boldsymbol{H})$ while the Hotelling Lawley test statistic for the one way MANOVA test with $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_p$ is $U = tr(\boldsymbol{W}^{-1}\boldsymbol{B}_T)$. Rupasinghe Arachchige Don (2018) showed that these two test statistics are the the same for the above $\boldsymbol{X}$ by showing that $\boldsymbol{B}_T = \boldsymbol{H}$.

Theorem 7.2. For the one way MANOVA test using $\boldsymbol{A}$ as defined below Theorem 7.1, let the Hotelling Lawley trace statistic $U = tr(\boldsymbol{W}^{-1}\boldsymbol{B}_T)$. Then

$$(n-p)U = t_0 = [\boldsymbol{AT} - \boldsymbol{\theta}_0]^T \left[\boldsymbol{A}\ diag\left(\frac{\hat{\boldsymbol{\Sigma}}}{n_1}, \frac{\hat{\boldsymbol{\Sigma}}}{n_2}, ..., \frac{\hat{\boldsymbol{\Sigma}}}{n_p}\right)\ \boldsymbol{A}^T\right]^{-1} [\boldsymbol{AT} - \boldsymbol{\theta}_0].$$

Hence if the $\boldsymbol{\Sigma}_i \equiv \boldsymbol{\Sigma}$ and $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_p$ is true, then $(n-p)U = t_0 \xrightarrow{D} \chi^2_{m(p-1)}$.

Proof. Let $\boldsymbol{B}$ and $\boldsymbol{X}$ be as in Remark 7.2. Let $\boldsymbol{L} = [\boldsymbol{0}\ \ \boldsymbol{I}_{p-1}]$ be an $s \times p$ matrix with $s = p-1$. For this choice of $\boldsymbol{X}$, $U = tr(\boldsymbol{W}^{-1}\boldsymbol{B}_T) = tr(\boldsymbol{W}^{-1}\boldsymbol{H})$ by Remark 7.4. Hence by Theorem 6.21,

$$(n-p)U = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T [\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]. \quad (7.11)$$

Now $vec([\boldsymbol{L}\hat{\boldsymbol{B}}]^T) = \boldsymbol{w} = \boldsymbol{AT}$ of Equation (7.3) with $T_i = \overline{\boldsymbol{y}}_i$. Then

$$t_0 = \boldsymbol{w}^T \left(\frac{\hat{\boldsymbol{\Sigma}}\boldsymbol{w}}{n}\right)^{-1} \boldsymbol{w}$$

where

$$\frac{\hat{\boldsymbol{\Sigma}}\boldsymbol{w}}{n} = \boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T \otimes \hat{\boldsymbol{\Sigma}}$$

is given by Equation (7.4) with each $\hat{\boldsymbol{\Sigma}}_i$ replaced by $\hat{\boldsymbol{\Sigma}}$. Thus $t_0 =$

$$[vec([\boldsymbol{L}\hat{\boldsymbol{B}}]^T)]^T [(\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1} \otimes \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}][vec([\boldsymbol{L}\hat{\boldsymbol{B}}]^T)]. \quad (7.12)$$

Then $t_0 = (n-p)U$ by Equation (7.9) with $\boldsymbol{G} = \boldsymbol{L}\hat{\boldsymbol{B}}$. □

Hence the one way MANOVA test is a special case of Equation (7.2) where $\boldsymbol{\theta}_0 = \boldsymbol{0}$ and $\hat{\boldsymbol{\Sigma}}_i \equiv \hat{\boldsymbol{\Sigma}}$, but then Theorem 7.1 only holds if H_0 is true and $\boldsymbol{\Sigma}_i \equiv \boldsymbol{\Sigma}$. Note that the large sample theory of Theorem 7.1 is trivial compared to the large sample theory of $(n-p)U$ given in Theorem 7.2. Fujikoshi (2002) showed $(n-m-p-1)U \xrightarrow{D} \chi^2_{m(p-1)}$ while $(n-p)U \xrightarrow{D} \chi^2_{m(p-1)}$ by Theorem 7.2 if H_0 is true under the common covariance matrix assumption. There is no

contradiction since $(m+1)U \xrightarrow{P} 0$ as the $n_i \to \infty$. Note the $\boldsymbol{A}$ is $m(p-1) \times mp$.

For tests corresponding to Theorem 7.1, we will use bootstrap with the prediction region method confidence region of Chapter 5 to test H_0 when $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{W}}$ or the $\hat{\boldsymbol{\Sigma}}_i$ are unknown or difficult to estimate. To bootstrap the test $H_0 : \boldsymbol{A\nu} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{A\nu} \neq \boldsymbol{\theta}_0$, use $Z_n = \boldsymbol{AT}$. Take a sample of size n_j with replacement from the n_j cases for each group for $j = 1, 2, ..., p$ to obtain T_j^* and $\boldsymbol{T}_1^*$. Repeat B times to obtain $\boldsymbol{T}_1^*, ..., \boldsymbol{T}_B^*$. Then $Z_i^* = \boldsymbol{AT}_i^*$ for $i = 1, ..., B$. We will illustrate this method with the analog for the one way MANOVA test for $H_0 : \boldsymbol{A\theta} = \boldsymbol{0}$ which is equivalent to $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_p$, where $\boldsymbol{0}$ is an $r \times 1$ vector of zeroes with $r = m(p-1)$. Then $Z_n = \boldsymbol{AT} = \boldsymbol{w}$ given by Equation (7.3). Hence the $m(p-1) \times 1$ vector $Z_i^* = \boldsymbol{AT}_i^* = ((T_1^* - T_p^*)^T, ..., (T_{p-1}^* - T_p^*)^T)^T$ where T_j is a multivariate location estimator (such as the sample mean, coordinatewise median, or trimmed mean), applied to the cases in the jth treatment group. The prediction region method fails to reject H_0 if $\boldsymbol{0}$ is in the resulting confidence region.

We may need $B \geq 50m(p-1)$, $n \geq (m+p)^2$, and $n_i \geq 40m$. If the n_i are not large, the one way MANOVA test can be regarded as a regularized estimator, and can perform better than the tests that do not assume equal population covariance matrices. See the simulations in Rupasinghe Arachchige Don and Olive (2019).

If $H_0 : \boldsymbol{A\nu} = \boldsymbol{\theta}_0$ is true and if the $\boldsymbol{\Sigma}_i \equiv \boldsymbol{\Sigma}$ for $i = 1, ..., p$, then

$$t_0 = [\boldsymbol{AT} - \boldsymbol{\theta}_0]^T \left[\boldsymbol{A}\ diag\left(\frac{\hat{\boldsymbol{\Sigma}}}{n_1}, \frac{\hat{\boldsymbol{\Sigma}}}{n_2}, ..., \frac{\hat{\boldsymbol{\Sigma}}}{n_p} \right) \boldsymbol{A}^T \right]^{-1} [\boldsymbol{AT} - \boldsymbol{\theta}_0] \xrightarrow{D} \chi^2_r.$$

If H_0 is true but the $\boldsymbol{\Sigma}_i$ are not equal, we may be able to get a bootstrap cutoff by using

$$t_{0i}^* = [\boldsymbol{AT}_i^* - \boldsymbol{AT}]^T \left[\boldsymbol{A}\ diag\left(\frac{\hat{\boldsymbol{\Sigma}}}{n_1}, \frac{\hat{\boldsymbol{\Sigma}}}{n_2}, ..., \frac{\hat{\boldsymbol{\Sigma}}}{n_p} \right) \boldsymbol{A}^T \right]^{-1} [\boldsymbol{AT}_i^* - \boldsymbol{AT}] =$$

$$D^2_{\boldsymbol{AT}_i^*}\left(\boldsymbol{AT}, \boldsymbol{A}\ diag\left(\frac{\hat{\boldsymbol{\Sigma}}}{n_1}, \frac{\hat{\boldsymbol{\Sigma}}}{n_2}, ..., \frac{\hat{\boldsymbol{\Sigma}}}{n_p} \right) \boldsymbol{A}^T \right).$$

7.4 Bootstrap Tests

This section follows Rajapaksha and Olive (2024) closely. Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where a $g \times 1$ statistic T_n satisfies $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{\Sigma})$. If $\hat{\boldsymbol{\Sigma}}^{-1} \xrightarrow{P} \boldsymbol{\Sigma}^{-1}$ and H_0 is true, then

$$D_n^2 = D_{\boldsymbol{\theta}_0}^2(T_n, \hat{\boldsymbol{\Sigma}}/n) = n(T_n - \boldsymbol{\theta}_0)^T \hat{\boldsymbol{\Sigma}}^{-1}(T_n - \boldsymbol{\theta}_0) \xrightarrow{D} \boldsymbol{u}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{u} \sim \chi_g^2$$

as $n \to \infty$. Then a Wald type test rejects H_0 at significance level δ if $D_n^2 > \chi_{g,1-\delta}^2$ where $P(X \leq \chi_{g,1-\delta}^2) = 1 - \delta$ if $X \sim \chi_g^2$, a chi-square distribution with g degrees of freedom.

It is common to implement a Wald type test using

$$D_n^2 = D_{\boldsymbol{\theta}_0}^2(T_n, \boldsymbol{C}_n/n) = n(T_n - \boldsymbol{\theta}_0)^T \boldsymbol{C}_n^{-1}(T_n - \boldsymbol{\theta}_0) \xrightarrow{D} \boldsymbol{u}^T \boldsymbol{C}^{-1} \boldsymbol{u}$$

as $n \to \infty$ if H_0 is true, where the $g \times g$ symmetric positive definite matrix $\boldsymbol{C}_n \xrightarrow{P} \boldsymbol{C} \neq \boldsymbol{\Sigma}$. Hence $\boldsymbol{C}_n$ is the wrong dispersion matrix, and $\boldsymbol{u}^T \boldsymbol{C}^{-1} \boldsymbol{u}$ does not have a χ_g^2 distribution when H_0 is true. Often $\boldsymbol{C}_n$ is a regularized estimator of $\boldsymbol{\Sigma}$, or $\boldsymbol{C}_n^{-1}$ is a regularized estimator of the precision matrix $\boldsymbol{\Sigma}^{-1}$, such as $\boldsymbol{C}_n = diag(\hat{\boldsymbol{\Sigma}})$ or $\boldsymbol{C}_n = \boldsymbol{I}_g$, the $g \times g$ identity matrix. Another example is $\boldsymbol{C}_n = \boldsymbol{S}_p$, where $\boldsymbol{S}_p$ is a pooled covariance matrix, and it is assumed that the p groups have the same covariance matrix $\boldsymbol{\Sigma}$. When this assumption is violated, $\boldsymbol{C}_n$ is usually not a consistent estimator of $\boldsymbol{\Sigma}$. When the bootstrap is used, often $\boldsymbol{C}_n = n\boldsymbol{S}_T^*$ where $\boldsymbol{S}_T^*$ is the sample covariance matrix of the bootstrap sample $T_1^*, ..., T_B^*$. The assumption that $n\boldsymbol{S}_T^*$ is a consistent estimator of $\boldsymbol{\Sigma}$ is strong. See, for example, Machado and Parente (2005).

The BR and PR confidence regions can be used since if $\boldsymbol{C}_n^{-1} \xrightarrow{P} \boldsymbol{C}^{-1}$, then $D_j^2 \xrightarrow{D} D^2 = \boldsymbol{u}^T \boldsymbol{C}^{-1} \boldsymbol{u}$, then and (5.34) and (5.35) are large sample confidence regions. If $\boldsymbol{C}_n^{-1}$ is "not too ill conditioned," then $D_j^2 \approx \boldsymbol{u}^T \boldsymbol{C}_n^{-1} \boldsymbol{u}$ for large n, and the confidence regions (5.34) and (5.35) will have coverage near $1 - \delta$.

If $H_0 : \boldsymbol{A}\boldsymbol{\nu} = \boldsymbol{\theta}_0$ is true, if the $\boldsymbol{\Sigma}_i \equiv \boldsymbol{\Sigma}$ for $i = 1, ..., p$, and if $\hat{\boldsymbol{\Sigma}}$ is a consistent estimator of $\boldsymbol{\Sigma}$, then by Theorem 7.1

$$t_0 = [\boldsymbol{A}\boldsymbol{T} - \boldsymbol{\theta}_0]^T \left[\boldsymbol{A}\ diag\left(\frac{\hat{\boldsymbol{\Sigma}}}{n_1}, \frac{\hat{\boldsymbol{\Sigma}}}{n_2}, ..., \frac{\hat{\boldsymbol{\Sigma}}}{n_p} \right) \boldsymbol{A}^T \right]^{-1} [\boldsymbol{A}\boldsymbol{T} - \boldsymbol{\theta}_0] \xrightarrow{D} \chi_r^2.$$

If H_0 is true but the $\boldsymbol{\Sigma}_i$ are not equal, then we get a bootstrap cutoff by using

$$t_{0i}^* = [\boldsymbol{A}\boldsymbol{T}_i^* - \boldsymbol{A}\boldsymbol{T}]^T \left[\boldsymbol{A}\ diag\left(\frac{\hat{\boldsymbol{\Sigma}}}{n_1}, \frac{\hat{\boldsymbol{\Sigma}}}{n_2}, ..., \frac{\hat{\boldsymbol{\Sigma}}}{n_p} \right) \boldsymbol{A}^T \right]^{-1} [\boldsymbol{A}\boldsymbol{T}_i^* - \boldsymbol{A}\boldsymbol{T}] =$$

$$D_{\boldsymbol{A}\boldsymbol{T}_i^*}^2 \left(\boldsymbol{A}\boldsymbol{T}, \boldsymbol{A}\ diag\left(\frac{\hat{\boldsymbol{\Sigma}}}{n_1}, \frac{\hat{\boldsymbol{\Sigma}}}{n_2}, ..., \frac{\hat{\boldsymbol{\Sigma}}}{n_p} \right) \boldsymbol{A}^T \right).$$

Let $F_0 = t_0/r$. Then we can get a bootstrap cutoff using $F_{0i}^* = t_{0i}^*/r$. For $T_i = \overline{\boldsymbol{y}}_i$, let $\hat{\boldsymbol{\Sigma}}$ be the usual pooled covariance matrix estimator.

The Wald-type tests with $\boldsymbol{C}_n = \boldsymbol{I}_g = \boldsymbol{C}$ often performed fairly well with the nonparametric bootstrap in that the simulated level of the test tended to be closer to the nominal level for samller sample sizes n_i than methods that used other choices for $\boldsymbol{C}_n$. A drawback of the tests that use $\boldsymbol{C}_n = \boldsymbol{I}_g$ is that the volume of the confidence region, which is a hypersphere, can be quite large. Alternative choices of $\boldsymbol{C}_n$ tend to result in confidence regions (5.34) and (5.35) that are hyperellipsoids.

Remark 7.3. It may be interesting to replace the nonparametric bootstrap by the m out of n bootstrap = subsampling = delete d jackknife.

7.5 Summary

1) The **multivariate linear model** $\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$ has $m \geq 2$ response variables $Y_1, ..., Y_m$ and p predictor variables $x_1, x_2, ..., x_p$. The ith case is $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T) = (x_{i1}, x_{i2}, ..., x_{ip}, Y_{i1}, ..., Y_{im})$. If a constant $x_{i1} = 1$ is in the model, then x_{i1} could be omitted from the case. The model is written in matrix form as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = (\sigma_{ij})$ for $k = 1, ..., n$. Also $E(\boldsymbol{e}_i) = \boldsymbol{0}$ while $\text{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij}\boldsymbol{I}_n$ for $i, j = 1, ..., m$. Then $\boldsymbol{B}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are unknown matrices of parameters to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{X}\boldsymbol{B}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T\boldsymbol{\beta}_j$.

The data matrix $\boldsymbol{W} = [\boldsymbol{X} \;\; \boldsymbol{Z}]$ except usually the first column $\boldsymbol{1}$ of $\boldsymbol{X}$ is omitted if $x_{i,1} \equiv 1$. The $n \times m$ matrix

$$\boldsymbol{Z} = \begin{bmatrix} Y_{1,1} & Y_{1,2} & \dots & Y_{1,m} \\ Y_{2,1} & Y_{2,2} & \dots & Y_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n,1} & Y_{n,2} & \dots & Y_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}_1 & \boldsymbol{Y}_2 & \dots & \boldsymbol{Y}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_n^T \end{bmatrix}.$$

The $n \times p$ matrix

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}$$

where often $\boldsymbol{v}_1 = \boldsymbol{1}$.

The $p \times m$ matrix

$$\boldsymbol{B} = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \dots & \beta_{1,m} \\ \beta_{2,1} & \beta_{2,2} & \dots & \beta_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p,1} & \beta_{p,2} & \dots & \beta_{p,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \dots & \boldsymbol{\beta}_m \end{bmatrix}.$$

The $n \times m$ matrix

$$\boldsymbol{E} = \begin{bmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \dots & \epsilon_{1,m} \\ \epsilon_{2,1} & \epsilon_{2,2} & \dots & \epsilon_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n,1} & \epsilon_{n,2} & \dots & \epsilon_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \dots & \boldsymbol{e}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{\epsilon}_1^T \\ \vdots \\ \boldsymbol{\epsilon}_n^T \end{bmatrix}.$$

2) The univariate linear model is $Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i = \boldsymbol{\beta}^T\boldsymbol{x}_i + e_i$ for $i = 1, \ldots, n$. In matrix notation, these n equations become $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, where $\boldsymbol{Y}$ is an $n \times 1$ vector of response variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors.

3) Each response variable in a multivariate linear model follows a univariate linear model $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where it is assumed that $E(\boldsymbol{e}_j) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$.

4) In a MANOVA model, $\boldsymbol{y}_k = \boldsymbol{B}^T\boldsymbol{x}_k + \boldsymbol{\epsilon}_k$ for $k = 1, ..., n$ is written in matrix form as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = (\sigma_{ij})$ for $k = 1, ..., n$. Each response variable in a MANOVA model follows an ANOVA model $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ for $j = 1, ..., m$ where it is assumed that $E(\boldsymbol{e}_j) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}_j) = \sigma_{jj}\boldsymbol{I}_n$.

5) The **one way MANOVA** model is as above where $\boldsymbol{Y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{e}_j$ is a one way ANOVA model for $j = 1, ..., m$. Check the model by making m response and residual plots and a DD plot of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$.

6) The one way MANOVA model is a generalization of the Hotelling's T^2 test from 2 groups to $p \geq 2$ groups, assumed to have different means but a common covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. Want to test $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_p$. This model is a multivariate linear model so there are m response variables $Y_1, ..., Y_m$ measured for each group. Each Y_i follows a one way ANOVA model for $i = 1, ..., m$.

7) For the one way MANOVA model, make a DD plot of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$ where $i = 1, ..., n$. Use the plot to check whether the $\boldsymbol{\epsilon}_i$ follow a multivariate normal distribution or some other elliptically contoured distribution. We want $n \geq (m + p)^2$ and $n_i \geq 10m$.

8) For the one way MANOVA model, write the data as Y_{ijk} where $i = 1, ..., p$ and $j = 1, ..., n_i$. So k corresponds to the kth variable Y_k for $k = 1, ..., m$. Then $\hat{Y}_{ijk} = \hat{\mu}_{ik} = \overline{Y}_{iok}$ for $i = 1, ..., p$. So for the kth variable, the means $\mu_{1k}, ..., \mu_{pk}$ are of interest. The residuals are $r_{ijk} = Y_{ijk} - \hat{Y}_{ijk}$. For each variable Y_k make a response plot of $\overline{Y}_{iok}$ versus Y_{ijk} and a residual plot of $\overline{Y}_{iok}$ versus r_{ijk}. Both plots will consist of p dot plots of n_i cases located at the $\overline{Y}_{iok}$. The dot plots should follow the identity line in the response plot and the horizontal $r = 0$ line in the residual plot for each of the m response variables $Y_1, ..., Y_m$. For each variable Y_k, let R_{ik} be the range of the ith dot plot. If each $n_i \geq 5$, we want $\max(R_{1k}, ..., R_{pk}) \leq 2\min(R_{1k}, ..., R_{pk})$. The

one way MANOVA model may be reasonable for the test in point 9) if the m response and residual plots satisfy the above graphical checks.

9) The four steps of the one way MANOVA test follow.
i) State the hypotheses $H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_p$ and H_1 : not H_0.
ii) Get t_0 from output.
iii) Get pval from output.
iv) State whether you reject H_0 or fail to reject H_0. If pval $\leq \alpha$, reject H_0 and conclude that not all of the p treatment means are equal. If pval $> \alpha$, fail to reject H_0 and conclude that all p treatment means are equal or that there is not enough evidence to conclude that not all of the p treatment means are equal. Give a nontechnical sentence as the conclusion, if possible. As a textbook convention, use $\alpha = 0.05$ if α is not given.

10) The one way MANOVA test assumes that the p treatment groups or populations have the same covariance matrix: $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_p$, but the test has some resistance to this assumption. See points 6) and 8).

7.6 Complements

Useful papers for one way MANOVA models include Rupasinghe Arachchige Don and Olive (2019), Rupasinghe Arachchige Don and Pelawa Watagoda (2018), and Rajapaksha and Olive (2024).

7.7 Problems

Also see Problems 3.10, 3.19, 6.9, and 6.11.

7.1. Let $\boldsymbol{\Sigma}_i$ be the nonsingular population covariance matrix of the ith treatment group or population. To simplify the large sample theory, assume $n_i = \pi_i n$ where $0 < \pi_i < 1$ and $\sum_{i=1}^{3} \pi_i = 1$. Let T_i be a multivariate location estimator such that
$\sqrt{n_i}(T_i - \boldsymbol{\mu}_i) \xrightarrow{D} N_m(\mathbf{0}, \boldsymbol{\Sigma}_i)$, and $\sqrt{n}(T_i - \boldsymbol{\mu}_i) \xrightarrow{D} N_m\left(\mathbf{0}, \frac{\boldsymbol{\Sigma}_i}{\pi_i}\right)$ for $i = 1, 2, 3$.
Assume the T_i are independent.

Then

$$\sqrt{n}\begin{bmatrix} T_1 - \boldsymbol{\mu}_1 \\ T_2 - \boldsymbol{\mu}_2 \\ T_3 - \boldsymbol{\mu}_3 \end{bmatrix} \xrightarrow{D} \boldsymbol{u}.$$

a) Find the distribution of $\boldsymbol{u}$.
b) Suggest an estimator $\hat{\pi}_i$ of π_i.

Chapter 8
Robust Statistics

This chapter considers large sample theory for robust statistics. Robust estimators of multivariate location and dispersion are useful for outlier detection and for developing robust regression estimators. This chapter follows Olive (2008, 2017b, 2022c) closely.

Definition 8.1 An **outlier** corresponds to a case that is far from the bulk of the data.

8.1 The Location Model

The location model is

$$Y_i = \mu + e_i, \quad i = 1, \ldots, n \tag{8.1}$$

where $e_1, ..., e_n$ are error random variables, often iid with zero mean. The location model is used when there is one variable Y, such as height, of interest. The location model is a special case of the multiple linear regression model and of the multivariate location and dispersion model, where there are p variables $x_1, ..., x_p$ of interest, such as height and weight if $p = 2$.

The location model is often summarized by obtaining point estimates and confidence intervals for a location parameter and a scale parameter. Assume that there is a sample $Y_1, \ldots, Y_n$ of size n where the Y_i are iid from a distribution with median MED(Y), mean $E(Y)$, and variance $V(Y)$ if they exist. The location parameter μ is often the population mean or median while the scale parameter is often the population standard deviation $\sqrt{V(Y)}$. The ith *case* is Y_i.

Four important statistics for the location model are the sample mean, median, variance, and the median absolute deviation (MAD). Let $Y_1, \ldots, Y_n$ be the random sample; i.e., assume that $Y_1, ..., Y_n$ are iid. The sample

mean is a measure of location and estimates the population mean (expected value) $\mu = E(Y)$. The *sample mean* $\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$. The *sample variance* $S_n^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1} = \frac{\sum_{i=1}^{n} Y_i^2 - n(\overline{Y})^2}{n-1}$, and the *sample standard deviation* $S_n = \sqrt{S_n^2}$.

If the data set $Y_1, ..., Y_n$ is arranged in ascending order from smallest to largest and written as $Y_{(1)} \leq \cdots \leq Y_{(n)}$, then $Y_{(i)}$ is the ith order statistic and the $Y_{(i)}$'s are called the *order statistics*. If the data $Y_1 = 1, Y_2 = 4, Y_3 = 2, Y_4 = 5$, and $Y_5 = 3$, then $\overline{Y} = 3$, $Y_{(i)} = i$ for $i = 1, ..., 5$ and $\text{MED}(n) = 3$ where the sample size $n = 5$. The sample median is a measure of location while the sample standard deviation is a measure of spread. The sample mean and standard deviation are vulnerable to outliers, while the sample median and MAD, defined below, are outlier resistant.

Definition 8.2. The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \text{ if n is odd,} \tag{8.2}$$

$$\text{MED}(n) = \frac{Y_{(n/2)} + Y_{((n/2)+1)}}{2} \text{ if n is even.}$$

The notation $\text{MED}(n) = \text{MED}(n, Y_i) = \text{MED}(Y_1, ..., Y_n)$ will also be used.

Definition 8.3. The *sample median absolute deviation* is

$$\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|, \ i = 1, \ldots, n). \tag{8.3}$$

Since $\text{MAD}(n) = \text{MAD}(n, Y_i)$ is the median of n distances, at least half of the observations are within a distance $\text{MAD}(n)$ of $\text{MED}(n)$ and at least half of the observations are a distance of $\text{MAD}(n)$ or more away from $\text{MED}(n)$. Like the standard deviation, $\text{MAD}(n)$ is a measure of spread.

Example 8.1. Let the data be $1, 2, 3, 4, 5, 6, 7, 8, 9$. Then $\text{MED}(n) = 5$ and $\text{MAD}(n) = 2 = \text{MED}\{0, 1, 1, 2, 2, 3, 3, 4, 4\}$.

The population median $\text{MED}(Y)$ and the population median absolute deviation $\text{MAD}(Y)$ are important quantities of a distribution.

Definition 8.4. The *population median* is any value $\text{MED}(Y)$ such that

$$P(Y \leq \text{MED}(Y)) \geq 0.5 \text{ and } P(Y \geq \text{MED}(Y)) \geq 0.5. \tag{8.4}$$

Definition 8.5. The *population median absolute deviation* is

$$\text{MAD}(Y) = \text{MED}(|Y - \text{MED}(Y)|). \tag{8.5}$$

$\text{MED}(Y)$ is a measure of location while $\text{MAD}(Y)$ is a measure of scale. The median is the middle value of the distribution. Since $\text{MAD}(Y)$ is the me-

Table 8.1 Some commonly used notation.

population	sample
$E(Y)$, μ, θ	$\overline{Y}_n$, $E(n)$ $\hat{\mu}$, $\hat{\theta}$
MED(Y), M	MED(n), $\hat{M}$
VAR(Y), σ^2	VAR(n), S^2, $\hat{\sigma}^2$
SD(Y), σ	SD(n), S, $\hat{\sigma}$
MAD(Y)	MAD(n)
IQR(Y)	IQR(n)

dian distance from MED(Y), at least half of the mass is inside [MED(Y) − MAD(Y), MED(Y) + MAD(Y)] and at least half of the mass of the distribution is outside of the interval (MED(Y) − MAD(Y), MED(Y) + MAD(Y)). In other words, MAD(Y) is any value such that

$$P(Y \in [\text{MED}(Y) - \text{MAD}(Y), \text{MED}(Y) + \text{MAD}(Y)]) \geq 0.5,$$

$$\text{and } P(Y \in (\text{MED}(Y) - \text{MAD}(Y), \text{MED}(Y) + \text{MAD}(Y)) \ ) \leq 0.5.$$

Definition 8.6. The *sample interquantile range* IQR$(n) = Y_{(\lceil 0.75n \rceil)} - Y_{(\lceil 0.25n \rceil)}$. The *population interquantile range* IQR$(Y) = y_{0.75} - y_{0.25}$ where $P(Y \leq y_\alpha) = \alpha$ if y_α is a continuity point of the cdf $F_Y(y)$.

Notation is needed in order to distinguish between population quantities, random quantities, and observed quantities. For population quantities, capital letters like $E(Y)$ and MAD(Y) will often be used while the estimators will often be denoted by MED(n), MAD(n), MED$(Y_i, i = 1, ..., n)$, or MED$(Y_1, \ldots, Y_n)$. The random sample will be denoted by $Y_1, \ldots, Y_n$. Sometimes the observed sample will be fixed and lower case letters will be used. For example, the observed sample may be denoted by $y_1, ..., y_n$ while the estimates may be denoted by med(n), mad(n), or $\overline{y}_n$. Table 8.1 summarizes some of this notation.

Definition 8.7. Let $f_Y(y)$ be the pdf of Y. Then the family of pdfs $f_W(w) = f_Y(w - \mu)$ indexed by the *location parameter* μ, $-\infty < \mu < \infty$, is the *location family* for the random variable $W = \mu + Y$ with *standard pdf* $f_Y(y)$.

Definition 8.8. Let $f_Y(y)$ be the pdf of Y. Then the family of pdfs $f_W(w) = (1/\sigma) f_Y(w/\sigma)$ indexed by the *scale parameter* $\sigma > 0$, is the *scale family* for the random variable $W = \sigma Y$ with *standard pdf* $f_Y(y)$.

Definition 8.9. Let $f_Y(y)$ be the pdf of Y. Then the family of pdfs $f_W(w) = (1/\sigma) f_Y((w - \mu)/\sigma)$ indexed by the *location and scale parameters* μ, $-\infty < \mu < \infty$, and $\sigma > 0$, is the *location–scale family* for the random variable $W = \mu + \sigma Y$ with *standard pdf* $f_Y(y)$.

Finding MED(Y) and MAD(Y) for symmetric distributions and location–scale families is made easier by the following theorem. Let $F(y_\alpha) = P(Y \leq y_\alpha) = \alpha$ for $0 < \alpha < 1$ where the cdf $F(y) = P(Y \leq y)$. Let $D = \text{MAD}(Y)$, $M = \text{MED}(Y) = y_{0.5}$ and $U = y_{0.75}$.

Theorem 8.1. a) If $W = a + bY$, then $\text{MED}(W) = a + b\text{MED}(Y)$ and $\text{MAD}(W) = |b|\text{MAD}(Y)$.

b) If Y has a pdf that is continuous and positive on its support and symmetric about μ, then $\text{MED}(Y) = \mu$ and $\text{MAD}(Y) = y_{0.75} - \text{MED}(Y)$. Find $M = \text{MED}(Y)$ by solving the equation $F(M) = 0.5$ for M, and find U by solving $F(U) = 0.75$ for U. Then $D = \text{MAD}(Y) = U - M$.

c) Suppose that W is from a location–scale family with standard pdf $f_Y(y)$ that is continuous and positive on its support. Then $W = \mu + \sigma Y$ where $\sigma > 0$. First find M by solving $F_Y(M) = 0.5$. After finding M, find D by solving $F_Y(M + D) - F_Y(M - D) = 0.5$. Then $\text{MED}(W) = \mu + \sigma M$ and $\text{MAD}(W) = \sigma D$.

Proof sketch. a) Assume the probability density function of Y is continuous and positive on its support. Assume $b > 0$. Then

$$1/2 = P[Y \leq \text{MED}(Y)] = P[a + bY \leq a + b\text{MED}(Y)] = P[W \leq \text{MED}(W)].$$

$$\begin{aligned} 1/2 &= P[\text{MED}(Y) - \text{MAD}(Y) \leq Y \leq \text{MED}(Y) + \text{MAD}(Y)] \\ &= P[a + b\text{MED}(Y) - b\text{MAD}(Y) \leq a + bY \leq a + b\text{MED}(Y) + b\text{MAD}(Y)] \\ &= P[\text{MED}(W) - b\text{MAD}(Y) \leq W \leq \text{MED}(W) + b\text{MAD}(Y)] \\ &= P[\text{MED}(W) - \text{MAD}(W) \leq W \leq \text{MED}(W) + \text{MAD}(W)]. \end{aligned}$$

The proofs of b) and c) are similar. □

Application 8.1. *The MAD Method:* In analogy with the method of moments, *robust point estimators* can be obtained by solving $\text{MED}(n) = \text{MED}(Y)$ and $\text{MAD}(n) = \text{MAD}(Y)$. In particular, the location and scale parameters of a location–scale family can often be estimated robustly using $c_1\text{MED}(n)$ and $c_2\text{MAD}(n)$ where c_1 and c_2 are appropriate constants.

Estimators that use order statistics are common. The shorth estimator of Section 4.1 was used for prediction and confidence intervals.

Definition 8.10. Consider intervals that contain c_n cases: $[Y_{(1)}, Y_{(c_n)}]$, $[Y_{(2)}, Y_{(c_n+1)}], ..., [Y_{(n-c_n+1)}, Y_{(n)}]$. Denote the set of c_n cases in the ith interval by J_i, for $i = 1, 2, ..., n - c_n + 1$. Often $c_n = \lfloor n/2 \rfloor + 1$.

i) Let the shorth(c_n) estimator $= [Y_{(s)}, Y_{(s+c_n-1)}]$ be the shortest such interval. Then the *least median of squares estimator* $\text{LMS}(c_n)$ is $(Y_{(s)} + Y_{(s+c_n-1)})/2$, the midpoint of the shorth(c_n) interval. The LMS estimator is also called the *least quantile of squares estimator* $\text{LQS}(c_n)$.

ii) Compute the sample mean and sample variance $(\overline{Y}_{J_i}, S^2_{J_i})$ of the c_n cases in the ith interval. The *minimum covariance determinant* estimator MCD(c_n) estimator $(\overline{Y}_{MCD}, S^2_{MCD})$ is equal to the $(\overline{Y}_{J_j}, S^2_{J_j})$ with the smallest $S^2_{J_i}$. The *least trimmed sum of squares estimator* is LTS(c_n) $= \overline{Y}_{MCD}$.

iii) Compute the sample median M_{J_i} of the c_n cases in the ith interval. Let $Q_{LTA}(M_{J_i}) = \sum_{j \in J_i} |y_j - M_{J_i}|$. The *least trimmed sum of absolute deviations estimator* LTA(c_n) is equal to the M_{J_j} with the smallest $Q_{LTA}(M_{J_i})$.

8.1.1 Robust Confidence Intervals

In this subsection, large sample confidence intervals (CIs) for the sample median and 25% trimmed mean are given. Theory is given later in Section 8.1. The following confidence interval provides some resistance to gross outliers while being very simple to compute. The standard error SE(MED(n)) is due to Bloch and Gastwirth (1968), but the degrees of freedom $p \approx \lceil \sqrt{n} \rceil$) is motivated by the confidence interval for the trimmed mean. Let $\lfloor x \rfloor$ denote the "greatest integer function" (e.g., $\lfloor 7.7 \rfloor = 7$). Let $\lceil x \rceil$ denote the smallest integer greater than or equal to x (e.g., $\lceil 7.7 \rceil = 8$).

Warning: Closed intervals should be used instead of open intervals: $a \pm b = [a - b, a + b]$.

Application 8.2: inference with the sample median. Let $U_n = n - L_n$ where $L_n = \lfloor n/2 \rfloor - \lceil \sqrt{n/4} \rceil$ and use

$$SE(\text{MED}(n)) = 0.5(Y_{(U_n)} - Y_{(L_n+1)}).$$

Let $p = U_n - L_n - 1$. Then a $100(1-\alpha)\%$ confidence interval for the population median is

$$\text{MED}(n) \pm t_{p,1-\alpha/2} SE(\text{MED}(n)). \tag{8.6}$$

Warning. This CI is easy to compute by hand, but tends to be long with undercoverage if $n < 100$. See Baszczyńska and Pekasiewicz (2010) for two competitors that work better. We recommend bootstrap confidence intervals for the population median.

The trimmed mean is also useful, and we recommend the 25% trimmed mean. Let $\lfloor x \rfloor$ denote the "greatest integer function" (e.g., $\lfloor 7.7 \rfloor = 7$).

Definition 8.11. The symmetrically trimmed mean or the δ *trimmed mean*

$$T_n = T_n(L_n, U_n) = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} Y_{(i)} \tag{8.7}$$

where $L_n = \lfloor n\delta \rfloor$ and $U_n = n - L_n$. If $\delta = 0.25$, say, then the δ trimmed mean is called the 25% trimmed mean.

The $(\delta, 1-\gamma)$ *trimmed mean* uses $L_n = \lfloor n\delta \rfloor$ and $U_n = \lfloor n\gamma \rfloor$.

The trimmed mean is estimating a truncated mean μ_T. Assume that Y has a probability density function $f_Y(y)$ that is continuous and positive on its support. Let y_δ be the number satisfying $P(Y \leq y_\delta) = \delta$. Then

$$\mu_T = \frac{1}{1-2\delta} \int_{y_\delta}^{y_{1-\delta}} y f_Y(y) dy. \tag{8.8}$$

Notice that the 25% trimmed mean is estimating

$$\mu_T = \int_{y_{0.25}}^{y_{0.75}} 2y f_Y(y) dy.$$

To perform inference, find $d_1, ..., d_n$ where

$$d_i = \begin{cases} Y_{(L_n+1)}, & i \leq L_n \\ Y_{(i)}, & L_n + 1 \leq i \leq U_n \\ Y_{(U_n)}, & i \geq U_n + 1. \end{cases}$$

Then the Winsorized variance is the sample variance $S_n^2(d_1, ..., d_n)$ of $d_1, ..., d_n$, and the scaled Winsorized variance

$$V_{SW}(L_n, U_n) = \frac{S_n^2(d_1, ..., d_n)}{([U_n - L_n]/n)^2}. \tag{8.9}$$

The standard error (SE) of T_n is $SE(T_n) = \sqrt{V_{SW}(L_n, U_n)/n}$.

Application 8.3: inference with the δ trimmed mean. A large sample 100 $(1-\alpha)$% confidence interval (CI) for μ_T is

$$T_n \pm t_{p,1-\frac{\alpha}{2}} SE(T_n) \tag{8.10}$$

where $P(t_p \leq t_{p,1-\frac{\alpha}{2}}) = 1 - \alpha/2$ if t_p is from a t distribution with $p = U_n - L_n - 1$ degrees of freedom. This interval is the classical t–interval when $\delta = 0$, but $\delta = 0.25$ gives a robust CI.

Example 8.2. Let the data be 6, 9, 9, 7, 8, 9, 9, 7. Assume the data came from a symmetric distribution with mean μ, and find a 95% CI for μ.

Solution. When computing small examples by hand, the steps are to sort the data from smallest to largest value, find n, L_n, U_n, $Y_{(L_n+1)}$, $Y_{(U_n)}$, p, MED(n) and SE(MED(n)). After finding $t_{p,1-\alpha/2}$, plug the relevant quantities into the formula for the CI. The sorted data are 6, 7, 7, 8, 9, 9, 9, 9. Thus MED$(n) = (8+9)/2 = 8.5$. Since $n = 8$, $L_n = \lfloor 4 \rfloor - \lceil \sqrt{2} \rceil = 4 - \lceil 1.414 \rceil = 4 - 2 = 2$ and $U_n = n - L_n = 8 - 2 = 6$. Hence $SE(\text{MED}(n)) = 0.5(Y_{(6)} - Y_{(3)}) = 0.5 * (9 - 7) = 1$. The degrees of free-

dom $p = U_n - L_n - 1 = 6 - 2 - 1 = 3$. The cutoff $t_{3,0.975} = 3.182$. Thus the 95% CI for MED(Y) is

$$\text{MED}(n) \pm t_{3,0.975} SE(\text{MED}(n))$$

$= 8.5 \pm 3.182(1) = [5.318, 11.682]$. The classical t–interval uses $\overline{Y} = (6 + 7 + 7 + 8 + 9 + 9 + 9 + 9)/8$ and $S_n^2 = (1/7)[(\sum_{i=1}^{n} Y_i^2) - 8(8^2)] = (1/7)[(522 - 8(64)] = 10/7 \approx 1.4286$, and $t_{7,0.975} \approx 2.365$. Hence the 95% CI for μ is $8 \pm 2.365(\sqrt{1.4286/8}) = [7.001, 8.999]$. Notice that the t-cutoff $= 2.365$ for the classical interval is less than the t-cutoff $= 3.182$ for the median interval and that $SE(\overline{Y}) < SE(\text{MED}(n))$. The parameter μ is between 1 and 9 since the test scores are integers between 1 and 9. Hence for this example, the t–interval is considerably superior to the overly long median interval.

Example 8.3. In the last example, what happens if the 6 becomes 66 and a 9 becomes 99?

Solution. Then the ordered data are 7, 7, 8, 9, 9, 9, 66, 99. Hence MED$(n) = 9$. Since L_n and U_n only depend on the sample size, they take the same values as in the previous example and $SE(\text{MED}(n)) = 0.5(Y_{(6)} - Y_{(3)}) = 0.5 * (9 - 8) = 0.5$. Hence the 95% CI for MED(Y) is MED$(n) \pm t_{3,0.975} SE(\text{MED}(n)) = 9 \pm 3.182(0.5) = [7.409, 10.591]$. Notice that with discrete data, it is possible to drive $SE(\text{MED}(n))$ to 0 with a few outliers if n is small. The classical confidence interval $\overline{Y} \pm t_{7,0.975} S/\sqrt{n}$ blows up and is equal to $[-2.955, 56.455]$.

8.1.2 Some Two Stage Trimmed Means

Robust estimators are often obtained by applying the sample mean to a sequence of consecutive order statistics. The sample median, trimmed mean, metrically trimmed mean, and two stage trimmed means are examples. For the trimmed mean given in Definition 8.11 and for the Winsorized mean, defined below, the proportion of cases trimmed and the proportion of cases covered are fixed.

Definition 8.12. Using the same notation as in Definition 8.11, the *Winsorized mean*

$$W_n = W_n(L_n, U_n) = \frac{1}{n}[L_n Y_{(L_n+1)} + \sum_{i=L_n+1}^{U_n} Y_{(i)} + (n - U_n) Y_{(U_n)}]. \quad (8.11)$$

Definition 8.13. A *randomly trimmed mean*

$$R_n = R_n(L_n, U_n) = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} Y_{(i)} \tag{8.12}$$

where $L_n < U_n$ are integer valued random variables. $U_n - L_n$ of the cases are *covered* by the randomly trimmed mean while $n - U_n + L_n$ of the cases are trimmed.

Definition 8.14. The *metrically trimmed mean* (also called the Huber type skipped mean) M_n is the sample mean of the cases inside the interval

$$[\hat{\theta}_n - k_1 D_n,\ \hat{\theta}_n + k_2 D_n]$$

where $\hat{\theta}_n$ is a location estimator, D_n is a scale estimator, $k_1 \geq 1$, and $k_2 \geq 1$.

The proportions of cases covered and trimmed by randomly trimmed means such as the metrically trimmed mean are now random. Typically MED(n) and MAD(n) are used for $\hat{\theta}_n$ and D_n, respectively. The amount of trimming will depend on the distribution of the data. For example, if M_n uses $k_1 = k_2 = 5.2$ and the data is normal (Gaussian), about 1% of the data will be trimmed while if the data is Cauchy, about 12% of the data will be trimmed. Hence the upper and lower trimming points estimate lower and upper population percentiles $L(F)$ and $U(F)$ and change with the distribution F.

Two stage estimators are frequently used in robust statistics. Often the initial estimator used in the first stage has good resistance properties but has a low asymptotic relative efficiency or no convenient formula for the SE. Ideally, the estimator in the second stage will have resistance similar to the initial estimator but will be efficient and easy to use. The metrically trimmed mean M_n with tuning parameter $k_1 = k_2 \equiv k = 6$ will often be the initial estimator for the two stage trimmed means. That is, retain the cases that fall in the interval

$$[\text{MED}(n) - 6\text{MAD}(n), \text{MED}(n) + 6\text{MAD}(n)].$$

Let $L(M_n)$ be the number of observations that fall to the left of MED$(n) - k_1$ MAD(n) and let $n - U(M_n)$ be the number of observations that fall to the right of MED$(n) + k_2$ MAD(n). When $k_1 = k_2 \equiv k \geq 1$, at least half of the cases will be covered. Consider the set of 51 trimming proportions in the set $C = \{0, 0.01, 0.02, ..., 0.49, 0.50\}$. Alternatively, the coarser set of 6 trimming proportions $C = \{0, 0.01, 0.1, 0.25, 0.40, 0.49\}$ may be of interest. The greatest integer function (e.g. $\lfloor 7.7 \rfloor = 7$) is used in the following definitions.

Definition 8.15. Consider the smallest proportion $\alpha_{o,n} \in C$ such that $\alpha_{o,n} \geq L(M_n)/n$ and the smallest proportion $1 - \beta_{o,n} \in C$ such that $1 - \beta_{o,n} \geq 1 - (U(M_n)/n)$. Let $\alpha_{M,n} = \max(\alpha_{o,n}, 1 - \beta_{o,n})$. Then the *two stage*

symmetrically trimmed mean $T_{S,n}$ is the $\alpha_{M,n}$ trimmed mean. Hence $T_{S,n}$ is a randomly trimmed mean with $L_n = \lfloor n \ \ \alpha_{M,n} \rfloor$ and $U_n = n - L_n$. If $\alpha_{M,n} = 0.50$, then use $T_{S,n} = \text{MED}(n)$.

Definition 8.16. As in the previous definition, consider the smallest proportion $\alpha_{o,n} \in C$ such that $\alpha_{o,n} \geq L(M_n)/n$ and the smallest proportion $1 - \beta_{o,n} \in C$ such that $1 - \beta_{o,n} \geq 1 - (U(M_n)/n)$. Then the *two stage asymmetrically trimmed mean* $T_{A,n}$ is the $(\alpha_{o,n}, 1 - \beta_{o,n})$ trimmed mean. Hence $T_{A,n}$ is a randomly trimmed mean with $L_n = \lfloor n \ \ \alpha_{o,n} \rfloor$ and $U_n = \lfloor n \ \ \beta_{o,n} \rfloor$. If $\alpha_{o,n} = 1 - \beta_{o,n} = 0.5$, then use $T_{A,n} = \text{MED}(n)$.

Example 8.4. These two stage trimmed means are almost as easy to compute as the classical trimmed mean, and no knowledge of the unknown parameters is needed to do inference. First, order the data and find the number of cases $L(M_n)$ less than $\text{MED}(n) - k_1\text{MAD}(n)$ and the number of cases $n - U(M_n)$ greater than $\text{MED}(n) + k_2\text{MAD}(n)$. (These are the cases trimmed by the metrically trimmed mean M_n, but M_n need not be computed.) Next, convert these two numbers into percentages and round both percentages up to the nearest integer. For $T_{S,n}$ find the maximum of the two percentages. For example, suppose that there are $n = 205$ cases and M_n trims the smallest 15 cases and the largest 20 cases. Then $L(M_n)/n = 0.073$ and $1 - (U(M_n)/n) = 0.0976$. Hence M_n trimmed the 7.3% smallest cases and the 9.76% largest cases, and $T_{S,n}$ is the 10% trimmed mean while $T_{A,n}$ is the (0.08, 0.10) trimmed mean.

Definition 8.17. The standard error SE_{RM} for the two stage trimmed means given in Definitions 8.11, 8.15, or 8.16 is

$$SE_{RM}(L_n, U_n) = \sqrt{V_{SW}(L_n, U_n)/n}$$

where the *scaled Winsorized variance* $V_{SW}(L_n, U_n) =$

$$\frac{[L_n Y_{(L_n+1)}^2 + \sum_{i=L_n+1}^{U_n} Y_{(i)}^2 + (n - U_n)Y_{(U_n)}^2] - \ n \ [W_n(L_n, U_n)]^2}{(n-1)[(U_n - L_n)/n]^2}. \quad (8.13)$$

Remark 8.1. A simple method for computing $V_{SW}(L_n, U_n)$ has the following steps. First, find $d_1, ..., d_n$ where

$$d_i = \begin{cases} Y_{(L_n+1)}, & i \leq L_n \\ Y_{(i)}, & L_n + 1 \leq i \leq U_n \\ Y_{(U_n)}, & i \geq U_n + 1. \end{cases}$$

Then the Winsorized variance is the sample variance $S_n^2(d_1, ..., d_n)$ of $d_1, ..., d_n$, and the scaled Winsorized variance

$$V_{SW}(L_n, U_n) = \frac{S_n^2(d_1, ..., d_n)}{([U_n - L_n]/n)^2}. \tag{8.14}$$

Notice that the SE given in Definition 8.17 is the SE for the δ trimmed mean where L_n and U_n are fixed constants rather than random.

Application 8.4. Let T_n be the two stage (symmetrically or) asymmetrically trimmed mean that trims the L_n smallest cases and the $n - U_n$ largest cases. Then for the one and two sample procedures described in Section 5.1, use the one sample standard error $\mathrm{SE}_{RM}(L_n, U_n)$ given in Definition 8.17 and the t_p distribution where the degrees of freedom $p = U_n - L_n - 1$.

The CIs and tests for the δ trimmed mean and two stage trimmed means given by Applications 8.3 and 8.4 are very similar once L_n has been computed. For example, a large sample 100 $(1 - \alpha)\%$ confidence interval (CI) for μ_T is

$$[T_n - t_{U_n - L_n - 1, 1-\frac{\alpha}{2}} SE_{RM}(L_n, U_n), T_n + t_{U_n - L_n - 1, 1-\frac{\alpha}{2}} SE_{RM}(L_n, U_n)] \tag{8.15}$$

where $P(t_p \leq t_{p,1-\frac{\alpha}{2}}) = 1 - \alpha/2$ if t_p is from a t distribution with p degrees of freedom. Section 8.1.6 provides the asymptotic theory for the δ and two stage trimmed means and shows that μ_T is the mean of a truncated distribution. Next Examples 8.2 and 8.3 are repeated using the intervals based on the two stage trimmed means instead of the median.

Example 8.5. Let the data be 6, 9, 9, 7, 8, 9, 9, 7. Assume the data came from a symmetric distribution with mean μ, and find a 95% CI for μ.

Solution. If $T_{A,n}$ or $T_{S,n}$ is used with the metrically trimmed mean that uses $k = k_1 = k_2$, e.g. $k = 6$, then $\mu_T(a, b) = \mu$. When computing small examples by hand, it is convenient to sort the data:
6, 7, 7, 8, 9, 9, 9, 9.
Thus $\mathrm{MED}(n) = (8 + 9)/2 = 8.5$. The ordered residuals $Y_{(i)} - \mathrm{MED}(n)$ are
-2.5, -1.5, -1.5, 0.5, 0.5, 0.5, 0.5, 0.5.
Find the absolute values and sort them to get
0.5, 0.5, 0.5, 0.5, 0.5, 1.5, 1.5, 2.5.
Then $\mathrm{MAD}(n) = 0.5$, $\mathrm{MED}(n) - 6MAD(n) = 5.5$, and $\mathrm{MED}(n) + 6MAD(n) = 11.5$. Hence no cases are trimmed by the metrically trimmed mean, i.e. $L(M_n) = 0$ and $U(M_n) = n = 8$. Thus $L_n = \lfloor 8(0) \rfloor = 0$, and $U_n = n - L_n = 8$. Since no cases are trimmed by the two stage trimmed means, the robust interval will have the same endpoints as the classical t–interval. To see this, note that $M_n = T_{S,n} = T_{A,n} = \overline{Y} = (6 + 7 + 7 + 8 + 9 + 9 + 9 + 9)/8 = 8 = W_n(L_n, U_n)$. Now $V_{SW}(L_n, U_n) = (1/7)[\sum_{i=1}^{n} Y_{(i)}^2 - 8(8^2)]/[8/8]^2 = (1/7)[(522 - 8(64)] = 10/7 \approx 1.4286$, and $t_{7,0.975} \approx 2.365$. Hence the 95% CI for μ is $8 \pm 2.365(\sqrt{1.4286/8}) = [7.001, 8.999]$.

Example 8.6. In the last example, what happens if a 6 becomes 66 and a 9 becomes 99? Use $k = 6$ and $T_{A,n}$. Then the ordered data are
7, 7, 8, 9, 9, 9, 66, 99.

Thus MED$(n) = 9$ and MAD$(n) = 1.5$. With $k = 6$, the metrically trimmed mean M_n trims the two values 66 and 99. Hence the left and right trimming proportions of the metrically trimmed mean are 0.0 and $0.25 = 2/8$, respectively. These numbers are also the left and right trimming proportions of $T_{A,n}$ since after converting these proportions into percentages, both percentages are integers. Thus $L_n = \lfloor 0 \rfloor = 0$, $U_n = \lfloor 0.75(8) \rfloor = 6$ and the two stage asymmetrically trimmed mean trims 66 and 99. So $T_{A,n} = 49/6 \approx 8.1667$. To compute the scaled Winsorized variance, use Remark 8.3 to find that the d_i's are
7, 7, 8, 9, 9, 9, 9, 9
and

$$V_{SW} = \frac{S_n^2(d_1, ..., d_8)}{[(6-0)/8]^2} \approx \frac{0.8393}{.5625} \approx 1.4921.$$

Hence the robust confidence interval is $8.1667 \pm t_{5,0.975}\sqrt{1.4921/8} \approx 8.1667 \pm 1.1102 \approx [7.057, 9.277]$. The classical confidence interval $\overline{Y} \pm t_{n-1,0.975} S/\sqrt{n}$ blows up and is equal to $[-2.955, 56.455]$.

Example 8.7. Use $k = 6$ and $T_{A,n}$ to compute a robust CI using the 87 heights from the Buxton (1920) data that includes 5 outliers. The mean height is $\overline{Y} = 1598.862$ while $T_{A,n} = 1695.22$. The classical 95% CI is [1514.206,1683.518] and is more than five times as long as the robust 95% CI which is [1679.907,1710.532]. In this example the five outliers can be corrected. For the corrected data, no cases are trimmed and the robust and classical estimators have the same values. The results are $\overline{Y} = 1692.356 = T_{A,n}$ and the robust and classical 95% CIs are both [1678.595,1706.118]. Note that the outliers did not have much affect on the robust confidence interval.

8.1.3 Asymptotics for Two Stage Trimmed Means

Large sample theory is very important for understanding robust statistics. Truncated and Winsorized random variables are important because they simplify the asymptotic theory of robust estimators. Let Y be a random variable with continuous cdf F and let $\alpha = F(a) < F(b) = \beta$. Thus α is the *left trimming proportion* and $1 - \beta$ is the *right trimming proportion.* Let $F(a-) = P(Y < a)$. (Refer to Section 1.8 for the notation used below.)

Definition 8.18. The *truncated random variable* $Y_T \equiv Y_T(a, b)$ with *truncation points* a and b has cdf

$$F_{Y_T}(y|a, b) = G(y) = \frac{F(y) - F(a-)}{F(b) - F(a-)} \tag{8.16}$$

for $a \leq y \leq b$. Also G is 0 for $y < a$ and G is 1 for $y > b$. The mean and variance of Y_T are

$$\mu_T = \mu_T(a,b) = \int_{-\infty}^{\infty} y dG(y) = \frac{\int_a^b y dF(y)}{\beta - \alpha} \tag{8.17}$$

and

$$\sigma_T^2 = \sigma_T^2(a,b) = \int_{-\infty}^{\infty} (y - \mu_T)^2 dG(y) = \frac{\int_a^b y^2 dF(y)}{\beta - \alpha} - \mu_T^2.$$

See Cramér (1946, p. 247).

Definition 8.19. The *Winsorized random variable*

$$Y_W = Y_W(a,b) = \begin{cases} a, & Y \leq a \\ Y, & a \leq Y \leq b \\ b, & Y \geq b. \end{cases}$$

If the cdf of $Y_W(a,b) = Y_W$ is F_W, then

$$F_W(y) = \begin{cases} 0, & y < a \\ F(a), & y = a \\ F(y), & a < y < b \\ 1, & y \geq b. \end{cases}$$

Since Y_W is a mixture distribution with a point mass at a and at b, the mean and variance of Y_W are

$$\mu_W = \mu_W(a,b) = \alpha a + (1-\beta)b + \int_a^b y dF(y)$$

and

$$\sigma_W^2 = \sigma_W^2(a,b) = \alpha a^2 + (1-\beta)b^2 + \int_a^b y^2 dF(y) - \mu_W^2.$$

Definition 8.20. The *quantile function*

$$F_Q^{-1}(t) = Q(t) = \inf\{y : F(y) \geq t\}. \tag{8.18}$$

The *sample* ρ *quantile* $\hat{\xi}_{n,\rho} = Y_{(\lceil n\rho \rceil)} = \hat{y}_\rho$. The *population quantile* $y_\rho = \pi_\rho = \xi_\rho = Q(\rho)$ where $0 < \rho < 1$.

Warning: Software often uses a slightly different definition of the sample quantile then the one given in Definition 8.20.

Note that $Q(t)$ is the left continuous inverse of F and if F is strictly increasing and continuous, then F has an inverse F^{-1} and $F^{-1}(t) = Q(t)$. The following conditions on the cdf are used.

Regularity Conditions. (R1) Let $Y_1, \ldots, Y_n$ be iid with cdf F.
(R2) Let F be continuous and strictly increasing at $a = Q(\alpha)$ and $b = Q(\beta)$.

The following theorem is proved in Bickel (1965), Stigler (1973), and Shorack and Wellner (1986, p. 678-679). The α trimmed mean is asymptotically equivalent to the $(\alpha, 1-\alpha)$ trimmed mean. Let T_n be the $(\alpha, 1-\beta)$ trimmed mean. Theorem 8.3 shows that the standard error SE_{RM} given in the previous section is estimating the appropriate asymptotic standard deviation of T_n.

Theorem 8.2. If conditions (R1) and (R2) hold and if $0 < \alpha < \beta < 1$, then

$$\sqrt{n}(T_n - \mu_T(a,b)) \xrightarrow{D} N\left[0, \frac{\sigma^2_W(a,b)}{(\beta-\alpha)^2}\right]. \tag{8.19}$$

Theorem 8.3: Shorack and Wellner (1986, p. 680). Assume that regularity conditions (R1) and (R2) hold and that

$$\frac{L_n}{n} \xrightarrow{P} \alpha \text{ and } \frac{U_n}{n} \xrightarrow{P} \beta. \tag{8.20}$$

Then

$$V_{SW}(L_n, U_n) \xrightarrow{P} \frac{\sigma^2_W(a,b)}{(\beta-\alpha)^2}.$$

Since $L_n = \lfloor n\alpha \rfloor$ and $U_n = n - L_n$ (or $L_n = \lfloor n\alpha \rfloor$ and $U_n = \lfloor n\beta \rfloor$) satisfy the above lemma, the standard error SE_{RM} can be used for both trimmed means and two stage trimmed means: $\mathrm{SE}_{RM}(L_n, U_n) = \sqrt{V_{SW}(L_n, U_n)/n}$ where the *scaled Winsorized variance* $V_{SW}(L_n, U_n) =$

$$\frac{[L_n Y^2_{(L_n+1)} + \sum_{i=L_n+1}^{U_n} Y^2_{(i)} + (n-U_n)Y^2_{(U_n)}] - \ n\ [W_n(L_n, U_n)]^2}{(n-1)[(U_n - L_n)/n]^2}.$$

Again L_n is the number of cases trimmed to the left and $n - U_n$ is the number of cases trimmed to the right by the trimmed mean.

The following notation will be useful for finding the asymptotic distribution of the two stage trimmed means. Let $a = \mathrm{MED}(Y) - k\mathrm{MAD}(Y)$ and $b = \mathrm{MED}(Y) + k\mathrm{MAD}(Y)$ where $\mathrm{MED}(Y)$ and $\mathrm{MAD}(Y)$ are the population median and median absolute deviation respectively. Let $\alpha = F(a-) = P(Y < a)$ and let $\alpha_o \in C = \{0, 0.01, 0.02, ..., 0.49, 0.50\}$ be the smallest value in C such that $\alpha_o \geq \alpha$. Similarly, let $\beta = F(b)$ and let $1 - \beta_o \in C$ be the smallest value in the index set C such that $1 - \beta_o \geq 1 - \beta$. Let $\alpha_o = F(a_o-)$, and let $\beta_o = F(b_o)$. Recall that $L(M_n)$ is the number of cases trimmed to the left and that $n - U(M_n)$ is the number of cases trimmed to the right by the metrically trimmed mean M_n. Let $\alpha_{o,n} \equiv \hat{\alpha}_o$ be the smallest value in C such that $\alpha_{o,n} \geq L(M_n)/n$, and let $1 - \beta_{o,n} \equiv 1 - \hat{\beta}_o$ be the smallest value in C such that $1 - \beta_{o,n} \geq 1 - (U(M_n)/n)$. Then the robust estimator $T_{A,n}$ is the $(\alpha_{o,n}, 1 - \beta_{o,n})$ trimmed mean while $T_{S,n}$ is the $\max(\alpha_{o,n}, 1 - \beta_{o,n})100\%$

trimmed mean. The following theorem is useful for showing that $T_{A,n}$ is asymptotically equivalent to the $(\alpha_o, 1-\beta_o)$ trimmed mean and that $T_{S,n}$ is asymptotically equivalent to the $\max(\alpha_o, 1-\beta_o)$ trimmed mean. One proof of Theorem 8.5 is to show that $T_{A,n}$ and $T_{S,n}$ are model selection estimators where the probability $T_{A,n}$ selects the $(\alpha_o, 1-\beta_o)$ trimmed mean and the probability that $T_{S,n}$ selects the $\max(\alpha_o, 1-\beta_o)$ trimmed mean goes to one.

Theorem 8.4: Shorack and Wellner (1986, p. 682-683). Let F have a strictly positive and continuous derivative in some neighborhood of $\text{MED}(Y) \pm k\text{MAD}(Y)$. Assume that

$$\sqrt{n}(MED(n) - MED(Y)) = O_P(1) \tag{8.21}$$

and

$$\sqrt{n}(MAD(n) - MAD(X)) = O_P(1). \tag{8.22}$$

Then

$$\sqrt{n}(\frac{L(M_n)}{n} - \alpha) = O_P(1) \tag{8.23}$$

and

$$\sqrt{n}(\frac{U(M_n)}{n} - \beta) = O_P(1). \tag{8.24}$$

Theorem 8.5. Let $Y_1, ..., Y_n$ be iid from a distribution with cdf F that has a strictly positive and continuous pdf f on its support. Let $\alpha_M = \max(\alpha_o, 1-\beta_o) \leq 0.49$, $\beta_M = 1 - \alpha_M$, $a_M = F^{-1}(\alpha_M)$, and $b_M = F^{-1}(\beta_M)$. Assume that α and $1-\beta$ are not elements of $C = \{0, 0.01, 0.02, ..., 0.50\}$. Then

$$\sqrt{n}[T_{A,n} - \mu_T(a_o, b_o)] \xrightarrow{D} N\left[0, \frac{\sigma^2_W(a_o, b_o)}{(\beta_o - \alpha_o)^2}\right],$$

and

$$\sqrt{n}[T_{S,n} - \mu_T(a_M, b_M)] \xrightarrow{D} N\left[0, \frac{\sigma^2_W(a_M, b_M)}{(\beta_M - \alpha_M)^2}\right].$$

Proof. The first result follows from Theorem 8.2 if the probability that $T_{A,n}$ is the $(\alpha_o, 1-\beta_o)$ trimmed mean goes to one as n tends to infinity. This condition holds if $L(M_n)/n \xrightarrow{D} \alpha$ and $U(M_n)/n \xrightarrow{D} \beta$. But these conditions follow from Theorem 8.4. The proof for $T_{S,n}$ is similar. □

8.1.4 Asymptotic Theory for the MAD

Let $\text{MD}(n) = \text{MED}(|Y_i - \text{MED}(Y)|,\ i = 1, \ldots, n)$. Since $\text{MD}(n)$ is a median and convergence results for the median are well known, see for example Serfling (1980, p. 74-77) or Theorem 2.6, it is simple to prove conver-

gence results for MAD(n). Typically MED$(n) =$ MED$(Y) + O_P(n^{-1/2})$ and MAD$(n) =$ MAD$(Y) + O_P(n^{-1/2})$.

Theorem 8.6. If MED$(n) =$ MED$(Y) + O_P(n^{-\delta})$ and MD$(n) =$ MAD$(Y) + O_P(n^{-\delta})$, then MAD$(n) =$ MAD$(Y) + O_P(n^{-\delta})$.

Proof. Let $W_i = |Y_i - \text{MED}(n)|$ and let $V_i = |Y_i - \text{MED}(Y)|$. Then

$$W_i = |Y_i - \text{MED}(Y) + \text{MED}(Y) - \text{MED}(n)| \le V_i + |\text{MED}(Y) - \text{MED}(n)|,$$

and

$$\text{MAD}(n) = \text{MED}(W_1, \ldots, W_n) \le \text{MED}(V_1, \ldots, V_n) + |\text{MED}(Y) - \text{MED}(n)|.$$

Similarly

$$V_i = |Y_i - \text{MED}(n) + \text{MED}(n) - \text{MED}(Y)| \le W_i + |\text{MED}(n) - \text{MED}(Y)|$$

and thus

$$\text{MD}(n) = \text{MED}(V_1, \ldots, V_n) \le \text{MED}(W_1, \ldots, W_n) + |\text{MED}(Y) - \text{MED}(n)|.$$

Combining the two inequalities shows that

$$\text{MD}(n) - |\text{MED}(Y) - \text{MED}(n)| \le \text{MAD}(n) \le \text{MD}(n) + |\text{MED}(Y) - \text{MED}(n)|,$$

or

$$|\text{MAD}(n) - \text{MD}(n)| \le |\text{MED}(n) - \text{MED}(Y)|. \tag{8.25}$$

Adding and subtracting MAD(Y) to the left hand side shows that

$$|\text{MAD}(n) - \text{MAD}(Y) - O_P(n^{-\delta})| = O_P(n^{-\delta}) \tag{8.26}$$

and the result follows. □

The main point of the following theorem is that the joint distribution of MED(n) and MAD(n) is asymptotically normal. Hence the limiting distribution of MED$(n) + k$MAD(n) is also asymptotically normal for any constant k. The parameters of the covariance matrix are quite complex and hard to estimate. The assumptions of f used in Theorem 8.7 guarantee that MED(Y) and MAD(Y) are unique.

Theorem 8.7: Falk (1997). Let the cdf F of Y be continuous near and differentiable at MED$(Y) = F^{-1}(1/2)$ and MED$(Y) \pm$MAD(Y). Assume that $f = F'$, $f(F^{-1}(1/2)) > 0$, and $A \equiv f(F^{-1}(1/2) - \text{MAD}(Y)) + f(F^{-1}(1/2) + \text{MAD}(Y)) > 0$. Let $C \equiv f(F^{-1}(1/2) - \text{MAD}(Y)) - f(F^{-1}(1/2) + \text{MAD}(Y))$, and let $B \equiv C^2 + 4Cf(F^{-1}(1/2))[1 - F(F^{-1}(1/2) - \text{MAD}(Y)) - F(F^{-1}(1/2) + \text{MAD}(Y))]$. Then

$$\sqrt{n}\left(\begin{pmatrix}\text{MED}(n)\\ \text{MAD}(n)\end{pmatrix} - \begin{pmatrix}\text{MED}(Y)\\ \text{MAD}(Y)\end{pmatrix}\right) \xrightarrow{D}$$

$$N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_M^2 & \sigma_{M,D}\\ \sigma_{M,D} & \sigma_D^2\end{pmatrix}\right) \tag{8.27}$$

where

$$\sigma_M^2 = \frac{1}{4f^2(F^{-1}(\frac{1}{2}))},\ \sigma_D^2 = \frac{1}{4A^2}(1 + \frac{B}{f^2(F^{-1}(\frac{1}{2}))}),$$

and

$$\sigma_{M,D} = \frac{1}{4Af(F^{-1}(\frac{1}{2}))}(1 - 4F(F^{-1}(\frac{1}{2}) + \text{MAD}(Y)) + \frac{C}{f(F^{-1}(\frac{1}{2}))}).$$

Determining whether the population median and mad are unique can be useful. Recall that $F(y) = P(Y \le y)$ and $F(y-) = P(Y < y)$. The median is unique unless there is a flat spot at $F^{-1}(0.5)$, that is, unless there exist a and b with $a < b$ such that $F(a) = F(b) = 0.5$. If $\text{MED}(Y)$ is unique, then $\text{MAD}(Y)$ is unique unless F has flat spots at both $F^{-1}(\text{MED}(Y) - \text{MAD}(Y))$ and $F^{-1}(\text{MED}(Y) + \text{MAD}(Y))$. Moreover, $\text{MAD}(Y)$ is unique unless there exist $a_1 < a_2$ and $b_1 < b_2$ such that $F(a_1) = F(a_2)$, $F(b_1) = F(b_2)$,

$$P(a_i \le Y \le b_i) = F(b_i) - F(a_i-) \ge 0.5,$$

and

$$P(Y \le a_i) + P(Y \ge b_i) = F(a_i) + 1 - F(b_i-) \ge 0.5$$

for $i = 1, 2$. The following theorem gives some simple bounds for $\text{MAD}(Y)$.

Theorem 8.8. Assume $\text{MED}(Y)$ and $\text{MAD}(Y)$ are unique. a) Then

$$\min\{\text{MED}(Y) - F^{-1}(0.25), F^{-1}(0.75) - \text{MED}(Y)\} \le \text{MAD}(Y) \le$$

$$\max\{\text{MED}(Y) - F^{-1}(0.25), F^{-1}(0.75) - \text{MED}(Y)\}. \tag{8.28}$$

b) If Y is symmetric about $\mu = F^{-1}(0.5)$, then the three terms in a) are equal.
c) If the distribution is symmetric about zero, then $\text{MAD}(Y) = F^{-1}(0.75)$.
d) If Y is symmetric and continuous with a finite second moment, then

$$\text{MAD}(Y) \le \sqrt{2\text{VAR}(Y)}.$$

e) Suppose $Y \in [a, b]$. Then

$$0 \le \text{MAD}(Y) \le m = \min\{\text{MED}(Y) - a, b - \text{MED}(Y)\} \le (b - a)/2,$$

and the inequalities are sharp.

Proof. a) This result follows since half the mass is between the upper and lower quartiles and the median is between the two quartiles.

b) and c) are corollaries of a).

d) This inequality holds by Chebyshev's inequality, since

$$P(\ |Y - E(Y)| \geq \text{MAD}(Y)\ ) = 0.5 \geq P(\ |Y - E(Y)| \geq \sqrt{2\text{VAR}(Y)}\ ),$$

and $E(Y) = \text{MED}(Y)$ for symmetric distributions with finite second moments.

e) Note that if $\text{MAD}(Y) > m$, then either $\text{MED}(Y) - \text{MAD}(Y) < a$ or $\text{MED}(Y) + \text{MAD}(Y) > b$. Since at least half of the mass is between a and $\text{MED}(Y)$ and between $\text{MED}(Y)$ and b, this contradicts the definition of $\text{MAD}(Y)$. To see that the inequalities are sharp, note that if at least half of the mass is at some point $c \in [a, b]$, than $\text{MED}(Y) = c$ and $\text{MAD}(Y) = 0$. If each of the points a, b, and c has $1/3$ of the mass where $a < c < b$, then $\text{MED}(Y) = c$ and $\text{MAD}(Y) = m$. □

Many other results for $\text{MAD}(Y)$ and $\text{MAD}(n)$ are possible. For example, note that Theorem 8.8 b) implies that when Y is symmetric, $\text{MAD}(Y) = F^{-1}(3/4) - \mu$ and $F(\mu + \text{MAD}(Y)) = 3/4$. Also note that $\text{MAD}(Y)$ and the interquartile range $IQR(Y)$ are related by

$$2\text{MAD}(Y) = IQR(Y) \equiv y_{0.75} - y_{0.25}$$

when Y is symmetric.

8.1.5 Truncated Distributions

Truncated distributions can be used to simplify the asymptotic theory of robust estimators of location and regression. This subsection is useful when the underlying distribution is exponential, double exponential, normal, or Cauchy.

Definitions 8.18 and 8.19 defined the truncated random variable $Y_T(a, b)$ and the Winsorized random variable $Y_W(a, b)$. Let Y have cdf F and let the truncated random variable $Y_T(a, b)$ have the cdf $F_{T(a,b)}$. The following lemma illustrates the relationship between the means and variances of $Y_T(a, b)$ and $Y_W(a, b)$. Note that $Y_W(a, b)$ is a mixture of $Y_T(a, b)$ and two point masses at a and b. Let $c = \mu_T(a, b) - a$ and $d = b - \mu_T(a, b)$.

Theorem 8.9. Let $a = \mu_T(a, b) - c$ and $b = \mu_T(a, b) + d$. Then

a) $\mu_W(a, b) = \mu_T(a, b) - \alpha c + (1 - \beta)d$, and

b) $\sigma^2_W(a, b) = (\beta - \alpha)\sigma^2_T(a, b) + (\alpha - \alpha^2)c^2 + [(1 - \beta) - (1 - \beta)^2]d^2 + 2\alpha(1 - \beta)cd$.

c) If $\alpha = 1 - \beta$ then

$$\sigma_W^2(a,b) = (1-2\alpha)\sigma_T^2(a,b) + (\alpha-\alpha^2)(c^2+d^2) + 2\alpha^2 cd.$$

d) If $c = d$ then

$$\sigma_W^2(a,b) = (\beta-\alpha)\sigma_T^2(a,b) + [\alpha - \alpha^2 + 1 - \beta - (1-\beta)^2 + 2\alpha(1-\beta)]d^2.$$

e) If $\alpha = 1-\beta$ and $c = d$, then $\mu_W(a,b) = \mu_T(a,b)$ and

$$\sigma_W^2(a,b) = (1-2\alpha)\sigma_T^2(a,b) + 2\alpha d^2.$$

Proof. We will prove b) since its proof contains the most algebra. Now

$$\sigma_W^2 = \alpha(\mu_T - c)^2 + (\beta-\alpha)(\sigma_T^2 + \mu_T^2) + (1-\beta)(\mu_T + d)^2 - \mu_W^2.$$

Collecting terms shows that

$$\sigma_W^2 = (\beta-\alpha)\sigma_T^2 + (\beta - \alpha + \alpha + 1 - \beta)\mu_T^2 + 2[(1-\beta)d - \alpha c]\mu_T$$
$$+\alpha c^2 + (1-\beta)d^2 - \mu_W^2.$$

From a),

$$\mu_W^2 = \mu_T^2 + 2[(1-\beta)d - \alpha c]\mu_T + \alpha^2 c^2 + (1-\beta)^2 d^2 - 2\alpha(1-\beta)cd,$$

and we find that

$$\sigma_W^2 = (\beta-\alpha)\sigma_T^2 + (\alpha-\alpha^2)c^2 + [(1-\beta) - (1-\beta)^2]d^2 + 2\alpha(1-\beta)cd. \quad \square$$

The Truncated Exponential Distribution

Let Y be a (one sided) truncated exponential $TEXP(\lambda, b)$ random variable. Then the pdf of Y is

$$f_Y(y|\lambda, b) = \frac{\frac{1}{\lambda}e^{-y/\lambda}}{1 - \exp(-\frac{b}{\lambda})}$$

for $0 < y \leq b$ where $\lambda > 0$. Let $b = k\lambda$, and let

$$c_k = \int_0^{k\lambda} \frac{1}{\lambda} e^{-y/\lambda} dy = 1 - e^{-k}.$$

Next we will find the first two moments of $Y \sim TEXP(\lambda, b = k\lambda)$ for $k > 0$.

Theorem 8.10. If Y is $TEXP(\lambda, b = k\lambda)$ for $k > 0$, then

$$a)\ E(Y) = \lambda \left[\frac{1 - (k+1)e^{-k}}{1 - e^{-k}} \right],$$

and

$$b)\ E(Y^2) = 2\lambda^2 \left[\frac{1 - \frac{1}{2}(k^2 + 2k + 2)e^{-k}}{1 - e^{-k}} \right].$$

See Problem 8.6 for a related result.

Proof. a) Note that

$$c_k E(Y) = \int_0^{k\lambda} \frac{y}{\lambda} e^{-y/\lambda} dy = -ye^{-y/\lambda}|_0^{k\lambda} + \int_0^{k\lambda} e^{-y/\lambda} dy$$

(use integration by parts). So

$$c_k E(Y) = -k\lambda e^{-k} + (-\lambda e^{-y/\lambda})|_0^{k\lambda} = -k\lambda e^{-k} + \lambda(1 - e^{-k}).$$

Hence

$$E(Y) = \lambda \left[\frac{1 - (k+1)e^{-k}}{1 - e^{-k}} \right].$$

b) Note that

$$c_k E(Y^2) = \int_0^{k\lambda} \frac{y^2}{\lambda} e^{-y/\lambda} dy.$$

Since

$$\frac{d}{dy}[-(y^2 + 2\lambda y + 2\lambda^2)e^{-y/\lambda}] = \frac{1}{\lambda} e^{-y/\lambda}(y^2 + 2\lambda y + 2\lambda^2) - e^{-y/\lambda}(2y + 2\lambda)$$

$$= y^2 \frac{1}{\lambda} e^{-y/\lambda},$$

we have $c_k E(Y^2) = [-(y^2 + 2\lambda y + 2\lambda^2)e^{-y/\lambda}]_0^{k\lambda} =$
$-(k^2\lambda^2 + 2\lambda^2 k + 2\lambda^2)e^{-k} + 2\lambda^2$. So the result follows. □

Since as $k \to \infty$, $E(Y) \to \lambda$, and $E(Y^2) \to 2\lambda^2$, we have $\text{VAR}(Y) \to \lambda^2$. If $k = 9\log(2) \approx 6.24$, then $E(Y) \approx .998\lambda$, and $E(Y^2) \approx 0.95(2\lambda^2)$.

The Truncated Double Exponential Distribution

Suppose that X is a double exponential $DE(\mu, \lambda)$ random variable. Then $\text{MED}(X) = \mu$ and $\text{MAD}(X) = \log(2)\lambda$. Let $c = k\log(2)$, and let the truncation points $a = \mu - k\text{MAD}(X) = \mu - c\lambda$ and $b = \mu + kMAD(X) = \mu + c\lambda$. Let $X_T(a, b) \equiv Y$ be the truncated double exponential $TDE(\mu, \lambda, a, b)$ random variable. Then for $a \le y \le b$, the pdf of Y is

$$f_Y(y|\mu, \lambda, a, b) = \frac{1}{2\lambda(1 - \exp(-c))} \exp(-|y - \mu|/\lambda).$$

Theorem 8.11. a) $E(Y) = \mu$.

$$b)\ \mathrm{VAR}(Y) = 2\lambda^2 \left[\frac{1 - \frac{1}{2}(c^2 + 2c + 2)e^{-c}}{1 - e^{-c}} \right].$$

Proof. a) follows by symmetry and b) follows from Theorem 8.10 b) since $\mathrm{VAR}(Y) = E[(Y - \mu)^2] = E(W_T^2)$ where W_T is $TEXP(\lambda, b = c\lambda)$. $\square$

As $c \to \infty$, $\mathrm{VAR}(Y) \to 2\lambda^2$. If $k = 9$, then $c = 9\log(2) \approx 6.24$ and $\mathrm{VAR}(Y) \approx 0.95(2\lambda^2)$.

The Truncated Normal Distribution

Now if X is $N(\mu, \sigma^2)$ then let Y be a truncated normal $TN(\mu, \sigma^2, a, b)$ random variable. Then $f_Y(y) = \dfrac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} I_{[a,b]}(y)$ where Φ is the standard normal cdf. The indicator function

$$I_{[a,b]}(y) = 1 \ \text{ if } \ \mathrm{a} \leq \mathrm{y} \leq \mathrm{b}$$

and is zero otherwise. Let ϕ be the standard normal pdf.

Theorem 8.12. $E(Y) = \mu + \left[\dfrac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} \right] \sigma$, and

$$V(Y) = \sigma^2 \left[1 + \frac{(\frac{a-\mu}{\sigma})\phi(\frac{a-\mu}{\sigma}) - (\frac{b-\mu}{\sigma})\phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} \right] - \sigma^2 \left[\frac{\phi(\frac{a-\mu}{\sigma}) - \phi(\frac{b-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})} \right]^2.$$

(See Johnson and Kotz 1970a, p. 83.)

Proof. Let $c = \dfrac{1}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}$.

Then $E(Y) = \int_a^b y f_Y(y) dy$. Hence

$$\frac{1}{c} E(Y) = \int_a^b \frac{y}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy$$

$$= \int_a^b \left(\frac{y-\mu}{\sigma}\right) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy \ + \ \frac{\mu}{\sigma} \frac{1}{\sqrt{2\pi}} \int_a^b \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy$$

$$= \int_a^b \left(\frac{y-\mu}{\sigma}\right) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy \ + \ \mu \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) dy.$$

Note that the integrand of the last integral is the pdf of a $N(\mu, \sigma^2)$ distribution. Let $z = (y - \mu)/\sigma$. Thus $dz = dy/\sigma$, and $E(Y)/c =$

$$\int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \sigma \frac{z}{\sqrt{2\pi}} e^{-z^2/2} dz + \frac{\mu}{c} = \frac{\sigma}{\sqrt{2\pi}} (-e^{-z^2/2}) \Big|_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} + \frac{\mu}{c}.$$

Multiplying both sides by c gives the expectation result.

$$E(Y^2) = \int_a^b y^2 f_Y(y)dy.$$

Hence

$$\frac{1}{c}E(Y^2) = \int_a^b \frac{y^2}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)dy$$

$$= \sigma \int_a^b \left(\frac{y^2}{\sigma^2} - \frac{2\mu y}{\sigma^2} + \frac{\mu^2}{\sigma^2}\right)\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)dy$$

$$+\sigma \int_a^b \frac{2y\mu - \mu^2}{\sigma^2}\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)dy$$

$$= \sigma \int_a^b \left(\frac{y-\mu}{\sigma}\right)^2 \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)dy + 2\frac{\mu}{c}E(Y) - \frac{\mu^2}{c}.$$

Let $z = (y-\mu)/\sigma$. Then $dz = dy/\sigma$, $dy = \sigma dz$, and $y = \sigma z + \mu$. Hence

$$\frac{E(Y^2)}{c} = 2\frac{\mu}{c}E(Y) - \frac{\mu^2}{c} + \sigma \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \sigma \frac{z^2}{\sqrt{2\pi}} e^{-z^2/2}dz.$$

Next integrate by parts with $w = z$ and $dv = ze^{-z^2/2}dz$. Then $E(Y^2)/c =$

$$2\frac{\mu}{c}E(Y) - \frac{\mu^2}{c} + \frac{\sigma^2}{\sqrt{2\pi}}[(-ze^{-z^2/2})|_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} + \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} e^{-z^2/2}dz]$$

$$= 2\frac{\mu}{c}E(Y) - \frac{\mu^2}{c} + \sigma^2\left[\left(\frac{a-\mu}{\sigma}\right)\phi\left(\frac{a-\mu}{\sigma}\right) - \left(\frac{b-\mu}{\sigma}\right)\phi\left(\frac{b-\mu}{\sigma}\right) + \frac{1}{c}\right].$$

Using

$$\text{VAR}(Y) = c\frac{1}{c}E(Y^2) - (E(Y))^2$$

gives the result. □

Theorem 8.13. Let Y be $TN(\mu, \sigma^2, a = \mu - k\sigma, b = \mu + k\sigma)$. Then $E(Y) = \mu$ and $V(Y) = \sigma^2\left[1 - \frac{2k\phi(k)}{2\Phi(k) - 1}\right]$.

Proof. Use the symmetry of ϕ, the fact that $\Phi(-x) = 1 - \Phi(x)$, and the above lemma to get the result. □

Examining $V(Y)$ for several values of k shows that the $TN(\mu, \sigma^2, a = \mu - k\sigma, b = \mu + k\sigma)$ distribution does not change much for $k > 3.0$. See Table 8.2.

The Truncated Cauchy Distribution

Table 8.2 Variances for Several Truncated Normal Distributions

k	$V(Y)$
2.0	$0.774\sigma^2$
2.5	$0.911\sigma^2$
3.0	$0.973\sigma^2$
3.5	$0.994\sigma^2$
4.0	$0.999\sigma^2$

If X is a Cauchy $C(\mu, \sigma)$ random variable, then $\text{MED}(X) = \mu$ and $\text{MAD}(X) = \sigma$. If Y is a truncated Cauchy $TC(\mu, \sigma, \mu - a\sigma, \mu + b\sigma)$ random variable, then

$$f_Y(y) = \frac{1}{\tan^{-1}(b) + \tan^{-1}(a)} \; \frac{1}{\sigma[1 + (\frac{y-\mu}{\sigma})^2]}$$

for $\mu - a\sigma < y < \mu + b\sigma$. For the following theorem, see Johnson and Kotz (1970a, p. 162) and Dahiya, Staneski and Chaganty (2001).

Theorem 8.14. a)

$$E(Y) = \mu + \sigma \left(\frac{\log(1 + b^2) - \log(1 + a^2)}{2[\tan^{-1}(b) + \tan^{-1}(a)]} \right), \quad \text{and}$$

$$V(Y) = \sigma^2 \left[\frac{b + a - \tan^{-1}(b) - \tan^{-1}(a)}{\tan^{-1}(b) + \tan^{-1}(a)} - \left(\frac{\log(1 + b^2) - \log(1 + a^2)}{\tan^{-1}(b) + \tan^{-1}(a)} \right)^2 \right].$$

b) If $a = b$, then $E(Y) = \mu$, and $V(Y) = \sigma^2 \left[\frac{b - \tan^{-1}(b)}{\tan^{-1}(b)} \right]$.

8.1.6 Asymptotic Variances for Trimmed Means

The truncated distributions will be useful for finding the asymptotic variances of trimmed and two stage trimmed means. Assume that Y is from a symmetric location–scale family with parameters μ and σ and that the truncation points are $a = \mu - z\sigma$ and $b = \mu + z\sigma$. Recall that for the trimmed mean T_n,

$$\sqrt{n}(T_n - \mu_T(a, b)) \xrightarrow{D} N \left[0, \frac{\sigma_W^2(a, b)}{(\beta - \alpha)^2} \right].$$

Since the family is symmetric and the truncation is symmetric, $\alpha = F(a) = 1 - \beta$ and $\mu_T(a, b) = \mu$.

Definition 8.21. Let $Y_1, ..., Y_n$ be iid random variables and let $D_n \equiv D_n(Y_1, ..., Y_n)$ be an estimator of a parameter μ_D such that

$$\sqrt{n}(D_n - \mu_D) \xrightarrow{D} N(0, \sigma_D^2).$$

Then the *asymptotic variance* of $\sqrt{n}(D_n - \mu_D)$ is σ_D^2 and the *asymptotic variance* (AV) of D_n is σ_D^2/n. If S_D^2 is a consistent estimator of σ_D^2, then the (asymptotic) *standard error* (SE) of D_n is $S_D/\sqrt{n}$.

Remark 8.2. In the literature, usually either σ_D^2 or σ_D^2/n is called the asymptotic variance of D_n. The parameter σ_D^2 is a function of both the estimator D_n and the underlying distribution F of Y_1. Frequently $n\text{VAR}(D_n)$ converges in distribution to σ_D^2, but not always. See Staudte and Sheather (1990, p. 51) and Lehmann (1999, p. 232).

Example 8.8. If $Y_1, ..., Y_n$ are iid from a distribution with mean μ and variance σ^2, then by the central limit theorem,

$$\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Recall that $\text{VAR}(\overline{Y}_n) = \sigma^2/n = AV(\overline{Y}_n)$ and that the standard error $SE(\overline{Y}_n) = S_n/\sqrt{n}$ where S_n^2 is the sample variance.

Remark 8.3. Returning to the trimmed mean T_n where Y is from a symmetric location–scale family, take $\mu = 0$ since the asymptotic variance does not depend on μ. Then

$$n\ AV(T_n) = \frac{\sigma_W^2(a,b)}{(\beta-\alpha)^2} = \frac{\sigma_T^2(a,b)}{1-2\alpha} + \frac{2\alpha(F^{-1}(\alpha))^2}{(1-2\alpha)^2}.$$

See, for example, Bickel (1965). This formula is useful since the variance of the truncated distribution $\sigma_T^2(a,b)$ has been computed for several distributions in the previous subsection.

Definition 8.22. An estimator D_n is a *location and scale equivariant estimator* if $D_n(\alpha + \beta Y_1, ..., \alpha + \beta Y_n) = \alpha + \beta D_n(Y_1, ..., Y_n)$ where α and β are arbitrary real constants.

Remark 8.4. Many location estimators such as the sample mean, sample median, trimmed mean, metrically trimmed mean, and two stage trimmed means are equivariant. Let $Y_1, ..., Y_n$ be iid from a distribution with cdf $F_Y(y)$ and suppose that D_n is an equivariant estimator of $\mu_D \equiv \mu_D(F_Y) \equiv \mu_D(F_Y(y))$. If $X_i = \alpha + \beta Y_i$ where $\beta \neq 0$, then the cdf of X is $F_X(y) = F_Y((y-\alpha)/\beta)$. Suppose that

$$\mu_D(F_X) \equiv \mu_D[F_Y(\frac{y-\alpha}{\beta})] = \alpha + \beta\mu_D[F_Y(y)]. \tag{8.29}$$

Let $D_n(\boldsymbol{Y}) \equiv D_n(Y_1, ..., Y_n)$. If $\sqrt{n}[D_n(\boldsymbol{Y}) - \mu_D(F_Y(y))] \xrightarrow{D} N(0, \sigma_D^2)$, then

$$\sqrt{n}[D_n(\boldsymbol{X}) - \mu_D(F_X)] = \sqrt{n}[\alpha + \beta D_n(\boldsymbol{Y}) - (\alpha + \beta\mu_D(F_Y))] \xrightarrow{D} N(0, \beta^2\sigma_D^2).$$

This result is especially useful when F is a cdf from a location–scale family with parameters μ and σ. In this case, Equation (8.29) holds when μ_D is the population mean, population median, and the population truncated mean with truncation points $a = \mu - z_1\sigma$ and $b = \mu + z_2\sigma$ (the parameter estimated by trimmed and two stage trimmed means).

Refer to the notation for two stage trimmed means below Theorem 8.3. Then from Theorem 8.5,

$$\sqrt{n}[T_{A,n} - \mu_T(a_o, b_o)] \xrightarrow{D} N\left[0, \frac{\sigma_W^2(a_o, b_o)}{(\beta_o - \alpha_o)^2}\right],$$

and

$$\sqrt{n}[T_{S,n} - \mu_T(a_M, b_M)] \xrightarrow{D} N\left[0, \frac{\sigma_W^2(a_M, b_M)}{(\beta_M - \alpha_M)^2}\right].$$

If the distribution of Y is symmetric then $T_{A,n}$ and $T_{S,n}$ are asymptotically equivalent. It is important to note that no knowledge of the unknown distribution and parameters is needed to compute the two stage trimmed means and their standard errors.

The next three theorems find the asymptotic variance for trimmed and two stage trimmed means when the underlying distribution is normal, double exponential and Cauchy, respectively. Assume $a = \text{MED}(Y) - k\text{MAD}(Y)$ and $b = \text{MED}(Y) + k\text{MAD}(Y)$.

Theorem 8.15. Suppose that Y comes from a normal $N(\mu, \sigma^2)$ distribution. Let $\Phi(x)$ be the cdf and let $\phi(x)$ be the density of the standard normal. Then for the α trimmed mean,

$$n\ AV = \left(\frac{1 - \frac{2z\phi(z)}{2\Phi(z)-1}}{1-2\alpha} + \frac{2\alpha z^2}{(1-2\alpha)^2}\right)\sigma^2 \tag{8.30}$$

where $\alpha = \Phi(-z)$, and $z = k\Phi^{-1}(0.75)$. For the two stage estimators, round 100α up to the nearest integer J. Then use $\alpha_J = J/100$ and $z_J = -\Phi^{-1}(\alpha_J)$ in Equation (8.30).

Proof. If Y follows the normal $N(\mu, \sigma^2)$ distribution, then $a = \mu - k\text{MAD}(Y)$ and $b = \mu + k\text{MAD}(Y)$ where $\text{MAD}(Y) = \Phi^{-1}(0.75)\sigma$. It is enough to consider the standard N(0,1) distribution since $n\ AV(T_n, N(\mu, \sigma^2)) = \sigma^2\ n\ AV(T_n, N(0,1))$. If $a = -z$ and $b = z$, then by Theorem 8.13,

$$\sigma_T^2(a, b) = 1 - \frac{2z\phi(z)}{2\Phi(z) - 1}.$$

Use Remark 8.3 with $z = k\Phi^{-1}(0.75)$, and $\alpha = \Phi(-z)$ to get Equation (8.30).

Theorem 8.16. Suppose that Y comes from a double exponential DE(0,1) distribution. Then for the α trimmed mean,

$$n\ AV = \frac{\frac{2-(z^2+2z+2)e^{-z}}{1-e^{-z}}}{1-2\alpha} + \frac{2\alpha z^2}{(1-2\alpha)^2} \tag{8.31}$$

where $z = k\log(2)$ and $\alpha = 0.5\exp(-z)$. For the two stage estimators, round 100α up to the nearest integer J. Then use $\alpha_J = J/100$ and let $z_J = -\log(2\alpha_J)$.

Proof Sketch. For the $DE(0,1)$ distribution, $\mathrm{MAD}(Y) = \log(2)$. If the DE(0,1) distribution is truncated at $-z$ and z, then use Remark 8.3 with

$$\sigma_T^2(-z,z) = \frac{2-(z^2+2z+2)e^{-z}}{1-e^{-z}}.$$

Theorem 8.17. Suppose that Y comes from a Cauchy (0,1) distribution. Then for the α trimmed mean,

$$n\ AV = \frac{z-\tan^{-1}(z)}{(1-2\alpha)\tan^{-1}(z)} + \frac{2\alpha(\tan[\pi(\alpha-\frac{1}{2})])^2}{(1-2\alpha)^2} \tag{8.32}$$

where $z = k$ and

$$\alpha = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(z).$$

For the two stage estimators, round 100α up to the nearest integer J. Then use $\alpha_J = J/100$ and let $z_J = \tan[\pi(\alpha_J - 0.5)]$.

Proof Sketch. For the $C(0,1)$ distribution, $\mathrm{MAD}(Y) = 1$. If the C(0,1) distribution is truncated at $-z$ and z, then use Remark 8.3 with

$$\sigma_T^2(-z,z) = \frac{z-\tan^{-1}(z)}{\tan^{-1}(z)}.$$

Next we give a theorem for the metrically trimmed mean M_n. Lopuhaä (1999) shows the following result. Suppose $(\hat{\boldsymbol{\mu}}_n, \boldsymbol{C}_n)$ is an estimator of multivariate location and dispersion. Suppose that the iid data follow an elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution. Let $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ be the classical estimator applied to the set J of cases with squared Mahalanobis distances $D_i^2(\hat{\boldsymbol{\mu}}_n, \boldsymbol{C}_n) \le k^2$. Under regularity conditions, if $(\hat{\boldsymbol{\mu}}_n, \boldsymbol{C}_n) \xrightarrow{P} (\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^δ where $0 < \delta \le 0.5$, then $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J) \xrightarrow{P} (\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ with the same rate n^δ where $s > 0$ and $d > 0$ are some constants. See Section 8.2 for discussion of the above quantities.

In the univariate setting with $p = 1$, let $\hat{\theta}_n = \hat{\mu}_n$ and let $D_n^2 = \boldsymbol{C}_n$ where D_n is an estimator of scale. Suppose the classical estimator $(\overline{Y}_J, S_J^2) \equiv (\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ is applied to the set J of cases with $\hat{\theta}_n - kD_n \leq Y_i \leq \hat{\theta}_n + kD_n$. Hence $\overline{Y}_J$ is the metrically trimmed mean M_n with $k_1 = k_2 \equiv k$. See Definition 8.14.

The population quantity estimated by $(\overline{Y}_J, S_J^2)$ is the truncated mean and variance $(\mu_T(a,b), \sigma_T^2(a,b))$ of Definition 8.18 where $\hat{\theta}_n - kD_n \xrightarrow{P} a$ and $\hat{\theta}_n + kD_n \xrightarrow{P} b$. In the theorem below, the pdf corresponds to an elliptically contoured distribution with $p = 1$ and $\boldsymbol{\Sigma} = \tau^2$. Each pdf corresponds to a location scale family with location parameter μ and scale parameter τ. Note that $(\hat{\theta}_n, D_n) = (\text{MED}(n), \text{MAD}(n))$ results in a $\sqrt{n}$ consistent estimator (M_n, S_J^2).

Assumption E1: Suppose $Y_1, ..., Y_n$ are iid from an $EC_1(\mu, \tau^2, g)$ distribution with pdf

$$f(y) = \frac{c}{\tau} \; g\left[\left(\frac{y-\mu}{\tau}\right)^2\right]$$

where g is continuously differentiable with finite 4th moment $\int y^4 g(y^2)dy < \infty$, $c > 0$ is some constant, $\tau > 0$ where y and μ are real.

Theorem 8.18. Let M_n be the metrically trimmed mean with $k_1 = k_2 \equiv k$. Assume (E1) holds. If $(\hat{\theta}_n, D_n^2) \xrightarrow{P} (\mu, s\tau^2)$ with rate n^δ for some constant $s > 0$ where $0 < \delta \leq 0.5$, then $(M_n, S_J^2) \xrightarrow{P} (\mu, \sigma_T^2(a,b))$ with the same rate n^δ.

Proof. The result is a special case of Lopuhaä (1999) which shows that $(M_n, S_J^2) \xrightarrow{P} (\mu, d\tau^2)$ with rate n^δ. Since $k_1 = k_2 = k$, $d\tau^2 = \sigma_T^2(a,b)$. □

Note that the classical estimator applied to the set $\tilde{J}$ of cases Y_i between a and b is a $\sqrt{n}$ consistent estimator of $(\mu_T(a,b), \sigma_T^2(a,b))$. Consider the set J of cases with $\text{MED}(n) - k\text{MAD}(n) \leq Y_i \leq \text{MED}(N) + k\text{MAD}(n)$. By Theorem 8.4 sets $\tilde{J}$ and J differ primarily in neighborhoods of a and b. This result leads to the following conjecture.

Conjecture 8.1. If $Y_1, ..., Y_n$ are iid from a distribution with a pdf that is positive in neighborhoods of a and b, and if $\hat{\theta}_n - k_1 D_n \xrightarrow{P} a$ and $\hat{\theta}_n + k_2 D_n \xrightarrow{P} b$ at rate $n^{0.5}$, then $(M_n, S_J^2) \xrightarrow{P} (\mu_T(a,b), \sigma_T^2(a,b))$ with rate $n^{0.5}$.

8.2 The Multivariate Location and Dispersion Model

The multivariate location and dispersion (MLD) model is a special case of the multivariate linear model, just like the location model is a special case of the

multiple linear regression model. Robust estimators of multivariate location and dispersion are useful for detecting outliers in the predictor variables and for developing an outlier resistant multiple linear regression estimator.

The practical, highly outlier resistant, $\sqrt{n}$ consistent FCH, RFCH, and RMVN estimators of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ are developed along with proofs. The RFCH and RMVN estimators are reweighted versions of the FCH estimator. Olive (2017b) shows why competing "robust estimators" fail to work, are impractical, or are not yet backed by theory. The RMVN and RFCH sets are defined and will be used for outlier detection and to create practical robust methods of multiple linear regression and multivariate linear regression. Many more applications are given in Olive (2017b).

Warning: This section contains many acronyms, abbreviations, and estimator names such as FCH, RFCH, and RMVN. Often the acronyms start with the added letter A, C, F, or R: A stands for *algorithm*, C for *concentration*, F for estimators that use a *fixed* number of trial fits, and R for *reweighted*.

Definition 8.23. The multivariate location and dispersion model is

$$\boldsymbol{Y}_i = \boldsymbol{\mu} + \boldsymbol{e}_i, \quad i = 1, \ldots, n \tag{8.33}$$

where $\boldsymbol{e}_1, ..., \boldsymbol{e}_n$ are $p \times 1$ error random vectors, often iid with zero mean and covariance matrix $\text{Cov}(\boldsymbol{e}) = \text{Cov}(\boldsymbol{Y}) = \boldsymbol{\Sigma}_{\boldsymbol{Y}} = \boldsymbol{\Sigma}_{\boldsymbol{e}}$.

Note that the location model is a special case of the MLD model with $p = 1$. If $E(\boldsymbol{e}) = \boldsymbol{0}$, then $E(\boldsymbol{Y}) = \boldsymbol{\mu}$. A $p \times p$ dispersion matrix is a symmetric matrix that measures the spread of a random vector. Covariance and correlation matrices are dispersion matrices. One way to get a robust estimator of multivariate location is to stack the marginal estimators of location into a vector. The coordinatewise median $\text{MED}(\boldsymbol{W})$ is an example. The sample mean $\overline{\boldsymbol{x}}$ also stacks the marginal estimators into a vector, but is not outlier resistant.

Let $\boldsymbol{\mu}$ be a $p \times 1$ location vector and $\boldsymbol{\Sigma}$ a $p \times p$ symmetric dispersion matrix. Because of symmetry, the first row of $\boldsymbol{\Sigma}$ has p distinct unknown parameters, the second row has $p-1$ distinct unknown parameters, the third row has $p-2$ distinct unknown parameters, ..., and the pth row has one distinct unknown parameter for a total of $1+2+\cdots+p = p(p+1)/2$ unknown parameters. Since $\boldsymbol{\mu}$ has p unknown parameters, an estimator $(T, \boldsymbol{C})$ of multivariate location and dispersion, needs to estimate $p(p+3)/2$ unknown parameters when there are p random variables.

The sample covariance or sample correlation matrices estimate these parameters very efficiently since $\boldsymbol{\Sigma} = (\sigma_{ij})$ where σ_{ij} is a population covariance or correlation. These quantities can be estimated with the sample covariance or correlation taking two variables X_i and X_j at a time. Note that there are $p(p+1)/2$ pairs that can be chosen from p random variables $X_1, ..., X_p$. See

Definition 4.5 for the sample mean $\overline{\boldsymbol{x}}$, the sample covariance matrix $\boldsymbol{S}$, and the sample correlation matrix $\boldsymbol{R}$.

Rule of thumb 8.1. For the classical estimators of multivariate location and dispersion, $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ or $(\overline{\boldsymbol{z}} = \boldsymbol{0}, \boldsymbol{R})$, we want $n \geq 10p$. We want $n \geq 20p$ for the robust MLD estimators (FCH, RFCH, or RMVN) described later in this section.

8.2.1 Affine Equivariance

Before defining an important equivariance property, some notation is needed. Assume that the data is collected in an $n \times p$ data matrix $\boldsymbol{W}$. Let $\boldsymbol{B} = \boldsymbol{1}\boldsymbol{b}^T$ where $\boldsymbol{1}$ is an $n \times 1$ vector of ones and $\boldsymbol{b}$ is a $p \times 1$ constant vector. Hence the ith row of $\boldsymbol{B}$ is $\boldsymbol{b}_i^T \equiv \boldsymbol{b}^T$ for $i = 1, ..., n$. For such a matrix $\boldsymbol{B}$, consider the affine transformation $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}$ where $\boldsymbol{A}$ is any nonsingular $p \times p$ matrix. An affine transformation changes $\boldsymbol{x}_i$ to $\boldsymbol{z}_i = \boldsymbol{A}\boldsymbol{x}_i + \boldsymbol{b}$ for $i = 1, ..., n$, and affine equivariant multivariate location and dispersion estimators change in natural ways.

Definition 8.24. The multivariate location and dispersion estimator $(T, \boldsymbol{C})$ is *affine equivariant* if

$$T(\boldsymbol{Z}) = T(\boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}) = \boldsymbol{A}T(\boldsymbol{W}) + \boldsymbol{b}, \tag{8.34}$$

$$\text{and} \quad \boldsymbol{C}(\boldsymbol{Z}) = \boldsymbol{C}(\boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}) = \boldsymbol{A}\boldsymbol{C}(\boldsymbol{W})\boldsymbol{A}^T. \tag{8.35}$$

The following theorem shows that the Mahalanobis distances are invariant under affine transformations. See Rousseeuw and Leroy (1987, pp. 252-262) for similar results. Thus if $(T, \boldsymbol{C})$ is affine equivariant, so is $(T, D^2_{(c_n)}(T, \boldsymbol{C})\ \boldsymbol{C})$ where $D^2_{(j)}(T, \boldsymbol{C})$ is the jth order statistic of the D_i^2.

Theorem 8.19. If $(T, \boldsymbol{C})$ is affine equivariant, then

$$D_i^2(\boldsymbol{W}) \equiv D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = D_i^2(T(\boldsymbol{Z}), \boldsymbol{C}(\boldsymbol{Z})) \equiv D_i^2(\boldsymbol{Z}). \tag{8.36}$$

Proof. Since $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{A}^T + \boldsymbol{B}$ has ith row $\boldsymbol{z}_i^T = \boldsymbol{x}_i^T\boldsymbol{A}^T + \boldsymbol{b}^T$,

$$D_i^2(\boldsymbol{Z}) = [\boldsymbol{z}_i - T(\boldsymbol{Z})]^T\boldsymbol{C}^{-1}(\boldsymbol{Z})[\boldsymbol{z}_i - T(\boldsymbol{Z})]$$

$$= [\boldsymbol{A}(\boldsymbol{x}_i - T(\boldsymbol{W}))]^T[\boldsymbol{A}\boldsymbol{C}(\boldsymbol{W})\boldsymbol{A}^T]^{-1}[\boldsymbol{A}(\boldsymbol{x}_i - T(\boldsymbol{W}))]$$

$$= [\boldsymbol{x}_i - T(\boldsymbol{W})]^T\boldsymbol{C}^{-1}(\boldsymbol{W})[\boldsymbol{x}_i - T(\boldsymbol{W})] = D_i^2(\boldsymbol{W}). \ \square$$

Definition 8.25. For MLD, an *elemental set* $J = \{m_1, ..., m_{p+1}\}$ is a set of $p+1$ cases drawn without replacement from the data set of n cases. The elemental fit $(T_J, \boldsymbol{C}_J) = (\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ is the sample mean and the sample covariance matrix computed from the cases in the elemental set.

If the data are iid, then the elemental fit gives an unbiased but inconsistent estimator of $(E(\boldsymbol{x}), \mathrm{Cov}(\boldsymbol{x}))$. Note that the elemental fit uses the smallest sample size $p+1$ such that $\boldsymbol{S}_J$ is nonsingular if the data are in "general position" defined in Definition 8.27.

8.2.2 Breakdown

This subsection gives a standard definition of breakdown for estimators of multivariate location and dispersion. The following notation will be useful. Let $\boldsymbol{W}$ denote the $n \times p$ data matrix with ith row $\boldsymbol{x}_i^T$ corresponding to the ith case. Let $\boldsymbol{w}_1, ...\boldsymbol{w}_n$ be the contaminated data after d_n of the $\boldsymbol{x}_i$ have been replaced by arbitrarily bad contaminated cases. Let $\boldsymbol{W}_d^n$ denote the $n \times p$ data matrix with ith row $\boldsymbol{w}_i^T$. Then the contamination fraction is $\gamma_n = d_n/n$. Let $(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}))$ denote an estimator of multivariate location and dispersion where the $p \times 1$ vector $T(\boldsymbol{W})$ is an estimator of location and the $p \times p$ symmetric positive semidefinite matrix $\boldsymbol{C}(\boldsymbol{W})$ is an estimator of dispersion.

Theorem 8.20. Let $\boldsymbol{B} > 0$ be a $p \times p$ symmetric matrix with eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), ..., (\lambda_p, \boldsymbol{e}_p)$ where $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p > 0$ and the orthonormal eigenvectors satisfy $\boldsymbol{e}_i^T \boldsymbol{e}_i = 1$ while $\boldsymbol{e}_i^T \boldsymbol{e}_j = 0$ for $i \neq j$. Let $\boldsymbol{d}$ be a given $p \times 1$ vector and let $\boldsymbol{a}$ be an arbitrary nonzero $p \times 1$ vector.

a) $\max_{\boldsymbol{a} \neq \boldsymbol{0}} \dfrac{\boldsymbol{a}^T \boldsymbol{d}\boldsymbol{d}^T \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{B}\boldsymbol{a}} = \boldsymbol{d}^T \boldsymbol{B}^{-1} \boldsymbol{d}$ where the max is attained for $\boldsymbol{a} = c\boldsymbol{B}^{-1}\boldsymbol{d}$ for any constant $c \neq 0$. Note that the numerator $= (\boldsymbol{a}^T \boldsymbol{d})^2$.

b) $\max_{\boldsymbol{a} \neq \boldsymbol{0}} \dfrac{\boldsymbol{a}^T \boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{a}} = \max_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T \boldsymbol{B}\boldsymbol{a} = \lambda_1$ where the max is attained for $\boldsymbol{a} = \boldsymbol{e}_1$.

c) $\min_{\boldsymbol{a} \neq \boldsymbol{0}} \dfrac{\boldsymbol{a}^T \boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{a}} = \min_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T \boldsymbol{B}\boldsymbol{a} = \lambda_p$ where the min is attained for $\boldsymbol{a} = \boldsymbol{e}_p$.

d) $\max_{\boldsymbol{a} \perp \boldsymbol{e}_1, ..., \boldsymbol{e}_k} \dfrac{\boldsymbol{a}^T \boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{a}} = \max_{\|\boldsymbol{a}\|=1, \boldsymbol{a} \perp \boldsymbol{e}_1, ..., \boldsymbol{e}_k} \boldsymbol{a}^T \boldsymbol{B}\boldsymbol{a} = \lambda_{k+1}$ where the max is attained for $\boldsymbol{a} = \boldsymbol{e}_{k+1}$ for $k = 1, 2, ..., p-1$.

e) Let $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ be the observed sample mean and sample covariance matrix where $\boldsymbol{S} > 0$. Then $\max_{\boldsymbol{a} \neq \boldsymbol{0}} \dfrac{n\boldsymbol{a}^T(\overline{\boldsymbol{x}} - \boldsymbol{\mu})(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{S}\boldsymbol{a}} = n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}) = T^2$ where the max is attained for $\boldsymbol{a} = c\boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu})$ for any constant $c \neq 0$.

f) Let $\boldsymbol{A}$ be a $p \times p$ symmetric matrix. Let $\boldsymbol{C} > 0$ be a $p \times p$ symmetric matrix. Then $\max_{\boldsymbol{a} \neq \boldsymbol{0}} \dfrac{\boldsymbol{a}^T \boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T \boldsymbol{C}\boldsymbol{a}} = \lambda_1(\boldsymbol{C}^{-1}\boldsymbol{A})$, the largest eigenvalue of $\boldsymbol{C}^{-1}\boldsymbol{A}$. The

value of $\boldsymbol{a}$ that achieves the max is the eigenvector $\boldsymbol{g}_1$ of $\boldsymbol{C}^{-1}\boldsymbol{A}$ corresponding to $\lambda_1(\boldsymbol{C}^{-1}\boldsymbol{A})$. Similarly $\min_{\boldsymbol{a}\neq\boldsymbol{0}} \frac{\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{a}}{\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a}} = \lambda_p(\boldsymbol{C}^{-1}\boldsymbol{A})$, the smallest eigenvalue of $\boldsymbol{C}^{-1}\boldsymbol{A}$. The value of $\boldsymbol{a}$ that achieves the min is the eigenvector $\boldsymbol{g}_p$ of $\boldsymbol{C}^{-1}\boldsymbol{A}$ corresponding to $\lambda_p(\boldsymbol{C}^{-1}\boldsymbol{A})$.

Proof Sketch. See Johnson and Wichern (1988, pp. 64-65, 184). For a), note that $\text{rank}(\boldsymbol{C}^{-1}\boldsymbol{A}) = 1$, where $\boldsymbol{C} = \boldsymbol{B}$ and $\boldsymbol{A} = \boldsymbol{d}\boldsymbol{d}^T$, since $\text{rank}(\boldsymbol{C}^{-1}\boldsymbol{A})$ $= \text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{d}) = 1$. Hence $\boldsymbol{C}^{-1}\boldsymbol{A}$ has one nonzero eigenvalue eigenvector pair $(\lambda_1, \boldsymbol{g}_1)$. Since

$$(\lambda_1 = \boldsymbol{d}^T\boldsymbol{B}^{-1}\boldsymbol{d}, \boldsymbol{g}_1 = \boldsymbol{B}^{-1}\boldsymbol{d})$$

is a nonzero eigenvalue eigenvector pair for $\boldsymbol{C}^{-1}\boldsymbol{A}$, and $\lambda_1 > 0$, the result follows by f).

Note that b) and c) are special cases of f) with $\boldsymbol{A} = \boldsymbol{B}$ and $\boldsymbol{C} = \boldsymbol{I}$.

Note that e) is a special case of a) with $\boldsymbol{d} = (\overline{\boldsymbol{x}} - \boldsymbol{\mu})$ and $\boldsymbol{B} = \boldsymbol{S}$.

(Also note that $(\lambda_1 = (\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T\boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}), \boldsymbol{g}_1 = \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}))$ is a nonzero eigenvalue eigenvector pair for the rank 1 matrix $\boldsymbol{C}^{-1}\boldsymbol{A}$ where $\boldsymbol{C} = \boldsymbol{S}$ and $\boldsymbol{A} = (\overline{\boldsymbol{x}} - \boldsymbol{\mu})(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T$.)

For f), see Mardia et al. (1979, p. 480). □

From Theorem 8.20, if $\boldsymbol{C}(\boldsymbol{W}_d^n) > 0$, then $\max_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T\boldsymbol{C}(\boldsymbol{W}_d^n)\boldsymbol{a} = \lambda_1$ and $\min_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^T\boldsymbol{C}(\boldsymbol{W}_d^n)\boldsymbol{a} = \lambda_p$. A high breakdown dispersion estimator $\boldsymbol{C}$ is positive definite if the amount of contamination is less than the breakdown value. Since $\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{a} = \sum_{i=1}^p \sum_{j=1}^p c_{ij}a_ia_j$, the largest eigenvalue λ_1 is bounded as $\boldsymbol{W}_d^n$ varies iff $\boldsymbol{C}(\boldsymbol{W}_d^n)$ is bounded as $\boldsymbol{W}_d^n$ varies.

Definition 8.26. The *breakdown value* of the multivariate location estimator T at $\boldsymbol{W}$ is

$$B(T, \boldsymbol{W}) = \min\left\{\frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} \|T(\boldsymbol{W}_d^n)\| = \infty\right\}$$

where the supremum is over all possible corrupted samples $\boldsymbol{W}_d^n$ and $1 \leq d_n \leq n$. Let $\lambda_1(\boldsymbol{C}(\boldsymbol{W})) \geq \cdots \geq \lambda_p(\boldsymbol{C}(\boldsymbol{W})) \geq 0$ denote the eigenvalues of the dispersion estimator applied to data $\boldsymbol{W}$. The estimator $\boldsymbol{C}$ breaks down if the smallest eigenvalue can be driven to zero or if the largest eigenvalue can be driven to ∞. Hence the *breakdown value* of the dispersion estimator is

$$B(\boldsymbol{C}, \boldsymbol{W}) = \min\left\{\frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} \max\left[\frac{1}{\lambda_p(\boldsymbol{C}(\boldsymbol{W}_d^n))}, \lambda_1(\boldsymbol{C}(\boldsymbol{W}_d^n))\right] = \infty\right\}.$$

Definition 8.27. Let γ_n be the breakdown value of $(T, \boldsymbol{C})$. *High breakdown (HB) statistics* have $\gamma_n \to 0.5$ as $n \to \infty$ if the (uncontaminated) clean

data are in *general position*: no more than p points of the clean data lie on any $(p-1)$-dimensional hyperplane. Estimators are *zero breakdown* if $\gamma_n \to 0$ and *positive breakdown* if $\gamma_n \to \gamma > 0$ as $n \to \infty$.

Note that if the number of outliers is less than the number needed to cause breakdown, then $\|T\|$ is bounded and the eigenvalues are bounded away from 0 and ∞. Also, the bounds do not depend on the outliers but do depend on the estimator $(T, \boldsymbol{C})$ and on the clean data $\boldsymbol{W}$.

The following result shows that a multivariate location estimator T basically "breaks down" if the d outliers can make the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|)$ arbitrarily large where $\boldsymbol{w}_i^T$ is the ith row of $\boldsymbol{W}_d^n$. Thus a multivariate location estimator T will not break down if T can not be driven out of some ball of (possibly huge) radius r about the origin. For an affine equivariant estimator, the largest possible breakdown value is $n/2$ or $(n+1)/2$ for n even or odd, respectively. Hence in the proof of the following result, we could replace $d_n < d_T$ by $d_n < \min(n/2, d_T)$.

Theorem 8.21. Fix n. If nonequivariant estimators (that may have a breakdown value of greater than 1/2) are excluded, then a multivariate location estimator has a breakdown value of d_T/n iff $d_T = d_{T,n}$ is the smallest number of arbitrarily bad cases that can make the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|)$ arbitrarily large.

Proof. Suppose the multivariate location estimator T satisfies $\|T(\boldsymbol{W}_d^n)\| \leq M$ for some constant M if $d_n < d_T$. Note that for a fixed data set $\boldsymbol{W}_d^n$ with ith row $\boldsymbol{w}_i$, the median Euclidean distance $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|) \leq \max_{i=1,\ldots,n} \|\boldsymbol{x}_i - T(\boldsymbol{W}_d^n)\| \leq \max_{i=1,\ldots,n} \|\boldsymbol{x}_i\| + M$ if $d_n < d_T$. Similarly, suppose $\text{MED}(\|\boldsymbol{w}_i - T(\boldsymbol{W}_d^n)\|) \leq M$ for some constant M if $d_n < d_T$, then $\|T(\boldsymbol{W}_d^n)\|$ is bounded if $d_n < d_T$. □

Since the coordinatewise median $\text{MED}(\boldsymbol{W})$ is a HB estimator of multivariate location, it is also true that a multivariate location estimator T will not break down if T can not be driven out of some ball of radius r about $\text{MED}(\boldsymbol{W})$. Hence $(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ is a HB estimator of MLD.

If a high breakdown estimator $(T, \boldsymbol{C}) \equiv (T(\boldsymbol{W}_d^n), \boldsymbol{C}(\boldsymbol{W}_d^n))$ is evaluated on the contaminated data $\boldsymbol{W}_d^n$, then the location estimator T is contained in some ball about the origin of radius r, and $0 < a < \lambda_p \leq \lambda_1 < b$ where the constants a, r, and b depend on the clean data and $(T, \boldsymbol{C})$, but not on $\boldsymbol{W}_d^n$ if the number of outliers d_n satisfies $0 \leq d_n < n\gamma_n < n/2$ where the breakdown value $\gamma_n \to 0.5$ as $n \to \infty$.

The following theorem will be used to show that if the classical estimator $(\overline{\boldsymbol{X}}_B, \boldsymbol{S}_B)$ is applied to $c_n \approx n/2$ cases contained in a ball about the origin of radius r where r depends on the clean data but not on $\boldsymbol{W}_d^n$, then $(\overline{\boldsymbol{X}}_B, \boldsymbol{S}_B)$ is a high breakdown estimator.

Theorem 8.22. If the classical estimator $(\overline{\boldsymbol{X}}_B, \boldsymbol{S}_B)$ is applied to c_n cases that are contained in some bounded region where $p+1 \leq c_n \leq n$, then the maximum eigenvalue λ_1 of $\boldsymbol{S}_B$ is bounded.

Proof. The largest eigenvalue of a $p \times p$ matrix $\boldsymbol{A}$ is bounded above by $p \max |a_{i,j}|$ where $a_{i,j}$ is the (i,j) entry of $\boldsymbol{A}$. See Datta (1995, p. 403). Denote the c_n cases by $\boldsymbol{z}_1, ..., \boldsymbol{z}_{c_n}$. Then the (i,j)th element $a_{i,j}$ of $\boldsymbol{A} = \boldsymbol{S}_B$ is

$$a_{i,j} = \frac{1}{c_n - 1} \sum_{m=1}^{c_n} (z_{i,m} - \overline{z}_i)(z_{j,m} - \overline{z}_j).$$

Hence the maximum eigenvalue λ_1 is bounded. □

The determinant $det(\boldsymbol{S}) = |\boldsymbol{S}|$ of $\boldsymbol{S}$ is known as the *generalized sample variance*. Consider the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq D^2_{(c_n)}\} \tag{8.37}$$

where $D^2_{(c_n)}$ is the c_nth smallest squared Mahalanobis distance based on $(T, \boldsymbol{C})$. This hyperellipsoid contains the c_n cases with the smallest D^2_i. Suppose $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}_M, b\, \boldsymbol{S}_M)$ is the sample mean and scaled sample covariance matrix applied to some subset of the data where $b > 0$. The classical, RFCH, and RMVN estimators satisfy this assumption. For $h > 0$, the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}$$

has volume equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p \sqrt{det(\boldsymbol{C})} = \frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p b^{p/2} \sqrt{det(\boldsymbol{S}_M)}.$$

If $h^2 = D^2_{(c_n)}$, then the volume is proportional to the square root of the determinant $|\boldsymbol{S}_M|^{1/2}$, and this volume will be positive unless extreme degeneracy is present among the c_n cases. See Johnson and Wichern (1988, pp. 103-104).

8.2.3 The Concentration Algorithm

Concentration algorithms are widely used since impractical brand name estimators, such as the MCD estimator given in Definition 8.28, take too long to compute. The concentration algorithm, defined in Definition 8.29, use K starts and attractors. A *start* is an initial estimator, and an *attractor* is an estimator obtained by refining the start. For example, let the start be the classical estimator $(\overline{\boldsymbol{x}}, \boldsymbol{S})$. Then the attractor could be the classical estimator $(T_1, \boldsymbol{C}_1)$ applied to the half set of cases with the smallest Mahalanobis

distances. This concentration algorithm uses one concentration step, but the process could be iterated for k concentration steps, producing an estimator $(T_k, \boldsymbol{C}_k)$

If more than one attractor is used, then some criterion is needed to select which of the K attractors is to be used in the final estimator. If each attractor $(T_{k,j}, \boldsymbol{C}_{k,j})$ is the classical estimator applied to $c_n \approx n/2$ cases, then the minimum covariance determinant (MCD) criterion is often used: choose the attractor that has the minimum value of $det(\boldsymbol{C}_{k,j})$ where $j = 1, ..., K$.

The remainder of this section will explain the concentration algorithm, explain why the MCD criterion is useful but can be improved, provide some theory for practical robust multivariate location and dispersion estimators, and show how the set of cases used to compute the recommended RMVN or RFCH estimator can be used to create outlier resistant regression estimators. The RMVN and RFCH estimators are reweighted versions of the practical FCH estimator, given in Definition 8.32.

Definition 8.28. Consider the subset J_o of $c_n \approx n/2$ observations whose sample covariance matrix has the lowest determinant among all $C(n, c_n)$ subsets of size c_n. Let T_{MCD} and $\boldsymbol{C}_{MCD}$ denote the sample mean and sample covariance matrix of the c_n cases in J_o. Then the *minimum covariance determinant* MCD(c_n) estimator is $(T_{MCD}(\boldsymbol{W}), \boldsymbol{C}_{MCD}(\boldsymbol{W}))$.

Here

$$C(n,i) = \binom{n}{i} = \frac{n!}{i!\ \ (n-i)!}$$

is the binomial coefficient.

The MCD estimator is a high breakdown (HB) estimator, and the value $c_n = \lfloor (n + p + 1)/2 \rfloor$ is often used as the default. The MCD estimator is the pair

$$(\hat{\beta}_{LTS}, Q_{LTS}(\hat{\beta}_{LTS})/(c_n - 1))$$

in the location model where LTS stands for the least trimmed sum of squares estimator. See Definition 8.10. The population analog of the MCD estimator is closely related to the hyperellipsoid of highest concentration that contains $c_n/n \approx$ half of the mass. The MCD estimator is a $\sqrt{n}$ consistent HB asymptotically normal estimator for $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ where a_{MCD} is some positive constant when the data $\boldsymbol{x}_i$ are iid from a large class of distributions. See Cator and Lopuhaä (2010, 2012) who extended some results of Butler et al. (1993).

Computing robust covariance estimators can be very expensive. For example, to compute the exact MCD(c_n) estimator (T_{MCD}, C_{MCD}), we need to consider the $C(n, c_n)$ subsets of size c_n. Woodruff and Rocke (1994, p. 893) noted that if 1 billion subsets of size 101 could be evaluated per second, it would require 10^{33} millenia to search through all $C(200, 101)$ subsets if the sample size $n = 200$. See Section 8.8 for the MCD complexity.

Hence algorithm estimators will be used to approximate the robust estimators. Elemental sets are the key ingredient for both *basic resampling* and *concentration* algorithms.

Definition 8.29. Suppose that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are $p \times 1$ vectors of observed data. For the multivariate location and dispersion model, an *elemental set* J is a set of $p+1$ cases. An elemental start is the sample mean and sample covariance matrix of the data corresponding to J. In a *concentration algorithm,* let $(T_{-1,j}, \boldsymbol{C}_{-1,j})$ be the jth start (not necessarily elemental) and compute all n Mahalanobis distances $D_i(T_{-1,j}, \boldsymbol{C}_{-1,j})$. At the next iteration, the classical estimator $(T_{0,j}, \boldsymbol{C}_{0,j}) = (\overline{\boldsymbol{x}}_{0,j}, \boldsymbol{S}_{0,j})$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest distances. This iteration can be continued for k *concentration steps* resulting in the sequence of estimators $(T_{-1,j}, \boldsymbol{C}_{-1,j}), (T_{0,j}, \boldsymbol{C}_{0,j}), ..., (T_{k,j}, \boldsymbol{C}_{k,j})$. The result of the iteration $(T_{k,j}, \boldsymbol{C}_{k,j})$ is called the jth *attractor*. If K_n starts are used, then $j = 1, ..., K_n$. The *concentration attractor*, $(T_A, \boldsymbol{C}_A)$, is the attractor chosen by the algorithm. The attractor is used to obtain the final estimator. A common choice is the attractor that has the smallest determinant $det(\boldsymbol{C}_{k,j})$. The *basic resampling algorithm* estimator is a special case where $k = -1$ so that the attractor is the start: $(\overline{\boldsymbol{x}}_{k,j}, \boldsymbol{S}_{k,j}) = (\overline{\boldsymbol{x}}_{-1,j}, \boldsymbol{S}_{-1,j})$.

This concentration algorithm is a simplified version of the algorithms given by Rousseeuw and Van Driessen (1999) and Hawkins and Olive (1999a). Using $k = 10$ concentration steps often works well. The following proposition is useful and shows that $det(\boldsymbol{S}_{0,j})$ tends to be greater than the determinant of the attractor $det(\boldsymbol{S}_{k,j})$.

Theorem 8.23: Rousseeuw and Van Driessen (1999, p. 214). Suppose that the classical estimator $(\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$ is computed from c_n cases and that the n Mahalanobis distances $D_i \equiv D_i(\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$ are computed. If $(\overline{\boldsymbol{x}}_{t+1,j}, \boldsymbol{S}_{t+1,j})$ is the classical estimator computed from the c_n cases with the smallest Mahalanobis distances D_i, then $det(\boldsymbol{S}_{t+1,j}) \leq det(\boldsymbol{S}_{t,j})$ with equality iff $(\overline{\boldsymbol{x}}_{t+1,j}, \boldsymbol{S}_{t+1,j}) = (\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$.

Starts that use a consistent initial estimator could be used. K_n is the number of starts and k is the number of concentration steps used in the algorithm. Suppose the algorithm estimator uses some criterion to choose an attractor as the final estimator where there are K attractors and K is fixed, e.g. $K = 500$, so K does not depend on n. A crucial observation is that the theory of the algorithm estimator depends on the theory of the attractors, not on the estimator corresponding to the criterion.

For example, let $(\mathbf{0}, \boldsymbol{I}_p)$ and $(\mathbf{1}, diag(1, 3, ..., p))$ be the high breakdown attractors where $\mathbf{0}$ and $\mathbf{1}$ are the $p \times 1$ vectors of zeroes and ones. If the minimum determinant criterion is used, then the final estimator is $(\mathbf{0}, \boldsymbol{I}_p)$. Although the MCD criterion is used, the algorithm estimator does not have the same properties as the MCD estimator.

Hawkins and Olive (2002) showed that if K randomly selected elemental starts are used with concentration to produce the attractors, then the resulting estimator is inconsistent and zero breakdown if K and k are fixed and free of n. Note that each elemental start can be made to breakdown by changing one case. Hence the breakdown value of the final estimator is bounded by $K/n \to 0$ as $n \to \infty$. Note that the classical estimator computed from h_n randomly drawn cases is an inconsistent estimator unless $h_n \to \infty$ as $n \to \infty$. Thus the classical estimator applied to a randomly drawn elemental set of $h_n \equiv p + 1$ cases is an inconsistent estimator, so the K starts and the K attractors are inconsistent.

This theory shows that the Maronna et al. (2006, pp. 198-199) estimators that use $K = 500$ and one concentration step ($k = 0$) are inconsistent and zero breakdown. The following theorem is useful because it does not depend on the criterion used to choose the attractor.

Suppose there are K consistent estimators $(T_j, \boldsymbol{C}_j)$ of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ for some constant $a > 0$, each with the same rate n^δ. If $(T_A, \boldsymbol{C}_A)$ is an estimator obtained by choosing one of the K estimators, then $(T_A, \boldsymbol{C}_A)$ is a consistent estimator of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ with rate n^δ by Pratt (1959). See Theorem 2.18.

Theorem 8.24. Suppose the algorithm estimator chooses an attractor as the final estimator where there are K attractors and K is fixed.

i) If all of the attractors are consistent estimators of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$, then the algorithm estimator is a consistent estimator of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$.

ii) If all of the attractors are consistent estimators of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ with the same rate, e.g. n^δ where $0 < \delta \leq 0.5$, then the algorithm estimator is a consistent estimator of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ with the same rate as the attractors.

iii) If all of the attractors are high breakdown, then the algorithm estimator is high breakdown.

iv) Suppose the data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid and $P(\boldsymbol{x}_i = \boldsymbol{\mu}) < 1$. The elemental basic resampling algorithm estimator ($k = -1$) is inconsistent.

v) The elemental concentration algorithm is zero breakdown.

Proof. i) Choosing from K consistent estimators for $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$ results in a consistent estimator for of $(\boldsymbol{\mu}, a\ \boldsymbol{\Sigma})$, and ii) follows from Pratt (1959). iii) Let $\gamma_{n,i}$ be the breakdown value of the ith attractor if the clean data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are in general position. The breakdown value γ_n of the algorithm estimator can be no lower than that of the worst attractor: $\gamma_n \geq \min(\gamma_{n,1}, ..., \gamma_{n,K}) \to 0.5$ as $n \to \infty$.

iv) Let $(\overline{\boldsymbol{x}}_{-1,j}, \boldsymbol{S}_{-1,j})$ be the classical estimator applied to a randomly drawn elemental set. Then $\overline{\boldsymbol{x}}_{-1,j}$ is the sample mean applied to $p + 1$ iid cases. Hence $E(\boldsymbol{S}_j) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$, $E[\overline{\boldsymbol{x}}_{-1,j}] = E(\boldsymbol{x}) = \boldsymbol{\mu}$, and $\mathrm{Cov}(\overline{\boldsymbol{x}}_{-1,j}) = \mathrm{Cov}(\boldsymbol{x})/(p+1) = \boldsymbol{\Sigma}_{\boldsymbol{x}}/(p+1)$ assuming second moments. So the $(\overline{\boldsymbol{x}}_{-1,j}, \boldsymbol{S}_{-1,j})$ are identically distributed and inconsistent estimators of $(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. Even without second moments, there exists $\epsilon > 0$ such that $P(\|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\| > \epsilon) = \delta_\epsilon > 0$ where the probability, ϵ, and δ_ϵ do not depend on n since the distribution of $\overline{\boldsymbol{x}}_{-1,j}$ only depends on the distribution of the iid $\boldsymbol{x}_i$, not on n. Then

$P(\min_j \|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\| > \epsilon) = P(\text{all } \|\overline{\boldsymbol{x}}_{-1,\text{j}} - \boldsymbol{\mu}\| > \epsilon) \to \delta_\epsilon^{\text{K}} > 0$ as $n \to \infty$ where equality would hold if the $\overline{\boldsymbol{x}}_{-1,j}$ were iid. Hence the "best start" that minimizes $\|\overline{\boldsymbol{x}}_{-1,j} - \boldsymbol{\mu}\|$ is inconsistent.

v) The classical estimator with breakdown $1/n$ is applied to each elemental start. Hence $\gamma_n \leq K/n \to 0$ as $n \to \infty$. □

Since the FMCD estimator is a zero breakdown elemental concentration algorithm, the Hubert et al. (2008) claim that "MCD can be efficiently computed with the FAST-MCD estimator" is false. Suppose K is fixed, but at least one randomly drawn start is iterated to convergence so that k is not fixed. Then it is not known whether the attractors are inconsistent or consistent estimators, so it is not known whether FMCD is consistent. It is possible to produce consistent estimators if $K \equiv K_n$ is allowed to increase to ∞.

Remark 8.5. Let γ_o be the highest percentage of large outliers that an elemental concentration algorithm can detect reliably. For many data sets,

$$\gamma_o \approx \min\left(\frac{n - c_n}{n}, 1 - [1 - (0.2)^{1/K}]^{1/h}\right) 100\% \tag{8.38}$$

if n is large, $c_n \geq n/2$ and $h = p + 1$.

Proof. Suppose that the data set contains n cases with d outliers and $n - d$ clean cases. Suppose K elemental sets are chosen with replacement. If W_i is the number of outliers in the ith elemental set, then the W_i are iid hypergeometric$(d, n - d, h)$ random variables. Suppose that it is desired to find K such that the probability P(that at least one of the elemental sets is clean) $\equiv P_1 \approx 1 - \alpha$ where $0 < \alpha < 1$. Then $P_1 = 1-$ P(none of the K elemental sets is clean) $\approx 1 - [1 - (1 - \gamma)^h]^K$ by independence. If the contamination proportion γ is fixed, then the probability of obtaining at least one clean subset of size h with high probability (say $1 - \alpha = 0.8$) is given by $0.8 = 1 - [1 - (1 - \gamma)^h]^K$. Fix the number of starts K and solve this equation for γ. □

8.2.4 Theory for Practical Estimators

It is convenient to let the $\boldsymbol{x}_i$ be random vectors for large sample theory, but the $\boldsymbol{x}_i$ are fixed clean observed data vectors when discussing breakdown. This subsection presents the FCH estimator to be used along with the classical estimator. Recall from Definition 8.29 that a *concentration algorithm* uses K_n *starts* $(T_{-1,j}, \boldsymbol{C}_{-1,j})$. After finding $(T_{0,j}, \boldsymbol{C}_{0,j})$, each start is refined with k concentration steps, resulting in K_n *attractors* $(T_{k,j}, \boldsymbol{C}_{k,j})$, and the concentration attractor $(T_A, \boldsymbol{C}_A)$ is the attractor that optimizes the criterion.

Concentration algorithms include the *basic resampling algorithm* as a special case with $k = -1$. Using $k = 10$ concentration steps works well, and iterating until convergence is usually fast. The DGK estimator (Devlin et al. 1975, 1981) defined below is one example. The DGK estimator is affine equivariant since the classical estimator is affine equivariant and Mahalanobis distances are invariant under affine transformations by Theorem 8.19. This subsection will show that the Olive (2004a) MB estimator is a high breakdown estimator and that the DGK estimator is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$, the same quantity estimated by the MCD estimator. Both estimators use the classical estimator computed from $c_n \approx n/2$ cases. The breakdown point of the DGK estimator has been conjectured to be "at most $1/p$." See Rousseeuw and Leroy (1987, p. 254).

Definition 8.30. The *DGK estimator* $(T_{k,D}, \boldsymbol{C}_{k,D}) = (T_{DGK}, \boldsymbol{C}_{DGK})$ uses the classical estimator $(T_{-1,D}, \boldsymbol{C}_{-1,D}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ as the only start.

Definition 8.31. The *median ball (MB) estimator* $(T_{k,M}, \boldsymbol{C}_{k,M}) = (T_{MB}, \boldsymbol{C}_{MB})$ uses $(T_{-1,M}, \boldsymbol{C}_{-1,M}) = (\mathrm{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$ as the only start where $\mathrm{MED}(\boldsymbol{W})$ is the coordinatewise median. So $(T_{0,M}, \boldsymbol{C}_{0,M})$ is the classical estimator applied to the "half set" of data closest to $\mathrm{MED}(\boldsymbol{W})$ in Euclidean distance.

The proof of the following theorem implies that a high breakdown estimator $(T, \boldsymbol{C})$ has $\mathrm{MED}(D_i^2) \leq V$ and that the hyperellipsoid $\{\boldsymbol{x} | D_{\boldsymbol{x}}^2 \leq D_{(c_n)}^2\}$ that contains $c_n \approx n/2$ of the cases is in some ball about the origin of radius r, where V and r do not depend on the outliers even if the number of outliers is close to $n/2$. Also the attractor of a high breakdown estimator is a high breakdown estimator if the number of concentration steps k is fixed, e.g. $k = 10$. The theorem implies that the MB estimator $(T_{MB}, \boldsymbol{C}_{MB})$ is high breakdown.

Theorem 8.25. Suppose $(T, \boldsymbol{C})$ is a high breakdown estimator where $\boldsymbol{C}$ is a symmetric, positive definite $p \times p$ matrix if the contamination proportion d_n/n is less than the breakdown value. Then the concentration attractor $(T_k, \boldsymbol{C}_k)$ is a high breakdown estimator if the coverage $c_n \approx n/2$ and the data are in general position.

Proof. Following Leon (1986, p. 280), if $\boldsymbol{A}$ is a symmetric positive definite matrix with eigenvalues $\tau_1 \geq \cdots \geq \tau_p$, then for any nonzero vector $\boldsymbol{x}$,

$$0 < \|\boldsymbol{x}\|^2 \ \tau_p \leq \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \leq \|\boldsymbol{x}\|^2 \ \tau_1. \tag{8.39}$$

Let $\lambda_1 \geq \cdots \geq \lambda_p$ be the eigenvalues of $\boldsymbol{C}$. By (8.39),

$$\frac{1}{\lambda_1}\|\boldsymbol{x} - T\|^2 \leq (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \leq \frac{1}{\lambda_p}\|\boldsymbol{x} - T\|^2. \tag{8.40}$$

By (8.40), if the $D^2_{(i)}$ are the order statistics of the $D^2_i(T, \boldsymbol{C})$, then $D^2_{(i)} < V$ for some constant V that depends on the clean data but not on the outliers even if i and d_n are near $n/2$. (Note that $1/\lambda_p$ and $\text{MED}(\|\boldsymbol{x}_i - T\|^2)$ are both bounded for high breakdown estimators even for d_n near $n/2$.)

Following Johnson and Wichern (1988, pp. 50, 103), the boundary of the set $\{\boldsymbol{x}|D^2_{\boldsymbol{x}} \leq h^2\} = \{\boldsymbol{x}|(\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T) \leq h^2\}$ is a hyperellipsoid centered at T with axes of length $2h\sqrt{\lambda_i}$. Hence $\{\boldsymbol{x}|D^2_{\boldsymbol{x}} \leq D^2_{(c_n)}\}$ is contained in some ball about the origin of radius r where r does not depend on the number of outliers even for d_n near $n/2$. This is the set containing the cases used to compute $(T_0, \boldsymbol{C}_0)$. Since the set is bounded, T_0 is bounded and the largest eigenvalue $\lambda_{1,0}$ of $\boldsymbol{C}_0$ is bounded by Theorem 8.22. The determinant $det(\boldsymbol{C}_{MCD})$ of the HB minimum covariance determinant estimator satisfies $0 < det(\boldsymbol{C}_{MCD}) \leq det(\boldsymbol{C}_0) = \lambda_{1,0} \cdots \lambda_{p,0}$, and $\lambda_{p,0} > \inf det(\boldsymbol{C}_{MCD})/\lambda^{p-1}_{1,0} > 0$ where the infimum is over all possible data sets with $n - d_n$ clean cases and d_n outliers. Since these bounds do not depend on the outliers even for d_n near $n/2$, $(T_0, \boldsymbol{C}_0)$ is a high breakdown estimator. Now repeat the argument with $(T_0, \boldsymbol{C}_0)$ in place of $(T, \boldsymbol{C})$ and $(T_1, \boldsymbol{C}_1)$ in place of $(T_0, \boldsymbol{C}_0)$. Then $(T_1, \boldsymbol{C}_1)$ is high breakdown. Repeating the argument iteratively shows $(T_k, \boldsymbol{C}_k)$ is high breakdown. □

The following corollary shows that it is easy to find a subset J of $c_n \approx n/2$ cases such that the classical estimator $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ applied to J is a HB estimator of MLD.

Theorem 8.26. Let J consist of the c_n cases $\boldsymbol{x}_i$ such that $\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\| \leq \text{MED}(\|\boldsymbol{x}_i - \text{MED}(\boldsymbol{W})\|)$. Then the classical estimator $(\overline{\boldsymbol{x}}_J, \boldsymbol{S}_J)$ applied to J is a HB estimator of MLD.

To investigate the consistency and rate of robust estimators of multivariate location and dispersion, review Definitions 3.5 and 3.6.

The following assumption (E1) gives a class of distributions where we can prove that the new robust estimators are $\sqrt{n}$ consistent. Cator and Lopuhaä (2010, 2012) showed that MCD is consistent provided that the MCD functional is unique. Distributions where the functional is unique are called "unimodal," and rule out, for example, a spherically symmetric uniform distribution. Theorem 8.27 is crucial for theory and Theorem 8.28 shows that under (E1), both MCD and DGK are estimating $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$.

Assumption (E1): The $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid from a "unimodal" elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with nonsingular covariance matrix $\text{Cov}(\boldsymbol{x}_i)$ where g is continuously differentiable with finite 4th moment: $\int (\boldsymbol{x}^T\boldsymbol{x})^2 g(\boldsymbol{x}^T\boldsymbol{x}) d\boldsymbol{x} < \infty$.

Lopuhaä (1999) showed that if a start $(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$, then the classical estimator applied to the cases with $D^2_i(T, \boldsymbol{C}) \leq h^2$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ where $a, s > 0$ are

some constants. Affine equivariance is not used for $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. Also, the attractor and the start have the same rate. If the start is inconsistent, then so is the attractor. The weight function $I(D_i^2(T, \boldsymbol{C}) \leq h^2)$ is an indicator that is 1 if $D_i^2(T, \boldsymbol{C}) \leq h^2$ and 0 otherwise.

Theorem 8.27, Lopuhaä (1999). Assume the number of concentration steps k is fixed. a) If the start $(T, \boldsymbol{C})$ is inconsistent, then so is the attractor.

b) Suppose $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\boldsymbol{I}_p)$ with rate n^δ where $s > 0$ and $0 < \delta \leq 0.5$. Assume (E1) holds and $\boldsymbol{\Sigma} = \boldsymbol{I}_p$. Then the classical estimator $(T_0, \boldsymbol{C}_0)$ applied to the cases with $D_i^2(T, \boldsymbol{C}) \leq h^2$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{I}_p)$ with the same rate n^δ where $a > 0$.

c) Suppose $(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^δ where $s > 0$ and $0 < \delta \leq 0.5$. Assume (E1) holds. Then the classical estimator $(T_0, \boldsymbol{C}_0)$ applied to the cases with $D_i^2(T, \boldsymbol{C}) \leq h^2$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with the same rate n^δ where $a > 0$. The constant a depends on the positive constants s, h, p, and the elliptically contoured distribution, but does not otherwise depend on the consistent start $(T, \boldsymbol{C})$.

Let $\delta = 0.5$. Applying Theorem 8.27c) iteratively for a fixed number k of steps produces a sequence of estimators $(T_0, \boldsymbol{C}_0), ..., (T_k, \boldsymbol{C}_k)$ where $(T_j, \boldsymbol{C}_j)$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_j\boldsymbol{\Sigma})$ where the constants $a_j > 0$ depend on s, h, p, and the elliptically contoured distribution, but do not otherwise depend on the consistent start $(T, \boldsymbol{C}) \equiv (T_{-1}, \boldsymbol{C}_{-1})$.

The 4th moment assumption was used to simplify theory, but likely holds under 2nd moments. Affine equivariance is needed so that the attractor is affine equivariant, but probably is not needed to prove consistency.

Conjecture 8.2. Change the finite 4th moments assumption to a finite 2nd moments in assumption E1). Suppose $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^δ where $s > 0$ and $0 < \delta \leq 0.5$. Then the classical estimator applied to the cases with $D_i^2(T, \boldsymbol{C}) \leq h^2$ is a consistent estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with the same rate n^δ where $a > 0$.

Remark 8.6. To see that the Lopuhaä (1999) theory extends to concentration where the weight function uses $h^2 = D^2_{(c_n)}(T, \boldsymbol{C})$, note that $(T, \tilde{\boldsymbol{C}}) \equiv (T, D^2_{(c_n)}(T, \boldsymbol{C})\ \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, b\boldsymbol{\Sigma})$ where $b > 0$ is derived in (8.42), and weight function $I(D_i^2(T, \tilde{\boldsymbol{C}}) \leq 1)$ is equivalent to the concentration weight function $I(D_i^2(T, \boldsymbol{C}) \leq D^2_{(c_n)}(T, \boldsymbol{C}))$. As noted above Theorem 8.19, $(T, \tilde{\boldsymbol{C}})$ is affine equivariant if $(T, \boldsymbol{C})$ is affine equivariant. Hence Lopuhaä (1999) theory applied to $(T, \tilde{\boldsymbol{C}})$ with $h = 1$ is equivalent to theory applied to affine equivariant $(T, \boldsymbol{C})$ with $h^2 = D^2_{(c_n)}(T, \boldsymbol{C})$.

If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ with rate n^δ where $0 < \delta \leq 0.5$, then $D^2(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) =$

$$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T[\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1} + s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)$$
$$= s^{-1}D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(n^{-\delta}). \tag{8.41}$$

Thus the sample percentiles of $D_i^2(T, \boldsymbol{C})$ are consistent estimators of the percentiles of $s^{-1}D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Suppose $c_n/n \to \xi \in (0, 1)$ as $n \to \infty$, and let $D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the 100ξth percentile of the population squared distances. Then $D_{(c_n)}^2(T, \boldsymbol{C}) \xrightarrow{P} s^{-1}D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $b\boldsymbol{\Sigma} = s^{-1}D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})s\boldsymbol{\Sigma} = D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})\boldsymbol{\Sigma}$. Thus

$$b = D_\xi^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{8.42}$$

does not depend on $s > 0$ or $\delta \in (0, 0.5]$. □

Concentration applies the classical estimator to cases with $D_i^2(T, \boldsymbol{C}) \leq D_{(c_n)}^2(T, \boldsymbol{C})$. Let $c_n \approx n/2$ and

$$b = D_{0.5}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

be the population median of the population squared distances. By Remark 8.6, if $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ then $(T, \tilde{\boldsymbol{C}}) \equiv (T, D_{(c_n)}^2(T, \boldsymbol{C})\ \boldsymbol{C})$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, b\boldsymbol{\Sigma})$, and $D_i^2(T, \tilde{\boldsymbol{C}}) \leq 1$ is equivalent to $D_i^2(T, \boldsymbol{C}) \leq D_{(c_n)}^2(T, \boldsymbol{C})$). Hence Lopuhaä (1999) theory applied to $(T, \tilde{\boldsymbol{C}})$ with $h = 1$ is equivalent to theory applied to the concentration estimator using the affine equivariant estimator $(T, \boldsymbol{C}) \equiv (T_{-1}, \boldsymbol{C}_{-1})$ as the start. Since b does not depend on s, concentration produces a sequence of estimators $(T_0, \boldsymbol{C}_0), ..., (T_k, \boldsymbol{C}_k)$ where $(T_j, \boldsymbol{C}_j)$ is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ where the constant $a > 0$ is the same for $j = 0, 1, ..., k$.

Theorem 8.28 shows that $a = a_{MCD}$ where $\xi = 0.5$. Hence concentration with a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^δ as a start results in a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ with rate n^δ. This result can be applied iteratively for a finite number of concentration steps. Hence DGK is a $\sqrt{n}$ consistent affine equivariant estimator of the same quantity that MCD is estimating. It is not known if the results hold if concentration is iterated to convergence. For multivariate normal data, $D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \chi_p^2$.

Theorem 8.28. Assume that (E1) holds and that $(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ with rate n^δ where the constants $s > 0$ and $0 < \delta \leq 0.5$. Then the classical estimator $(\overline{\boldsymbol{x}}_{t,j}, \boldsymbol{S}_{t,j})$ computed from the $c_n \approx n/2$ of cases with the smallest distances $D_i(T, \boldsymbol{C})$ is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ with the same rate n^δ.

Proof. By Remark 8.6, the estimator is a consistent affine equivariant estimator of $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ with rate n^δ. By the remarks above, a will be the same for any consistent affine equivariant estimator of $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$ and a does not depend on $s > 0$ or $\delta \in (0, 0.5]$. Hence the result follows if $a = a_{MCD}$. The MCD

estimator is a $\sqrt{n}$ consistent affine equivariant estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$ by Cator and Lopuhaä (2010, 2012). If the MCD estimator is the start, then it is also the attractor by Theorem 8.23 which shows that concentration does not increase the MCD criterion. Hence $a = a_{MCD}$. □

Next we define the easily computed robust $\sqrt{n}$ consistent FCH estimator, so named since it is fast, consistent, and uses a high breakdown attractor. The FCH and MBA estimators use the $\sqrt{n}$ consistent DGK estimator $(T_{DGK}, \boldsymbol{C}_{DGK})$ and the high breakdown MB estimator $(T_{MB}, \boldsymbol{C}_{MB})$ as attractors.

Definition 8.32. Let the "median ball" be the hypersphere containing the "half set" of data closest to MED($\boldsymbol{W}$) in Euclidean distance. The *FCH estimator* uses the MB attractor if the DGK location estimator T_{DGK} is outside of the median ball, and the attractor with the smallest determinant, otherwise. Let $(T_A, \boldsymbol{C}_A)$ be the attractor used. Then the estimator $(T_{FCH}, \boldsymbol{C}_{FCH})$ takes $T_{FCH} = T_A$ and

$$\boldsymbol{C}_{FCH} = \frac{\text{MED}(D_i^2(T_A, \boldsymbol{C}_A))}{\chi^2_{p,0.5}} \boldsymbol{C}_A \tag{8.43}$$

where $\chi^2_{p,0.5}$ is the 50th percentile of a chi–square distribution with p degrees of freedom.

Remark 8.7. The *MBA estimator* $(T_{MBA}, \boldsymbol{C}_{MBA})$ uses the attractor $(T_A, \boldsymbol{C}_A)$ with the smallest determinant. Hence the DGK estimator is used as the attractor if $det(\boldsymbol{C}_{DGK}) \leq det(\boldsymbol{C}_{MB})$, and the MB estimator is used as the attractor, otherwise. Then $T_{MBA} = T_A$ and $\boldsymbol{C}_{MBA}$ is computed using the right hand side of (8.43). The difference between the FCH and MBA estimators is that the FCH estimator also uses a location criterion to choose the attractor: if the DGK location estimator T_{DGK} has a greater Euclidean distance from MED($\boldsymbol{W}$) than half the data, then FCH uses the MB attractor. The FCH estimator only uses the attractor with the smallest determinant if $\|T_{DGK} - \text{MED}(\boldsymbol{W})\| \leq \text{MED}(D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p))$. Using the location criterion increases the outlier resistance of the FCH estimator for certain types of outliers. See Olive (2017b).

The following theorem shows the FCH estimator has good statistical properties. We conjecture that FCH is high breakdown. Note that the location estimator T_{FCH} is high breakdown and that $det(\boldsymbol{C}_{FCH})$ is bounded away from 0 and ∞ if the data is in general position, even if nearly half of the cases are outliers.

Theorem 8.29. T_{FCH} is high breakdown if the clean data are in general position. Suppose (E1) holds. If $(T_A, \boldsymbol{C}_A)$ is the DGK or MB attractor with the smallest determinant, then $(T_A, \boldsymbol{C}_A)$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, a_{MCD}\boldsymbol{\Sigma})$. Hence the MBA and FCH estimators are outlier resistant

$\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ where $c = u_{0.5}/\chi^2_{p,0.5}$, and $c = 1$ for multivariate normal data.

Proof. T_{FCH} is high breakdown since it is a bounded distance from MED($\boldsymbol{W}$) even if the number of outliers is close to $n/2$. Under (E1) the FCH and MBA estimators are asymptotically equivalent since $\|T_{DGK} - \text{MED}(\boldsymbol{W})\| \to 0$ in probability. The estimator satisfies $0 < det(\boldsymbol{C}_{MCD}) \leq det(\boldsymbol{C}_A) \leq det(\boldsymbol{C}_{0,M}) < \infty$ by Theorem 8.25 if up to nearly 50% of the cases are outliers. If the distribution is spherical about $\boldsymbol{\mu}$, then the result follows from Pratt (1959) and Theorem 8.23 since both starts are $\sqrt{n}$ consistent. Otherwise, the MB estimator $\boldsymbol{C}_{MB}$ is a biased estimator of $a_{MCD}\boldsymbol{\Sigma}$. But the DGK estimator $\boldsymbol{C}_{DGK}$ is a $\sqrt{n}$ consistent estimator of $a_{MCD}\boldsymbol{\Sigma}$ by Theorem 8.28 and $\|\boldsymbol{C}_{MCD} - \boldsymbol{C}_{DGK}\| = O_P(n^{-1/2})$. Thus the probability that the DGK attractor minimizes the determinant goes to one as $n \to \infty$, and $(T_A, \boldsymbol{C}_A)$ is asymptotically equivalent to the DGK estimator $(T_{DGK}, \boldsymbol{C}_{DGK})$.

Let $\boldsymbol{C}_F = \boldsymbol{C}_{FCH}$ or $\boldsymbol{C}_F = \boldsymbol{C}_{MBA}$. Let $P(U \leq u_\alpha) = \alpha$ where U is given by (1.62). Then the scaling in (8.43) makes $\boldsymbol{C}_F$ a consistent estimator of $c\boldsymbol{\Sigma}$ where $c = u_{0.5}/\chi^2_{p,0.5}$, and $c = 1$ for multivariate normal data. □

A standard method of reweighting can be used to produce the RMBA and RFCH estimators. RMVN uses a slightly modified method of reweighting so that RMVN gives good estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for multivariate normal data, even when certain types of outliers are present.

Definition 8.33. The *RFCH estimator* uses two standard reweighting steps. Let $(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1)$ be the classical estimator applied to the n_1 cases with $D_i^2(T_{FCH}, \boldsymbol{C}_{FCH}) \leq \chi^2_{p,0.975}$, and let

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{\text{MED}(D_i^2(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1))}{\chi^2_{p,0.5}} \tilde{\boldsymbol{\Sigma}}_1.$$

Then let $(T_{RFCH}, \tilde{\boldsymbol{\Sigma}}_2)$ be the classical estimator applied to the cases with $D_i^2(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1) \leq \chi^2_{p,0.975}$, and let

$$\boldsymbol{C}_{RFCH} = \frac{\text{MED}(D_i^2(T_{RFCH}, \tilde{\boldsymbol{\Sigma}}_2))}{\chi^2_{p,0.5}} \tilde{\boldsymbol{\Sigma}}_2.$$

RMBA and RFCH are $\sqrt{n}$ consistent estimators of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ by Lopuhaä (1999) where the weight function uses $h^2 = \chi^2_{p,0.975}$, but the two estimators use nearly 97.5% of the cases if the data is multivariate normal.

Definition 8.34. The *RMVN estimator* uses $(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1)$ and n_1 as above. Let $q_1 = \min\{0.5(0.975)n/n_1, 0.995\}$, and

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{\text{MED}(D_i^2(\hat{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\Sigma}}_1))}{\chi^2_{p,q_1}} \tilde{\boldsymbol{\Sigma}}_1.$$

Then let $(T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2)$ be the classical estimator applied to the n_2 cases with $D_i^2(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1)) \leq \chi^2_{p,0.975}$. Let $q_2 = \min\{0.5(0.975)n/n_2, 0.995\}$, and

$$\boldsymbol{C}_{RMVN} = \frac{\text{MED}(D_i^2(T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2))}{\chi^2_{p,q_2}} \tilde{\boldsymbol{\Sigma}}_2.$$

The RMVN estimator is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ by Lopuhaä (1999) where the weight function uses $h^2 = \chi^2_{p,0.975}$ and $d = u_{0.5}/\chi^2_{p,q}$ where $q_2 \to q$ in probability as $n \to \infty$. Here $0.5 \leq q < 1$ depends on the elliptically contoured distribution, but $q = 0.5$ and $d = 1$ for multivariate normal data.

Hubert et al. (2008, 2012) claim that FMCD computes the MCD estimator. This claim is trivially shown to be false in the following theorem.

Theorem 8.30. Neither FMCD nor Det-MCD compute the MCD estimator.

Proof. A necessary condition for an estimator to be the MCD estimator is that the determinant of the covariance matrix for the estimator be the smallest for every run in a simulation. Sometimes FMCD had the smaller determinant and sometimes Det-MCD had the smaller determinant in the simulations done by Hubert et al. (2012). □

The following theorem shows that it is very difficult to drive the determinant of the dispersion estimator from a concentration algorithm to zero.

Theorem 8.31. Consider the concentration and MCD estimators that both cover c_n cases. For multivariate data, if at least one of the starts is nonsingular, then the concentration attractor $\boldsymbol{C}_A$ is less likely to be singular than the high breakdown MCD estimator $\boldsymbol{C}_{MCD}$.

Proof. If all of the starts are singular, then the Mahalanobis distances cannot be computed and the classical estimator can not be applied to c_n cases. Suppose that at least one start was nonsingular. Then $\boldsymbol{C}_A$ and $\boldsymbol{C}_{MCD}$ are both sample covariance matrices applied to c_n cases, but by definition $\boldsymbol{C}_{MCD}$ minimizes the determinant of such matrices. Hence $0 \leq \det(\boldsymbol{C}_{MCD}) \leq \det(\boldsymbol{C}_A)$. □

Software

The `robustbase` library was downloaded from (www.r-project.org/#doc). The preface explains how to use the source command to get the `lspack` functions in *R* and how to download a library from *R*. Type the commands `library(MASS)` and `library(robustbase)` to compute the FMCD and OGK estimators with the `cov.mcd` and `covOGK` functions. To use Det-MCD instead of FMCD, change

```
out <- covMcd(x)  to out <- covMcd(x,nsamp="deterministic"),
```

but in Spring 2015 this change was more likely to cause errors.

The function function *covfch* computes FCH and RFCH, while *covrmvn* computes the RMVN and MB estimators. The function *covrmb* computes MB and RMB where RMB is like RMVN except the MB estimator is reweighted instead of FCH. Functions *covdgk*, *covmba*, and *rmba* compute the scaled DGK, MBA, and RMBA estimators. **Better programs would use MB if DGK causes an error.**

8.2.5 The RMVN and RFCH Sets

Both the RMVN and RFCH estimators compute the classical estimator $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ on some set U containing $n_U \geq n/2$ of the cases. Referring to Definition 8.33, for the RFCH estimator, $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U) = (T_{RFCH}, \tilde{\boldsymbol{\Sigma}}_2)$, and then $\boldsymbol{S}_U$ is scaled to form $\boldsymbol{C}_{RFCH}$. Referring to Definition 8.34, for the RMVN estimator, $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U) = (T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2)$, and then $\boldsymbol{S}_U$ is scaled to form $\boldsymbol{C}_{RMVN}$. See Definition 8.35. The RFCH set can be defined similarly.

Definition 8.35. Let the n_2 cases in Definition 8.34 be known as the *RMVN set* U. Hence $(T_{RMVN}, \tilde{\boldsymbol{\Sigma}}_2) = (\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ is the classical estimator applied to the RMVN set U, which can be regarded as the untrimmed data (the data not trimmed by ellipsoidal trimming) or the cleaned data. Also $\boldsymbol{S}_U$ is the unscaled estimated dispersion matrix while $\boldsymbol{C}_{RMVN}$ is the scaled estimated dispersion matrix.

Remark 8.8. Classical methods can be applied to the RMVN subset U to make robust methods. Under (E1), $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c_U \boldsymbol{\Sigma})$ for some constant $c_U > 0$ that depends on the underlying distribution of the iid $\boldsymbol{x}_i$. For a general estimator of multivariate location and dispersion $(T_A, \boldsymbol{C}_A)$, typically a reweight for efficiency step is performed, resulting in a set U such that the classical estimator $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ is the classical estimator applied to a set U. For example, use $U = \{\boldsymbol{x}_i | D_i^2(T_A, \boldsymbol{C}_A) \leq \chi^2_{p,0.975}\}$. Then the final estimator is $(T_F, \boldsymbol{C}_F) = (\overline{\boldsymbol{x}}_U, a\boldsymbol{S}_U)$ where scaling is done as in Equation (8.43) in an attempt to make $\boldsymbol{C}_F$ a good estimator of $\boldsymbol{\Sigma}$ if the iid data are from a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Then $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ can be shown to be a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c_U \boldsymbol{\Sigma})$ for a large class of distributions for the RMVN set, for the RFCH set, or if $(T_A, \boldsymbol{C}_A)$ is an affine equivariant $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c_A \boldsymbol{\Sigma})$ on a large class of distributions.

The two main ways to handle outliers are i) apply the multivariate method to the cleaned data, and ii) plug in robust estimators for classical estimators. Practical plug in robust estimators have rarely been shown to be $\sqrt{n}$ consistent and highly outlier resistant.

Using the RMVN or RFCH set U is simultaneously a plug in method and an objective way to clean the data such that the resulting robust method is

often backed by theory. This result is extremely useful computationally: find the RMVN set or RFCH set U, then apply the classical method to the cases in the set U. This procedure is often equivalent to using $(\overline{\boldsymbol{x}}_U, \boldsymbol{S}_U)$ as plug in estimators. The method can be applied if $n > 2(p+1)$ but may not work well unless $n > 20p$. The *lspack* function `getu` gets the RMVN set U as well as the case numbers corresponding to the cases in U.

The set U is a small volume hyperellipsoid containing at least half of the cases since concentration is used. The set U can also be regarded as the "untrimmed data": the data that was not trimmed by ellipsoidal trimming. Theory has been proved for a large class of elliptically contoured distributions, but it is conjectured that theory holds for a much wider class of distributions. See Olive (2017b, pp. 127-128).

Application 8.6. Outlier resistant regression: Let the ith case $\boldsymbol{w}_i = (Y_i, \boldsymbol{x}_i^T)^T$ where the continuous predictors from $\boldsymbol{x}_i$ are denoted by $\boldsymbol{u}_i$ for $i = 1, ..., n$. Find the RFCH or RMVN set from the $\boldsymbol{u}_i$, and then run the regression method on the n_U cases $\boldsymbol{w}_i$ corresponding to the set U indices $i_1, ...i_{n_U}$, where $n_U \geq n/2$. Since the response variable was not used to pick the cases, this regression method, conditional on n_U and on the n_U selected cases, has similar large sample theory to the classical regression method that uses all n cases. A similar technique can be used if there is more than one response variable.

Often the theory of the method applies to the cleaned data set since $\boldsymbol{y}$ was not used to pick the subset of the data. Efficiency can be much lower since n_u cases are used where $n/2 \leq n_u \leq n$, and the trimmed cases tend to be the "farthest" from the center of $\boldsymbol{w}$.

In R, assume Y is the vector of response variables, x is the data matrix of the predictors (often not including the trivial predictor), and w is the data matrix of the $\boldsymbol{w}_i$. Then the following R commands can be used to get the cleaned data set. We could use the `covmb2` set B instead of the RMVN set U computed from the $\boldsymbol{w}$ by replacing the command *getu(w)* by `getB(w)`.

```
indx <- getu(w)$indx  #often w = x
Yc <- Y[indx]
Xc <- x[indx,]
#example
indx <- getu(buxx)$indx
Yc <- buxy[indx]
Xc <- buxx[indx,]
outr <- lsfit(Xc,Yc)
MLRplot(Xc,Yc) #right click Stop twice
```

8.2.6 MLD Outlier Detection if $p > n$

Most outlier detection methods work best if $n \geq 20p$, but often data sets have $p > n$, and outliers are a major problem. One of the simplest outlier detection methods uses the Euclidean distances of the $\boldsymbol{x}_i$ from the coordinatewise median $D_i = D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Concentration type steps compute the weighted median MED_j: the coordinatewise median computed from the "half set" of cases $\boldsymbol{x}_i$ with $D_i^2 \leq \text{MED}(D_i^2(\text{MED}_{j-1}, \boldsymbol{I}_p))$ where $\text{MED}_0 = \text{MED}(\boldsymbol{W})$. We often used $j = 0$ (no concentration type steps) or $j = 9$. Let $D_i = D_i(\text{MED}_j, \boldsymbol{I}_p)$. Let $W_i = 1$ if $D_i \leq \text{MED}(D_1, ..., D_n) + k\text{MAD}(D_1, ..., D_n)$ where $k \geq 0$ and $k = 5$ is the default choice. Let $W_i = 0$, otherwise. Using $k \geq 0$ insures that at least half of the cases get weight 1. This weighting corresponds to the weighting that would be used in a one sided metrically trimmed mean (Huber type skipped mean) of the distances.

Definition 8.36. Let the *covmb2 set* B of at least $n/2$ cases correspond to the cases with weight $W_i = 1$. The cases not in set B get weight $W_i = 0$. Then the *covmb2* estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix applied to the cases in set B. Hence

$$T = \frac{\sum_{i=1}^n W_i \boldsymbol{x}_i}{\sum_{i=1}^n W_i} \text{ and } \boldsymbol{C} = \frac{\sum_{i=1}^n W_i (\boldsymbol{x}_i - T)(\boldsymbol{x}_i - T)^T}{\sum_{i=1}^n W_i \ - \ 1}.$$

Example 8.9. Let the clean data (nonoutliers) be i **1** for $i = 1, 2, 3, 4$, and 5 while the outliers are j **1** for $j = 16, 17, 18$, and 19. Here $n = 9$ and **1** is $p \times 1$. Making a plot of the data for $p = 2$ may be useful. Then the coordinatewise median $\text{MED}_0 = \text{MED}(\boldsymbol{W}) = 5$ **1**. The median Euclidean distance of the data is the Euclidean distance of 5 **1** from 1 **1** = the Euclidean distance of 5 **1** from 9 **1**. The *median ball* is the hypersphere centered at the coordinatewise median with radius $r = \text{MED}(D_i(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p), \ i = 1, ..., n)$ that tends to contain $(n+1)/2$ of the cases if n is odd. Hence the clean data are in the median ball and the outliers are outside of the median ball. The coordinatewise median of the cases with the 5 smallest distances is the coordinatewise median of the clean data: $\text{MED}_1 = 3$ **1**. Then the median Euclidean distance of the data from MED_1 is the Euclidean distance of 3 **1** from 1 **1** = the Euclidean distance of 3 **1** from 5 **1**. Again the clean cases are the cases with the 5 smallest Euclidean distances. Hence $\text{MED}_j = 3$ **1** for $j \geq 1$. For $j \geq 1$, if $\boldsymbol{x}_i = j$ **1**, then $D_i = |j - 3|\sqrt{p}$. Thus $D_{(1)} = 0$, $D_{(2)} = D_{(3)} = \sqrt{p}$, and $D_{(4)} = D_{(5)} = 2\sqrt{p}$. Hence $\text{MED}(D_1, ..., D_n) = D_{(5)} = 2\sqrt{p} = \text{MAD}(D_1, ..., D_n)$ since the median distance of the D_i from $D_{(5)}$ is $2\sqrt{p} - 0 = 2\sqrt{p}$. Note that the 5 smallest absolute distances $|D_i - D_{(5)}|$ are $0, 0, \sqrt{p}, \sqrt{p}$, and $2\sqrt{p}$. Hence $W_i = 1$ if $D_i \leq 2\sqrt{p} + 10\sqrt{p} = 12\sqrt{p}$. The clean data get weight 1 while the outliers get weight 0 since the smallest distance D_i for the outliers is the Euclidean distance of 3 **1** from 16 **1** with a $D_i = \|16 \ \mathbf{1} - 3 \ \mathbf{1}\| = 13\sqrt{p}$. Hence the `covmb2` estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix

of the clean data. **Note that the distance for the outliers to get zero weight is proportional to the square root of the dimension $\sqrt{p}$.**

Application 8.7. Outlier resistant regression: Let the ith case $\boldsymbol{w}_i = (Y_i, \boldsymbol{x}_i^T)^T$ where the continuous predictors from $\boldsymbol{x}_i$ are denoted by $\boldsymbol{u}_i$ for $i = 1, ..., n$. Apply the `covmb2` estimator to the $\boldsymbol{u}_i$, and then run the regression method on the m cases $\boldsymbol{w}_i$ corresponding to the `covmb2` set B indices $i_1, ...i_m$, where $m \geq n/2$.

The `covmb2` estimator can also be used for $n > p$. The `covmb2` estimator attempts to give a robust dispersion estimator that reduces the bias by using a big ball about MED_j instead of a ball that contains half of the cases. The *lspack* function `getB` gives the set B of cases that got weight 1 along with the index `indx` of the case numbers that got weight 1.

8.3 Resistant Multiple Linear Regression

Consider the multiple linear regression model, written in matrix form as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Some good outlier resistant regression estimators are `rmreg2` from Section 8.5, the `hbreg` estimator from Section 8.4, and the Olive (2005) MBA and trimmed views estimators described below. Also apply a multiple linear regression method such as OLS or lasso to the cases corresponding to the RFCH, RMVN, or `covmb2` set applied to the continuous predictors. See Applications 8.6 and 8.7.

Resistant estimators are often created by computing several trial fits $\boldsymbol{b}_i$ that are estimators of $\boldsymbol{\beta}$. Then a criterion is used to select the trial fit to be used in the resistant estimator. Suppose $c \approx n/2$. The LMS(c) criterion is $Q_{LMS}(\boldsymbol{b}) = r^2_{(c)}(\boldsymbol{b})$ where $r^2_{(1)} \leq \cdots \leq r^2_{(n)}$ are the ordered squared residuals, and the LTS(c) criterion is $Q_{LTS}(\boldsymbol{b}) = \sum_{i=1}^{c} r^2_{(i)}(\boldsymbol{b})$. The LTA($c$) criterion is $Q_{LTA}(\boldsymbol{b}) = \sum_{i=1}^{c} |r(\boldsymbol{b})|_{(i)}$ where $|r(\boldsymbol{b})|_{(i)}$ is the ith ordered absolute residual. Three impractical high breakdown robust estimators are the Hampel (1975) least median of squares (LMS) estimator, the Rousseeuw (1984) least trimmed sum of squares (LTS) estimator, and the Hössjer (1991) least trimmed sum of absolute deviations (LTA) estimator. Also see Hawkins and Olive (1999ab). These estimators correspond to the $\hat{\boldsymbol{\beta}}_L \in \mathbb{R}^p$ that minimizes the corresponding criterion. LMS, LTA, and LTS have $O(n^p)$ or $O(n^{p+1})$ complexity. See Bernholt (2005), Hawkins and Olive (1999b), Klouda (2015), and Mount et al. (2014). Estimators with $O(n^4)$ or higher complexity take too long to compute. LTS and LTA are $\sqrt{n}$ consistent while LMS has the lower $n^{1/3}$ rate. See Kim and Pollard (1990), Čížek (2006, 2008), and Mašíček (2004). If $c = n$, the LTS and LTA criteria are the OLS and L_1 criteria. See Olive (2008, 2017b: ch. 14) for more on these estimators.

A good resistant estimator is the Olive (2005) *median ball algorithm* (MBA or `mbareg`). The Euclidean distance of the ith vector of predictors $\boldsymbol{x}_i$ from the jth vector of predictors $\boldsymbol{x}_j$ is

$$D_i(\boldsymbol{x}_j) = D_i(\boldsymbol{x}_j, \boldsymbol{I}_p) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T(\boldsymbol{x}_i - \boldsymbol{x}_j)}.$$

For a fixed $\boldsymbol{x}_j$ consider the ordered distances $D_{(1)}(\boldsymbol{x}_j), ..., D_{(n)}(\boldsymbol{x}_j)$. Next, let $\hat{\boldsymbol{\beta}}_j(\alpha)$ denote the OLS fit to the $\min(p + 3 + \lfloor \alpha n/100 \rfloor, n)$ cases with the smallest distances where the approximate percentage of cases used is $\alpha \in \{1, 2.5, 5, 10, 20, 33, 50\}$. (Here $\lfloor x \rfloor$ is the greatest integer function so $\lfloor 7.7 \rfloor = 7$. The extra $p + 3$ cases are added so that OLS can be computed for small n and α.) This yields seven OLS fits corresponding to the cases with predictors closest to $\boldsymbol{x}_j$. A fixed number of K cases are selected at random without replacement to use as the $\boldsymbol{x}_j$. Hence $7K$ OLS fits are generated. We use $K = 7$ as the default. A robust criterion Q is used to evaluate the $7K$ fits and the OLS fit to all of the data. Hence $7K + 1$ OLS fits are generated and the MBA estimator is the fit that minimizes the criterion. The median squared residual is a good choice for Q.

Three ideas motivate this estimator. First, $\boldsymbol{x}$-outliers, which are outliers in the predictor space, tend to be much more destructive than Y-outliers which are outliers in the response variable. Suppose that the proportion of outliers is γ and that $\gamma < 0.5$. We would like the algorithm to have at least one "center" $\boldsymbol{x}_j$ that is not an outlier. The probability of drawing a center that is not an outlier is approximately $1 - \gamma^K > 0.99$ for $K \geq 7$ and this result is free of p. Secondly, by using the different percentages of coverages, for many data sets there will be a center and a coverage that contains no outliers. Third, by Theorem 2.28, the MBA estimator is a $\sqrt{n}$ consistent estimator of the same parameter vector $\boldsymbol{\beta}$ estimated by OLS under mild conditions.

Ellipsoidal trimming can be used to create resistant multiple linear regression (MLR) estimators. To perform ellipsoidal trimming, an estimator $(T, \boldsymbol{C})$ is computed and used to create the squared Mahalanobis distances D_i^2 for each vector of observed predictors $\boldsymbol{x}_i$. If the ordered distance $D_{(j)}$ is unique, then j of the $\boldsymbol{x}_i$'s are in the ellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \leq D_{(j)}^2\}. \tag{8.44}$$

The ith case $(Y_i, \boldsymbol{x}_i^T)^T$ is trimmed if $D_i > D_{(j)}$. Then an estimator of $\boldsymbol{\beta}$ is computed from the remaining cases. For example, if $j \approx 0.9n$, then about 10% of the cases are trimmed, and OLS or L_1 could be used on the cases that remain. Ellipsoidal trimming differs from using the RFCH, RMVN, or `covmb2` set since these sets use a random amount of trimming. (The ellipsoidal trimming technique can also be used for other regression models, and the theory of the regression method tends to apply to the method applied to

the cleaned data that was not trimmed since the response variables were not used to select the cases.)

Use ellipsoidal trimming on the RFCH, RMVN, or `covmb2` set applied to the continuous predictors to get a fit $\hat{\boldsymbol{\beta}}_C$. Then make a response and residual plot using all of the data, not just the cleaned data that was not trimmed.

The resistant trimmed views estimator combines ellipsoidal trimming and the response plot. First compute $(T, \boldsymbol{C})$ on the $\boldsymbol{x}_i$, perhaps using the RMVN estimator. Trim the $M\%$ of the cases with the largest Mahalanobis distances, and then compute the MLR estimator $\hat{\boldsymbol{\beta}}_M$ from the remaining cases. Use $M = 0$, 10, 20, 30, 40, 50, 60, 70, 80, and 90 to generate ten response plots of the fitted values $\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}_i$ versus Y_i using all n cases. (Fewer plots are used for small data sets if $\hat{\boldsymbol{\beta}}_M$ can not be computed for large M.) These plots are called "trimmed views."

Definition 8.37. The trimmed views (TV) estimator $\hat{\boldsymbol{\beta}}_{T,n}$ corresponds to the trimmed view where the bulk of the plotted points follow the identity line with smallest variance function, ignoring any outliers.

Example 8.10. For the Buxton (1920) data, *height* was the response variable while an intercept, *head length, nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five individuals, cases 61–65, were reported to be about 0.75 inches tall with head lengths well over five feet! OLS was used on the cases remaining after trimming, and Figure 7.18 shows four trimmed views corresponding to 90%, 70%, 40%, and 0% trimming. The OLS TV estimator used 70% trimming since this trimmed view was best. Since the vertical distance from a plotted point to the identity line is equal to the case's residual, the outliers had massive residuals for 90%, 70%, and 40% trimming. Notice that the OLS trimmed view with 0% trimming "passed through the outliers" since the cluster of outliers is scattered about the identity line.

The TV estimator $\hat{\boldsymbol{\beta}}_{T,n}$ has good statistical properties if an estimator with good statistical properties is applied to the cases $(\boldsymbol{X}_{M,n}, \boldsymbol{Y}_{M,n})$ that remain after trimming. Candidates include OLS, L_1, Huber's M–estimator, Mallows' GM–estimator, or the Wilcoxon rank estimator. See Rousseeuw and Leroy (1987, pp. 12-13, 150). The basic idea is that if an estimator with $O_P(n^{-1/2})$ convergence rate is applied to a set of $n_M \propto n$ cases, then the resulting estimator $\hat{\boldsymbol{\beta}}_{M,n}$ also has $O_P(n^{-1/2})$ rate provided that the response Y was not used to select the n_M cases in the set. If $\|\hat{\boldsymbol{\beta}}_{M,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$ for $M = 0, ..., 90$ then $\|\hat{\boldsymbol{\beta}}_{T,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$ by Theorem 2.28.

Let $\boldsymbol{X}_n = \boldsymbol{X}_{0,n}$ denote the full design matrix. Often when proving asymptotic normality of an MLR estimator $\hat{\boldsymbol{\beta}}_{0,n}$, it is assumed that

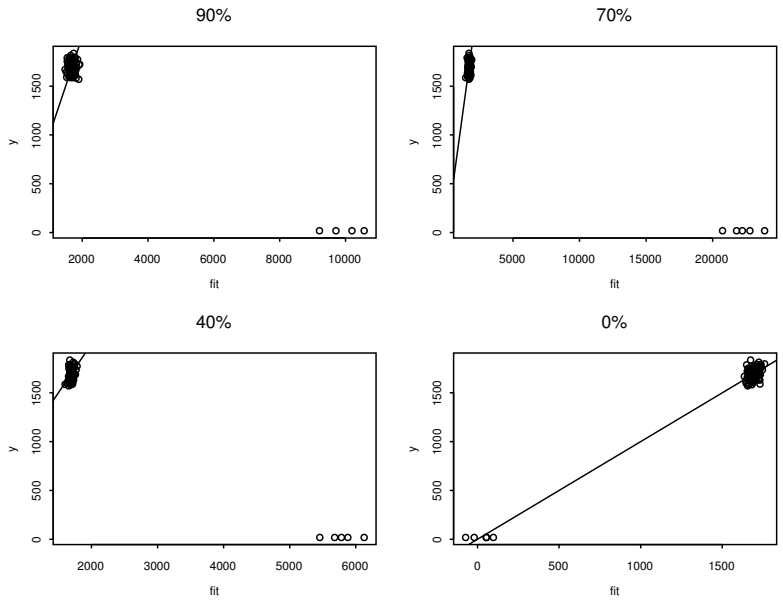

Fig. 8.1 4 Trimmed Views for the Buxton Data

$$\frac{\boldsymbol{X}_n^T \boldsymbol{X}_n}{n} \to \boldsymbol{W}^{-1}.$$

If $\hat{\boldsymbol{\beta}}_{0,n}$ has $O_P(n^{-1/2})$ rate and if for big enough n all of the diagonal elements of

$$\left(\frac{\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n}}{n} \right)^{-1}$$

are all contained in an interval $[0, B)$ for some $B > 0$, then $\|\hat{\boldsymbol{\beta}}_{M,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$.

The distribution of the estimator $\hat{\boldsymbol{\beta}}_{M,n}$ is especially simple when OLS is used and the errors are iid $N(0, \sigma^2)$. Then

$$\hat{\boldsymbol{\beta}}_{M,n} = (\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n})^{-1} \boldsymbol{X}_{M,n}^T \boldsymbol{Y}_{M,n} \sim N_p(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n})^{-1})$$

and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{M,n} - \boldsymbol{\beta}) \sim N_p(\boldsymbol{0}, \sigma^2 (\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n}/n)^{-1})$. This result does not imply that $\hat{\boldsymbol{\beta}}_{T,n}$ is asymptotically normal.

Warning: When $Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e$, MLR estimators tend to estimate the same slopes $\beta_2, ..., \beta_p$, but the constant β_1 tends to depend on the estimator unless the errors are symmetric. The MBA and trimmed views estimators do

estimate the same $\boldsymbol{\beta}$ as OLS asymptotically, but samples may need to be huge before the MBA and trimmed views estimates of the constant are close to the OLS estimate of the constant. See Olive (2017b, p. 444) for an explanation for why large sample sizes may be needed to estimate the constant.

Often practical "robust estimators" generate a sequence of K trial fits called *attractors*: $\boldsymbol{b}_1, ..., \boldsymbol{b}_K$. Then some criterion is evaluated and the attractor $\boldsymbol{b}_A$ that minimizes the criterion is used in the final estimator.

Definition 8.38. For MLR, an *elemental set* J is a set of p cases drawn with replacement from the data set of n cases. The elemental fit is the OLS estimator $\hat{\boldsymbol{\beta}}_{J_i} = (\boldsymbol{X}_{J_i}^T \boldsymbol{X}_{J_i})^{-1} \boldsymbol{X}_{J_i}^T \boldsymbol{Y}_{J_i} = \boldsymbol{X}_{J_i}^{-1} \boldsymbol{Y}_{J_i}$ applied to the cases corresponding to the elemental set provided that the inverse of $\boldsymbol{X}_{J_i}$ exists. In a *concentration algorithm*, let $\boldsymbol{b}_{0,j}$ be the jth start, not necessarily elemental, and compute all n residuals $r_i(\boldsymbol{b}_{0,j}) = Y_i - \boldsymbol{x}_i^T \boldsymbol{b}_{0,j}$. At the next iteration, the OLS estimator $\boldsymbol{b}_{1,j}$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest squared residuals $r_i^2(\boldsymbol{b}_{0,j})$. This iteration can be continued for k steps resulting in the sequence of estimators $\boldsymbol{b}_{0,j}, \boldsymbol{b}_{1,j}, ..., \boldsymbol{b}_{k,j}$. Then $\boldsymbol{b}_{k,j}$ is the jth *attractor* for $j = 1, ..., K$. Then the attractor $\boldsymbol{b}_A$ that minimizes the LTS criterion is used in the final estimator. Using $k = 10$ concentration steps often works well, and the basic resampling algorithm is a special case with $k = 0$, i.e., the attractors are the starts. Such an algorithm is called a CLTS concentration algorithm or CLTS.

Remark 8.9. Consider drawing K elemental sets $J_1, ..., J_K$ with replacement to use as starts. For multivariate location and dispersion, use the attractor with the smallest MCD criterion to get the final estimator. For multiple linear regression, use the attractor with the smallest LMS, LTA, or LTS criterion to get the final estimator. For $500 \leq K \leq 3000$ and p not much larger than 5, the elemental set algorithm is very good for detecting certain "outlier configurations," including i) a mixture of two regression hyperplanes that cross in the center of the data cloud for MLR (not an outlier configuration since outliers are far from the bulk of the data) and ii) a cluster of outliers that can often be placed close enough to the bulk of the data so that an MB, RFCH, or RMVN DD plot can not detect the outliers. However, the outlier resistance of elemental algorithms decreases rapidly as p increases.

Suppose the data set has n cases where d are outliers and $n-d$ are "clean" (not outliers). The the outlier proportion $\gamma = d/n$. Suppose that K elemental sets are chosen with replacement and that it is desired to find K such that the probability P(that at least one of the elemental sets is clean) $\equiv P_1 \approx 1-\alpha$ where $\alpha = 0.05$ is a common choice. Then $P_1 = 1-$ P(none of the K elemental sets is clean) $\approx 1-[1-(1-\gamma)^p]^K$ by independence. Hence $\alpha \approx [1-(1-\gamma)^p]^K$ or

$$K \approx \frac{\log(\alpha)}{\log([1-(1-\gamma)^p])} \approx \frac{\log(\alpha)}{-(1-\gamma)^p} \tag{8.45}$$

using the approximation $\log(1-x) \approx -x$ for small x. Since $\log(0.05) \approx -3$, if $\alpha = 0.05$, then $K \approx \dfrac{3}{(1-\gamma)^p}$. Frequently a clean subset is wanted even if the contamination proportion $\gamma \approx 0.5$. Then for a 95% chance of obtaining at least one clean elemental set, $K \approx 3\ (2^p)$ elemental sets need to be drawn. If the start passes through an outlier, so does the attractor. For concentration algorithms for multivariate location and dispersion, if the start passes through a cluster of outliers, sometimes the attractor would be clean. See Olive (2017b: pp. 114-117).

Notice that the number of subsets K needed to obtain a clean elemental set with high probability is an exponential function of the number of predictors p but is free of n. Hawkins and Olive (2002) showed that if K is fixed and free of n, then the resulting elemental or concentration algorithm (that uses k concentration steps), is inconsistent and zero breakdown. See Theorem 8.39. Nevertheless, many practical estimators tend to use a value of K that is free of both n and p (e.g. $K = 500$ or $K = 3000$). Such algorithms include ALMS = FLMS = `lmsreg` and ALTS = FLTS = `ltsreg`. The "A" denotes that an algorithm was used. The "F" means that a fixed number of trial fits (K elemental fits) was used and the criterion (LMS or LTS) was used to select the trial fit used in the final estimator.

To examine the outlier resistance of such inconsistent zero breakdown estimators, fix both K and the contamination proportion γ and then find the largest number of predictors p that can be in the model such that the probability of finding at least one clean elemental set is high. Given K and γ, P(at least one of K subsamples is clean) $= 0.95 \approx$
$1 - [1 - (1-\gamma)^p]^K$. Thus the largest value of p satisfies $\dfrac{3}{(1-\gamma)^p} \approx K$, or

$$p \approx \left\lfloor \frac{\log(3/K)}{\log(1-\gamma)} \right\rfloor \tag{8.46}$$

if the sample size n is very large. Again $\lfloor x \rfloor$ is the greatest integer function: $\lfloor 7.7 \rfloor = 7$.

Theorem 8.32. Let $h = p$ be the number of randomly selected cases in an elemental set, and let γ_o be the highest percentage of massive outliers that a resampling algorithm can detect reliably. If n is large, then

$$\gamma_o \approx \min\left(\frac{n-c}{n}, 1 - [1 - (0.2)^{1/K}]^{1/h}\right) 100\%. \tag{8.47}$$

Proof. As in Remark 8.5, if the contamination proportion γ is fixed, then the probability of obtaining at least one clean subset of size h with high probability (say $1 - \alpha = 0.8$) is given by $0.8 = 1 - [1 - (1-\gamma)^h]^K$. Fix the number of starts K and solve this equation for γ. □

The value of γ_o depends on $c \geq n/2$ and h. To maximize γ_o, take $c \approx n/2$ and $h = p$. For example, with $K = 500$ starts, $n > 100$, and $h = p \leq 20$ the resampling algorithm should be able to detect up to 24% outliers provided every clean start is able to at least partially separate inliers (clean cases) from outliers. However, if $h = p = 50$, this proportion drops to 11%.

8.4 Robust Regression

This section will consider the breakdown of a regression estimator and then develop the practical high breakdown `hbreg` estimator.

8.4.1 MLR Breakdown and Equivariance

Breakdown and equivariance properties have received considerable attention in the literature. Several of these properties involve transformations of the data, and are discussed below. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the original data, then the vector of the coefficient estimates is

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) = T(\boldsymbol{X}, \boldsymbol{Y}), \tag{8.48}$$

the vector of predicted values is

$$\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{Y}}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}), \tag{8.49}$$

and the vector of residuals is

$$\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}. \tag{8.50}$$

If the design matrix $\boldsymbol{X}$ is transformed into $\boldsymbol{W}$ and the vector of dependent variables $\boldsymbol{Y}$ is transformed into $\boldsymbol{Z}$, then $(\boldsymbol{W}, \boldsymbol{Z})$ is the new data set.

Definition 8.39. Regression Equivariance: Let $\boldsymbol{u}$ be any $p \times 1$ vector. Then $\widehat{\boldsymbol{\beta}}$ is regression equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{u}) = T(\boldsymbol{X}, \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{u}) = T(\boldsymbol{X}, \boldsymbol{Y}) + \boldsymbol{u} = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) + \boldsymbol{u}. \tag{8.51}$$

Hence if $\boldsymbol{W} = \boldsymbol{X}$ and $\boldsymbol{Z} = \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{u}$, then $\widehat{\boldsymbol{Z}} = \widehat{\boldsymbol{Y}} + \boldsymbol{X}\boldsymbol{u}$ and $\boldsymbol{r}(\boldsymbol{W}, \boldsymbol{Z}) = \boldsymbol{Z} - \widehat{\boldsymbol{Z}} = \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y})$. Note that the residuals are invariant under this type of transformation, and note that if $\boldsymbol{u} = -\widehat{\boldsymbol{\beta}}$, then regression equivariance implies that we should not find any linear structure if we regress the residuals on $\boldsymbol{X}$.

Definition 8.40. Scale Equivariance: Let c be any scalar. Then $\widehat{\boldsymbol{\beta}}$ is scale equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, c\boldsymbol{Y}) = T(\boldsymbol{X}, c\boldsymbol{Y}) = cT(\boldsymbol{X}, \boldsymbol{Y}) = c\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \tag{8.52}$$

Hence if $\boldsymbol{W} = \boldsymbol{X}$ and $\boldsymbol{Z} = c\boldsymbol{Y}$, then $\widehat{\boldsymbol{Z}} = c\widehat{\boldsymbol{Y}}$ and $\boldsymbol{r}(\boldsymbol{X}, c\boldsymbol{Y}) = c\ \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y})$. Scale equivariance implies that if the Y_i's are stretched, then the fits and the residuals should be stretched by the same factor.

Definition 8.41. Affine Equivariance: Let $\boldsymbol{A}$ be any $p \times p$ nonsingular matrix. Then $\widehat{\boldsymbol{\beta}}$ is affine equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = T(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = \boldsymbol{A}^{-1}T(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \tag{8.53}$$

Hence if $\boldsymbol{W} = \boldsymbol{X}\boldsymbol{A}$ and $\boldsymbol{Z} = \boldsymbol{Y}$, then $\widehat{\boldsymbol{Z}} = \boldsymbol{W}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) =$ $\boldsymbol{X}\boldsymbol{A}\boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) = \widehat{\boldsymbol{Y}}$, and $\boldsymbol{r}(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = \boldsymbol{Z} - \widehat{\boldsymbol{Z}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y})$. Note that both the predicted values and the residuals are invariant under an affine transformation of the predictor variables.

Definition 8.42. Permutation Invariance: Let $\boldsymbol{P}$ be an $n \times n$ permutation matrix. Then $\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{I}_n$ where $\boldsymbol{I}_n$ is an $n \times n$ identity matrix and the superscript T denotes the transpose of a matrix. Then $\widehat{\boldsymbol{\beta}}$ is permutation invariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{P}\boldsymbol{X}, \boldsymbol{P}\boldsymbol{Y}) = T(\boldsymbol{P}\boldsymbol{X}, \boldsymbol{P}\boldsymbol{Y}) = T(\boldsymbol{X}, \boldsymbol{Y}) = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \tag{8.54}$$

Hence if $\boldsymbol{W} = \boldsymbol{P}\boldsymbol{X}$ and $\boldsymbol{Z} = \boldsymbol{P}\boldsymbol{Y}$, then $\widehat{\boldsymbol{Z}} = \boldsymbol{P}\widehat{\boldsymbol{Y}}$ and $\boldsymbol{r}(\boldsymbol{P}\boldsymbol{X}, \boldsymbol{P}\boldsymbol{Y}) =$ $\boldsymbol{P}\ \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y})$. If an estimator is not permutation invariant, then swapping rows of the $n \times (p+1)$ augmented matrix $(\boldsymbol{X}, \boldsymbol{Y})$ will change the estimator. Hence the case number is important. If the estimator is permutation invariant, then the position of the case in the data cloud is of primary importance. Resampling algorithms are not permutation invariant because permuting the data causes different subsamples to be drawn.

Remark 8.10. OLS has the above invariance properties, but most Statistical Learning alternatives such as lasso and ridge regression do not have all four properties. Hence Remark 6.2 is used to fit the data with $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. Then obtain $\hat{\boldsymbol{\beta}}$ from $\hat{\boldsymbol{\eta}}$.

The remainder of this subsection gives a standard definition of breakdown and then shows that if the median absolute residual is bounded in the presence of high contamination, then the regression estimator has a high breakdown value. The following notation will be useful. Let $\boldsymbol{W}$ denote the data matrix where the ith row corresponds to the ith case. For regression, $\boldsymbol{W}$ is the $n \times (p+1)$ matrix with ith row $(\boldsymbol{x}_i^T, Y_i)$. Let $\boldsymbol{W}_d^n$ denote the data matrix where any d_n of the cases have been replaced by arbitrarily bad contaminated

cases. Then the contamination fraction is $\gamma \equiv \gamma_n = d_n/n$, and the breakdown value of $\hat{\boldsymbol{\beta}}$ is the smallest value of γ_n needed to make $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large.

Definition 8.43. Let $1 \leq d_n \leq n$. If $T(\boldsymbol{W})$ is a $p \times 1$ vector of regression coefficients, then the *breakdown value* of T is

$$B(T, \boldsymbol{W}) = \min\left\{ \frac{d_n}{n} : \sup_{\boldsymbol{W}_d^n} \|T(\boldsymbol{W}_d^n)\| = \infty \right\}$$

where the supremum is over all possible corrupted samples $\boldsymbol{W}_d^n$.

Definition 8.44. *High breakdown* regression estimators have $\gamma_n \to 0.5$ as $n \to \infty$ if the clean (uncontaminated) data are in *general position*: any p clean cases give a unique estimate of $\boldsymbol{\beta}$. Estimators are *zero breakdown* if $\gamma_n \to 0$ and *positive breakdown* if $\gamma_n \to \gamma > 0$ as $n \to \infty$.

The following result greatly simplifies some breakdown proofs and shows that a regression estimator basically breaks down if the median absolute residual MED$(|r_i|)$ can be made arbitrarily large. The result implies that if the breakdown value ≤ 0.5, breakdown can be computed using the median absolute residual MED$(|r_i|(\boldsymbol{W}_d^n))$ instead of $\|T(\boldsymbol{W}_d^n)\|$. Similarly $\hat{\boldsymbol{\beta}}$ is high breakdown if the median squared residual or the c_nth largest absolute residual $|r_i|_{(c_n)}$ or squared residual $r^2_{(c_n)}$ stay bounded under high contamination where $c_n \approx n/2$. Note that $\|\hat{\boldsymbol{\beta}}\| \equiv \|\hat{\boldsymbol{\beta}}(\boldsymbol{W}_d^n)\| \leq M$ for some constant M that depends on T and $\boldsymbol{W}$ but not on the outliers if the number of outliers d_n is less than the smallest number of outliers needed to cause breakdown.

Theorem 8.33. If the breakdown value ≤ 0.5, computing the breakdown value using the median absolute residual MED$(|r_i|(\boldsymbol{W}_d^n))$ instead of $\|T(\boldsymbol{W}_d^n)\|$ is asymptotically equivalent to using Definition 8.43.

Proof. Consider any contaminated data set $\boldsymbol{W}_d^n$ with ith row $(\boldsymbol{w}_i^T, Z_i)^T$. If the regression estimator $T(\boldsymbol{W}_d^n) = \hat{\boldsymbol{\beta}}$ satisfies $\|\hat{\boldsymbol{\beta}}\| \leq M$ for some constant M if $d < d_n$, then the median absolute residual MED$(|Z_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{w}_i|)$ is bounded by $\max_{i=1,\ldots,n} |Y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i| \leq \max_{i=1,\ldots,n}[|Y_i| + \sum_{j=1}^p M|x_{i,j}|]$ if $d_n < n/2$.

If the median absolute residual is bounded by M when $d < d_n$, then $\|\hat{\boldsymbol{\beta}}\|$ is bounded provided fewer than half of the cases line on the hyperplane (and so have absolute residual of 0), as shown next. Now suppose that $\|\hat{\boldsymbol{\beta}}\| = \infty$. Since the absolute residual is the vertical distance of the observation from the hyperplane, the absolute residual $|r_i| = 0$ if the ith case lies on the regression hyperplane, but $|r_i| = \infty$ otherwise. Hence MED$(|r_i|) = \infty$ if fewer than half of the cases lie on the regression hyperplane. This will occur unless the proportion of outliers $d_n/n > (n/2 - q)/n \to 0.5$ as $n \to \infty$ where q is the number of "good" cases that lie on a hyperplane of lower dimension than p.

In the literature it is usually assumed that the original data are in *general position*: $q = p - 1$. □

Suppose that the clean data are in general position and that the number of outliers is less than the number needed to make the median absolute residual and $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large. If the $\boldsymbol{x}_i$ are fixed, and the outliers are moved up and down by adding a large positive or negative constant to the Y values of the outliers, then for high breakdown (HB) estimators, $\hat{\boldsymbol{\beta}}$ and MED$(|r_i|)$ stay bounded where the bounds depend on the clean data $\boldsymbol{W}$ but not on the outliers even if the number of outliers is nearly as large as $n/2$. Thus if the $|Y_i|$ values of the outliers are large enough, the $|r_i|$ values of the outliers will be large.

If the Y_i's are fixed, arbitrarily large $\boldsymbol{x}$-outliers tend to drive the slope estimates to 0, not ∞. If both $\boldsymbol{x}$ and Y can be varied, then a cluster of outliers can be moved arbitrarily far from the bulk of the data but may still have small residuals. For example, move the outliers along the regression hyperplane formed by the clean cases.

If the $(\boldsymbol{x}_i^T, Y_i)$ are in general position, then the contamination could be such that $\hat{\boldsymbol{\beta}}$ passes exactly through $p - 1$ "clean" cases and d_n "contaminated" cases. Hence $d_n + p - 1$ cases could have absolute residuals equal to zero with $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large (but finite). Nevertheless, if T possesses reasonable equivariant properties and $\|T(\boldsymbol{W}_d^n)\|$ is replaced by the median absolute residual in the definition of breakdown, then the two breakdown values are asymptotically equivalent. (If $T(\boldsymbol{W}) \equiv \boldsymbol{0}$, then T is neither regression nor affine equivariant. The breakdown value of T is one, but the median absolute residual can be made arbitrarily large if the contamination proportion is greater than $n/2$.)

If the Y_i's are fixed, arbitrarily large $\boldsymbol{x}$-outliers will rarely drive $\|\hat{\boldsymbol{\beta}}\|$ to ∞. The $\boldsymbol{x}$-outliers can drive $\|\hat{\boldsymbol{\beta}}\|$ to ∞ if they can be constructed so that the estimator is no longer defined, e.g. so that $\boldsymbol{X}^T\boldsymbol{X}$ is nearly singular. The examples following some results on norms may help illustrate these points.

Definition 8.45. Let $\boldsymbol{y}$ be an $n \times 1$ vector. Then $\|\boldsymbol{y}\|$ is a *vector norm* if
vn1) $\|\boldsymbol{y}\| \geq 0$ for every $\boldsymbol{y} \in \mathbb{R}^n$ with equality iff $\boldsymbol{y}$ is the zero vector,
vn2) $\|a\boldsymbol{y}\| = |a| \, \|\boldsymbol{y}\|$ for all $\boldsymbol{y} \in \mathbb{R}^n$ and for all scalars a, and
vn3) $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^n$.

Definition 8.46. Let $\boldsymbol{G}$ be an $n \times p$ matrix. Then $\|\boldsymbol{G}\|$ is a *matrix norm* if
mn1) $\|\boldsymbol{G}\| \geq 0$ for every $n \times p$ matrix $\boldsymbol{G}$ with equality iff $\boldsymbol{G}$ is the zero matrix,
mn2) $\|a\boldsymbol{G}\| = |a| \, \|\boldsymbol{G}\|$ for all scalars a, and
mn3) $\|\boldsymbol{G} + \boldsymbol{H}\| \leq \|\boldsymbol{G}\| + \|\boldsymbol{H}\|$ for all $n \times p$ matrices $\boldsymbol{G}$ and $\boldsymbol{H}$.

Example 8.11. The *q-norm* of a vector $\boldsymbol{y}$ is $\|\boldsymbol{y}\|_q = (|y_1|^q + \cdots + |y_n|^q)^{1/q}$. In particular, $\|\boldsymbol{y}\|_1 = |y_1| + \cdots + |y_n|$, the *Euclidean norm* $\|\boldsymbol{y}\|_2 = \sqrt{y_1^2 + \cdots + y_n^2}$, and $\|\boldsymbol{y}\|_\infty = \max_i \ |y_i|$. Given a matrix $\boldsymbol{G}$ and

a vector norm $\|\boldsymbol{y}\|_q$ the *q-norm* or *subordinate matrix norm* of matrix $\boldsymbol{G}$ is $\|\boldsymbol{G}\|_q = \max_{\boldsymbol{y} \neq \boldsymbol{0}} \frac{\|\boldsymbol{G}\boldsymbol{y}\|_q}{\|\boldsymbol{y}\|_q}$. It can be shown that the *maximum column sum norm* $\|\boldsymbol{G}\|_1 = \max_{1 \leq j \leq p} \sum_{i=1}^{n} |g_{ij}|$, the *maximum row sum norm* $\|\boldsymbol{G}\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{p} |g_{ij}|$, and the *spectral norm* $\|\boldsymbol{G}\|_2 = \sqrt{\text{maximum eigenvalue of } \boldsymbol{G}^T\boldsymbol{G}}$. The *Frobenius norm*

$$\|\boldsymbol{G}\|_F = \sqrt{\sum_{j=1}^{p} \sum_{i=1}^{n} |g_{ij}|^2} = \sqrt{\text{trace}(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})}.$$

Several useful results involving matrix norms will be used. First, for any subordinate matrix norm, $\|\boldsymbol{G}\boldsymbol{y}\|_q \leq \|\boldsymbol{G}\|_q \, \|\boldsymbol{y}\|_q$. Let $J = J_m = \{m_1, ..., m_p\}$ denote the p cases in the mth elemental fit $\boldsymbol{b}_J = \boldsymbol{X}_J^{-1}\boldsymbol{Y}_J$. Then for any elemental fit $\boldsymbol{b}_J$ (suppressing $q = 2$),

$$\|\boldsymbol{b}_J - \boldsymbol{\beta}\| = \|\boldsymbol{X}_J^{-1}(\boldsymbol{X}_J\boldsymbol{\beta} + \boldsymbol{e}_J) - \boldsymbol{\beta}\| = \|\boldsymbol{X}_J^{-1}\boldsymbol{e}_J\| \leq \|\boldsymbol{X}_J^{-1}\| \, \|\boldsymbol{e}_J\|. \quad (8.55)$$

The following results (Golub and Van Loan 1989, pp. 57, 80) on the Euclidean norm are useful. Let $0 \leq \sigma_p \leq \sigma_{p-1} \leq \cdots \leq \sigma_1$ denote the singular values of $\boldsymbol{X}_J = (x_{mi,j})$. Then

$$\|\boldsymbol{X}_J^{-1}\| = \frac{\sigma_1}{\sigma_p \|\boldsymbol{X}_J\|}, \quad (8.56)$$

$$\max_{i,j} |x_{mi,j}| \leq \|\boldsymbol{X}_J\| \leq p \, \max_{i,j} |x_{mi,j}|, \text{ and} \quad (8.57)$$

$$\frac{1}{p \, \max_{i,j} |x_{mi,j}|} \leq \frac{1}{\|\boldsymbol{X}_J\|} \leq \|\boldsymbol{X}_J^{-1}\|. \quad (8.58)$$

From now on, unless otherwise stated, we will use the spectral norm as the matrix norm and the Euclidean norm as the vector norm.

Example 8.12. Suppose the response values Y are near 0. Consider the fit from an elemental set: $\boldsymbol{b}_J = \boldsymbol{X}_J^{-1}\boldsymbol{Y}_J$ and examine Equations (8.56), (8.57), and (8.58). Now $\|\boldsymbol{b}_J\| \leq \|\boldsymbol{X}_J^{-1}\| \, \|\boldsymbol{Y}_J\|$, and *since x-outliers make $\|\boldsymbol{X}_J\|$ large, x-outliers tend to drive $\|\boldsymbol{X}_J^{-1}\|$ and $\|\boldsymbol{b}_J\|$ towards zero not towards ∞.* The x-outliers may make $\|\boldsymbol{b}_J\|$ large if they can make the trial design $\|\boldsymbol{X}_J\|$ nearly singular. Notice that Euclidean norm $\|\boldsymbol{b}_J\|$ can easily be made large if one or more of the elemental response variables is driven far away from zero.

Example 8.13. Without loss of generality, assume that the clean Y's are contained in an interval $[a, f]$ for some a and f. Assume that the regression

model contains an intercept β_1. Then there exists an estimator $\hat{\boldsymbol{\beta}}_M$ of $\boldsymbol{\beta}$ such that $\|\hat{\boldsymbol{\beta}}_M\| \leq \max(|a|, |f|)$ if $d_n < n/2$.

Proof. Let $\text{MED}(n) = \text{MED}(Y_1, ..., Y_n)$ and $\text{MAD}(n) = \text{MAD}(Y_1, ..., Y_n)$. Take $\hat{\boldsymbol{\beta}}_M = (\text{MED}(n), 0, ..., 0)^T$. Then $\|\hat{\boldsymbol{\beta}}_M\| = |\text{MED}(n)| \leq \max(|a|, |f|)$. Note that the median absolute residual for the fit $\hat{\boldsymbol{\beta}}_M$ is equal to the median absolute deviation $\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|, i = 1, ..., n) \leq f - a$ if $d_n < \lfloor (n+1)/2 \rfloor$. □

Note that $\hat{\boldsymbol{\beta}}_M$ is a poor high breakdown estimator of $\boldsymbol{\beta}$ and $\hat{Y}_i(\hat{\boldsymbol{\beta}}_M)$ tracks the Y_i very poorly. If the data are in general position, a high breakdown regression estimator is an estimator which has a bounded median absolute residual even when close to half of the observations are arbitrary. Rousseeuw and Leroy (1987, pp. 29, 206) conjectured that high breakdown regression estimators can not be computed cheaply, and that if the algorithm is also affine equivariant, then the complexity of the algorithm must be at least $O(n^p)$. The following theorem shows that these two conjectures are false.

Theorem 8.34. If the clean data are in general position and the model has an intercept, then a scale and affine equivariant high breakdown estimator $\hat{\boldsymbol{\beta}}_w$ can be found by computing OLS on the set of cases that have $Y_i \in [\text{MED}(Y_1, ..., Y_n) \pm w \text{ MAD}(Y_1, ..., Y_n)]$ where $w \geq 1$ (so at least half of the cases are used).

Proof. Note that $\hat{\boldsymbol{\beta}}_w$ is obtained by computing OLS on the set J of the n_j cases which have

$$Y_i \in [\text{MED}(Y_1, ..., Y_n) \pm w\text{MAD}(Y_1, ..., Y_n)] \equiv [\text{MED}(n) \pm w\text{MAD}(n)]$$

where $w \geq 1$ (to guarantee that $n_j \geq n/2$). Consider the estimator $\hat{\boldsymbol{\beta}}_M = (\text{MED}(n), 0, ..., 0)^T$ which yields the predicted values $\hat{Y}_i \equiv \text{MED}(n)$. The squared residual $r_i^2(\hat{\boldsymbol{\beta}}_M) \leq (w \text{ MAD}(n))^2$ if the ith case is in J. Hence the weighted LS fit $\hat{\boldsymbol{\beta}}_w$ is the OLS fit to the cases in J and has

$$\sum_{i \in J} r_i^2(\hat{\boldsymbol{\beta}}_w) \leq n_j (w \text{ MAD}(n))^2.$$

Thus

$$\text{MED}(|r_1(\hat{\boldsymbol{\beta}}_w)|, ..., |r_n(\hat{\boldsymbol{\beta}}_w)|) \leq \sqrt{n_j}\ w \text{ MAD}(n) < \sqrt{n}\ w \text{ MAD}(n) < \infty.$$

Thus the estimator $\hat{\boldsymbol{\beta}}_w$ has a median absolute residual bounded by $\sqrt{n}\ w \text{ MAD}(Y_1, ..., Y_n)$. Hence $\hat{\boldsymbol{\beta}}_w$ is high breakdown, and it is affine equivariant since the design is not used to choose the observations. It is scale equivariant since for constant $c = 0$, $\hat{\boldsymbol{\beta}}_w = \mathbf{0}$, and for $c \neq 0$ the set of

cases used remains the same under scale transformations and OLS is scale equivariant. □

Note that if w is huge and $\text{MAD}(n) \neq 0$, then the high breakdown estimator $\hat{\boldsymbol{\beta}}_w$ and $\hat{\boldsymbol{\beta}}_{OLS}$ will be the same for most data sets. Thus high breakdown estimators can be very nonrobust. Even if $w = 1$, the HB estimator $\hat{\boldsymbol{\beta}}_w$ only resists large Y outliers.

An ALTA concentration algorithm uses the L_1 estimator instead of OLS in the concentration step and uses the LTA criterion. Similarly an ALMS concentration algorithm uses the L_∞ estimator and the LMS criterion.

Theorem 8.35. If the clean data are in general position and if a high breakdown start is added to an ALTA, ALTS, or ALMS concentration algorithm, then the resulting estimator is HB.

Proof. Concentration reduces (or does not increase) the corresponding HB criterion that is based on $c_n \geq n/2$ absolute residuals, so the median absolute residual of the resulting estimator is bounded as long as the criterion applied to the HB estimator is bounded. □

For example, consider the LTS(c_n) criterion. Suppose the ordered squared residuals from the high breakdown mth start $\boldsymbol{b}_{0m}$ are obtained. If the data are in general position, then $Q_{LTS}(\boldsymbol{b}_{0m})$ is bounded even if the number of outliers d_n is nearly as large as $n/2$. Then $\boldsymbol{b}_{1m}$ is simply the OLS fit to the cases corresponding to the c_n smallest squared residuals $r^2_{(i)}(\boldsymbol{b}_{0m})$ for $i = 1, ..., c_n$. Denote these cases by $i_1, ..., i_{c_n}$. Then $Q_{LTS}(\boldsymbol{b}_{1m}) =$

$$\sum_{i=1}^{c_n} r^2_{(i)}(\boldsymbol{b}_{1m}) \leq \sum_{j=1}^{c_n} r^2_{i_j}(\boldsymbol{b}_{1m}) \leq \sum_{j=1}^{c_n} r^2_{i_j}(\boldsymbol{b}_{0m}) = \sum_{j=1}^{c_n} r^2_{(i)}(\boldsymbol{b}_{0m}) = Q_{LTS}(\boldsymbol{b}_{0m})$$

where the second inequality follows from the definition of the OLS estimator. Hence concentration steps reduce or at least do not increase the LTS criterion. If $c_n = (n+1)/2$ for n odd and $c_n = 1+n/2$ for n even, then the LTS criterion is bounded iff the median squared residual is bounded.

Theorem 8.35 can be used to show that the following two estimators are high breakdown. The estimator $\hat{\boldsymbol{\beta}}_B$ is the high breakdown attractor used by the $\sqrt{n}$ consistent high breakdown `hbreg` estimator of Definition 8.48.

Definition 8.47. Make an OLS fit to the $c_n \approx n/2$ cases whose Y values are closest to the $\text{MED}(Y_1, ..., Y_n) \equiv \text{MED}(n)$ and use this fit as the start for concentration. Define $\hat{\boldsymbol{\beta}}_B$ to be the attractor after k concentration steps. Define $\boldsymbol{b}_{k,B} = 0.9999\hat{\boldsymbol{\beta}}_B$.

Theorem 8.36. If the clean data are in general position, then $\hat{\boldsymbol{\beta}}_B$ and $\boldsymbol{b}_{k,B}$ are high breakdown regression estimators.

Proof. The start can be taken to be $\hat{\boldsymbol{\beta}}_w$ with $w = 1$ from Theorem 8.34. Since the start is high breakdown, so is the attractor $\hat{\boldsymbol{\beta}}_B$ by Theorem 8.35. Multiplying a HB estimator by a positive constant does not change the breakdown value, so $\boldsymbol{b}_{k,B}$ is HB. □

The following result shows that it is easy to make a HB estimator that is asymptotically equivalent to a consistent estimator on a large class of iid zero mean symmetric error distributions, although the outlier resistance of the HB estimator is poor. The following result may not hold if $\hat{\boldsymbol{\beta}}_C$ estimates $\boldsymbol{\beta}_C$ and $\hat{\boldsymbol{\beta}}_{LMS}$ estimates $\boldsymbol{\beta}_{LMS}$ where $\boldsymbol{\beta}_C \neq \boldsymbol{\beta}_{LMS}$. Then $\boldsymbol{b}_{k,B}$ could have a smaller median squared residual than $\hat{\boldsymbol{\beta}}_C$ even if there are no outliers. The two parameter vectors could differ because the constant term is different if the error distribution is not symmetric. For a large class of symmetric error distributions, $\boldsymbol{\beta}_{LMS} = \boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_C \equiv \boldsymbol{\beta}$, then the ratio $\text{MED}(r_i^2(\hat{\boldsymbol{\beta}}))/\text{MED}(r_i^2(\boldsymbol{\beta})) \to 1$ as $n \to \infty$ for any consistent estimator of $\boldsymbol{\beta}$. The estimator below has two attractors, $\hat{\boldsymbol{\beta}}_C$ and $\boldsymbol{b}_{k,B}$, and the probability that the final estimator $\hat{\boldsymbol{\beta}}_D$ is equal to $\hat{\boldsymbol{\beta}}_C$ goes to one under the strong assumption that the error distribution is such that both $\hat{\boldsymbol{\beta}}_C$ and $\hat{\boldsymbol{\beta}}_{LMS}$ are consistent estimators of $\boldsymbol{\beta}$.

Theorem 8.37. Assume the clean data are in general position, and that the LMS estimator is a consistent estimator of $\boldsymbol{\beta}$. Let $\hat{\boldsymbol{\beta}}_C$ be any practical consistent estimator of $\boldsymbol{\beta}$, and let $\hat{\boldsymbol{\beta}}_D = \hat{\boldsymbol{\beta}}_C$ if $\text{MED}(r_i^2(\hat{\boldsymbol{\beta}}_C)) \leq \text{MED}(r_i^2(\boldsymbol{b}_{k,B}))$. Let $\hat{\boldsymbol{\beta}}_D = \boldsymbol{b}_{k,B}$, otherwise. Then $\hat{\boldsymbol{\beta}}_D$ is a HB estimator that is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$.

Proof. The estimator is HB since the median squared residual of $\hat{\boldsymbol{\beta}}_D$ is no larger than that of the HB estimator $\boldsymbol{b}_{k,B}$. Since $\hat{\boldsymbol{\beta}}_C$ is consistent, $\text{MED}(r_i^2(\hat{\boldsymbol{\beta}}_C)) \to \text{MED}(e^2)$ in probability where $\text{MED}(e^2)$ is the population median of the squared error e^2. Since the LMS estimator is consistent, the probability that $\hat{\boldsymbol{\beta}}_C$ has a smaller median squared residual than the biased estimator $\hat{\boldsymbol{\beta}}_{k,B}$ goes to 1 as $n \to \infty$. Hence $\hat{\boldsymbol{\beta}}_D$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$. □

The elemental concentration and elemental resampling algorithms use K elemental fits where K is a fixed number that does not depend on the sample size n, e.g. $K = 500$. See Definitions 8.29 and 8.38. Note that an estimator can not be consistent for θ unless the number of randomly selected cases goes to ∞, except in degenerate situations. The following theorem shows the widely used elemental estimators are zero breakdown estimators. (If $K = K_n \to \infty$, then the elemental estimator is zero breakdown if $K_n = o(n)$. A necessary condition for the elemental basic resampling estimator to be consistent is $K_n \to \infty$.)

Theorem 8.38: a) The elemental basic resampling algorithm estimators are inconsistent. b) The elemental concentration and elemental basic resampling algorithm estimators are zero breakdown.

Proof: a) Note that you can not get a consistent estimator by using Kh randomly selected cases since the number of cases Kh needs to go to ∞ for consistency except in degenerate situations.

b) Contaminating all Kh cases in the K elemental sets shows that the breakdown value is bounded by $Kh/n \to 0$, so the estimator is zero breakdown. □

8.4.2 A Practical High Breakdown Consistent Estimator

Olive and Hawkins (2011) showed that the practical `hbreg` estimator is a high breakdown $\sqrt{n}$ consistent robust estimator that is asymptotically equivalent to the least squares estimator for many error distributions. This subsection follows Olive (2017b, pp. 420-423).

The outlier resistance of the `hbreg` estimator is not very good, but roughly comparable to the best of the practical "robust regression" estimators available in *R* packages as of 2022. The estimator is of some interest since it proved that practical high breakdown consistent estimators are possible. Other practical regression estimators that claim to be high breakdown and consistent appear to be zero breakdown because they use the zero breakdown elemental concentration algorithm. See Theorem 8.38.

The following theorem is powerful because it does not depend on the criterion used to choose the attractor. Suppose there are K consistent estimators $\hat{\boldsymbol{\beta}}_j$ of $\boldsymbol{\beta}$, each with the same rate n^δ. If $\hat{\boldsymbol{\beta}}_A$ is an estimator obtained by choosing one of the K estimators, then $\hat{\boldsymbol{\beta}}_A$ is a consistent estimator of $\boldsymbol{\beta}$ with rate n^δ by Pratt (1959). See Theorem 2.18.

Theorem 8.39. Suppose the algorithm estimator chooses an attractor as the final estimator where there are K attractors and K is fixed.

i) If all of the attractors are consistent, then the algorithm estimator is consistent.

ii) If all of the attractors are consistent with the same rate, e.g., n^δ where $0 < \delta \leq 0.5$, then the algorithm estimator is consistent with the same rate as the attractors.

iii) If all of the attractors are high breakdown, then the algorithm estimator is high breakdown.

Proof. i) Choosing from K consistent estimators results in a consistent estimator, and ii) follows from Pratt (1959). iii) Let $\gamma_{n,i}$ be the breakdown value of the ith attractor if the clean data are in general position. The breakdown value γ_n of the algorithm estimator can be no lower than that of the worst attractor: $\gamma_n \geq \min(\gamma_{n,1}, ..., \gamma_{n,K}) \to 0.5$ as $n \to \infty$. □

The consistency of the algorithm estimator changes dramatically if K is fixed but the start size $h = h_n = g(n)$ where $g(n) \to \infty$. In particular, if K starts with rate $n^{1/2}$ are used, the final estimator also has rate $n^{1/2}$. The drawback to these algorithms is that they may not have enough outlier resistance. Notice that the basic resampling result below is free of the criterion.

Theorem 8.40. Suppose $K_n \equiv K$ starts are used and that all starts have subset size $h_n = g(n) \uparrow \infty$ as $n \to \infty$. Assume that the estimator applied to the subset has rate n^δ.
i) For the h_n-set basic resampling algorithm, the algorithm estimator has rate $[g(n)]^\delta$.
ii) Under regularity conditions (e.g. given by He and Portnoy 1992), the k–step CLTS estimator has rate $[g(n)]^\delta$.

Proof. i) The $h_n = g(n)$ cases are randomly sampled without replacement. Hence the classical estimator applied to these $g(n)$ cases has rate $[g(n)]^\delta$. Thus all K starts have rate $[g(n)]^\delta$, and the result follows by Pratt (1959). ii) By He and Portnoy (1992), all K attractors have $[g(n)]^\delta$ rate, and the result follows by Pratt (1959). □

Remark 8.11. Theorem 8.33 shows that $\hat{\boldsymbol{\beta}}$ is HB if the median absolute or squared residual (or $|r(\hat{\boldsymbol{\beta}})|_{(c_n)}$ or $r^2_{(c_n)}$ where $c_n \approx n/2$) stays bounded under high contamination. Let $Q_L(\hat{\boldsymbol{\beta}}_H)$ denote the LMS, LTS, or LTA criterion for an estimator $\hat{\boldsymbol{\beta}}_H$; therefore, the estimator $\hat{\boldsymbol{\beta}}_H$ is high breakdown if and only if $Q_L(\hat{\boldsymbol{\beta}}_H)$ is bounded for d_n near $n/2$ where $d_n < n/2$ is the number of outliers. The concentration operator refines an initial estimator by successively reducing the LTS criterion. If $\hat{\boldsymbol{\beta}}_F$ refers to the final estimator (attractor) obtained by applying concentration to some starting estimator $\hat{\boldsymbol{\beta}}_H$ that is high breakdown, then since $Q_{LTS}(\hat{\boldsymbol{\beta}}_F) \leq Q_{LTS}(\hat{\boldsymbol{\beta}}_H)$, applying concentration to a high breakdown start results in a high breakdown attractor. See Theorem 8.35.

High breakdown estimators are, however, not necessarily useful for detecting outliers. Suppose $\gamma_n < 0.5$. On the one hand, if the $\boldsymbol{x}_i$ are fixed, and the outliers are moved up and down parallel to the Y axis, then for high breakdown estimators, $\hat{\boldsymbol{\beta}}$ and MED($|r_i|$) will be bounded. Thus if the $|Y_i|$ values of the outliers are large enough, the $|r_i|$ values of the outliers will be large, suggesting that the high breakdown estimator is useful for outlier detection. On the other hand, if the Y_i's are fixed at any values and the $\boldsymbol{x}$ values perturbed, sufficiently large $\boldsymbol{x}$-outliers tend to drive the slope estimates to 0, not ∞. For many estimators, including LTS, LMS, and LTA, a cluster of Y outliers can be moved arbitrarily far from the bulk of the data but still, by perturbing their $\boldsymbol{x}$ values, have arbitrarily small residuals.

Our practical high breakdown procedure is made up of three components.
1) A practical estimator $\hat{\boldsymbol{\beta}}_C$ that is consistent for clean data. Suitable choices would include the full-sample OLS and L_1 estimators.
2) A practical estimator $\hat{\boldsymbol{\beta}}_A$ that is effective for outlier identification. Suitable choices include the `mbareg`, `rmreg2`, `lmsreg`, or FLTS estimators.
3) A practical high-breakdown estimator such as $\hat{\boldsymbol{\beta}}_B$ from Definition 8.47 with $k = 10$.

By selecting one of these three estimators according to the features each of them uncovers in the data, we may inherit some of the good properties of each of them.

Definition 8.48. The `hbreg` estimator $\hat{\boldsymbol{\beta}}_H$ is defined as follows. Pick a constant $a > 1$ and set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_C$. If $aQ_L(\hat{\boldsymbol{\beta}}_A) < Q_L(\hat{\boldsymbol{\beta}}_C)$, set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_A$. If $aQ_L(\hat{\boldsymbol{\beta}}_B) < \min[Q_L(\hat{\boldsymbol{\beta}}_C), aQ_L(\hat{\boldsymbol{\beta}}_A)]$, set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_B$.

That is, find the smallest of the three scaled criterion values $Q_L(\hat{\boldsymbol{\beta}}_C)$, $aQ_L(\hat{\boldsymbol{\beta}}_A)$, $aQ_L(\hat{\boldsymbol{\beta}}_B)$. According to which of the three estimators attains this minimum, set $\hat{\boldsymbol{\beta}}_H$ to $\hat{\boldsymbol{\beta}}_C, \hat{\boldsymbol{\beta}}_A$, or $\hat{\boldsymbol{\beta}}_B$ respectively.

Large sample theory for `hbreg` is simple and given in the following theorem. Let $\hat{\boldsymbol{\beta}}_L$ be the LMS, LTS, or LTA estimator that minimizes the criterion Q_L. Note that the impractical estimator $\hat{\boldsymbol{\beta}}_L$ is never computed. The following theorem shows that $\hat{\boldsymbol{\beta}}_H$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$ on a large class of zero mean finite variance symmetric error distributions. Thus if $\hat{\boldsymbol{\beta}}_C$ is $\sqrt{n}$ consistent or asymptotically efficient, so is $\hat{\boldsymbol{\beta}}_H$. Notice that $\hat{\boldsymbol{\beta}}_A$ does not need to be consistent. This point is crucial since `lmsreg` is not consistent and it is not known whether FLTS is consistent. The clean data are in *general position* if any p clean cases give a unique estimate of $\hat{\boldsymbol{\beta}}$.

Theorem 8.41. Assume the clean data are in general position, and suppose that both $\hat{\boldsymbol{\beta}}_L$ and $\hat{\boldsymbol{\beta}}_C$ are consistent estimators of $\boldsymbol{\beta}$ where the regression model contains a constant. Then the `hbreg` estimator $\hat{\boldsymbol{\beta}}_H$ is high breakdown and asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$.

Proof. Since the clean data are in general position and $Q_L(\hat{\boldsymbol{\beta}}_H) \leq aQ_L(\hat{\boldsymbol{\beta}}_B)$ is bounded for γ_n near 0.5, the `hbreg` estimator is high breakdown. Let $Q_L^* = Q_L$ for LMS and $Q_L^* = Q_L/n$ for LTS and LTA. As $n \to \infty$, consistent estimators $\hat{\boldsymbol{\beta}}$ satisfy $Q_L^*(\hat{\boldsymbol{\beta}}) - Q_L^*(\boldsymbol{\beta}) \to 0$ in probability. Since LMS, LTS, and LTA are consistent and the minimum value is $Q_L^*(\hat{\boldsymbol{\beta}}_L)$, it follows that $Q_L^*(\hat{\boldsymbol{\beta}}_C) - Q_L^*(\hat{\boldsymbol{\beta}}_L) \to 0$ in probability, while $Q_L^*(\hat{\boldsymbol{\beta}}_L) < aQ_L^*(\hat{\boldsymbol{\beta}})$ for any estimator $\hat{\boldsymbol{\beta}}$. Thus with probability tending to one as $n \to \infty$, $Q_L(\hat{\boldsymbol{\beta}}_C) < a \min(Q_L(\hat{\boldsymbol{\beta}}_A), Q_L(\hat{\boldsymbol{\beta}}_B))$. Hence $\hat{\boldsymbol{\beta}}_H$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$. □

Remark 8.12. i) Let $\hat{\boldsymbol{\beta}}_C = \hat{\boldsymbol{\beta}}_{OLS}$. Then `hbreg` is asymptotically equivalent to OLS when the errors e_i are iid from a large class of zero mean finite variance symmetric distributions, including the $N(0, \sigma^2)$ distribution, since the probability that `hbreg` uses OLS instead of $\hat{\boldsymbol{\beta}}_A$ or $\hat{\boldsymbol{\beta}}_B$ goes to one as $n \to \infty$.

ii) The above theorem proves that practical high breakdown estimators with 100% asymptotic Gaussian efficiency exist; however, such estimators are not necessarily good.

iii) The theorem holds when both $\hat{\boldsymbol{\beta}}_L$ and $\hat{\boldsymbol{\beta}}_C$ are consistent estimators of $\boldsymbol{\beta}$, for example, when the iid errors come from a large class or zero mean finite variance symmetric distributions. For asymmetric distributions, $\hat{\boldsymbol{\beta}}_C$ estimates $\boldsymbol{\beta}_C$ and $\hat{\boldsymbol{\beta}}_L$ estimates $\boldsymbol{\beta}_L$ where the constants usually differ. The theorem holds for some distributions that are not symmetric because of the penalty a. As $a \to \infty$, the class of asymmetric distributions where the theorem holds greatly increases, but the outlier resistance decreases rapidly as a increases for $a > 1.4$.

iv) The default `hbreg` estimator used OLS, `mbareg`, and $\hat{\boldsymbol{\beta}}_B$ with $a = 1.4$ and the LTA criterion. For the simulated data with symmetric error distributions, $\hat{\boldsymbol{\beta}}_B$ appeared to give biased estimates of the slopes. However, for the simulated data with right skewed error distributions, $\hat{\boldsymbol{\beta}}_B$ appeared to give good estimates of the slopes but not the constant estimated by OLS, and the probability that the `hbreg` estimator selected $\hat{\boldsymbol{\beta}}_B$ appeared to go to one.

v) Both MBA and OLS are $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$, even for a large class of skewed distributions. Using $\hat{\boldsymbol{\beta}}_A = \hat{\boldsymbol{\beta}}_{MBA}$ and removing $\hat{\boldsymbol{\beta}}_B$ from the `hbreg` estimator results in a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ when $\hat{\boldsymbol{\beta}}_C =$ OLS is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$, but massive sample sizes were still needed to get good estimates of the constant for skewed error distributions. For skewed distributions, if OLS needed $n = 1000$ to estimate the constant well, `mbareg` might need $n >$ one million to estimate the constant well.

vi) The outlier resistance of `hbreg` is not especially good.

The family of `hbreg` estimators is enormous and depends on i) the practical high breakdown estimator $\hat{\boldsymbol{\beta}}_B$, ii) $\hat{\boldsymbol{\beta}}_C$, iii) $\hat{\boldsymbol{\beta}}_A$, iv) a, and v) the criterion Q_L. Note that the theory needs the error distribution to be such that both $\hat{\boldsymbol{\beta}}_C$ and $\hat{\boldsymbol{\beta}}_L$ are consistent. Sufficient conditions for LMS, LTS, and LTA to be consistent are rather strong. To have reasonable sufficient conditions for the `hbreg` estimator to be consistent, $\hat{\boldsymbol{\beta}}_C$ should be consistent under weak conditions. Hence OLS is a good choice that results in 100% asymptotic Gaussian efficiency.

We suggest using the LTA criterion since in simulations, `hbreg` behaved like $\hat{\boldsymbol{\beta}}_C$ for smaller sample sizes than those needed by the LTS and LMS criteria. We want a near 1 so that `hbreg` has outlier resistance similar to $\hat{\boldsymbol{\beta}}_A$, but we want a large enough so that `hbreg` performs like $\hat{\boldsymbol{\beta}}_C$ for moderate n on clean data. Simulations suggest that $a = 1.4$ is a reasonable choice.

The default `hbreg` program from *linmodpack* uses the $\sqrt{n}$ consistent outlier resistant estimator `mbareg` as $\hat{\boldsymbol{\beta}}_A$.

There are at least three reasons for using $\hat{\boldsymbol{\beta}}_B$ as the high breakdown estimator. First, $\hat{\boldsymbol{\beta}}_B$ is high breakdown and simple to compute. Second, the fitted values roughly track the bulk of the data. Lastly, although $\hat{\boldsymbol{\beta}}_B$ has rather poor outlier resistance, $\hat{\boldsymbol{\beta}}_B$ does perform well on several outlier configurations where some common alternatives fail.

As implemented in *lspack*, the `hbreg` estimator is a practical $\sqrt{n}$ consistent high breakdown estimator that appears to perform like OLS for moderate n if the errors are unimodal and symmetric, and to have outlier resistance comparable to competing practical "outlier resistant" estimators.

8.5 The Robust rmreg2 Estimator

The robust multivariate linear regression estimator `rmreg2` is the classical multivariate linear regression estimator applied to the RMVN set when RMVN is computed from the vectors $\boldsymbol{u}_i = (x_{i2}, ..., x_{ip}, Y_{i1}, ..., Y_{im})^T$ for $i = 1, ..., n$. Hence $\boldsymbol{u}_i$ is the ith case with $x_{i1} = 1$ deleted. This regression estimator has considerable outlier resistance, and is one of the most outlier resistant practical robust regression estimator for the $m = 1$ multiple linear regression case. The `rmreg2` estimator has been shown to be consistent if the $\boldsymbol{u}_i$ are iid from a large class of elliptically contoured distributions, which is a much stronger assumption than having iid error vectors $\boldsymbol{\epsilon}_i$.

Let $\boldsymbol{x} = (1, \boldsymbol{u}^T)^T$and let $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_2^T)^T = (\alpha, \boldsymbol{\eta}^T)^T$. Now for multivariate linear regression, $\hat{\boldsymbol{\beta}}_j = (\hat{\alpha}_j, \hat{\boldsymbol{\eta}}_j^T)^T$ where $\hat{\alpha}_j = \overline{Y}_j - \hat{\boldsymbol{\eta}}_j^T \overline{\boldsymbol{u}}$ and $\hat{\boldsymbol{\eta}}_j = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y_j}$. Let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{uy}} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{w}_i - \overline{\boldsymbol{w}})(\boldsymbol{y}_i - \overline{\boldsymbol{y}})^T$ which has jth column $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Y_j}$ for $j = 1, ..., m$. Let

$$\boldsymbol{v} = \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{y} \end{pmatrix}, \quad E(\boldsymbol{v}) = \boldsymbol{\mu}_{\boldsymbol{v}} = \begin{pmatrix} E(\boldsymbol{u}) \\ E(\boldsymbol{y}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{u}} \\ \boldsymbol{\mu}_{\boldsymbol{y}} \end{pmatrix}, \quad \text{and} \quad \text{Cov}(\boldsymbol{v}) = \boldsymbol{\Sigma}_{\boldsymbol{v}} =$$

$$\begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{uu}} & \boldsymbol{\Sigma}_{\boldsymbol{uy}} \\ \boldsymbol{\Sigma}_{\boldsymbol{yu}} & \boldsymbol{\Sigma}_{\boldsymbol{yy}} \end{pmatrix}.$$

Let the vector of constants be $\boldsymbol{\alpha}^T = (\alpha_1, ..., \alpha_m)$ and the matrix of slope vectors $\boldsymbol{B}_S = \begin{bmatrix} \boldsymbol{\eta}_1 & \boldsymbol{\eta}_2 & \dots & \boldsymbol{\eta}_m \end{bmatrix}$. Then the population least squares coefficient matrix is

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{\alpha}^T \\ \boldsymbol{B}_S \end{pmatrix}$$

where $\boldsymbol{\alpha} = \boldsymbol{\mu}_{\boldsymbol{y}} - \boldsymbol{B}_S^T \boldsymbol{\mu}_{\boldsymbol{u}}$ and $\boldsymbol{B}_S = \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{uy}}$ where $\boldsymbol{\Sigma}_{\boldsymbol{u}} = \boldsymbol{\Sigma}_{\boldsymbol{uu}}$.

If the $\boldsymbol{u}_i$ are iid with nonsingular covariance matrix Cov($\boldsymbol{u}$), the least squares estimator

$$\hat{\boldsymbol{B}} = \begin{pmatrix} \hat{\boldsymbol{\alpha}}^T \\ \hat{\boldsymbol{B}}_S \end{pmatrix}$$

where $\hat{\boldsymbol{\alpha}} = \overline{\boldsymbol{y}} - \hat{\boldsymbol{B}}_S^T \overline{\boldsymbol{u}}$ and $\hat{\boldsymbol{B}}_S = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{uy}}$. The least squares multivariate linear regression estimator can be calculated by computing the classical estimator $(\overline{\boldsymbol{v}}, \boldsymbol{S}_{\boldsymbol{v}}) = (\overline{\boldsymbol{v}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{v}})$ of multivariate location and dispersion on the $\boldsymbol{v}_i$, and then plug in the results into the formulas for $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{B}}_S$.

Let $(T, \boldsymbol{C}) = (\tilde{\boldsymbol{\mu}}_{\boldsymbol{v}}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}})$ be a robust estimator of multivariate location and dispersion. If $\tilde{\boldsymbol{\mu}}_{\boldsymbol{v}}$ is a consistent estimator of $\boldsymbol{\mu}_{\boldsymbol{v}}$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}}$ is a consistent estimator of $c\, \boldsymbol{\Sigma}_{\boldsymbol{v}}$ for some constant $c > 0$, then a robust estimator of multivariate linear regression is the plug in estimator $\tilde{\boldsymbol{\alpha}} = \tilde{\boldsymbol{\mu}}_{\boldsymbol{y}} - \tilde{\boldsymbol{B}}_S^T \tilde{\boldsymbol{\mu}}_{\boldsymbol{u}}$ and $\tilde{\boldsymbol{B}}_S = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{uy}}$.

For the `rmreg2` estimator, $(T, \boldsymbol{C})$ is the classical estimator applied to the RMVN set when RMVN is applied to vectors $\boldsymbol{v}_i$ for $i = 1, ..., n$ (could use $(T, \boldsymbol{C})$ = RMVN estimator since the scaling does not matter for this application). Then $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}_{\boldsymbol{v}}, c\, \boldsymbol{\Sigma}_{\boldsymbol{v}})$ if the $\boldsymbol{v}_i$ are iid from a large class of $EC_d(\boldsymbol{\mu}_{\boldsymbol{v}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}, g)$ distributions where $d = m+p-1$. Thus the classical and robust estimators of multivariate linear regression are both $\sqrt{n}$ consistent estimators of $\boldsymbol{B}$ if the $\boldsymbol{v}_i$ are iid from a large class of elliptically contoured distributions. This assumption is very strong, but the robust estimator is useful for detecting outliers. It seems likely that the estimator is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ under mild conditions where the parameter vector $\boldsymbol{\beta}$ is not, in general, the parameter vector estimated by OLS. When there are categorical predictors or the joint distribution of $\boldsymbol{v}$ is not elliptically contoured, it is possible that the robust estimator is bad and very different from the good classical least squares estimator. The *lspack* function `rmreg2` computes the `rmreg2` estimator and produces the response and residual plots.

8.6 Summary

1) For the location model, the sample mean $\overline{Y} = \dfrac{\sum_{i=1}^n Y_i}{n}$, the sample variance $S_n^2 = \dfrac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1}$, and the sample standard deviation $S_n = \sqrt{S_n^2}$. If the data $Y_1, ..., Y_n$ is arranged in ascending order from smallest to largest and written as $Y_{(1)} \leq \cdots \leq Y_{(n)}$, then $Y_{(i)}$ is the ith order statistic and the $Y_{(i)}$'s are called the *order statistics*. The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \quad \text{if n is odd,}$$

$$\text{MED}(n) = \frac{Y_{(n/2)} + Y_{((n/2)+1)}}{2} \quad \text{if n is even.}$$

The notation $\text{MED}(n) = \text{MED}(Y_1, ..., Y_n)$ will also be used. The *sample median absolute deviation* is $\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|,\ i = 1, \ldots, n)$.

2) Suppose the multivariate data has been collected into an $n \times p$ matrix

$$\boldsymbol{W} = \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}.$$

The *coordinatewise median* $\text{MED}(\boldsymbol{W}) = (\text{MED}(X_1), ..., \text{MED}(X_p))^T$ where $\text{MED}(X_i)$ is the sample median of the data in column i corresponding to variable X_i. The **sample mean** $\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = (\overline{X}_1, ..., \overline{X}_p)^T$ where $\overline{X}_i$ is the sample mean of the data in column i corresponding to variable X_i. The **sample covariance matrix**

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T = (S_{ij}).$$

That is, the ij entry of $\boldsymbol{S}$ is the sample covariance S_{ij}. The *classical estimator of multivariate location and dispersion* is $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$.

3) Let $(T, \boldsymbol{C}) = (T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}))$ be an estimator of multivariate location and dispersion. The ith *Mahalanobis distance* $D_i = \sqrt{D_i^2}$ where the ith *squared Mahalanobis distance* is $D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) =$
$(\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W}))$.

4) The squared Euclidean distances of the $\boldsymbol{x}_i$ from the coordinatewise median is $D_i^2 = D_i^2(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Concentration type steps compute the weighted median MED_j: the coordinatewise median computed from the cases $\boldsymbol{x}_i$ with $D_i^2 \leq \text{MED}(D_i^2(\text{MED}_{j-1}, \boldsymbol{I}_p))$ where $\text{MED}_0 = \text{MED}(\boldsymbol{W})$. Often used $j = 0$ (no concentration type steps) or $j = 9$. Let $D_i = D_i(\text{MED}_j, \boldsymbol{I}_p)$. Let $W_i = 1$ if $D_i \leq \text{MED}(D_1, ..., D_n) + k\text{MAD}(D_1, ..., D_n)$ where $k \geq 0$ and $k = 5$ is the default choice. Let $W_i = 0$, otherwise.

5) Let the *covmb2 set* B of at least $n/2$ cases correspond to the cases with weight $W_i = 1$. Then the *covmb2* estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix applied to the cases in set B. Hence

$$T = \frac{\sum_{i=1}^{n} W_i \boldsymbol{x}_i}{\sum_{i=1}^{n} W_i} \text{ and } \boldsymbol{C} = \frac{\sum_{i=1}^{n} W_i (\boldsymbol{x}_i - T)(\boldsymbol{x}_i - T)^T}{\sum_{i=1}^{n} W_i \ - \ 1}.$$

The function `ddplot5` plots the Euclidean distances from the coordinatewise median versus the Euclidean distances from the `covmb2` location estimator. Typically the plotted points in this DD plot cluster about the identity line, and outliers appear in the upper right corner of the plot with a gap between the bulk of the data and the outliers.

8.7 Complements

Nearly all of the literature for high breakdown regression and high breakdown multivariate location and dispersion has massive errors: i) the estimators that have large sample theory tend to be impractical to compute, while ii) estimators that are practical to compute tend to be inconsistent and zero breakdown, or have no proven large sample theory. See Hawkins and Olive (2002). Read Olive (2008, 2017b, 2022c) for practical robust statistics backed by some large sample theory. Sections 8.2 and 8.4 showed that getting large sample theory for practical estimators is very difficult.

Location Model: The two stage trimmed means are due to Olive (2001). The confidence interval for the population median appears in Olive (2017b). Huber and Ronchetti (2009) is useful for other estimators.

Robust MLD

For the FCH, RFCH, and RMVN estimators, see Olive and Hawkins (2010), Olive (2017b, ch. 4), and Zhang et al. (2012). See Olive (2017b, p. 120) for the `covmb2` estimator.

The fastest estimators of multivariate location and dispersion that have been shown to be both consistent and high breakdown are the minimum covariance determinant (MCD) estimator with $O(n^v)$ complexity where $v = 1 + p(p+3)/2$ and possibly an all elemental subset estimator of He and Wang (1997). See Bernholt and Fischer (2004). The minimum volume ellipsoid (MVE) complexity is far higher, and **for $p > 2$ there may be no known method for computing** S, τ, projection based, and constrained M estimators. For some depth estimators, like the Stahel-Donoho estimator, the exact algorithm of Liu and Zuo (2014) appears to take too long if $p \geq 6$ and $n \geq 100$, and simulations may need $p \leq 3$. It is possible to compute the MCD and MVE estimators for $p = 4$ and $n = 100$ in a few hours using branch and bound algorithms (like estimators with $O(100^4)$ complexity). See Agulló (1996, 1998) and Pesch (1999). These algorithms take too long if both $p \geq 5$ and $n \geq 100$. Simulations may need $p \leq 2$. Two stage estimators such as the MM estimator, that need an initial high breakdown consistent estimator, take longer to compute than the initial estimator. Rousseeuw (1984) introduced the MCD and MVE estimators. See Maronna et al. (2006, ch. 6) for descriptions and references.

Estimators with complexity higher than $O[(n^3+n^2p+np^2+p^3)\log(n)]$ take too long to compute and will rarely be used. Reyen et al. (2009) simulated the OGK and the Olive (2004a) median ball algorithm (MBA) estimators for $p = 100$ and n up to 50000, and noted that the OGK complexity is $O[p^3 + np^2\log(n)]$ while that of MBA is $O[p^3 + np^2 + np\log(n)]$. FCH, RMBA, and RMVN have the same complexity as MBA. FMCD has the same complexity as FCH, but FCH is roughly 100 to 200 times faster.

Robust Regression

For the `hbreg` estimator, see Olive and Hawkins (2011) and Olive (2017b, ch. 14). Robust regression estimators have unsatisfactory outlier resistance

and large sample theory. The `hbreg` estimator is fast and high breakdown, but does not provide an adequate remedy for outliers, and the symmetry condition for consistency is too strong. OLS response and residual plots are useful for detecting multiple linear regression outliers.

Many of the robust statistics for the location model are practical to compute, outlier resistant, and backed by theory. See Huber and Ronchetti (2009). A few estimators of multivariate location and dispersion, such as the coordinatewise median, are practical to compute, outlier resistant, and backed by theory.

For practical estimators for MLR and MCD, `hbreg` and FCH appear to be the only estimators proven to be consistent (for a large class of symmetric error distributions and for a large class of EC distributions, respectively) with some breakdown theory (T_{FCH} is HB). Perhaps all other "robust statistics" for MLR and MLD that have been shown to be both consistent and high breakdown are impractical to compute for $p > 4$: the impractical "brand name" estimators have at least $O(n^p)$ complexity, while the practical estimators used in the software for the "brand name estimators" have not been shown to be both high breakdown and consistent. See Theorems 8.30 and 8.38, Hawkins and Olive (2002), Olive (2008, 2017b), Hubert et al. (2002), and Maronna and Yohai (2002). Huber and Ronchetti (2009, pp. xiii, 8-9, 152-154, 196-197) suggested that high breakdown regression estimators do not provide an adequate remedy for the ill effects of outliers, that their statistical and computational properties are not adequately understood, that high breakdown estimators "break down for all except the smallest regression problems by failing to provide a timely answer!" and that "there are no known high breakdown point estimators of regression that are demonstrably stable."

A large number of impractical high breakdown regression estimators have been proposed, including LTS, LMS, LTA, S, LQD, τ, constrained M, repeated median, cross checking, one step GM, one step GR, t-type, and regression depth estimators. See Rousseeuw and Leroy (1987) and Maronna et al. (2019). The practical algorithms used in the software use a brand name criterion to evaluate a fixed number of trial fits and should be denoted as an F-brand name estimator such as FLTS. Two stage estimators, such as the MM estimator, that need an initial consistent high breakdown estimator often have the same breakdown value and consistency rate as the initial estimator. These estimators are typically implemented with a zero breakdown inconsistent initial estimator and hence are zero breakdown with zero efficiency.

Maronna and Yohai (2015) used OLS and 500 elemental sets as the 501 trial fits to produce an FS estimator used as the initial estimator for an FMM estimator. Since the 501 trial fits are zero breakdown, so is the FS estimator. Since the FMM estimator has the same breakdown as the initial estimator, the FMM estimator is zero breakdown. For regression, they show that the FS estimator is consistent on a large class of zero mean finite variance

symmetric distributions. Consistency follows since the elemental fits and OLS are unbiased estimators of $\boldsymbol{\beta}_{OLS}$ but an elemental fit is an OLS fit to p cases. Hence the elemental fits are very variable, and the probability that the OLS fit has a smaller S-estimator criterion than a randomly chosen elemental fit (or K randomly chosen elemental fits) goes to one as $n \to \infty$. (OLS and the S-estimator are both $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$, so the ratio of their criterion values goes to one, and the S-estimator minimizes the criterion value.) Hence the FMM estimator is asymptotically equivalent to the MM estimator that has the smallest criterion value for a large class of iid zero mean finite variance symmetric error distributions. This FMM estimator is asymptotically equivalent to the FMM estimator that uses OLS as the initial estimator. When the error distribution is skewed the S-estimator and OLS population constant are not the same, and the probability that an elemental fit is selected is close to one for a skewed error distribution as $n \to \infty$. (The OLS estimator $\hat{\boldsymbol{\beta}}$ gets very close to $\boldsymbol{\beta}_{OLS}$ while the elemental fits are highly variable unbiased estimators of $\boldsymbol{\beta}_{OLS}$, so one of the elemental fits is likely to have a constant that is closer to the S-estimator constant while still having good slope estimators.) Hence the FS estimator is inconsistent, and the FMM estimator is likely inconsistent for skewed distributions. No practical method is known for computing a $\sqrt{n}$ consistent FS or FMM estimator that has the same breakdown and maximum bias function as the S or MM estimator that has the smallest S or MM criterion value.

8.8 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

8.1. Use Theorem 2.6 to find the limiting distribution of $\sqrt{n}(\text{MED}(n) - \text{MED}(Y))$.

8.2. The interquartile range $\text{IQR}(n) = \hat{\xi}_{n,0.75} - \hat{\xi}_{n,0.25}$ and is a popular estimator of scale. Use Theorem 3.11 to show that

$$\sqrt{n}\frac{1}{2}(IQR(n) - IQR(Y)) \xrightarrow{D} N(0, \sigma_A^2)$$

where

$$\sigma_A^2 = \frac{1}{64}\left[\frac{3}{[f(\xi_{3/4})]^2} - \frac{2}{f(\xi_{3/4})f(\xi_{1/4})} + \frac{3}{[f(\xi_{1/4})]^2}\right].$$

8.3*. Let F be the $N(0,1)$ cdf. Show that the ARE of the sample median $\text{MED}(n)$ with respect to the sample mean $\overline{Y}_n$ is $ARE \approx 0.64$.

8.4*. Let F be the $DE(0,1)$ cdf. Show that the ARE of the sample median MED(n) with respect to the sample mean $\overline{Y}_n$ is $ARE \approx 2.0$.

8.5. If Y is $TEXP(\lambda, b = k\lambda)$ for $k > 0$, show that

$$a) \quad E(Y) = \lambda \left[1 - \frac{k}{e^k - 1} \right].$$

$$b) \quad E(Y^2) = 2\lambda^2 \left[1 - \frac{(0.5k^2 + k)}{e^k - 1} \right].$$

Chapter 9
Time Series

9.1 ARMA Time Series

This section reviews ARMA time series models. We will use the R software notation and write a moving average parameter θ with a positive sign. Many references and software will write the model with a negative sign for the moving average parameters.

Definition 9.1. a) A *moving average* MA(q) times series is

$$Y_t = \tau + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} + e_t$$

where $\theta_q \neq 0$.
b) An *autoregressive* AR(p) times series is

$$Y_t = \tau + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t$$

where $\phi_p \neq 0$.
c) An *autoregressive moving average* ARMA(p, q) times series is

$$Y_t = \tau + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} + e_t \tag{9.1}$$

where $\theta_q \neq 0$ and $\phi_p \neq 0$.

The results in this chapter also apply to a time series X_t that follows an ARIMA(p, d, q) model with known d if the differenced time series model Y_t follows an ARMA(p, q) model. See Box and Jenkins (1976) for more on these models. We will assume that the e_t are independent and identically distributed (iid) with zero mean and variance σ^2. The observed time series is $\{Y_t\} = Y_1, ..., Y_n$.

We usually want the ARMA(p, q) model to be weakly stationary, causal, and invertible. Let $Z_t = Y_t - \mu$ where $\mu = E(Y_t)$ if $\{Y_t\}$ is weakly stationary. Then the causal property implies that $Z_t = \sum_{j=1}^{\infty} \psi_j e_{t-j} + e_t$, which is an MA($\infty$) representation, where the $\psi_j \to 0$ rapidly as $j \to \infty$. Invertibil-

ity implies that $Z_t = \sum_{j=1}^{\infty} \chi_j Z_{t-j} + e_t$, which is an AR($\infty$) representation, where the $\chi_j \to 0$ rapidly as $j \to \infty$. We will make the usual assumption that the AR(∞) and MA(∞) parameters are square summable. Thus if the ARMA(p, q) model is weakly stationary, causal, and invertible, then Y_t depends almost entirely on nearby lags of Y_t and e_t, not on the distant past.

9.2 Large Sample theory

Some notation and preliminary results are needed. The Gaussian maximum likelihood estimator (GMLE) will be used. The Yule Walker and least squares estimators will also be used for AR(p) models. Let the r_i be the m (one step ahead) residuals where often $m = n$ or $m = n-p$. Under regularity conditions,

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{m} r_i^2}{m - p - q - c} \tag{9.2}$$

is a consistent estimator of σ^2 where often $c = 0$ or $c = 1$. See Granger and Newbold (1977, p. 85) and Pankratz (1983, p. 206). Let $\hat{\sigma}^2$ be the estimator of σ^2 produced by the time series model, and let $\gamma_k = Cov(Y_t, Y_{t-k})$. Let

$$\boldsymbol{\Gamma}_n = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n-1} & \gamma_{n-2} & \cdots & \gamma_0 \end{bmatrix}.$$

The following large sample theorem for the AR(p) model is due to Mann and Wald (1943). Also see McElroy and Politis (2020, p. 333) and Anderson (1971, pp. 210-217). For large sample theory for MA and ARMA models, see Hannan (1973), Kreiss (1985), and Yao and Brockwell (2006).

Remark 9.1. *There is a strong regularity condition for the GMLE of the ARMA(p, q) model.* Assume the ARMA(p_S, q_S) model is the true model.
a) If both $p > p_S$ and $q > q_S$, then the GMLE is not a consistent estimator. See Chan, Ling, and Yau (2020) and Hannan (1980).
b) The GMLE for the ARMA(p, q) model needs to satisfy $p \geq p_S$ and $q \geq q_S$ with either $p = p_S$ or $q = q_S$ for the model $\hat{\boldsymbol{\beta}}$ to be a consistent estimator of $\boldsymbol{\beta}$. Pötscher (1990) showed how to estimate $\max(p_S, q_S)$ consistently.

Theorem 9.1 Let the iid zero mean e_i have variance σ^2, and let the time series have mean $E(Y_t) = \mu$.

a) Let $Y_1, ..., Y_n$ be a weakly stationary and invertible AR(p) time series, and let $\boldsymbol{\beta} = (\phi_1, ..., \phi_p)$. Let $\hat{\boldsymbol{\beta}}$ be the Yule Walker estimator of $\boldsymbol{\beta}$. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}) \tag{9.3}$$

where $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\beta}) = \sigma^2 \boldsymbol{\Gamma}_p^{-1}$. Equation (9.3) also holds under mild regularity conditions for the least squares estimator, and the GMLE of $\boldsymbol{\beta}$.

b) Let $Y_1, ..., Y_n$ be a weakly stationary, causal, and invertible MA(q) time series, and let $\boldsymbol{\beta} = (\theta_1, ..., \theta_q)$. Let $\hat{\boldsymbol{\beta}}$ be the GMLE. Under regularity conditions,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_q(\mathbf{0}, \boldsymbol{V}) \tag{9.4}$$

where $\boldsymbol{V}$ is given, for example, by McElroy and Politis (2022, pp. 340-341).

c) Let $Y_1, ..., Y_n$ be a weakly stationary, causal, and invertible ARMA(p, q) time series, and let $\boldsymbol{\beta} = (\phi_1, ..., \phi_p, \theta_1, ..., \theta_q)$ with $g = p + q$. Let $\hat{\boldsymbol{\beta}}$ be the GMLE. Under regularity conditions,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{V}) \tag{9.5}$$

where $\boldsymbol{V}$ depends on the autocorrelation function and σ^2.

The main point of Theorem 9.1 is that the theory can hold even if the e_t are not iid $N(0, \sigma^2)$. The basic idea for the GMLE is that $\{Y_t\}$ satisfies an AR(∞) model which is approximately an AR(p_y) model, and the large sample theory for the AR(p_y) model depends on the zero mean error distribution through σ^2 by Theorem 9.1 a). See Anderson (1971: ch. 5, 1977). When the e_t are iid $N(0, \sigma^2)$, $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\beta}) = \boldsymbol{I}_1^{-1}(\boldsymbol{\beta})$, the inverse information matrix. Then for the AR(p) model, $\boldsymbol{V}(\boldsymbol{\phi}) = \sigma^2 \boldsymbol{\Gamma}_p^{-1}(\boldsymbol{\phi}) = \boldsymbol{I}_1^{-1}(\boldsymbol{\phi})$. See Box and Jenkins (1976, p. 241) and McElroy and Politis (2020, pp. 340-344).

9.3 Inference after Model Selection

This section considers model selection where it is assumed that it is known that the model is ARMA, AR, or MA, but the order needs to be determined. For ARMA model selection, let the full model be an ARMA(p_{max}, q_{max}) model. For AR model selection $q_{max} = 0$, while for MA model selection $p_{max} = 0$. Granger and Newbold (1977, p. 178) suggested using $p_{max} = 13$ for AR model selection, and we may use $p_{max} = q_{max} = 5$ for ARMA model selection, and $q_{max} = 13$ for MA model selection. For ARMA model selection, there are $J = (p_{max} + 1)(q_{max} + 1)$ ARMA(p, q) submodels where p ranges from 0 to p_{max} and q ranges from 0 to q_{max}. For AR and MA model selection there are $J = p_{max} + 1$ and $J = q_{max} + 1$ submodels, respectively. Assume the true (optimal) model is an ARMA(p_S, q_S) model with $p_S \leq p_{max}$ and $q_S \leq q_{max}$. Let the selected model I be an ARMA(p_I, q_I) model. Then the model underfits unless $p_I \geq p_S$ and $q_I \geq q_S$. For AR model selection, the probability of underfitting goes to 0 if the Akaike (1973) AIC, Schwartz (1978) BIC, or Hurvich and Tsai (1989) AIC_C criterion are used. See Hannan and Quinn (1979) and Shibata (1976).

Remark 9.2: Are Statisticians crazy? Similar results for ARMA models were given by Hannan (1980) and Hannan and Kavalieris (1984). Also see Huang et al. (2022). However, in simulations with $(k_{max}+1)^2$ possible ARMA(p,q) models where $p,q = 0,1,...,k_{max}$, ARMA model selection with AIC, BIC, and AIC_C was unable to select models satisfying Remark 9.1 (and Remark 9.3) with probability close to 1. Underfitting was common for BIC. AIC and AIC_C often underfit for $n \leq 500$, and often failed to satisfy Remark 9.1 for $n > 5000$. The Pötscher (1990) model selection procedure restricted the search to ARMA(k,k) models for $k = 0,1,...,k_{max}$ with a BIC type criterion, and this procedure worked fairly well for $n \geq 600$ (Chan, Ling, and Yau (2020) suggested $n \geq 1000$).

More notation is needed for model selection. Let the full model be the AR(p_{max}), MA(q_{max}), or ARMA(p_{max}, q_{max}) model. Let $\boldsymbol{\beta}$ be a $b \times 1$ vector. For ARMA model selection, let $\boldsymbol{\beta} = (\boldsymbol{\phi}^T, \boldsymbol{\theta}^T)^T = (\phi_1, ..., \phi_{p_{max}}, \theta_1, ..., \theta_{q_{max}})^T$ with $b = p_{max} + q_{max}$. For AR model selection, let $\boldsymbol{\beta} = (\phi_1, ..., \phi_{p_{max}})^T$ with $b = p_{max}$, and for MA model selection, let $\boldsymbol{\beta} = (\theta_1, ..., \theta_{q_{max}})^T$ with $b = q_{max}$. Hence $\boldsymbol{\beta} = (\beta_1, ..., \beta_{p_{max}}, \beta_{p_{max}+1}, ..., \beta_{p_{max}+q_{max}})^T$. Let $S = \{1, ..., p_S, p_{max}+1, ..., p_{max}+q_S\}$ index the true ARMA(p_S, q_S) model. If $S = \emptyset$ is the empty set, then the time series random variables $Y_1, ..., Y_n$ are iid. Let $I = \{1, ..., p_I, p_{max}+1, ..., p_{max}+q_I\}$ index the ARMA(p_I, q_I) model. Let $\hat{\boldsymbol{\beta}}_{I,0}$ be a $b \times 1$ estimator of $\boldsymbol{\beta}$ which is a obtained by padding $\hat{\boldsymbol{\beta}}_I$ with zeroes. If $\boldsymbol{\beta}_I = (\phi_1, ..., \phi_{p_I}, \theta_1, ..., \theta_{q_I})^T$, then $\hat{\boldsymbol{\beta}}_{I,0} = (\hat{\phi}_1, ..., \hat{\phi}_{p_I}, 0, .., 0, \hat{\theta}_1, ..., \hat{\theta}_{q_I}, 0, ..., 0)^T$. If $q_I = 0$, then $\hat{\boldsymbol{\beta}}_{I,0} = (\hat{\phi}_1, ..., \hat{\phi}_{p_I}, 0, .., 0)^T$. If $p_I = 0$ then $\hat{\boldsymbol{\beta}}_{I,0} = (0, ..., .., 0, \hat{\theta}_1, ..., \hat{\theta}_{q_I}, 0, ..., 0)^T$. If $I = \emptyset$ with $p_I = q_I = 0$, then define $\hat{\boldsymbol{\beta}}_{I,0} = \mathbf{0}$, the $b \times 1$ vector of zeroes. The submodel I underfits unless $S \subseteq I$.

For example, if $p_{max} = q_{max} = 5$, then $S = \{1, 6, 7\}$ corresponds to the ARMA(1,2) model, and $I = \{1, 6, 7, 8\}$ corresponds to the ARMA(1,3) model. Then $\hat{\boldsymbol{\beta}}_S = (\hat{\phi}_1, \hat{\theta}_1, \hat{\theta}_2)^T$, $\hat{\boldsymbol{\beta}}_{S,0} = (\hat{\phi}_1, 0, 0, 0, 0, \hat{\theta}_1, \hat{\theta}_2, 0, 0, 0)^T$, and $\hat{\boldsymbol{\beta}}_{I,0} = (\hat{\phi}_1, 0, 0, 0, 0, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, 0, 0)^T$.

The model I_{min} corresponds to the model that minimizes the AIC, AIC_C, or BIC criterion. Then the model selection estimator $\hat{\boldsymbol{\beta}}_{MS} = \hat{\boldsymbol{\beta}}_{I_{min},0}$. With this notation, the ARMA time series model selection theory developed in this chapter is very similar to the variable selection theory for regression models, such as multiple linear regression and generalized linear models, developed in Chapter 6.

Assume $\hat{\boldsymbol{\beta}}_{MS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn} = P(I_{min} = I_k)$ for $k = 1, ..., J$. Let $\hat{\boldsymbol{\beta}}_{MIX}$ be a random vector with a mixture distribution of the $\hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities equal to π_{kn}. Hence $\hat{\boldsymbol{\beta}}_{MIX} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with the same probabilities π_{kn} of the model selection estimator $\hat{\boldsymbol{\beta}}_{MS}$, but the I_k are randomly selected. The large sample theory for $\hat{\boldsymbol{\beta}}_{MIX}$ is useful for explaining that of $\hat{\boldsymbol{\beta}}_{MS}$ and for bootstrap confidence regions. Note that $\hat{\boldsymbol{\beta}}_{MIX}$ can not be computed since the π_{kn} are unknown. For mixture distributions, see Section 1.8.

Inference will consider bootstrap hypothesis testing with confidence intervals (CIs) and regions. Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a known $g \times 1$ vector.

Next we extend the Chapter 6 theory for variable selection estimators to time series model selection estimators. Suppose the full model is as in Section 9.1 and that if $S \subseteq I_j$ where the dimension of I_j is a_j, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\mathbf{0}, \boldsymbol{V}_j)$ where $\boldsymbol{V}_j$ is the covariance matrix of the asymptotic multivariate normal distribution. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_b(\mathbf{0}, \boldsymbol{V}_{j,0}) \tag{9.6}$$

where $\boldsymbol{V}_{j,0}$ adds columns and rows of zeros corresponding to the β_i not indexed by I_j, and $\boldsymbol{V}_{j,0}$ is singular unless I_j corresponds to the full model.

The first assumption in Theorem 9.2 is $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Then the model selection estimator corresponding to I_{min} underfits with probability going to zero. The assumption also requires $p_S \leq p_{max}$ and $q_S \leq q_{max}$. Then the assumption on $\boldsymbol{u}_{jn}$ in Theorem 9.2 may be reasonable by (9.6) since $S \subseteq I_j$ for each π_j, and since $\hat{\boldsymbol{\beta}}_{MIX}$ uses random selection. These two assumption may be reasonable if the Pötscher (1990) model selection procedure is used. (Need $P[I_{min}$ satisfies Remark 9.1 b) $] \to 1$ as $n \to \infty$.) The proofs of Theorems 9.2 and 9.3 are the same as those of Theorems 6.19 and 6.20.

Theorem 9.2 Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{MIX} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive π_k by π_j. Assume $\boldsymbol{u}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_b(\mathbf{0}, \boldsymbol{V}_{j,0})$. a) Then

$$\boldsymbol{u}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \tag{9.7}$$

where the cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$. Thus $\boldsymbol{u}$ is a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j, $E(\boldsymbol{u}) = \mathbf{0}$, and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \sum_j \pi_j \boldsymbol{V}_{j,0}$.

b) Let $\boldsymbol{A}$ be a $g \times b$ full rank matrix with $1 \leq g \leq b$. Then

$$\boldsymbol{v}_n = \boldsymbol{A}\boldsymbol{u}_n = \sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} = \boldsymbol{v} \tag{9.8}$$

where $\boldsymbol{v}$ has a mixture distribution of the $\boldsymbol{v}_j = \boldsymbol{A}\boldsymbol{u}_j \sim N_g(\mathbf{0}, \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T)$ with probabilities π_j.

c) The estimator $\hat{\boldsymbol{\beta}}_{MS}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$. Hence $\sqrt{n}(\hat{\boldsymbol{\beta}}_{MS} - \boldsymbol{\beta}) = O_P(1)$.

d) If $\pi_a = 1$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{SEL} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \sim N_b(\mathbf{0}, \boldsymbol{V}_{a,0})$ where SEL is MS or MIX.

Proof. a) Since $\boldsymbol{u}_n$ has a mixture distribution of the $\boldsymbol{u}_{kn}$ with probabilities π_{kn}, the cdf of $\boldsymbol{u}_n$ is $F_{\boldsymbol{u}_n}(\boldsymbol{t}) = \sum_k \pi_{kn} F_{\boldsymbol{u}_{kn}}(\boldsymbol{t}) \to F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$ at continuity points of the $F_{\boldsymbol{u}_j}(\boldsymbol{t})$ as $n \to \infty$.

b) Since $\boldsymbol{u}_n \xrightarrow{D} \boldsymbol{u}$, then $\boldsymbol{A}\boldsymbol{u}_n \xrightarrow{D} \boldsymbol{A}\boldsymbol{u}$.
c) The result follows since selecting from a finite number K of $\sqrt{n}$ consistent estimators (even on a set that goes to one in probability) results in a $\sqrt{n}$ consistent estimator by Pratt (1959).
d) If $\pi_a = 1$, there is no selection bias, asymptotically. The result also follows by Pötscher (1991, Lemma 1). □

Theorem 9.2 can be used to justify prediction intervals after model selection. See Section 9.4. Typically the mixture distribution is not asymptotically normal unless a $\pi_a = 1$ (e.g. if S is the full model). Theorem 9.2d) is useful for *model selection consistency* where $\pi_a = \pi_S = 1$ if $P(I_{min} = S) \to 1$ as $n \to \infty$. See Hannan (1980) and Claeskens and Hjort (2008) for references.

The following subscript notation is useful. Subscripts before the MIX are used for subsets of $\hat{\boldsymbol{\beta}}_{MIX} = (\hat{\beta}_1, ..., \hat{\beta}_b)^T$. Let $\hat{\boldsymbol{\beta}}_{i,MIX} = \hat{\beta}_i$. Similarly, if $I = \{i_1, ..., i_a\}$, then $\hat{\boldsymbol{\beta}}_{I,MIX} = (\hat{\beta}_{i_1}, ..., \hat{\beta}_{i_a})^T$. Subscripts after MIX denote the ith vector from a sample $\hat{\boldsymbol{\beta}}_{MIX,1}, ..., \hat{\boldsymbol{\beta}}_{MIX,B}$. Similar notation is used for other estimators such as $\hat{\boldsymbol{\beta}}_{MS}$. The subscript 0 is still used for zero padding. We may use $FULL$ to denote the full model $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{FULL}$.

Note that both $\sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta})$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{MS} - \boldsymbol{\beta})$ are selecting from the $\boldsymbol{u}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta})$ and asymptotically from the $\boldsymbol{u}_j$. The random selection for $\hat{\boldsymbol{\beta}}_{MIX}$ does not change the distribution of $\boldsymbol{u}_{jn}$, but selection bias does change the distribution of the selected $\boldsymbol{u}_{jn}$ and $\boldsymbol{u}_j$ to that of $\boldsymbol{w}_{jn}$ and $\boldsymbol{w}_j$. The assumption that $\boldsymbol{w}_{jn} \xrightarrow{D} \boldsymbol{w}_j$ may not be mild. The proof for Equation (9.9) is the same as that for (9.8).

Theorem 9.3 Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{MS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive π_k by π_j. Assume $\boldsymbol{w}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w}_j$. Then

$$\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{MS} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w} \tag{9.9}$$

where the cdf of $\boldsymbol{w}$ is $F_{\boldsymbol{w}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{w}_j}(\boldsymbol{t})$. Thus $\boldsymbol{w}$ is a mixture distribution of the $\boldsymbol{w}_j$ with probabilities π_j.

9.4 Bootstrapping ARMA time series model selection estimators

Remark 9.3. If the true model is the ARMA(p_S, q_S) model, then the ARMA(p_F, q_F) full model needs to satisfy $p_F \geq p_S$ and $q_F \geq q_S$ with either $p_F = p_S$ or $q_F = q_S$ for the full model $\hat{\boldsymbol{\beta}}_F$ to be a consistent estimator of $\boldsymbol{\beta}_F$.

For the bootstrap, we will ignore τ and build the bootstrap time series data set $\{Y_t^*\}$ sequentially. Assume the full model satisfies Remark 9.3. Fit

the full model to get the $\hat{\phi}_k$ and $\hat{\theta}_j$. Let

$$Y_t^* = \sum_{k=1}^{p_{max}} \hat{\phi}_k Y_{t-k}^* + e_t^*,$$

$$Y_t^* = \sum_{k=1}^{q_{max}} \hat{\theta}_k e_{t-k}^* + e_t^*,$$

or

$$Y_t^* = \sum_{k=1}^{p_{max}} \hat{\phi}_k Y_{t-k}^* + \sum_{k=1}^{q_{max}} \hat{\theta}_k e_{t-k}^* + e_t^*$$

for $t = 1, ..., n$. The ARMA and AR bootstrap may use a block of initial values $(Y_{-p+1}^*, ..., Y_0^*)^T = (Y_{j+1}, Y_{j+2}, ..., Y_{j+p})^T$ randomly selected from $Y_1, ..., Y_n$. For the *parametric bootstrap*, the e_t^* are iid $N(0, \hat{\sigma}^2)$ where $\hat{\sigma}^2$ is the estimate from fitting the full model with (p_{max}, q_{max}). For the *residual bootstrap*, assume the full model produces m residuals $r_1, ..., r_m$. Often $m = n$ or $m = n - p_{max}$. Refer to Equation (9.2) with (p, q) replaced by (p_{max}, q_{max}) and $b = p_{max} + q_{max}$. Let

$$\hat{e}_j = \sqrt{\frac{m}{m - b - c}} \, (r_j - \overline{r})$$

for $j = 1, ..., m$. Let the e_t^* be obtained by sampling with replacement from the $\hat{e}_j$. With respect to this bootstrap distribution, the e_t^* are iid with $E(e_t^*) = 0$ and $V(e_t^*) \approx \tilde{\sigma}^2$.

The following bootstrap algorithm produces pairs $(\hat{\boldsymbol{\beta}}_{MS,i}^*, \hat{\boldsymbol{\beta}}_{MIX,i}^*)$ for $i = 1, ..., B$ where the possible submodels I_k are selected with probabilities ρ_{kn} by the bootstrap model selection estimator. Then this bootstrap algorithm bootstraps both $\hat{\boldsymbol{\beta}}_{MS}$ and $\hat{\boldsymbol{\beta}}_{MIX}$ with $\pi_{kn} = \rho_{kn}$.

1) Generate a bootstrap time series data set $\{Y_i^*\}_{1,1} = \{Y_1^*, ..., Y_n^*\}_{1,1}$. Instead of computing the full model, use model selection to compute $\hat{\boldsymbol{\beta}}_{MS,1}^* = \hat{\boldsymbol{\beta}}_{I_1,0}^* = \hat{\boldsymbol{\beta}}_{I_1,0}^*(\{Y_i^*\}_{1,1})$.

2) Draw another bootstrap data set $\{Y_i^*\}_{1,2}$ and fit model I_1 from step 1) to get $\hat{\boldsymbol{\beta}}_{MIX,1}^* = \hat{\boldsymbol{\beta}}_{I_1,0}^*(\{Y_i^*\}_{1,2})$. (Selection bias is avoided since I_1 is selected before generating $\{Y_i^*\}_{1,2}$.)

3) Repeat B times to get the bootstrap samples $\hat{\boldsymbol{\beta}}_{MS,1}^*, ..., \hat{\boldsymbol{\beta}}_{MS,B}^*$ and $\hat{\boldsymbol{\beta}}_{MIX,1}^*, ..., \hat{\boldsymbol{\beta}}_{MIX,B}^*$.

Following McElroy and Politis (2020, pp. 438-439), consider a weakly stationary and invertible time series $Y_1, ..., Y_n$ where the e_t are iid with mean 0 and variance σ^2. A companion process uses ϵ_t that are iid with mean 0 and variance $\hat{\sigma}^2$. Both the residual bootstrap and parametric bootstrap produce companion processes $\{Y_t^*\}$. The residual bootstrap for an AR(p_{max}) model

is closely related to the sieve bootstrap for AR(p) and AR(∞) models. See McElroy and Politis (2020, pp. 430, 434).

It is important to note that for the parametric bootstrap, **we are not assuming** that the e_t are iid $N(0, \sigma^2)$. The following theorem is for bootstrapping the ARMA(p_F, q_F) full model. Assume the full model satisfies Remark 9.3.

Theorem 9.4 Assume the time series is such that Theorem 9.1 holds. Then $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_b(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\beta}))$ if the GMLE is used with the parametric bootstrap. This result also holds for the AR(p) model if the Yule Walker or least squares estimator is used with the parametric bootstrap or the residual bootstrap.

Proof. On a set A of probability going to one as $n \to \infty$, $Y_1^*, ..., Y_n^*$ with $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_n$ satisfies Theorem 9.1. Hence if n is fixed and the time series $Y_1^*, ..., Y_m^*$ is generated with $\hat{\boldsymbol{\beta}}_n$, then on the set A the estimator $\hat{\boldsymbol{\beta}}^*$ satisfies $\sqrt{m}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}_n) \xrightarrow{D} N_b(\mathbf{0}, \boldsymbol{V}(\hat{\boldsymbol{\beta}}_n))$ as $m \to \infty$. Since $\boldsymbol{V}(\hat{\boldsymbol{\beta}}) \xrightarrow{P} \boldsymbol{V}(\boldsymbol{\beta})$ if $\hat{\boldsymbol{\beta}}_n \xrightarrow{P} \boldsymbol{\beta}$ as $n \to \infty$, it follows that $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}_n) \xrightarrow{D} N_b(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\beta}))$ as $n \to \infty$. □

The basic idea is that for the parametric bootstrap, $Y_1^*, ..., Y_n^*$ satisfies the Gaussian time series model with $\hat{\boldsymbol{\beta}}_n$ as the parameter vector and $\hat{\boldsymbol{\beta}}_n$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$. Hence the Gaussian time series $Y_1^*, ..., Y_n^*$ with $\hat{\boldsymbol{\beta}}_n$ will be weakly stationary, causal, and invertible on a set A going to one in probability. Since $\hat{\boldsymbol{\beta}}_n$ depends on n, convergence along a triangular array needs to be used. Bootstrap results such as Theorem 9.4 are rather rare in the time series literature. Bühlmann (1994) has such a result for the AR(p) model.

9.5 Prediction Intervals

See Welagedara, Haile, and Olive (2024).

9.6 The Random Walk

A random walk (with drift) $Y_t = Y_{t-1} + e_t$ where the e_t are independent and identically distributed (iid). Suppose there is a sample $Y_1, ..., Y_n$ and we want a prediction interval (PI) for Y_{n+h}. Then $Y_t = Y_{t-2} + e_{t-1} + e_t = Y_{t-h} + e_{t-h+1} + \cdots + e_t = Y_0 + e_1 + \cdots + e_t$, or $Y_{n+h} = Y_n + e_{n+1} + e_{n+2} + \cdots + e_{n+h} = Y_n + \epsilon_{n,h}$. Let $e_j = Y_j - Y_{j-1}$ for $j = 2, ..., n$. Divide $e_2, ..., e_n$ into blocks of length h and let ϵ_i be the sum of the e_i in each block. Hence $\epsilon_1 = e_2 + \cdots + e_{h+1}$, $\epsilon_2 = e_{h+2} + \cdots + e_{2h+1}$, and $\epsilon_i = e_{(i-1)h+2} + e_{(i-1)h+3} + \cdots + e_{(i-1)h+h+1}$ for $i = 1, ..., m = \lfloor n/h \rfloor$. These ϵ_i are iid from

the same distribution as $\epsilon_{n,h}$. The same decomposition can be made for a vector valued random walk, $\boldsymbol{Y}_t = \boldsymbol{Y}_{t-1} + \boldsymbol{e}_t$, where the vectors are $p \times 1$. Thus $\boldsymbol{\epsilon}_i = \boldsymbol{e}_{(i-1)h+2} + \boldsymbol{e}_{(i-1)h+3} + \cdots + \boldsymbol{e}_{(i-1)h+h+1}$ for $i = 1, ..., m$.

The random walk can be written as $Y_t = Y_0 + \sum_{i=1}^{t} e_i$ where $Y_0 = y_0$ is often a constant. A stochastic process $\{N(t) : t \geq 0\}$ is a counting process if $N(t)$ counts the total number of events that occurred in time interval $(0, t]$. Let e_n be the interarrival time or waiting time between the $(n-1)$th and nth events counted by the process, $n \geq 1$. If the nonnegative e_i are iid with $P(e_i = 0) < 1$, then $\{N(t), t \geq 0\}$ is a *renewal process*. Let $Y_n = \sum_{i=1}^{n} e_i$ = the time of occurrence of the nth event = waiting time until the nth event. Then Y_n is a random walk with $Y_0 = y_0 = 0$. Let the expected value $E(e_i) = \mu > 0$. Then $E(Y_n) = n\mu$ and the variance $V(Y_n) = nV(e_i)$ if $V(e_i)$ exists. A Poisson process with rate λ is a renewal process where the e_i are iid exponential EXP(λ) with $E(e_i) = 1/\lambda$. See Ross (2014) for the Poisson process and renewal process. Given $Y_1, ..., Y_n$, then n events have occurred, and the 1-step ahead PI is for the time until the next event, the 2-step ahead PI is for the time until the next 2 events, and the h-step ahead PI is for the time for the next h events.

For forecasting, predict the test data $Y_{n+1}, ..., Y_{n+L}$ given the past training data $Y_1, ..., Y_n$. A large sample $100(1-\delta)\%$ prediction interval for Y_{n+h} is $[L_n, U_n]$ where the coverage $P(L_n \leq Y_{n+h} \leq U_n) = 1 - \delta_n$ is eventually bounded below by $1 - \delta$ as $n \to \infty$. We often want $1 - \delta_n \to 1 - \delta$ as $n \to \infty$. A large sample $100(1-\delta)\%$ PI is asymptotically optimal if it has the shortest asymptotic length: the length of $[L_n, U_n]$ converges to $U_s - L_s$ as $n \to \infty$ where $[L_s, U_s]$ is the population shorth: the shortest interval covering at least $100(1-\delta)\%$ of the mass.

For a large sample $100(1-\delta)\%$ PI, the nominal coverage is $100(1-\delta)\%$. Undercoverage occurs if the actual coverage is below the nominal coverage. For example, if the actual coverage is 0.93 when $n = 100$, then for a large sample 95% PI, the undercoverage is 0.02 =2%.

The prediction intervals and regions for the random walks are simple. First consider the random walk $Y_t = Y_{t-1} + e_t$ where the e_t are iid. Find the ϵ_i for $i = 1, ..., m = \lfloor n/h \rfloor$. Assume $n \geq 50h$ and let $[L, U]$ be the shorth(c) PI (4.4) for a future value of ϵ_f based on $\epsilon_1, ..., \epsilon_m$ with $m \geq 50$. Then the large sample $100(1-\delta)\%$ PI for Y_{n+h} is $[Y_n + L, Y_n + U]$. This PI tends to be asymptotically optimal as along as the e_t are iid. This PI is equivalent to applying the shorth(c) PI (4.4) on $Y_n + \epsilon_1, ..., Y_n + \epsilon_m$. Other PIs can be used.

For the vector valued random walk $\boldsymbol{Y}_t = \boldsymbol{Y}_{t-1} + \boldsymbol{e}_t$, find $\boldsymbol{\epsilon}_{1,h}, ..., \boldsymbol{\epsilon}_{m,h}$. The large sample $100(1-\delta)\%$ nonparametric prediction region (4.11) for a future value $\boldsymbol{\epsilon}_{f,h}$ is

$$\{\boldsymbol{z} : (\boldsymbol{z} - \overline{\boldsymbol{\epsilon}})^T \boldsymbol{S}_h^{-1} (\boldsymbol{z} - \overline{\boldsymbol{\epsilon}}) \leq D^2_{(U_m)}\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{\epsilon}}, \boldsymbol{S}_h) \leq D^2_{(U_m)}\} \qquad (9.10)$$

where $\boldsymbol{S}_h$ is the sample covariance matrix of the $\boldsymbol{\epsilon}_{i,h}$ and $D_i^2 = (\boldsymbol{\epsilon}_{i,h} - \overline{\boldsymbol{\epsilon}})^T \boldsymbol{S}_h^{-1} (\boldsymbol{\epsilon}_{i,h} - \overline{\boldsymbol{\epsilon}})$. This prediction region is a hyperellipsoid centered at the sample mean $\overline{\boldsymbol{\epsilon}}$. The following large sample $100(1-\delta)\%$ prediction region for $\boldsymbol{Y}_{n+h}$ shifts the hyperellipsoid (9.10) to be centered at $\boldsymbol{Y}_n + \overline{\boldsymbol{\epsilon}}$:

$$\{\boldsymbol{z} : [\boldsymbol{z} - (\boldsymbol{Y}_n + \overline{\boldsymbol{\epsilon}})]^T \boldsymbol{S}_h^{-1} [\boldsymbol{z} - (\boldsymbol{Y}_n + \overline{\boldsymbol{\epsilon}})] \leq D^2_{(U_m)}\}. \tag{9.11}$$

Since $\boldsymbol{Y}_{n+h}$ has the same distribution as $\boldsymbol{Y}_n + \boldsymbol{\epsilon}_{f,h}$, $P(\boldsymbol{Y}_{n+h} \in (9.11)) = P(\boldsymbol{\epsilon}_{f,h} \in (9.10)) = 1 - \delta_n$ which is bounded below by $1 - \delta$, asymptotically. The prediction region (9.11) is equivalent to applying the nonparametric prediction region (4.11) to $\boldsymbol{Y}_n + \boldsymbol{\epsilon}_{1,h}, ..., \boldsymbol{Y}_n + \boldsymbol{\epsilon}_{m,h}$. The prediction region (9.11) is similar to the Olive (2018) prediction region for the multivariate regression model.

Since the $\boldsymbol{\epsilon}_i = \boldsymbol{\epsilon}_{i,h}$ are iid, alternative prediction intervals and regions, such as those in Chapter 4, could be used.

9.7 Summary

9.8 Complements

Theorems 9.2 and 9.3 are from Haile and Olive (2023). The random walk material is from Haile, Zhang, and Olive (2024).

9.9 Problems

9.1. A *moving average* MA(q) times series is

$$Y_t = \tau + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} + e_t$$

where $\theta_q \neq 0$. Assume that the e_t are independent and identically distributed (iid) with zero mean and variance σ^2. Here τ, and the θ_i are unknown parameters. The Y_t are identically distributed. $Y_k, Y_{q+k+1}, Y_{2q+k+3}, Y_{3q+k+4}, ...,$ are iid. That is, there are blocks of iid data in the time series starting at $k = 1, ..., q, q+1$.

a) Find $E(Y_t)$.

b) Suppose the time series is $Y_1, Y_2, ..., Y_{n=m(q+1)}$. Consider the first iid block,

$$Y_1, Y_{q+2}, Y_{2q+3}, Y_{3q+4}, ..., Y_{(m-1)q+m} = Z_1, Z_2, ..., Z_m.$$

What does $\overline{Z}$ estimate?

c) Similarly, it can be shown that the sample percentile of each iid block estimates the "population percentile." Hence the sample percentile of the entire time series is a consistent estimator of the population percentile, and the sample shorth of the entire time series is a consistent estimator of the "population shorth." Suppose the sample variance of each iid block estimates σ_Y^2. What does $\sum_{i=1}^{n}(Y_i - \overline{Y})^2/n$ estimate?

9.2. Suppose $X_1, ..., X_{n_1}$ are iid, $Z_1, ..., Z_{n_2}$ are iid, and that the X_i and Z_i are identically distributed but not necessarily independent. (The random variables do need have the same probability space.) Assume $n_i/n \to \pi_i$ where $0 < \pi_i < 1$, $\pi_1 + \pi_2 = 1$ and $n_1 + n_2 = n$. Assume the kth moment $E(X_i^k) = E(Z_i^k) = E(Y^k)$ exists for $k = 1, ..., m$ for some integer $m \geq 2$. Then $\sum_{i=1}^{n_1} X_i^k/n_1 \xrightarrow{P} E(Y^k)$ and $\sum_{i=1}^{n_2} Z_i^k/n_2 \xrightarrow{P} E(Y^k)$. (For example, $Y_1, ..., Y_n$ could follow an MA(1) time series, the $X_i = Y_i$ for i odd, and the $Z_i = Y_i$ for i even.)

a) $\sum_{i=1}^{n_1} X_i^k/n \xrightarrow{P} a$. Find a.

b) $\sum_{i=1}^{n_2} Z_i^k/n \xrightarrow{P} b$. Find b.

c) $(\sum_{i=1}^{n_1} X_i^k + \sum_{i=1}^{n_2} Z_i^k)/n \xrightarrow{P} c$. Find c.

Chapter 10
Graphical Diagnostics

10.1 1D Regression

From Chapter 6, in a **1D regression model**, Y is conditionally independent of $\boldsymbol{x}$ given the **sufficient predictor** $SP = h(\boldsymbol{x})$, written

$$Y \perp\!\!\!\perp \boldsymbol{x} | SP \quad \text{or} \quad \mathrm{Y} \perp\!\!\!\perp \boldsymbol{x} | \mathrm{h}(\boldsymbol{x}), \tag{10.1}$$

where the real valued function $h : \mathbb{R}^p \to \mathbb{R}$. The **estimated sufficient predictor** $\mathrm{ESP} = \hat{h}(\boldsymbol{x})$.

Definition 10.1. A **response plot** is a plot of the ESP versus Y. A *residual plot* is a plot of the ESP versus the residuals.

A response plot is also called an *estimated sufficient summary plot* (ESSP). A sufficient summary plot is a plot of SP versus Y. Hence if the ESP is a consistent estimator of the SP, then the response plot estimates the sufficient summary plot.

Notation: In this text, a plot of x versus Y will have x on the horizontal axis, and Y on the vertical axis. For the *additive error regression* model $Y = m(\boldsymbol{x}) + e$, the ith residual is $r_i = Y_i - \hat{m}(\boldsymbol{x}_i) = Y_i - \hat{Y}_i$ where $\hat{Y}_i = \hat{m}(\boldsymbol{x}_i)$ is the ith fitted value. The additive error regression model is a 1D regression model with sufficient predictor $SP = h(\boldsymbol{x}) = m(\boldsymbol{x})$.

For the additive error regression model, the response plot is a plot of $\hat{Y}$ versus Y where the *identity line* with unit slope and zero intercept is added as a visual aid. The residual plot is a plot of $\hat{Y}$ versus r. Assume the errors e_i are iid from a unimodal distribution that is not highly skewed. Then the plotted points should scatter about the identity line and the $r = 0$ line (the horizontal axis) with no other pattern if the fitted model (that produces $\hat{m}(\boldsymbol{x})$) is good.

10.2 Plots for MLR

Theorem 10.1. Suppose that the MLR estimator $\boldsymbol{b}$ of $\boldsymbol{\beta}$ is used to find the residuals $r_i \equiv r_i(\boldsymbol{b})$ and the fitted values $\widehat{Y}_i \equiv \widehat{Y}_i(\boldsymbol{b}) = \boldsymbol{x}_i^T\boldsymbol{b}$. Then in the response plot of $\widehat{Y}_i$ versus Y_i, the vertical deviations from the identity line (that has unit slope and zero intercept) are the residuals $r_i(\boldsymbol{b})$.

Proof. The identity line in the response plot is $Y = \boldsymbol{x}^T\boldsymbol{b}$. Hence the vertical deviation is $Y_i - \boldsymbol{x}_i^T\boldsymbol{b} = r_i(\boldsymbol{b})$. □

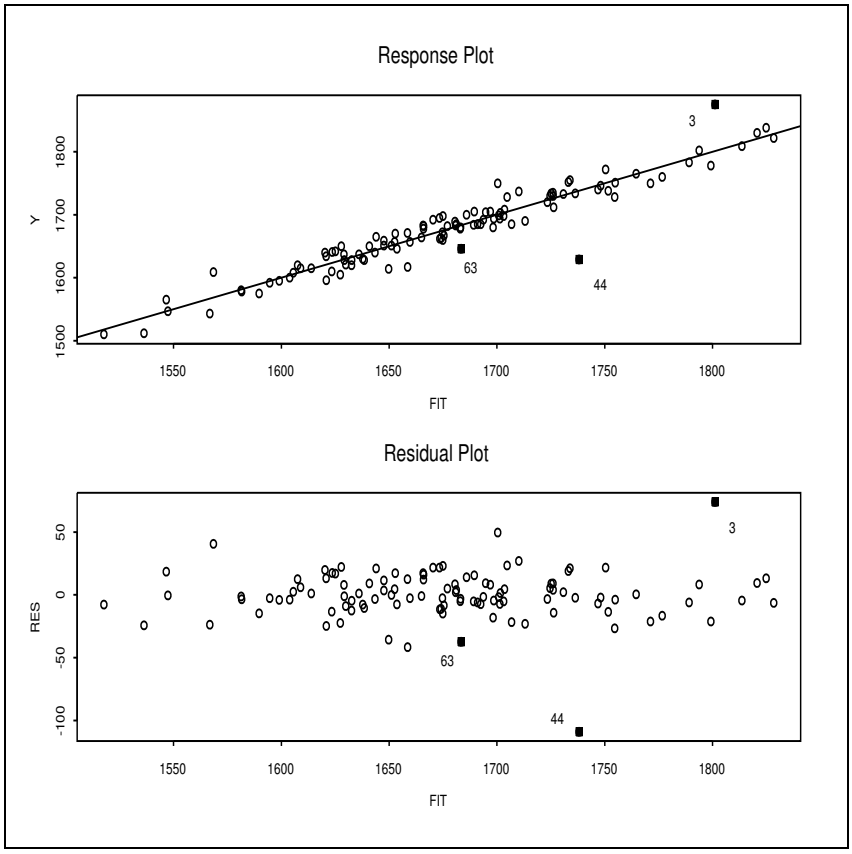

Fig. 10.1 Residual and Response Plots for the Tremearne Data

Example 10.1. Tremearne (1911) presents a data set of about 17 measurements on 115 people of Hausa nationality. We deleted 3 cases because of missing values and used *height* as the response variable Y. Along with a constant $x_{i,1} \equiv 1$, the five additional predictor variables used were *height*

when sitting, *height when kneeling*, *head length, nasal breadth*, and *span* (perhaps from left hand to right hand). Figure 6.1 presents the (ordinary) least squares (OLS) response and residual plots for this data set. These plots show that an MLR model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ should be a useful model for the data since the plotted points in the response plot are linear and follow the identity line while the plotted points in the residual plot follow the $r = 0$ line with no other pattern (except for a possible outlier marked 44). Note that many important acronyms, such as OLS and MLR, appear in Table 1.1.

To use the response plot to visualize the conditional distribution of $Y|\boldsymbol{x}^T\boldsymbol{\beta}$, use the fact that the fitted values $\hat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$. For example, suppose the height given fit = 1700 is of interest. Mentally examine the plot about a narrow vertical strip about fit = 1700, perhaps from 1685 to 1715. The cases in the narrow strip have a mean close to 1700 since they fall close to the identity line. Similarly, when the fit $= w$ for w between 1500 and 1850, the cases have heights near w, on average.

Cases 3, 44, and 63 are highlighted. The 3rd person was very tall while the 44th person was rather short. Beginners often label too many points as *outliers*: cases that lie far away from the bulk of the data. See Chapter 7. Mentally draw a box about the bulk of the data ignoring any outliers. Double the width of the box (about the identity line for the response plot and about the horizontal line for the residual plot). Cases outside of this imaginary doubled box are potential outliers. Alternatively, visually estimate the standard deviation of the residuals in both plots. In the residual plot look for residuals that are more than 5 standard deviations from the $r = 0$ line. In Figure 6.1, the standard deviation of the residuals appears to be around 10. Hence cases 3 and 44 are certainly worth examining.

The identity line can also pass through or near an outlier or a cluster of outliers. Then the outliers will be in the upper right or lower left of the response plot, and there will be a large gap between the cluster of outliers and the bulk of the data. Figure 6.1 was made with the following *R* commands, using *lspack* function `MLRplot` and the *major.lsp* data set from the text's webpage.

```
major <- matrix(scan(),nrow=112,ncol=7,byrow=T)
#copy and paste the data set, then press enter
major <- major[,-1]
X<-major[,-6]
Y <- major[,6]
MLRplot(X,Y) #left click the 3 highlighted cases,
#then right click Stop for each of the two plots
```

10.2.1 Plots for Variable Selection

Two important summaries for submodel I are $R^2(I)$, the proportion of the variability of Y explained by the nontrivial predictors in the model, and $MSE(I) = \hat{\sigma}_I^2$, the estimated error variance. Suppose that model I contains k predictors, including a constant. Since adding predictors does not decrease R^2, the adjusted $R_A^2(I)$ is often used, where

$$R_A^2(I) = 1 - (1 - R^2(I))\frac{n}{n-k} = 1 - MSE(I)\frac{n}{SST}.$$

See Seber and Lee (2003, pp. 400-401). Hence the model with the maximum $R_A^2(I)$ is also the model with the minimum $MSE(I)$.

For multiple linear regression, recall that if the candidate model of $\boldsymbol{x}_I$ has k terms (including the constant), then the partial F statistic for testing whether the $p-k$ predictor variables in $\boldsymbol{x}_O$ can be deleted is

$$F_I = \frac{SSE(I) - SSE}{(n-k)-(n-p)} / \frac{SSE}{n-p} = \frac{n-p}{p-k}\left[\frac{SSE(I)}{SSE} - 1\right]$$

where SSE is the error sum of squares from the full model, and SSE(I) is the error sum of squares from the candidate submodel. An extremely important criterion for variable selection is the C_p criterion.

Definition 10.2.

$$C_p(I) = \frac{SSE(I)}{MSE} + 2k - n = (p-k)(F_I - 1) + k$$

where MSE is the error mean square for the full model.

Note that when H_0 is true, $(p-k)(F_I-1)+k \xrightarrow{D} \chi^2_{p-k} + 2k - p$ for a large class of iid error distributions. Minimizing $C_p(I)$ is equivalent to minimizing $MSE\ [C_p(I)] = SSE(I) + (2k-n)MSE = \boldsymbol{r}^T(I)\boldsymbol{r}(I) + (2k-n)MSE$. The following theorem helps explain why C_p is a useful criterion and suggests that for subsets I with k terms, submodels with $C_p(I) \leq \min(2k, p)$ are especially interesting. Olive and Hawkins (2005) show that this interpretation of C_p can be generalized to 1D regression models with a linear predictor $\boldsymbol{\beta}^T\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{\beta}$, such as generalized linear models. Denote the residuals and fitted values from the *full model* by $r_i = Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} = Y_i - \hat{Y}_i$ and $\hat{Y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ respectively. Similarly, let $\hat{\boldsymbol{\beta}}_I$ be the estimate of $\boldsymbol{\beta}_I$ obtained from the regression of Y on $\boldsymbol{x}_I$ and denote the corresponding residuals and fitted values by $r_{I,i} = Y_i - \boldsymbol{x}_{I,i}^T\hat{\boldsymbol{\beta}}_I$ and $\hat{Y}_{I,i} = \boldsymbol{x}_{I,i}^T\hat{\boldsymbol{\beta}}_I$ where $i = 1, ..., n$.

Theorem 10.2. Suppose that a numerical variable selection method suggests several submodels with k predictors, including a constant, where $2 \leq k \leq p$.

a) The model I that minimizes $C_p(I)$ maximizes $\mathrm{corr}(r, r_I)$.

b) $C_p(I) \leq 2k$ implies that $\mathrm{corr}(\mathrm{r}, \mathrm{r_I}) \geq \sqrt{1 - \frac{\mathrm{p}}{\mathrm{n}}}$.

c) As $\mathrm{corr}(r, r_I) \to 1$,

$$\mathrm{corr}(\boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}, \boldsymbol{x}_{\mathrm{I}}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{I}}) = \mathrm{corr}(\mathrm{ESP}, \mathrm{ESP(I)}) = \mathrm{corr}(\hat{\mathrm{Y}}, \hat{\mathrm{Y}}_{\mathrm{I}}) \to 1.$$

Proof. These results are a corollary of Theorem 4.2 below. □

Remark 10.1. Consider the model I_i that deletes the predictor x_i. Then the model has $k = p - 1$ predictors including the constant, and the test statistic is t_i where

$$t_i^2 = F_{I_i}.$$

Using Definition 4.2 and $C_p(I_{full}) = p$, it can be shown that

$$C_p(I_i) = C_p(I_{full}) + (t_i^2 - 2).$$

Using the screen $C_p(I) \leq \min(2k, p)$ suggests that the predictor x_i should not be deleted if

$$|t_i| > \sqrt{2} \approx 1.414.$$

If $|t_i| < \sqrt{2}$ then the predictor can probably be deleted since C_p decreases. The literature suggests using the $C_p(I) \leq k$ screen, but this screen eliminates too many potentially useful submodels.

More generally, it can be shown that $C_p(I) \leq 2k$ iff

$$F_I \leq \frac{p}{p-k}.$$

Now k is the number of terms in the model I including a constant while $p-k$ is the number of terms set to 0. As $k \to 0$, the partial F test will reject Ho: $\boldsymbol{\beta}_O = \mathbf{0}$ (i.e. say that the full model should be used instead of the submodel I) unless F_I is not much larger than 1. If p is very large and $p - k$ is very small, then the partial F test will tend to suggest that there is a model I that is about as good as the full model even though model I deletes $p - k$ predictors.

Definition 10.3. The "fit–fit" or *FF plot* is a plot of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$ while a "residual–residual" or *RR plot* is a plot $r_{I,i}$ versus r_i. A *response plot* is a plot of $\hat{Y}_{I,i}$ versus Y_i. An *EE plot* is a plot of ESP(I) versus ESP. For MLR, the EE and FF plots are equivalent.

Six graphs will be used to compare the full model and the candidate submodel: the FF plot, RR plot, the response plots from the full and submodel, and the residual plots from the full and submodel. These six plots will contain a great deal of information about the candidate subset provided that Equation (4.1) holds and that a good estimator (such as OLS) for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_I$ is used.

Application 10.1. To visualize whether a candidate submodel using predictors $\boldsymbol{x}_I$ is good, use the fitted values and residuals from the submodel and full model to make an RR plot of the $r_{I,i}$ versus the r_i and an FF plot of $\hat{Y}_{I,i}$ versus $\hat{Y}_i$. Add the OLS line to the RR plot and identity line to both plots as visual aids. The subset I is good if the plotted points cluster tightly about the identity line in *both plots*. In particular, the OLS line and the identity line should "nearly coincide" so that it is difficult to tell that the two lines intersect at the origin in the RR plot.

To verify that the six plots are useful for assessing variable selection, the following notation will be useful. Suppose that all submodels include a constant and that $\boldsymbol{X}$ is the full rank $n \times p$ design matrix for the full model. Let the corresponding vectors of OLS fitted values and residuals be $\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$ and $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$, respectively. Suppose that $\boldsymbol{X}_I$ is the $n \times k$ design matrix for the candidate submodel and that the corresponding vectors of OLS fitted values and residuals are $\hat{\boldsymbol{Y}}_I = \boldsymbol{X}_I(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{Y} = \boldsymbol{H}_I\boldsymbol{Y}$ and $\boldsymbol{r}_I = (\boldsymbol{I} - \boldsymbol{H}_I)\boldsymbol{Y}$, respectively.

A plot can be very useful if the OLS line can be compared to a reference line and if the OLS slope is related to some quantity of interest. Suppose that a plot of w versus z places w on the horizontal axis and z on the vertical axis. Then denote the OLS line by $\hat{z} = a + bw$. The following theorem shows that the plotted points in the FF, RR, and response plots will cluster about the identity line. Notice that the theorem is a property of OLS and holds even if the data does not follow an MLR model. Let $\text{corr}(x, y)$ denote the correlation between x and y.

Theorem 10.3. Suppose that every submodel contains a constant and that $\boldsymbol{X}$ is a full rank matrix.
Response Plot: i) If $w = \hat{Y}_I$ and $z = Y$ then the OLS line is the identity line.
ii) If $w = Y$ and $z = \hat{Y}_I$ then the OLS line has slope $b = [\text{corr}(Y, \hat{Y}_I)]^2 = R^2(I)$ and intercept $a = \overline{Y}(1 - R^2(I))$ where $\overline{Y} = \sum_{i=1}^n Y_i/n$ and $R^2(I)$ is the coefficient of multiple determination from the candidate model.
FF or EE Plot: iii) If $w = \hat{Y}_I$ and $z = \hat{Y}$ then the OLS line is the identity line. Note that $ESP(I) = \hat{Y}_I$ and $ESP = \hat{Y}$.
iv) If $w = \hat{Y}$ and $z = \hat{Y}_I$ then the OLS line has slope $b = [\text{corr}(\hat{Y}, \hat{Y}_I)]^2 = SSR(I)/SSR$ and intercept $a = \overline{Y}[1 - (SSR(I)/SSR)]$ where SSR is the regression sum of squares.

RR Plot: v) If $w = r$ and $z = r_I$ then the OLS line is the identity line.
vi) If $w = r_I$ and $z = r$ then $a = 0$ and the OLS slope $b = [\text{corr}(r, r_I)]^2$ and

$$\text{corr}(r, r_I) = \sqrt{\frac{SSE}{SSE(I)}} = \sqrt{\frac{n-p}{C_p(I) + n - 2k}} = \sqrt{\frac{n-p}{(p-k)F_I + n - p}}.$$

Proof: Recall that $\boldsymbol{H}$ and $\boldsymbol{H}_I$ are symmetric idempotent matrices and that $\boldsymbol{H}\boldsymbol{H}_I = \boldsymbol{H}_I$. The mean of OLS fitted values is equal to $\overline{Y}$ and the mean of OLS residuals is equal to 0. If the OLS line from regressing z on w is $\hat{z} = a + bw$, then $a = \overline{z} - b\overline{w}$ and

$$b = \frac{\sum(w_i - \overline{w})(z_i - \overline{z})}{\sum(w_i - \overline{w})^2} = \frac{SD(z)}{SD(w)}\text{corr}(z, w).$$

Also recall that the OLS line passes through the means of the two variables $(\overline{w}, \overline{z})$.

(*) Notice that the OLS slope from regressing z on w is equal to one if and only if the OLS slope from regressing w on z is equal to $[\text{corr}(z, w)]^2$.

i) The slope $b = 1$ if $\sum \hat{Y}_{I,i} Y_i = \sum \hat{Y}_{I,i}^2$. This equality holds since $\hat{\boldsymbol{Y}}_I^T \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{H}_I \boldsymbol{Y} = \hat{\boldsymbol{Y}}_I^T \hat{\boldsymbol{Y}}_I$. Since $b = 1$, $a = \overline{Y} - \overline{Y} = 0$.

ii) By (*), the slope

$$b = [\text{corr}(Y, \hat{Y}_I)]^2 = R^2(I) = \frac{\sum(\hat{Y}_{I,i} - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2} = SSR(I)/SSTO.$$

The result follows since $a = \overline{Y} - b\overline{Y}$.

iii) The slope $b = 1$ if $\sum \hat{Y}_{I,i} \hat{Y}_i = \sum \hat{Y}_{I,i}^2$. This equality holds since $\hat{\boldsymbol{Y}}^T \hat{\boldsymbol{Y}}_I = \boldsymbol{Y}^T \boldsymbol{H}\boldsymbol{H}_I \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{Y} = \hat{\boldsymbol{Y}}_I^T \hat{\boldsymbol{Y}}_I$. Since $b = 1$, $a = \overline{Y} - \overline{Y} = 0$.

iv) From iii),

$$1 = \frac{SD(\hat{Y})}{SD(\hat{Y}_I)}[\text{corr}(\hat{Y}, \hat{Y}_I)].$$

Hence

$$\text{corr}(\hat{Y}, \hat{Y}_I) = \frac{SD(\hat{Y}_I)}{SD(\hat{Y})}$$

and the slope

$$b = \frac{SD(\hat{Y}_I)}{SD(\hat{Y})}\text{corr}(\hat{Y}, \hat{Y}_I) = [\text{corr}(\hat{Y}, \hat{Y}_I)]^2.$$

Also the slope

$$b = \frac{\sum(\hat{Y}_{I,i} - \overline{Y})^2}{\sum(\hat{Y}_i - \overline{Y})^2} = SSR(I)/SSR.$$

The result follows since $a = \overline{Y} - b\overline{Y}$.

v) The OLS line passes through the origin. Hence $a = 0$. The slope $b = \boldsymbol{r}^T\boldsymbol{r}_I/\boldsymbol{r}^T\boldsymbol{r}$. Since $\boldsymbol{r}^T\boldsymbol{r}_I = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}_I)\boldsymbol{Y}$ and $(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}_I) = \boldsymbol{I} - \boldsymbol{H}$, the numerator $\boldsymbol{r}^T\boldsymbol{r}_I = \boldsymbol{r}^T\boldsymbol{r}$ and $b = 1$.

vi) Again $a = 0$ since the OLS line passes through the origin. From v),

$$1 = \sqrt{\frac{SSE(I)}{SSE}}[\text{corr}(r, r_I)].$$

Hence

$$\text{corr}(r, r_I) = \sqrt{\frac{SSE}{SSE(I)}}$$

and the slope

$$b = \sqrt{\frac{SSE}{SSE(I)}}[\text{corr}(r, r_I)] = [\text{corr}(r, r_I)]^2.$$

Algebra shows that

$$\text{corr}(r, r_I) = \sqrt{\frac{n-p}{C_p(I) + n - 2k}} = \sqrt{\frac{n-p}{(p-k)F_I + n - p}}. \quad \square$$

Remark 10.2. Let I_{min} be the model than minimizes $C_p(I)$ among the models I generated from the variable selection method such as forward selection. Assuming the the full model I_p is one of the models generated, then $C_p(I_{min}) \leq C_p(I_p) = p$, and $\text{corr}(r, r_{I_{min}}) \to 1$ as $n \to \infty$ by Theorem 4.2 vi). Referring to Equation (4.1), if $P(S \subseteq I_{min})$ does not go to 1 as $n \to \infty$, then the above correlation would not go to one. Hence $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$.

10.2.2 Plots for Response Transformations

10.3 Plots for GLMs and GAMs

10.4 Outlier Detection for the MLD Model

Now suppose the multivariate data has been collected into an $n \times p$ matrix

$$\boldsymbol{W} = \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_p \end{bmatrix}$$

where the ith row of $\boldsymbol{W}$ is the ith case $\boldsymbol{x}_i^T$ and the jth column $\boldsymbol{v}_j$ of $\boldsymbol{W}$ corresponds to n measurements of the jth random variable X_j for $j = 1, ..., p$. Hence the n rows of the data matrix $\boldsymbol{W}$ correspond to the n cases, while the p columns correspond to measurements on the p random variables $X_1, ..., X_p$. For example, the data may consist of n visitors to a hospital where the $p = 2$ variables *height* and *weight* of each individual were measured.

Definition 10.36. The *coordinatewise median* $\text{MED}(\boldsymbol{W}) = (\text{MED}(X_1), ..., \text{MED}(X_p))^T$ where $\text{MED}(X_i)$ is the sample median of the data in column i corresponding to variable X_i and $\boldsymbol{v}_i$.

Example 10.11. Let the data for X_1 be $1, 2, 3, 4, 5, 6, 7, 8, 9$ while the data for X_2 is $7, 17, 3, 8, 6, 13, 4, 2, 1$. Then $\text{MED}(\boldsymbol{W}) = (\text{MED}(X_1), \text{MED}(X_2))^T = (5, 6)^T$.

Definition 10.37: Rousseeuw and Van Driessen (1999). The *DD plot* is a plot of the classical Mahalanobis distances MD_i versus robust Mahalanobis distances RD_i.

The DD plot is used as a diagnostic for multivariate normality, elliptical symmetry, and for outliers. Assume that the data set consists of iid vectors from an $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with second moments. See Section 1.7 for notation. Then the classical sample mean and covariance matrix $(T_M, \boldsymbol{C}_M) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ is a consistent estimator for $(\boldsymbol{\mu}, c_{\boldsymbol{x}}\boldsymbol{\Sigma}) = (E(\boldsymbol{x}), \text{Cov}(\boldsymbol{x}))$. Assume that an alternative algorithm estimator $(T_A, \boldsymbol{C}_A)$ is a consistent estimator for $(\boldsymbol{\mu}, a_A\boldsymbol{\Sigma})$ for some constant $a_A > 0$. By scaling the algorithm estimator, the DD plot can be constructed to follow the identity line with unit slope and zero intercept. Let $(T_R, \boldsymbol{C}_R) = (T_A, \boldsymbol{C}_A/\tau^2)$ denote the scaled algorithm estimator where $\tau > 0$ is a constant to be determined. Notice that $(T_R, \boldsymbol{C}_R)$ is a valid estimator of location and dispersion. Hence the robust distances used in the DD plot are given by

$$\mathrm{RD}_i = \mathrm{RD}_i(T_R, \boldsymbol{C}_R) = \sqrt{(\boldsymbol{x}_i - T_R(\boldsymbol{W}))^T [\boldsymbol{C}_R(\boldsymbol{W})]^{-1} (\boldsymbol{x}_i - T_R(\boldsymbol{W}))}$$

$= \tau\ D_i(T_A, \boldsymbol{C}_A)$ for $i = 1, ..., n$.

The following theorem shows that if consistent estimators are used to construct the distances, then the DD plot will tend to cluster tightly about the line segment through $(0, 0)$ and $(\mathrm{MD}_{n,\alpha}, \mathrm{RD}_{n,\alpha})$ where $0 < \alpha < 1$ and $\mathrm{MD}_{n,\alpha}$ is the 100αth sample percentile of the MD_i. Nevertheless, the variability in the DD plot may increase with the distances. Let $K > 0$ be a constant, e.g. the 99th percentile of the χ^2_p distribution.

Theorem 10.32. Assume that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid observations from a distribution with parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a symmetric positive definite matrix. Let $a_j > 0$ and assume that $(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$ are consistent estimators of $(\boldsymbol{\mu}, a_j \boldsymbol{\Sigma})$ for $j = 1, 2$.

a) $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - \frac{1}{a_j} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = o_P(1)$.

b) Let $0 < \delta \leq 0.5$. If $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - (\boldsymbol{\mu}, a_j \boldsymbol{\Sigma}) = O_p(n^{-\delta})$ and $a_j \hat{\boldsymbol{\Sigma}}_j^{-1} - \boldsymbol{\Sigma}^{-1} = O_P(n^{-\delta})$, then

$$D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - \frac{1}{a_j} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = O_P(n^{-\delta}).$$

c) Let $D_{i,j} \equiv D_i(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$ be the ith Mahalanobis distance computed from $(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$. Consider the cases in the region $R = \{i | 0 \leq D_{i,j} \leq K,\ j = 1, 2\}$. Let r_n denote the correlation between $D_{i,1}$ and $D_{i,2}$ for the cases in R (thus r_n is the correlation of the distances in the "lower left corner" of the DD plot). Then $r_n \to 1$ in probability as $n \to \infty$.

Proof. Let B_n denote the subset of the sample space on which both $\hat{\boldsymbol{\Sigma}}_{1,n}$ and $\hat{\boldsymbol{\Sigma}}_{2,n}$ have inverses. Then $P(B_n) \to 1$ as $n \to \infty$.

a) and b): $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) =$

$$(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} - \frac{\boldsymbol{\Sigma}^{-1}}{a_j} + \hat{\boldsymbol{\Sigma}}_j^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)$$

$$= (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{-\boldsymbol{\Sigma}^{-1}}{a_j} + \hat{\boldsymbol{\Sigma}}_j^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) + (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)$$

$$= \frac{1}{a_j} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T (-\boldsymbol{\Sigma}^{-1} + a_j\ \hat{\boldsymbol{\Sigma}}_j^{-1}) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)\ +$$

$$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} \right) (\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)$$

$$= \frac{1}{a_j} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$$

$$+\frac{2}{a_j}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_j)+\frac{1}{a_j}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_j)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_j)$$

$$+\frac{1}{a_j}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j)^T[a_j\hat{\boldsymbol{\Sigma}}_j^{-1}-\boldsymbol{\Sigma}^{-1}](\boldsymbol{x}-\hat{\boldsymbol{\mu}}_j) \tag{10.2}$$

on B_n, and the last three terms are $o_P(1)$ under a) and $O_P(n^{-\delta})$ under b).

c) Following the proof of a), $D_j^2 \equiv D_{\boldsymbol{x}}^2(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) \xrightarrow{P} (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/a_j$ for fixed $\boldsymbol{x}$, and the result follows. □

The above result implies that a plot of the MD_i versus the $D_i(T_A, \boldsymbol{C}_A) \equiv D_i(A)$ will follow a line through the origin with some positive slope since if $\boldsymbol{x} = \boldsymbol{\mu}$, then both the classical and the algorithm distances should be close to zero. We want to find τ such that $\mathrm{RD}_i = \tau\ D_i(T_A, \boldsymbol{C}_A)$ and the DD plot of MD_i versus RD_i follows the identity line. By Theorem 10.32, the plot of MD_i versus $D_i(A)$ will follow the line segment defined by the origin $(0, 0)$ and the point of observed median Mahalanobis distances, $(\mathrm{med}(\mathrm{MD}_i), \mathrm{med}(D_i(A)))$. This line segment has slope

$$\mathrm{med}(D_i(A))/\mathrm{med}(\mathrm{MD}_i)$$

which is generally not one. By taking $\tau = \mathrm{med}(\mathrm{MD}_i)/\mathrm{med}(D_i(A))$, the plot will follow the identity line if $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ is a consistent estimator of $(\boldsymbol{\mu}, c_{\boldsymbol{x}}\boldsymbol{\Sigma})$ and if $(T_A, \boldsymbol{C}_A)$ is a consistent estimator of $(\boldsymbol{\mu}, a_A\boldsymbol{\Sigma})$. (Using the notation from Theorem 10.32, let $(a_1, a_2) = (c_{\boldsymbol{x}}, a_A)$.) The classical estimator is consistent if the population has a nonsingular covariance matrix. The algorithm estimators $(T_A, \boldsymbol{C}_A)$ from Theorem 8.29 are consistent on a large class of EC distributions that have a nonsingular covariance matrix, but tend to be biased for non–EC distributions. We recommend using RFCH or RMVN as the robust estimators in DD plots.

By replacing the observed median $\mathrm{med}(\mathrm{MD}_i)$ of the classical Mahalanobis distances with the target population analog, say MED, τ can be chosen so that the DD plot is *simultaneously* a diagnostic for elliptical symmetry and a diagnostic for the target EC distribution. That is, the plotted points follow the identity line if the data arise from a target EC distribution such as the multivariate normal distribution, but the points follow a line with non-unit slope if the data arise from an alternative EC distribution. In addition the DD plot can often detect departures from elliptical symmetry such as outliers, the presence of two groups, or the presence of a mixture distribution.

Example 10.12. We will use the multivariate normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution as the target. If the data are indeed iid MVN vectors, then the $(\mathrm{MD}_i)^2$ are asymptotically χ_p^2 random variables, and $\mathrm{MED} = \sqrt{\chi^2_{p,0.5}}$ where $\chi^2_{p,0.5}$ is the median of the χ_p^2 distribution. Since the target distribution is Gaussian, let

$$\mathrm{RD}_i = \frac{\sqrt{\chi^2_{p,0.5}}}{\mathrm{med}(D_i(A))} D_i(A) \quad \text{so that} \quad \tau = \frac{\sqrt{\chi^2_{\mathrm{p},0.5}}}{\mathrm{med}(\mathrm{D_i(A)})}. \tag{10.3}$$

Since every nonsingular estimator of multivariate location and dispersion defines a hyperellipsoid, the DD plot can be used to examine which points are in the robust hyperellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T_R)^T \boldsymbol{C}_R^{-1} (\boldsymbol{x} - T_R) \leq RD^2_{(h)}\} \tag{10.4}$$

where $RD^2_{(h)}$ is the hth smallest squared robust Mahalanobis distance, and which points are in a classical hyperellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - \overline{\boldsymbol{x}})^T \boldsymbol{S}^{-1} (\boldsymbol{x} - \overline{\boldsymbol{x}}) \leq MD^2_{(h)}\}. \tag{10.5}$$

In the DD plot, points below $RD_{(h)}$ correspond to cases that are in the hyperellipsoid given by Equation (10.19) while points to the left of $MD_{(h)}$ are in a hyperellipsoid determined by Equation (10.20). In particular, we can use the DD plot to examine which points are in the nonparametric prediction region (4.11).

Application 10.5. Consider the DD plot with RFCH or RMVN. The DD plot can be used *simultaneously* as a diagnostic for whether the data arise from a multivariate normal distribution or from another EC distribution with nonsingular covariance matrix. EC data will cluster about a straight line through the origin; MVN data in particular will cluster about the identity line. Thus the DD plot can be used to assess the success of numerical transformations towards elliptical symmetry. The DD plot can be used to detect multivariate outliers. Use the DD plot to detect outliers and leverage groups if $n \geq 10p$ for the predictor variables in regression.

Fig. 10.2 4 DD Plots

For this application, the RFCH and RMVN estimators may be best. For MVN data, the RD_i from the RFCH estimator tend to have a higher correlation with the MD_i from the classical estimator than the RD_i from the FCH estimator, and the `cov.mcd` estimator may be inconsistent.

Figure 10.12 shows the DD plots for 3 artificial data sets using `cov.mcd`. The DD plot for 200 $N_3(\mathbf{0}, \boldsymbol{I}_3)$ points shown in Figure 10.12a resembles the identity line. The DD plot for 200 points from the elliptically contoured distribution $0.6N_3(\mathbf{0}, \boldsymbol{I}_3) + 0.4N_3(\mathbf{0}, 25\ \boldsymbol{I}_3)$ in Figure 10.12b clusters about a line through the origin with a slope close to 2.0.

A *weighted DD plot* magnifies the lower left corner of the DD plot by omitting the cases with $\mathrm{RD}_i \geq \sqrt{\chi^2_{p,.975}}$. This technique can magnify features that are obscured when large RD_i's are present. If the distribution of $\boldsymbol{x}$ is EC with nonsingular $\boldsymbol{\Sigma}$, Theorem 8.32 implies that the correlation of the points in the weighted DD plot will tend to one and that the points will cluster about a line passing through the origin. For example, the plotted points in the weighted DD plot (not shown) for the non-MVN EC data of Figure 10.12b are highly correlated and still follow a line through the origin with a slope close to 2.0.

Figures 10.12c and 10.12d illustrate how to use the weighted DD plot. The ith case in Figure 10.12c is $(\exp(x_{i,1}), \exp(x_{i,2}), \exp(x_{i,3}))^T$ where $\boldsymbol{x}_i$ is the ith case in Figure 10.12a; i.e. the marginals follow a lognormal distribution. The plot does not resemble the identity line, correctly suggesting that the distribution of the data is not MVN; however, the correlation of the plotted points is rather high. Figure 10.12d is the weighted DD plot where cases with $\mathrm{RD}_i \geq \sqrt{\chi^2_{3,.975}} \approx 3.06$ have been removed. Notice that the correlation of the plotted points is not close to one and that the best fitting line in Figure 10.12d may not pass through the origin. These results suggest that the distribution of $\boldsymbol{x}$ is not EC.

Fig. 10.3 DD Plots for the Buxton Data

Example 10.13. Buxton (1920, pp. 232-5) gave 20 measurements of 88 men. We will examine whether the multivariate normal distribution is a reasonable model for the measurements *head length, nasal height, bigonal breadth,* and *cephalic index* where one case has been deleted due to missing values. Figure 10.13a shows the DD plot. Five head lengths were recorded to be around 5 feet and are massive outliers. Figure 10.13b is the DD plot computed after deleting these points and suggests that the multivariate normal distribution is reasonable. (The recomputation of the DD plot means that the plot is not a weighted DD plot which would simply omit the outliers and then rescale the vertical axis.)

```
library(MASS)
x <- cbind(buxy,buxx)
ddplot(x,type=3) #Figure 7.13a), right click Stop

zx <- x[-c(61:65),]
ddplot(zx,type=3) #Figure 7.13b), right click Stop
```

10.5 Summary

10.6 Complements

10.7 Problems

Chapter 11
More Results

11.1 Martingales

Martingales use conditional expectation.

Remark 11.1. It can be shown that $E(Z_n) \in \mathbb{R}$ for each n iff $E(|X_n|) < \infty$ for each n. Technically, the inequalities in a), b), and c) hold with probability one or ae, but this is always understood and usually omitted in this Section. In the following definition, $Z_n = g_n(X_1, ..., X_n)$ where the function g_n can depend on n.

Definition 11.1. Let $Z_1, Z_2, ...$ be a sequence of random variables such that Z_n is a function of the random variables $X_1, ..., X_n$. Assume $E(Z_n) \in \mathbb{R}$ for each $n \geq 1$.
a) The sequence $\{Z_n\}$ is a *martingale* if $E(Z_{n+1}|X_1, ..., X_n) = Z_n$.
b) The sequence $\{Z_n\}$ is a *submartingale* if $E(Z_{n+1}|X_1, ..., X_n) \geq Z_n$.
c) The sequence $\{Z_n\}$ is a *supermartingale* if $E(Z_{n+1}|X_1, ..., X_n) \leq Z_n$.

Example 11.1. Note that $Z_n = X_n$ is a function of $X_1, ..., X_n$. Assume $E(X_n) \in \mathbb{R}$ for each n.
a) The sequence $\{X_n\}$ is a *martingale* if $E(X_{n+1}|X_1, ..., X_n) = X_n$.
b) The sequence $\{X_n\}$ is a *submartingale* if $E(X_{n+1}|X_1, ..., X_n) \geq X_n$.
c) The sequence $\{X_n\}$ is a *supermartingale* if $E(X_{n+1}|X_1, ..., X_n) \leq X_n$.

Remark 11.2. a) Try to write Z_{n+1} as a function of Z_n and X_{n+1} to show that Definition 11.1 a) holds.
b) If $Z_n = g(X_1, ..., X_n)$ where g is a function, then $E(Z_n W|X_1, ..., X_n) = Z_n E(W|X_1, ..., X_n)$.
c) If $E(W)$ exists and W is independent of $X_1, ..., X_n$, then $E(W|X_1, ..., X_n) = E(W)$.

Example 11.2. Note that a martingale is both a submartingale and a supermartingale.

Remark 11.3. If $\{Z_n\}$ is a martingale, then $E(Z_{n+k}|X_1, ..., X_n) = Z_n$ for $n, k = 1, 2, ...$ with corresponding results for sub- and supermartingales.

Theorem 11.1. Submartingale Convergence Theorem. Let $\{Z_n\}$ be a submartingale. If $K = \sup_n E[|Z_n|] \leq \infty$, then $Z_n \xrightarrow{ae} Z$ where Z is a random variable with $E[|X|] \leq Z$.

Example 11.3. a) Let $X_1, X_2, ...$ be independent random variables with $E(X_k) = 0$ for $k = 1, 2, ...$ Let the sum $S_n = \sum_{i=1}^n X_i = X_1 + X_2 + \cdots + X_n$. Show that $\{S_n\}$ is a martingale.

Proof. $E(S_n) = 0$ for each n. Now $E(S_{n+1}|\mathcal{F}_n) = E(S_{n+1}|X_1, ..., X_n) = E(S_n + X_{n+1}|X_1, ..., X_n) = E(S_n|X_1, ..., X_n) + E(X_{n+1}|X_1, ..., X_n) = S_n + E(X_{n+1}) = S_n$ for each n since S_n is a function of $X_1, ..., X_n$ and X_{n+1} is independent of $X_1, ..., X_n$.

b) Let $X_1, X_2, ...$ be independent random variables with $E(X_k) = 0$ and finite variances $\sigma_k^2 = E(X_k^2)$ for $k = 1, 2, ...$ Let $s_n^2 = \sigma_1^2 + \cdots + \sigma_n^2$. Then $s_n^2 = V(S_n)$ where S_n is given in a). Let $Y_n = S_n^2 - s_n^2$. Show that $\{Y_n\}$ is a martingale.

Proof. $E(|Y_n|) \leq E(S_n^2) + s_n^2 = 2s_n^2 < \infty$ for each n. Hence $E(Y_n) \in \mathbb{R}$ by Remark 11.1. Now $E(S_{n+1}^2|X_1, ..., X_n) = E[(S_n + X_{n+1})^2|X_1, ..., X_n) = E(S_n^2 + 2S_nX_{n+1} + X_{n+1}^2|X_1, ..., X_n) = E(S_n^2|X_1, ..., X_n) + 2S_nE(X_{n+1}|X_1, ..., X_n) + E(X_{n+1}^2|X_1, ..., X_n) = S_n^2 + 2S_nE(X_{n+1}) + E(X_{n+1}^2) = S_n^2 + \sigma_{n+1}^2$. Thus $E(Y_{n+1}|X_1, ..., X_n) = E(S_{n+1}^2 - s_{n+1}^2|X_1, ..., X_n) = E(S_{n+1}^2|X_1, ..., X_n) - s_{n+1}^2 = S_n^2 + \sigma_{n+1}^2 - (s_n^2 + \sigma_{n+1}^2) = S_n^2 - s_n^2 = Y_n$ for each n.

c) Let $X_1, X_2...$ be independent nonnegative random variables with $E(X_k) = 1$ for $k = 1, 2, ...$ Let $Y_n = \prod_{i=1}^n X_i$. i) Show that $\{Y_n\}$ is a martingale. ii) Is there a random variable Y such that $Y_n \xrightarrow{ae} Y$?

Proof. i) $E(Y_n) = \prod_{i=1}^n E(X_n) = 1$. Now $E(Y_{n+1}|X_1, ..., X_n) = E(Y_nX_{n+1}|X_1, ..., X_n) = Y_nE(X_{n+1}|X_1, ..., X_n) = Y_nE(X_{n+1}) = Y_n$ for each n.
ii) Since the X_i are nonnegative, $E(|Y_n|) = E(Y_n) = \prod_{i=1}^n E(X_i) = 1 = K$. Thus by Theorem 11.1, there does exist RV Y such that $Y_n \xrightarrow{ae} Y$.

d) Let $X_1, X_2, ...$ be independent random variables with $E(X_k) = \mu_k$ for $k = 1, 2, ...$ Let $T_n = \sum_{k=1}^n (X_k - \mu_k)$. Show that $\{T_n\}$ is a martingale.

Proof. $E(T_n) = 0$. Now $E(T_{n+1}|X_1, ..., X_n) = E(T_n + X_{n+1} - \mu_{n+1}|X_1, ..., X_n) = T_n + E(X_{n+1}|X_1, ..., X_n) - \mu_{n+1} = T_n + E(X_{n+1}) - \mu_{n+1} = T_n$.

The following result is useful for proving Theorem 11.2. Let g be a one to one and onto function so that the inverse function g^{-1} exists. For example, g could be increasing, decreasing, convex, or concave (some texts add the adjective "strictly"). Then $E(W|X_1, ..., X_n) = E(W|g(X_1), ..., g(X_n))$ since X_i is known iff $g(X_i)$ is known.

Theorem 11.2. Suppose $\{X_n\}$ is a martingale.
a) If g is convex, then $\{Z_n = g(X_n)\}$ is a submartingale.
b) If g is concave, then $\{Z_n = g(X_n)\}$ is a supermartingale.

Proof. a) Using a version of Jensen's Inequality adapted to conditional expectations, $E(Z_{n+1}|Z_1, ..., Z_n) = E[g(X_{n+1})|g(X_1), ..., g(X_n)] =$

$E[g(X_{n+1})|X_1, ..., X_n] \geq g(E[X_{n+1}|X_1, ..., X_n]) = g(X_n) = Z_n$.
b) Note that g is concave iff $h = -g$ is convex. By a),
$E(-Z_{n+1}|-Z_1, ..., -Z_n) = E(-Z_{n+1}|Z_1, ..., Z_n) = E(-g(X_{n+1})|X_1, ..., X_n) \geq -g(E[X_{n+1}|X_1, ..., X_n]) = -g(X_n)$. Thus $-E(Z_{n+1}|Z_1, ..., Z_n) \geq -g(X_n)$ implies $E(Z_{n+1}|Z_1, ..., Z_n) \leq g(X_n) = Z_n$. □

The following theorem is taken from Sen and Singer (1993, p. 120). The X_i need not be independent and $\{T_n\}$ can be show to be a martingale with $E(T_n) = 0$ and $V(T_n) = E(T_n^2) = s_n^2$. Thus $Z_n = T_n/s_n$ is the zscore of T_n. The Lindeberg CLT of Theorem 2.40 needed rowwise independence in the triangular array. The v_k^2 are random variables that depend on $X_1, ..., X_{k-1}$.

Theorem 11.3, Martingale CLT: Let $X_1, X_2, ...$ be a sequence of random variables with $E(X_k) = 0$, $V(X_k) = E(X_k^2) = \sigma_k^2 < \infty$, and $E(X_k|X_1, ..., X_{k-1}) = 0$ with $X_0 = 0$. Let $T_n = \sum_{k=1}^n X_k$, $s_n^2 = \sum_{k=1}^n \sigma_k^2$, $v_k^2 = E(X_k^2|X_1, ..., X_{k-1})$, and $w_n^2 = \sum_{k=1}^n v_k^2$. If
a) $w_n^2/s_n^2 \xrightarrow{P} 1$ as $n \to \infty$, and
b) (Lindeberg's condition)

$$\lim_{n\to\infty} \sum_{k=1}^{n} \frac{E(X_k^2\ I[|X_k| \geq \epsilon s_n])}{s_n^2} = 0 \tag{11.1}$$

for any $\epsilon > 0$, then

$$Z_n = T_n/s_n \xrightarrow{D} N(0, 1).$$

The sequence $\{T_n\}$ is a martingale since $E(T_{n+1}|X_1, ..., X_n) = E(T_n + X_{n+1}|X_1, ..., X_n) = T_n$ since $E(X_{n+1}|X_1, ..., X_n) = 0$ by assumption with $k = n + 1$.

Remark 11.4. Let $T_n = \sum_{k=1}^n X_k$. Then $E(T_{n+1}|X_1, ..., X_n) = E(T_{n+1}|T_1, ..., T_n)$. To see that this result holds, note that if $X_1, ..., X_n$ are known, then $T_1, ..., T_n$ are known. If $T_1, ..., T_n$ are known then $T_1 = X_1$ and $T_k - T_{k-1} = X_k$ are known for $k = 2, ..., n$.

11.2 Hints and Solutions to Selected Problems

1.1. a) $X_2 \sim N(100, 6)$.

b)

$$\begin{pmatrix} X_1 \\ X_3 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 49 \\ 17 \end{pmatrix}, \begin{pmatrix} 3 & -1 \\ -1 & 4 \end{pmatrix} \right).$$

c) $X_1 \perp\!\!\!\perp X_4$ and $X_3 \perp\!\!\!\perp X_4$.

d)

$$\rho(X_1, X_2) = \frac{Cov(X_1, X_3)}{\sqrt{\text{VAR}(X_1)\text{VAR}(X_3)}} = \frac{-1}{\sqrt{3}\sqrt{4}} = -0.2887.$$

1.2. a) $Y|X \sim N(49, 16)$ since $Y \perp\!\!\!\perp X$. (Or use $E(Y|X) = \mu_Y + \Sigma_{12}\Sigma_{22}^{-1}(X - \mu_x) = 49 + 0(1/25)(X - 100) = 49$ and $\text{VAR}(Y|X) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 16 - 0(1/25)0 = 16$.)

b) $E(Y|X) = \mu_Y + \Sigma_{12}\Sigma_{22}^{-1}(X - \mu_x) = 49 + 10(1/25)(X - 100) = 9 + 0.4X$.

c) $\text{VAR}(Y|X) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 16 - 10(1/25)10 = 16 - 4 = 12$.

1.4 a)

$$N_2\left(\begin{pmatrix}3\\2\end{pmatrix}, \begin{pmatrix}3 & 1\\1 & 2\end{pmatrix}\right).$$

b) $X_2 \perp\!\!\!\perp X_4$ and $X_3 \perp\!\!\!\perp X_4$.

c) $\dfrac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{33}}} = \dfrac{1}{\sqrt{2}\sqrt{3}} = 1/\sqrt{6} = 0.4082$.

2.1. c) The histograms should become more like a normal distribution as n increases from 1 to 200. In particular, when $n = 1$ the histogram should be right skewed while for $n = 200$ the histogram should be nearly symmetric. Also the scale on the horizontal axis should decrease as n increases.

d) Now $\overline{Y} \sim N(0, 1/n)$. Hence the histograms should all be roughly symmetric, but the scale on the horizontal axis should be from about $-3/\sqrt{n}$ to $3/\sqrt{n}$.

2.3. a) $E(X) = \frac{3\theta}{\theta+1}$, thus
$\sqrt{n}(\overline{X} - E(X)) \xrightarrow{D} N(0, V(X))$, where
$V(X) = \frac{9\theta}{(\theta+2)(\theta+1)^2}$. Let $g(y) = \frac{y}{3-y}$, thus $g'(y) = \frac{3}{(3-y)^2}$. Using the delta method, $\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \frac{\theta(\theta+1)^2}{\theta+2})$.

b) It is asymptotically efficient if $\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \nu(\theta))$, where

$$\nu(\theta) = \frac{\frac{d}{d\theta}(\theta)}{-E(\frac{d^2}{d\theta^2} ln f(x|\theta))}.$$

But, $E((\frac{d^2}{d\theta^2} ln f(x|\theta)) = \frac{1}{\theta^2}$. Thus $\nu(\theta) = \theta^2 \neq \frac{\theta(\theta+1)^2}{\theta+2}$.

c) $\overline{X} \to \frac{3\theta}{\theta+1}$ in probability. Thus $T_n \to \theta$ in probability.

2.5. See Example 2.9.

2.7. a) See Example 2.8.

2.12. a) $Y_n \stackrel{D}{=} \sum_{i=1}^n X_i$ where the X_i are iid χ_1^2. Hence $E(X_i) = 1$ and $\text{Var}(X_i) = 2$. Thus by the CLT,

$$\sqrt{n}\left(\frac{Y_n}{n} - 1\right) \stackrel{D}{=} \sqrt{n}\left(\frac{\sum_{i=1}^n X_i}{n} - 1\right) \xrightarrow{D} N(0, 2).$$

b) Let $g(\theta) = \theta^3$. Then $g'(\theta) = 3\theta^2$, $g'(1) = 3$, and by the delta method,

$$\sqrt{n}\left[\left(\frac{Y_n}{n}\right)^3 - 1\right] \xrightarrow{D} N(0, 2(g'(1))^2) = N(0, 18).$$

2.13. $Y_i \sim bin(n = 1, F(x))$ since an indicator random variable Y_i takes on values 0 and 1, so $Y_i \sim bin(n = 1, p)$ with $p =$ to the probability of the indicator event: $p = P(X_i \leq x) = F(x)$.

a) $E(Y_i) = np = 1F(x) = F(X)$

b) $V(Y_i) = p(1-p) = F(x)[1 - F(x)]$

c) Then

$$\sqrt{n}\left(\hat{F}_n(x) - F(x)\right) \xrightarrow{D} N(0, F(x)[1 - F(x)])$$

by the CLT since $\hat{F}_n(x) = \overline{Y}_n$ and the Y_i are iid.

2.22. See the proof of Theorem 1.33.

2.26. a) See Example 2.1b.

b) See Example 2.3.

c) See Example 2.6.

2.27. a) By the CLT, $\sqrt{n}(\overline{X} - \lambda)/\sqrt{\lambda} \xrightarrow{D} N(0, 1)$. Hence $\sqrt{n}(\overline{X} - \lambda) \xrightarrow{D} N(0, \lambda)$.

b) Let $g(\lambda) = \lambda^3$ so that $g'(\lambda) = 3\lambda^2$ then $\sqrt{n}[(\overline{X})^3 - (\lambda)^3] \xrightarrow{D} N(0, \lambda[g'(\lambda)]^2) = N(0, 9\lambda^5)$.

2.28. a) $\overline{X}$ is a complete sufficient statistic. Also, we have $\frac{(n-1)S^2}{\sigma^2}$ has a chi square distribution with $df = n - 1$, thus since σ^2 is known the distribution of S^2 does not depend on μ, so S^2 is ancillary. Thus, by Basu's Theorem $\overline{X}$ and S^2 are independent.

b) by CLT (n is large) $\sqrt{n}(\overline{X} - \mu)$ has approximately normal distribution with mean 0 and variance σ^2. Let $g(x) = x^3$, thus, $g^{'}(x) = 3x^2$. Using delta method $\sqrt{n}(g(\overline{X}) - g(\mu))$ goes in distribution to $N(0, \sigma^2(g^{'}(\mu))^2)$ or $\sqrt{n}(\overline{X}^3 - \mu^3)$ goes in distribution to $N(0, \sigma^2(3\mu^2)^2)$.

2.29. a) According to the standard theorem, $\sqrt{n}(\hat{\theta}_n - \theta) \to N(0, 3)$.

b) $E(Y) = \theta, Var(Y) = \frac{\pi^2}{3}$, according to CLT we have $\sqrt{n}(\overline{Y}_n - \theta) \to N(0, \frac{\pi^2}{3})$.

c) $MED(Y) = \theta$, then $\sqrt{n}(MED(n) - \theta) \to N(0, \frac{1}{4\,f^2(MED(Y))})$ and $f(MED(Y)) = \frac{\exp(-(\theta-\theta))}{[1+\exp(-(\theta-\theta))]^2} = \frac{1}{4}$. Thus $\sqrt{n}(MED(n) - \theta) \to N(0, \frac{1}{4\,\frac{1}{16}}) \to \sqrt{n}(MED(n) - \theta) \to N(0, 4)$.

d) All three estimators are consistent, but $3 < \frac{\pi^2}{3} < 4$, therefore the estimator $\hat{\theta}_n$ is the best, and the estimator MED(n) is the worst.

2.31. a) $F_n(y) = 0.5 + 0.5y/n$ for $-n < y < n$, so $F(y) \equiv 0.5$.

b) No, since $F(y)$ is not a cdf.

2.32. a) $F_n(y) = y/n$ for $0 < y < n$, so $F(y) \equiv 0$.
b) No, since $F(y)$ is not a cdf.

2.33. a)

$$\sqrt{n}\left(\overline{Y} - \frac{1-\rho}{\rho}\right) \xrightarrow{D} N\left(0, \frac{1-\rho}{\rho^2}\right)$$

by the CLT.

c) The method of moments estimator of ρ is $\hat{\rho} = \frac{1}{1+\overline{Y}}$.

d) Let $g(\theta) = 1 + \theta$ so $g'(\theta) = 1$. Then by the delta method,

$$\sqrt{n}\left(g(\overline{Y}) - g(\frac{1-\rho}{\rho})\right) \xrightarrow{D} N\left(0, \frac{1-\rho}{\rho^2}1^2\right)$$

or

$$\sqrt{n}\left((1+\overline{Y}) - \frac{1}{\rho}\right) \xrightarrow{D} N\left(0, \frac{1-\rho}{\rho^2}\right).$$

This result could also be found with algebra since $1+\overline{Y}-\frac{1}{\rho} = \overline{Y}+1-\frac{1}{\rho} = \overline{Y} + \frac{\rho-1}{\rho} = \overline{Y} - \frac{1-\rho}{\rho}$.

e) $\overline{Y}$ is the method of moments estimator of $E(Y) = (1-\rho)/\rho$, so $1+\overline{Y}$ is the method of moments estimator of $1 + E(Y) = 1/\rho$.

2.34. a) $\sqrt{n}(\bar{X} - \mu)$ is approximately $N(0, \sigma^2)$. Define $g(x) = \frac{1}{x}$, $g'(x) = \frac{-1}{x^2}$. Using delta method, $\sqrt{n}(\frac{1}{\overline{X}} - \frac{1}{\mu})$ is approximately $N(0, \frac{\sigma^2}{\mu^4})$. Thus $1/\overline{X}$ is approximately $N(\frac{1}{\mu}, \frac{\sigma^2}{n\mu^4})$, provided $\mu \neq 0$.

b) Using part a)

$\frac{1}{\overline{X}}$ is asymptotically efficient for $\frac{1}{\mu}$ if

$$\frac{\sigma^2}{\mu^4} = \left[\frac{\left(\tau'(\mu)\right)^2}{E_\mu \left(\frac{\partial}{\partial \mu} \ln f(X/\mu) \right)^2} \right]$$

$$\tau(\mu) = \frac{1}{\mu}$$

$$\tau'(\mu) = \frac{-1}{\mu^2}$$

$$\ln f(x|\mu) = \frac{-1}{2} \ln 2\pi\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2}$$

$$E\left[\frac{\partial}{\partial \mu} \ln f(X/\mu) \right]^2 = \frac{E(X-\mu)^2}{\sigma^4}$$

$$= \frac{1}{\sigma^2}$$

Thus

$$\frac{\left(\tau'(\mu)\right)^2}{E_\mu \left[\frac{\partial}{\partial \mu} \ln f(X/\mu) \right]^2} = \frac{\sigma^2}{\mu^4}.$$

2.35. a) $E(Y^k) = 2\theta^k/(k+2)$ so $E(Y) = 2\theta/3$, $E(Y^2) = \theta^2/2$ and $V(Y) = \theta^2/18$. So $\sqrt{n}\left(\overline{Y} - \frac{2\theta}{3} \right) \xrightarrow{D} N\left(0, \frac{\theta^2}{18}\right)$ by the CLT.

b) Let $g(\tau) = \log(\tau)$ so $[g'(\tau)]^2 = 1/\tau^2$ where $\tau = 2\theta/3$. Then by the delta method,

$\sqrt{n}\left(\log(\overline{Y}) - \log\left(\frac{2\theta}{3}\right) \right) \xrightarrow{D} N\left(0, \frac{1}{8}\right)$.

c) $\hat{\theta}^k = \frac{k+2}{2n} \sum Y_i^k$.

2.36. a) $\sqrt{n}\left(\overline{Y} - \frac{r(1-\rho)}{\rho} \right) \xrightarrow{D} N\left(0, \frac{r(1-\rho)}{\rho^2}\right)$ by the CLT.

b) Let $\theta = r(1-\rho)/\rho$. Then

$$g(\theta) = \frac{r}{r + \frac{r(1-\rho)}{\rho}} = \frac{r\rho}{r\rho + r(1-\rho)} = \rho = c.$$

Now

$$g'(\theta) = \frac{-r}{(r+\theta)^2} = \frac{-r}{(r + \frac{r(1-\rho)}{\rho})^2} = \frac{-r\rho^2}{r^2}.$$

So

$$[g'(\theta)]^2 = \frac{r^2\rho^4}{r^4} = \frac{\rho^4}{r^2}.$$

Hence by the delta method

$$\sqrt{n}\ (\ g(\overline{Y}) - \rho\ ) \xrightarrow{D} N\left(0, \frac{r(1-\rho)}{\rho^2}\frac{\rho^4}{r^2}\right) = N\left(0, \frac{\rho^2(1-\rho)}{r}\right).$$

c) $\overline{Y} \stackrel{\text{set}}{=} r(1-\rho)/\rho$ or $\rho\overline{Y} = r - r\rho$ or $\rho\overline{Y} + r\rho = r$ or $\hat{\rho} = r/(r + \overline{Y})$.

2.37. a) By the CLT,

$$\sqrt{n}\ \left(\ \overline{X} - \frac{\theta}{2}\ \right) \xrightarrow{D} N\left(0, \frac{\theta^2}{12}\right).$$

b) Let $g(y) = y^2$. Then $g'(y) = 2y$ and by the delta method,

$$\sqrt{n}\ \left(\overline{X}^2 - (\frac{\theta}{2})^2\ \right) = \sqrt{n}\ \left(\overline{X}^2 - \frac{\theta^2}{4}\ \right) = \sqrt{n}\ \left(\ g(\overline{X}) - g(\frac{\theta}{2})\ \right) \xrightarrow{D}$$

$$N\left(0, \frac{\theta^2}{12}\ [g'(\frac{\theta}{2})]^2\right) = N\left(0, \frac{\theta^2}{12}\ \frac{4\theta^2}{4}\right) = N\left(0, \frac{\theta^4}{12}\right).$$

2.38. a) $E(X_i) = \beta/(\beta + \beta) = 1/2$ and $V(X_i) = \dfrac{\beta^2}{(2\beta)^2(2\beta+1)} = \dfrac{1}{4(2\beta+1)} = \dfrac{1}{8\beta+4}$. So

$$\sqrt{n}\left(\ \overline{X}_n - \frac{1}{2}\ \right) \xrightarrow{D} N\left(0, \frac{1}{8\beta+4}\right)$$

by the CLT.

b) Let $g(x) = \log(x)$. So $d = g(1/2) = \log(1/2)$. Now $g'(x) = 1/x$ and $(g'(x))^2 = 1/x^2$. So $(g'(1/2))^2 = 4$. So

$$\sqrt{n}(\ \log(\overline{X}_n) - \log(1/2)\ ) \xrightarrow{D} N\left(0, \frac{1}{8\beta+4}\ 4\right) = N\left(0, \frac{1}{2\beta+1}\right)$$

by the delta method.

2.92. a) The cdf $F_n(x)$ of X_n is

$$F_n(x) = \begin{cases} 0, & x \leq \frac{-1}{n} \\ \frac{nx}{2} + \frac{1}{2}, & \frac{-1}{n} \leq x \leq \frac{1}{n} \\ 1, & x \geq \frac{1}{n}. \end{cases}$$

Sketching $F_n(x)$ shows that it has a line segment rising from 0 at $x = -1/n$ to 1 at $x = 1/n$ and that $F_n(0) = 0.5$ for all $n \geq 1$. Examining the cases $x < 0$, $x = 0$ and $x > 0$ shows that as $n \to \infty$,

$$F_n(x) \to \begin{cases} 0, & x < 0 \\ \frac{1}{2} & x = 0 \\ 1, & x > 0. \end{cases}$$

Notice that if X is a random variable such that $P(X = 0) = 1$, then X has cdf

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0. \end{cases}$$

Since $x = 0$ is the only discontinuity point of $F_X(x)$ and since $F_n(x) \to F_X(x)$ for all continuity points of $F_X(x)$ (i.e. for $x \neq 0$),

$$X_n \xrightarrow{D} X.$$

b) $F_n(t) = t/n$ for $0 < t \leq n$ and $F_n(t) = 0$ for $t \leq 0$. Hence $\lim_{n\to\infty} F_n(t) = 0$ for $t \leq 0$. If $t > 0$ and $n > t$, then $F_n(t) = t/n \to 0$ as $n \to \infty$. Thus $\lim_{n\to\infty} F_n(t) = H(t) = 0$ for all t, and Y_n does not converge in distribution to any random variable Y since $H(t) \equiv 0$ is a continuous function but not a cdf.

2.93. If $X_n \sim U(a_n, b_n)$ with $a_n < b_n$, then

$$F_{X_n}(t) = \frac{t - a_n}{b_n - a_n}$$

for $a_n \leq t \leq b_n$, $F_{X_n}(t) = 0$ for $t \leq a_n$ and $F_{X_n}(t) = 1$ for $t \geq b_n$. On $[a_n, b_n]$, $F_{X_n}(t)$ is a line segment from $(a_n, 0)$ to $(b_n, 1)$ with slope $\dfrac{1}{b_n - a_n}$.

a) $F_{X_n}(t) \to H(t) \equiv 1 \ \ \forall t \in \mathbb{R}$. Since $H(t)$ is continuous but not a cdf, X_n does not converge in distribution to any RV X.

b) $F_{X_n}(t) \to H(t) \equiv 0 \ \ \forall t \in \mathbb{R}$. Since $H(t)$ is continuous but not a cdf, X_n does not converge in distribution to any RV X.

c)

$$F_{X_n}(t) \to F_X(t) = \begin{cases} 0 & t \leq a \\ \frac{t-a}{b-a} & a \leq t \leq b \\ 1 & t \geq b. \end{cases}$$

Hence $X_n \xrightarrow{D} X \sim U(a, b)$.

d)

$$F_{X_n}(t) \to \begin{cases} 0 & t < c \\ 1 & t > c. \end{cases}$$

Hence $X_n \xrightarrow{D} X$ where $P(X = c) = 1$. Hence X has a point mass distribution at c. (The behavior of $\lim_{n\to\infty} F_{X_n}(c)$ is not important, even if the limit does not exist.)

e)

$$F_{X_n}(t) = \frac{t + n}{2n} = \frac{1}{2} + \frac{t}{2n}$$

for $-n \leq t \leq n$. Thus $F_{X_n}(t) \to H(t) \equiv 0.5 \ \ \forall t \in \mathbb{R}$. Since $H(t)$ is continuous but not a cdf, X_n does not converge in distribution to any RV X.

f)

$$F_{X_n}(t) = \frac{t - c + \frac{1}{n}}{\frac{2}{n}} = \frac{1}{2} + \frac{n}{2}(t - c)$$

for $c - 1/n \leq t \leq c + 1/n$. Thus

$$F_{X_n}(t) \to H(t) = \begin{cases} 0 & t < c \\ 1/2 & t = c \\ 1 & t > c. \end{cases}$$

If X has the point mass at c, then

$$F_X(t) = \begin{cases} 0 & t < c \\ 1 & t \geq c. \end{cases}$$

Hence $t = c$ is the only discontinuity point of $F_X(t)$, and $H(t) = F_X(t)$ at all continuity points of $F_X(t)$. Thus $X_n \xrightarrow{D} X$ where $P(X = c) = 1$.

2.94. a) i) X_n is discrete and takes on two values with $E(X_n) = n\frac{1}{n}$ for all positive integers n. Hence $E[|X_n - 0|] = E(X_n) = 1 \ \ \forall n$ and X_n **does not satisfy** $X_n \xrightarrow{1} 0$.

ii) Let $\epsilon > 0$. Then

$$P[|X_n - 0| \geq \epsilon] \leq P(X_n = n) = \frac{1}{n} \to 0$$

as $n \to \infty$. Hence $X_n \xrightarrow{P} 0$.

iii) By ii) $X_n \xrightarrow{D} 0$.

b) i) X_n is discrete and takes on two values with

$$E[(X_n - 0)^2] = E(X_n^2) = \sum x^2 P(X_n = x) = 0^2(1 - \frac{1}{n}) + 1^2\frac{1}{n} = \frac{1}{n} \to 0$$

as $n \to \infty$. Hence $X_n \xrightarrow{2} 0$.

Since i) holds, so do ii), iii) and iv).

(Also note that

$$E[|X_n - 0|] = E(X_n) = \frac{1}{n} \to 0 \ \forall n.$$

Hence $X_n \xrightarrow{1} 0$.)

2.95. a) $E(X_i) = \beta/(\beta + \beta) = 1/2$ and $V(X_i) = \dfrac{\beta^2}{(2\beta)^2(2\beta + 1)} = \dfrac{1}{4(2\beta + 1)} = \dfrac{1}{8\beta + 4}$. So

$$\sqrt{n}\left(\overline{X}_n - \frac{1}{2} \right) \xrightarrow{D} N\left(0, \frac{1}{8\beta + 4}\right)$$

by the CLT.

b) Let $g(x) = \log(x)$. So $d = g(1/2) = \log(1/2)$. Now $g'(x) = 1/x$ and $(g'(x))^2 = 1/x^2$. So $(g'(1/2))^2 = 4$. So

$$\sqrt{n}(\ \log(\overline{X}_n) - \log(1/2)\ ) \xrightarrow{D} N\left(0, \frac{1}{8\beta+4}\ 4\right) = N\left(0, \frac{1}{2\beta+1}\right)$$

by the delta method.

2.96. a) We have $E[X] = \frac{3\theta}{2}$ and $Var(X) = \frac{\theta^2}{12}$.

Therefore $\sqrt{n}(\overline{X} - \frac{3\theta}{2}) \xrightarrow{D} N(0, \frac{\theta^2}{12})$ by the CLT.

b) Let $g(x) = \log(x)$ so $(g'(x))^2 = 1/x^2$ where $x = \frac{3\theta}{2}$. Then by using the delta method we have

$$\sqrt{n}\big(\log(\overline{X}) - \log(\frac{3\theta}{2})\big) \xrightarrow{D} N(0, \frac{1}{27}).$$

2.97. a) Let $Y_n \stackrel{D}{=} \sum_{i=1}^{n} X_i$, where X_i are iid Poisson(1), then by central limit theorem, we have

$$\sqrt{n}\ \left(\ \frac{Y_n}{n} -\ 1\ \right) \xrightarrow{D} N(0,1).$$

b) Let $g(t) = t^2, g'(t) = 2t \neq 0$. Using the Delta method, we have

$\sqrt{n}\ \left[\ \left(\frac{Y_n}{n}\right)^2 -\ 1\ \right] \xrightarrow{D} N(0, 1(2\cdot 1)^2) \sim N(0,4)$.

2.98. $E(Y) = 3\theta/2$ and $V(Y) = \theta^2/12$.

a) $\sqrt{n}(\overline{Y} - 3\theta/2) \xrightarrow{D} N(0, \theta^2/12)$ by the CLT

b) Let $g(\mu) = \mu^2$, $g'(\mu) = 2\mu$, and $g'(3\theta/2) = 3\theta$. Then by the delta method,

$$\sqrt{n}[\ (\overline{Y})^2 - g(3\theta/2)] \xrightarrow{D} N(0, [g'(3\theta/2)]^2\theta^2/12),\ \text{or}$$

$$\sqrt{n}\left[\ (\overline{Y})^2 - \frac{9\theta^2}{4}\right] \xrightarrow{D} N\left(0, \frac{9\theta^2\theta^2}{12}\right) \sim N\left(0, \frac{9\theta^4}{12}\right) \sim N\left(0, \frac{3\theta^4}{4}\right).$$

2.99. a) $\sqrt{n}(\overline{W} - \theta) \xrightarrow{D} N(0, \sigma_W^2)$ by the CLT.

b) Let $g(\theta) = \sqrt{\theta}$ with $g'(\theta) = 0.5\theta^{-0.5}$. Then $\sqrt{n}(\sqrt{\overline{W}} - \sqrt{\theta}) \xrightarrow{D} N(0, \sigma_W^2[g'(\theta)]^2) \sim N(0, 0.25\sigma_W^2/\theta)$ by the delta method provided $\theta > 0$.

11.3 Tables

Tabled values are F(0.95,k,d) where $P(F < F(0.95, k, d)) = 0.95$.
00 stands for ∞. Entries produced with the `qf(.95,k,d)` command in *R*.
The numerator degrees of freedom are k while the denominator degrees of freedom are d.

k	1	2	3	4	5	6	7	8	9	00
d										
1	161	200	216	225	230	234	237	239	241	254
2	18.5	19.0	19.2	19.3	19.3	19.3	19.4	19.4	19.4	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.41
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	1.84
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	1.62
00	3.84	3.00	2.61	2.37	2.21	2.10	2.01	1.94	1.88	1.00

Tabled values are $t_{\alpha,d}$ where $P(t < t_{\alpha,d}) = \alpha$ where t has a t distribution with d degrees of freedom. If $d > 29$ use the $N(0,1)$ cutoffs $d = Z = \infty$.

					alpha					pvalue
d	0.005	0.01	0.025	0.05	0.5	0.95	0.975	0.99	0.995	left tail
1	-63.66	-31.82	-12.71	-6.314	0	6.314	12.71	31.82	63.66	
2	-9.925	-6.965	-4.303	-2.920	0	2.920	4.303	6.965	9.925	
3	-5.841	-4.541	-3.182	-2.353	0	2.353	3.182	4.541	5.841	
4	-4.604	-3.747	-2.776	-2.132	0	2.132	2.776	3.747	4.604	
5	-4.032	-3.365	-2.571	-2.015	0	2.015	2.571	3.365	4.032	
6	-3.707	-3.143	-2.447	-1.943	0	1.943	2.447	3.143	3.707	
7	-3.499	-2.998	-2.365	-1.895	0	1.895	2.365	2.998	3.499	
8	-3.355	-2.896	-2.306	-1.860	0	1.860	2.306	2.896	3.355	
9	-3.250	-2.821	-2.262	-1.833	0	1.833	2.262	2.821	3.250	
10	-3.169	-2.764	-2.228	-1.812	0	1.812	2.228	2.764	3.169	
11	-3.106	-2.718	-2.201	-1.796	0	1.796	2.201	2.718	3.106	
12	-3.055	-2.681	-2.179	-1.782	0	1.782	2.179	2.681	3.055	
13	-3.012	-2.650	-2.160	-1.771	0	1.771	2.160	2.650	3.012	
14	-2.977	-2.624	-2.145	-1.761	0	1.761	2.145	2.624	2.977	
15	-2.947	-2.602	-2.131	-1.753	0	1.753	2.131	2.602	2.947	
16	-2.921	-2.583	-2.120	-1.746	0	1.746	2.120	2.583	2.921	
17	-2.898	-2.567	-2.110	-1.740	0	1.740	2.110	2.567	2.898	
18	-2.878	-2.552	-2.101	-1.734	0	1.734	2.101	2.552	2.878	
19	-2.861	-2.539	-2.093	-1.729	0	1.729	2.093	2.539	2.861	
20	-2.845	-2.528	-2.086	-1.725	0	1.725	2.086	2.528	2.845	
21	-2.831	-2.518	-2.080	-1.721	0	1.721	2.080	2.518	2.831	
22	-2.819	-2.508	-2.074	-1.717	0	1.717	2.074	2.508	2.819	
23	-2.807	-2.500	-2.069	-1.714	0	1.714	2.069	2.500	2.807	
24	-2.797	-2.492	-2.064	-1.711	0	1.711	2.064	2.492	2.797	
25	-2.787	-2.485	-2.060	-1.708	0	1.708	2.060	2.485	2.787	
26	-2.779	-2.479	-2.056	-1.706	0	1.706	2.056	2.479	2.779	
27	-2.771	-2.473	-2.052	-1.703	0	1.703	2.052	2.473	2.771	
28	-2.763	-2.467	-2.048	-1.701	0	1.701	2.048	2.467	2.763	
29	-2.756	-2.462	-2.045	-1.699	0	1.699	2.045	2.462	2.756	
Z	-2.576	-2.326	-1.960	-1.645	0	1.645	1.960	2.326	2.576	
CI						90%	95%		99%	
	0.995	0.99	0.975	0.95	0.5	0.05	0.025	0.01	0.005	right tail
	0.01	0.02	0.05	0.10	1	0.10	0.05	0.02	0.01	two tail

11.4 Summary

11.5 Complements

Sen and Singer (1993) and Woodroofe (1975) are good references for martingales. Ash (1972) and Billinglsey (1986) define martingales using σ-fields.

11.6 Problems

Abid, A.M. and Olive, D.J. (2025), "Some Simple High Dimensional One and Two Sample Tests," is at (http://parker.ad.siu.edu/Olive/pphd1samp.pdf).

Abuhassan, H., and Olive, D.J. (2008), "Inference for the Pareto, Half Normal and Related Distributions," unpublished manuscript, (http://parker.ad.siu.edu/Olive/pppar.pdf).

Agresti, A. (2002), *Categorical Data Analysis*, 2nd ed., Wiley, Hoboken, NJ.

Agresti, A. (2013), *Categorical Data Analysis*, 3rd ed., Wiley, Hoboken, NJ.

Agresti, A., and Caffo, B. (2000), "Simple and Effective Confidence Intervals for Proportions and Difference of Proportions Result by Adding Two Successes and Two Failures," *The American Statistician*, 54, 280-288.

Agresti, A., and Coull, B.A. (1998), "Approximate is Better than Exact for Interval Estimation of Binomial Parameters," *The American Statistician*, 52, 119-126.

Agulló, J. (1996), "Exact Iterative Computation of the Multivariate Minimum Volume Ellipsoid Estimator with a Branch and Bound Algorithm," in *Proceedings in Computational Statistics,* ed. Prat, A., Physica-Verlag, Heidelberg, 175-180.

Agulló, J. (1998), "Computing the Minimum Covariance Determinant Estimator," unpublished manuscript, Universidad de Alicante.

Akaike, H. (1973), "Information Theory as an Extension of the Maximum Likelihood Principle," in *Proceedings, 2nd International Symposium on Information Theory*, eds. Petrov, B.N., and Csakim, F., Akademiai Kiado, Budapest, 267-281.

Anderson, T.W. (1971), *The Statistical Analysis of Time Series*, Wiley, New York, NY.

Anderson, T.W. (1984), *An Introduction to Multivariate Statistical Analysis,* 2nd ed., Wiley, New York, NY.

Ash, R.B. (1972), *Real Analysis and Probability,* Academic Press, San Diego, CA.

Bain, L.J. (1978), *Statistical Analysis of Reliability and Life-Testing Models,* Marcel Dekker, New York, NY.

Basa, J., Cook, R.D., Forzani, L., and Marcos, M. (2024), "Asymptotic Distribution of One-Component Partial Least Squares Regression Estimators in High Dimensions," *The Canadian Journal of Statistics*, 52, 118-130.

Barker, L. (2002), "A Comparison of Nine Confidence Intervals for a Poisson Parameter When the Expected Number of Events ≤ 5," *The American Statistician*, 56, 85-89.

Barndorff-Nielsen, O. (1982), "Exponential Families," in *Encyclopedia of Statistical Sciences,* Vol. 2, eds. Kotz, S. and Johnson, N.L., Wiley, New York, NY, 587-596.

Basa, J., Cook, R.D., Forzani, L., and Marcos, M. (2024), "Asymptotic Distribution of One-Component Partial Least Squares Regression Estimators in High Dimensions," *The Canadian Journal of Statistics*, 52, 118-130.

Bassett, G.W., and Koenker, R.W. (1978), "Asymptotic Theory of Least Absolute Error Regression," *Journal of the American Statistical Association,* 73, 618-622.

Baszczyńska, A., and Pekasiewicz, D. (2010), "Selected Methods of Interval Estimation of the Median. The Analysis of Accuracy of Estimation," *ACTA Universitatis Lodziensis Folia Oeconomica*, 235, 21-30.

Becker, R.A., Chambers, J.M., and Wilks, A.R. (1988), *The New S Language: a Programming Environment for Data Analysis and Graphics,* Wadsworth and Brooks/Cole, Pacific Grove, CA.

Beran, R. (1990), "Calibrating Prediction Regions," *Journal of the American Statistical Association*, 85, 715-723.

Beran, R. (1993), "Probability-Centered Prediction Regions," *The Annals of Statistics*, 21, 1967-1981.

Berk, R. (1967), "Review 1922 of 'Invariance of Maximum Likelihood Estimators' by Peter W. Zehna," *Mathematical Reviews,* 33, 342-343.

Berk, R.H. (1972), "Consistency and Asymptotic Normality of MLE's for Exponential Models," *The Annals of Mathematical Statistics*, 43, 193-204.

Berndt, E.R., and Savin, N.E. (1977), "Conflict Among Criteria for Testing Hypotheses in the Multivariate Linear Regression Model," *Econometrika*, 45, 1263-1277.

Bernholt, T. (2005), "Computing the Least Median of Squares Estimator in Time $O(n^d)$," *Proceedings of ICCSA 2005*, LNCS, 3480, 697-706.

Bernholt, T., and Fischer, P. (2004), "The Complexity of Computing the MCD-Estimator," *Theoretical Computer Science*, 326, 383-398.

Bhatia, R., Elsner, L., and Krause, G. (1990), "Bounds for the Variation of the Roots of a Polynomial and the Eigenvalues of a Matrix," *Linear Algebra and Its Applications*, 142, 195-209.

Bickel, P.J. (1965), "On Some Robust Estimates of Location," *The Annals of Mathematical Statistics,* 36, 847-858.

Bickel, P.J., and Doksum, K.A. (1977), *Mathematical Statistics: Basic Ideas and Selected Topics,* 1st ed., Holden Day, Oakland, CA.

Bickel, P.J., and Doksum, K.A. (2007), *Mathematical Statistics: Basic Ideas and Selected Topics,* Vol. 1., 2nd ed., Updated Printing, Pearson Prentice Hall, Upper Saddle River, NJ.

Bickel, P.J., and Ren, J.–J. (2001), "The Bootstrap in Hypothesis Testing," in *State of the Art in Probability and Statistics: Festschrift for William R. van Zwet*, eds. de Gunst, M., Klaassen, C., and van der Vaart, A., The Institute of Mathematical Statistics, Hayward, CA, 91-112.

Billingsley, P. (1986), *Probability and Measure,* 2nd ed., Wiley, New York, NY.

Bloch, D.A., and Gastwirth, J.L. (1968), "On a Simple Estimate of the Reciprocal of the Density Function," *The Annals of Mathematical Statistics,* 39, 1083-1085.

Box, G.E.P, and Jenkins, G.M. (1976), *Time Series Analysis: Forecasting and Control*, revised ed., Holden-Day, Oakland, CA.

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24, 123-140.

Brown, L.D. (1986), *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*, Institute of Mathematical Statistics Lecture Notes – Monograph Series, IMS, Haywood, CA.

Brown, L.D., Cai, T.T., and DasGupta, A. (2001), "Interval Estimation for a Binomial Proportion," (with discussion), *Statistical Science*, 16, 101-133.

Brown, L.D., Cai, T.T., and DasGupta, A. (2002), "Confidence Intervals for a Binomial Proportion and Asymptotic Expansions," *The Annals of Statistics*, 30, 150-201.

Brown, M.B., and Forsythe, A.B. (1974a), "The ANOVA and Multiple Comparisons for Data with Heterogeneous Variances," *Biometrics*, 30, 719-724.

Brown, M.B., and Forsythe, A.B. (1974b), "The Small Sample Behavior of Some Statistics Which Test the Equality of Several Means," *Technometrics*, 16, 129-132.

Brownstein, N., and Pensky, M. (2008), "Application of Transformations in Parametric Inference," *Journal of Statistical Education*, 16 (online).

Büchlmann, P., and Yu, B. (2002), "Analyzing Bagging," *The Annals of Statistics*, 30, 927-961.

Budny, K. (2014), "A Generalization of Chebyshev's Inequality for Hilbert-Space-Valued Random Variables," *Statistics & Probability Letters,* 88, 62-65.

Buja, A., and Stuetzle, W. (2006), "Observations on Bagging," *Statistica Sinica*, 16, 323-352.

Burnham, K.P., and Anderson, D.R. (2002), *Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach*, 2nd ed., Springer, New York, NY.

Burnham, K.P., and Anderson, D.R. (2004), "Multimodel Inference Understanding AIC and BIC in Model Selection," *Sociological Methods & Research*, 33, 261-304.

Burr, D. (1994), "A Comparison of Certain Bootstrap Confidence Intervals in the Cox Model," *Journal of the American Statistical Association*, 89, 1290-1302.

Butler, R.W., Davies, P.L., and Jhun, M. (1993), "Asymptotics for the Minimum Covariance Determinant Estimator," *The Annals of Statistics,* 21, 1385-1400.

Buxton, L.H.D. (1920), "The Anthropology of Cyprus," *The Journal of the Royal Anthropological Institute of Great Britain and Ireland,* 50, 183-235.

Byrne, J., and Kabaila, P. (2005), "Comparison of Poisson Confidence Intervals," *Communications in Statistics: Theory and Methods*, 34, 545-556.

Casella, G., and Berger, R.L. (2002), *Statistical Inference,* 2nd ed., Duxbury, Belmont, CA.

Cator, E.A., and Lopuhaä, H.P. (2010), "Asymptotic Expansion of the Minimum Covariance Determinant Estimators," *Journal of Multivariate Analysis,* 101, 2372-2388.

Cator, E.A., and Lopuhaä, H.P. (2012), "Central Limit Theorem and Influence Function for the MCD Estimators at General Multivariate Distributions," *Bernoulli*, 18, 520-551.

Chan, N.H., Ling, S., and Yau, C.Y. (2020), "Lasso-Based Variable Selection of ARMA Models," *Statistica Sinica*, 30, 1925-1948.

Chang, J., and Hall, P. (2015), "Double Bootstrap Methods That Use a Single Double-Bootstrap Simulation," *Biometrika*, 102, 203-214.

Chang, J., and Olive, D.J. (2010), "OLS for 1D Regression Models," *Communications in Statistics: Theory and Methods*, 39, 1869-1882.

Charkhi, A., and Claeskens, G. (2018), "Asymptotic Post-Selection Inference for the Akaike Information Criterion," *Biometrika*, 105, 645-664.

Chen, M.H., and Shao, Q.M. (1999), "Monte Carlo Estimation of Bayesian Credible and HPD Intervals, *Journal of Computational and Graphical Statistics,* 8, 69-92.

Chen, S.X. (2016), "Peter Hall's Contributions to the Bootstrap," *The Annals of Statistics*, 44, 1821-1836.

Chen, X. (2011), "A New Generalization of Chebyshev Inequality for Random Vectors," see arXiv:0707.0805v2.

Chernoff, H. (1956), "Large-Sample Theory: Parametric Case," *The Annals of Mathematical Statistics*, 27, 1-22.

Chew, V. (1966), "Confidence, Prediction and Tolerance Regions for the Multivariate Normal Distribution," *Journal of the American Statistical Association,* 61, 605-617.

Chihara, L., and Hesterberg, T. (2011), *Mathematical Statistics with Resampling and R*, Wiley, Hoboken, NJ.

Chun, H., and Keleş, S. (2010), "Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Predictor Selection," *Journal of the Royal Statistical Society, B,* 72, 3-25.

Čížek, P. (2006), "Least Trimmed Squares Under Dependence," *Journal of Statistical Planning and Inference*, 136, 3967-3988.

Čížek, P. (2008), "General Trimmed Estimation: Robust Approach to Nonlinear and Limited Dependent Variable Models," *Econometric Theory*, 24, 1500-1529.

Claeskens, G., and Hjort, N.L. (2008), *Model Selection and Model Averaging*, Cambridge University Press, New York, NY.

Clarke, B.R. (1986), "Nonsmooth Analysis and Fréchet Differentiability of M Functionals," *Probability Theory and Related Fields*, 73, 137-209.

Clarke, B.R. (2000), "A Review of Differentiability in Relation to Robustness with an Application to Seismic Data Analysis," *Proceedings of the Indian National Science Academy, A*, 66, 467-482.

Cook, R.D. (1977), "Deletion of Influential Observations in Linear Regression," *Technometrics,* 19, 15-18.

Cook, R.D. (2018), *An Introduction to Envelopes: Dimension Reduction for Efficient Estimation in Multivariate Statistics,* Wiley, Hoboken, NJ.

Cook, R.D., and Forzani, L. (2018), "Big Data and Partial Least Squares Prediction," *The Canadian Journal of Statistics*, 46, 62-78.

Cook, R.D., and Forzani, L. (2019), "Partial Least Squares Prediction in High-Dimensional Regression," *The Annals of Statistics,* 47, 884-908.

Cook, R.D., and Forzani, L. (2024), *Partial Least Squares Regression: and Related Dimension Reduction Methods*, Chapman and Hall/CRC, Boca Raton, FL.

Cook, R.D., Helland, I.S., and Su, Z. (2013), "Envelopes and Partial Least Squares Regression," *Journal of the Royal Statistical Society, B*, 75, 851-877.

Cook, R.D., and Weisberg, S. (1999), *Applied Regression Including Computing and Graphics,* Wiley, New York, NY.

Cox, C. (1984), "An Elementary Introduction to Maximum Likelihood Estimations for Multinomial Models: Birch's Theorem and the Delta Method," *The American Statistician*, 38, 283-287.

Cox, D.R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society, B,* 34, 187-220.

Cox, D.R., and Hinkley, D.V. (1974), *Theoretical Statistics,* Chapman and Hall, London, UK.

Cramér, H. (1946), *Mathematical Methods of Statistics,* Princeton University Press, Princeton, NJ.

Crawley, M.J. (2005), *Statistics: an Introduction Using R,* Wiley, Hoboken, NJ.

Crawley, M.J. (2013), *The R Book,* 2nd ed., Wiley, Hoboken, NJ.

Dahiya, R.C., Staneski, P.G. and Chaganty, N.R. (2001), "Maximum Likelihood Estimation of Parameters of the Truncated Cauchy Distribution," *Communications in Statistics: Theory and Methods*, 30, 1737-1750.

DasGupta, A. (2008), *Asymptotic Theory of Statistics and Probability*, Springer, New York, NY.

Datta, B.N. (1995), *Numerical Linear Algebra and Applications,* Brooks/Cole Publishing Company, Pacific Grove, CA.

Davidson, J. (2021), *Stochastic Limit Theory: an Introduction for Econometricians*, 2nd ed., Oxford University Press, Oxford, UK.

DeGroot, M.H. (1975), *Probability and Statistics,* 1st ed., Addison-Wesley Publishing Company, Reading, MA.

Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1975), "Robust Estimation and Outlier Detection with Correlation Coefficients," *Biometrika,* 62, 531-545.

Devlin, S.J., Gnanadesikan, R., and Kettenring, J.R. (1981), "Robust Estimation of Dispersion Matrices and Principal Components," *Journal of the American Statistical Association,* 76, 354-362.

Eck, D.J. (2018), "Bootstrapping for Multivariate Linear Regression Models," *Statistics & Probability Letters*, 134, 141-149.

Efron, B. (1979), "Bootstrap Methods, Another Look at the Jackknife," *The Annals of Statistics*, 7, 1-26.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans,* SIAM, Philadelphia, PA.

Efron, B. (2014), "Estimation and Accuracy after Model Selection," (with discussion), *Journal of the American Statistical Association,* 109, 991-1007.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," (with discussion), *The Annals of Statistics,* 32, 407-451.

Efron, B., and Hastie, T. (2016), *Computer Age Statistical Inference*, Cambridge University Press, New York, NY.

Efron, B., and Tibshirani, R. (1986), "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Methods of Statistical Accuracy," (with discussion), *Statistical Science*, 1, 54-77.

Efron, B., and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, Chapman & Hall/CRC, New York, NY.

Efroymson, M.A. (1960), "Multiple Regression Analysis," in *Mathematical Methods for Digital Computers*, eds. Ralston, A., and Wilf, H.S., Wiley, New York, NY, 191-203.

Eicker, F. (1963), "Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions," *Annals of Mathematical Statistics*, 34, 447-456.

Einmahl, J.H.J., and Mason, D.M. (1992), "Generalized Quantile Processes," *The Annals of Statistics*, 20, 1062-1078.

Ewald, K., and Schneider, U. (2018), "Uniformly Valid Confidence Sets Based on the Lasso," *Electronic Journal of Statistics*, 12, 1358-1387.

Falk, M. (1997), "Asymptotic Independence of Median and MAD," *Statistics & Probability Letters,* 34, 341-345.

Fan, J., and Li, R. (2001), "Variable Selection Via Noncave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association,* 96, 1348-1360.

Ferguson, T.S. (1996), *A Course in Large Sample Theory*, Chapman & Hall, New York, NY.

Fernholtz, L.T. (1983), *von Mises Calculus for Statistical Functionals*, Springer, New York, NY.

Flury, B., and Riedwyl, H. (1988), *Multivariate Statistics: a Practical Approach*, Chapman & Hall, New York.

Fontana, M., Zeni, G., and Vantini, S. (2023), "Conformal Prediction: a Unified Review of Theory and New Challenges," *Bernoulli*, 29, 1-23.

Freedman, D.A. (1981), "Bootstrapping Regression Models," *The Annals of Statistics*, 9, 1218-1228.

Frey, J. (2013), "Data-Driven Nonparametric Prediction Intervals," *Journal of Statistical Planning and Inference*, 143, 1039-1048.

Friedman, J., Hastie, T., Hoefling, H., and Tibshirani, R. (2007), "Pathwise Coordinate Optimization," *Annals of Applied Statistics*, 1, 302-332.

Friedman, J., Hastie, T., Simon, N., and Tibshirani, R. (2015), *glmnet: Lasso and Elastic-net Regularized Generalized Linear Models, R* Package version 2.0, (http://cran.r-project.org/package=glmnet).

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models Via Coordinate Descent," *Journal of Statistical Software*, 33, 1-22.

Friedman, J.H., and Hall, P. (2007), "On Bagging and Nonlinear Estimation," *Journal of Statistical Planning and Inference,* 137, 669-683.

Fujikoshi, Y. (2002), "Asymptotic Expansions for the Distributions of Multivariate Basic Statistics and One-Way MANOVA Tests Under Nonnormality," *Journal of Statistical Planning and Inference*, 108, 263-282.

Garwood, F. (1936), "Fiducial Limits for the Poisson Distribution," *Biometrika*, 28, 437-442.

Geisser, S. (2006), *Modes of Parametric Inference*, Wiley, Hoboken, NJ.

Gill, R.D. (1989), "Non- and Semi-Parametric Maximum Likelihood Estimators and the von Mises Method, Part 1," *Scandinavian Journal of Statistics,* 16, 97-128.

Gladstone, R.J. (1905), "A Study of the Relations of the Brain to the Size of the Head," *Biometrika,* 4, 105-123.

Golub, G.H., and Van Loan, C.F. (1989), *Matrix Computations,* 2nd ed., John Hopkins University Press, Baltimore, MD.

Granger, C.W.J., and Newbold, P. (1977), *Forecasting Economic Time Series*, Academic Press, New York, NY.

Grosh, D. (1989), *A Primer of Reliability Theory*, Wiley, New York, NY.

Grübel, R. (1988), "The Length of the Shorth," *The Annals of Statistics,* 16, 619-628.

Gruber, M.H.J. (1998), *Improving Efficiency by Shrinkage: the James-Stein and Ridge Regression Estimators,* Marcel Dekker, New York, NY.

Guan, L. (2023), "Localized Conformal Prediction: a Generalized Inference Framework for Conformal Prediction," *Biometrika*, 110, 33-50.

Guenther, W.C. (1969), "Shortest Confidence Intervals," *The American Statistician,* 23, 22-25.

Gunst, R.F., and Mason, R.L. (1980), *Regression Analysis and Its Application: a Data Oriented Approach,* Marcel Dekker, New York, NY.

Haile, M.G., and Olive, D.J. (2024), "Bootstrapping ARMA Time Series Models after Model Selection," *Communications in Statistics: Theory and Methods*, 53, 8255-8270.

Haile, M.G., Zhang, L., and Olive, D.J. (2024), "Predicting Random Walks and a Data Splitting Prediction Region," *Stats*, 7, 23-33.

Hall, P. (1986), "On the Bootstrap and Confidence Intervals," *The Annals of Statistics*, 14, 1431-1452.

Hall, P. (1988), "Theoretical Comparisons of Bootstrap Confidence Intervals," (with discussion), *The Annals of Statistics*, 16, 927-985.

Hall, W.J., and Oakes, D. (2024), *A Course in the Large Sample Theory of Statistical Inference*, Chapman and Hall, CRC Press, Boca Raton, FL.

Hampel, F.R. (1975), "Beyond Location Parameters: Robust Concepts and Methods," *Bulletin of the International Statistical Institute,* 46, 375-382.

Hannan, E. J. (1973), "The Asymptotic Theory of Linear Time-Series Models," *Journal of Applied Probability*, 10, 130-145.

Hannan, E.J. (1980), "The Estimation of the Order of an ARMA Process," *The Annals of Statistics*, 8, 1071-1081.

Hannan, E.J., and Kavalieris, L. (1984), "A Method for Autoregressive-Moving Average Estimation," *Biometrika*, 71. 273-280.

Hannan, E.J., and Quinn, B.G. (1979), "The Determination of the Order of an Autoregression," *Journal of the Royal Statistical Society, B*, 41, 190-195.

Hastie, T.J., and Tibshirani, R.J. (1986), "Generalized Additive Models" (with discussion), *Statistical Science,* 1, 297-318.

Hastie, T.J., and Tibshirani, R.J. (1990), *Generalized Additive Models*, Chapman & Hall, London, UK.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd ed., Springer, New York, NY.

Hastie, T., Tibshirani, R., and Tibshirani, R. (2020), "Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons," *Statistical Science*, 35, 579-592.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015), *Statistical Learning with Sparsity: the Lasso and Generalizations*, CRC Press Taylor & Francis, Boca Raton, FL.

Hawkins, D.M., and Olive, D.J. (1999a), "Improved Feasible Solution Algorithms for High Breakdown Estimation," *Computational Statistics & Data Analysis,* 30, 1-11.

Hawkins, D.M., and Olive, D. (1999b), "Applications and Algorithms for Least Trimmed Sum of Absolute Deviations Regression," *Computational Statistics & Data Analysis,* 32, 119-134.

Hawkins, D.M., and Olive, D.J. (2002), "Inconsistency of Resampling Algorithms for High Breakdown Regression Estimators and a New Algorithm," (with discussion), *Journal of the American Statistical Association,* 97, 136-159.

He, X., and Portnoy, S. (1992), "Reweighted LS Estimators Converge at the Same Rate as the Initial Estimator," *The Annals of Statistics,* 20, 2161-2167.

He, X., and Wang, G. (1997), "Qualitative Robustness of S*- Estimators of Multivariate Location and Dispersion," *Statistica Neerlandica*, 51, 257-268.

Hebbler, B. (1847), "Statistics of Prussia," *Journal of the Royal Statistical Society, A*, 10, 154-186.

Helland, I.S. (1990), "Partial Least Squares Regression and Statistical Models," *Scandanavian Journal of Statistics*, 17, 97-114.

Henderson, H.V., and Searle, S.R. (1979), "Vec and Vech Operators for Matrices, with Some Uses in Jacobians and Multivariate Statistics," *The Canadian Journal of Statistics,* 7, 65-81.

Hesterberg, T., (2014), "What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum," available

from (http://arxiv.org/pdf/1411.5279v1.pdf). (An abbreviated version was published (2015), *The American Statistician*, 69, 371-386.)

Hillis, S.L., and Davis, C.S. (1994), "A Simple Justification of the Iterative Fitting Procedure for Generalized Linear Models," *The American Statistician*, 48, 288-289.

Hoel, P.G., Port, S.C., and Stone, C.J. (1971), *Introduction to Probability Theory,* Houghton Mifflin, Boston, MA.

Hoerl, A.E., and Kennard, R. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55-67.

Hössjer, O. (1991), *Rank-Based Estimates in the Linear Model with High Breakdown Point*, Ph.D. Thesis, Report 1991:5, Department of Mathematics, Uppsala University, Uppsala, Sweden.

Huang, H.H., Chan, N.H., Chen, K. and Ing, C.K. (2022), "Consistent Order Selection for ARFIMA Processes," *The Annals of Statistics*, 50, 1297-1319.

Huber, P.J., and Ronchetti, E.M. (2009), *Robust Statistics,* 2nd ed., Wiley, Hoboken, NJ.

Hubert, M., Rousseeuw, P.J., and Van Aelst, S. (2002), "Comment on 'Inconsistency of Resampling Algorithms for High Breakdown Regression and a New Algorithm' by D.M. Hawkins and D.J. Olive," *Journal of the American Statistical Association,* 97, 151-153.

Hubert, M., Rousseeuw, P.J., and Van Aelst, S. (2008), "High Breakdown Multivariate Methods," *Statistical Science*, 23, 92-119.

Hubert, M., Rousseeuw, P.J., and Verdonck, T. (2012), "A Deterministic Algorithm for Robust Location and Scatter," *Journal of Computational and Graphical Statistics,* 21, 618-637.

Hunter, D.R. (2014), *Notes for a Graduate-Level Course in Asymptotics for Statisticians*, available from (https://sites.math.rutgers.edu/∼sg1108/asymp1.pdf).

Hurvich, C., and Tsai, C.L. (1989), "Regression and Time Series Model Selection in Small Samples," *Biometrika*, 76, 297-307.

Hyndman, R.J. (1996), "Computing and Graphing Highest Density Regions," *The American Statistician,* 50, 120-126.

Hyndman, R.J., and Athanasopoulos, G. (2018), *Forecasting: Principles and Practice*, 2nd ed., OTexts: Melbourne, Australia. https://OTexts.org/fpp2/

Hyndman, R.J., and Khandakar, Y. (2008), "Automatic Time Series Forecasting: the Forecast Package for R." *Journal of Statistical Software*, 26, 1-22.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013, 2021), *An Introduction to Statistical Learning with Applications in R*, 1st and 2nd ed., Springer, New York, NY.

Jia, J., and Yu, B. (2010), "On Model Selection Consistency of the Elastic Net When $p >> n$," *Statistica Sinica*, 20, 595-611.

Jiang, J. (2022), *Large Sample Techniques for Statistics*, 2nd ed., Springer, New York, NY.

Jin, Y., and Olive, D.J. (2025), "Large Sample Theory for Some Ridge-Type Regression Estimators," is at (http://parker.ad.siu.edu/Olive/ppridgetype.pdf).

Johnson, M.E. (1987), *Multivariate Statistical Simulation,* Wiley, New York, NY.

Johnson, M.P., and Raven, P.H. (1973), "Species Number and Endemism, the Galápagos Archipelago Revisited," *Science*, 179, 893-895.

Johnson, N.J. (1978), "Modified t Tests and Confidence Intervals for Asymmetrical Populations," *Journal of the American Statistical Association*, 73, 536-544.

Johnson, N.L., and Kotz, S. (1970ab), *Distributions in Statistics: Continuous Univariate Distributions,* Vol. 1-2, Houghton Mifflin Company, Boston, MA.

Johnson, R.A., Ladella, J., and Liu, S.T. (1979), "Differential Relations, in the Original Parameters, Which Determine the First Two Moments of the Multi-Parameter Exponential Family," *The Annals of Statistics*, 7, 232-235.

Johnson, R.A., and Wichern, D.W. (1988), *Applied Multivariate Statistical Analysis,* 2nd ed., Prentice Hall, Englewood Cliffs, NJ.

Jones, H.L. (1946), "Linear Regression Functions with Neglected Variables," *Journal of the American Statistical Association,* 41, 356-369.

Kakizawa, Y. (2009), "Third-Order Power Comparisons for a Class of Tests for Multivariate Linear Hypothesis Under General Distributions," *Journal of Multivariate Analysis*, 100, 473-496.

Khattree, R., and Naik, D.N. (1999), *Applied Multivariate Statistics with SAS Software,* 2nd ed., SAS Institute, Cary, NC.

Kim, J., and Pollard, D. (1990), "Cube Root Asymptotics," *The Annals of Statistics,* 18, 191-219.

Klouda, K. (2015), "An Exact Polynomial Time Algorithm for Computing the Least Trimmed Squares Estimate," *Computational Statistics & Data Analysis*, 84, 27-40.

Knight, K., and Fu, W.J. (2000), "Asymptotics for Lasso-Type Estimators," *Annals of Statistics*, 28, 1356–1378.

Koenker, R.W. (2005), *Quantile Regression*, Cambridge University Press, Cambridge, UK.

Konietschke, F., Bathke, A.C., Harrar, S.W., and Pauly, M. (2015), "Parametric and Nonparametric Bootstrap Methods for General MANOVA," *Journal of Multivariate Analysis*, 140, 291-301.

Kreiss, J.P., (1985), "A Note on M-Estimation in Stationary ARMA Processes," *Statistics & Risk Modeling*, 3, 317-336.

Kshirsagar, A.M. (1972), *Multivariate Analysis*, Marcel Dekker, New York, NY.

Lai, T.L., Robbins, H., and Wei, C.Z. (1979), "Strong Consistency of Least Squares Estimates in Multiple Regression II," *Journal of Multivariate Analysis,* 9, 343-361.

Leeb, H., Pötscher, B.M., and Ewald, K. (2015), "On Various Confidence Intervals Post-Model-Selection," *Statistical Science*, 30, 216-227.

Lehmann, E.L. (1980), "Efficient Likelihood Estimators," *The American Statistician*, 34, 233-235.

Lehmann, E.L., (1983), *Theory of Point Estimation,* 1st ed., Wiley, New York, NY.

Lehmann, E.L. (1986), *Testing Statistical Hypotheses,* 2nd ed., Wiley, New York, NY.

Lehmann, E.L. (1999), *Elements of Large–Sample Theory*, Springer, New York, NY.

Lehmann, E.L., and Casella, G. (2003), *Theory of Point Estimation,* 2nd ed., Wiley, New York, NY.

Leon, S.J. (1986), *Linear Algebra with Applications*, 2nd ed., Macmillan Publishing Company, New York, NY.

Li, K.C., and Duan, N. (1989), "Regression Analysis Under Link Violation," *The Annals of Statistics,* 17, 1009-1052.

Liu, K. (1993), "A New Class of Biased Estimate in Linear Regression," *Communications in Statistics: Theory and Methods* 22, 393-402.

Liu, K. (2003), "Using Liu-Type Estimator to Combat Collinearity," *Communications in Statistics: Theory and Methods*, 32, 1009-1020.

Liu, X., and Zuo, Y. (2014), "Computing Projection Depth and Its Associated Estimators," *Statistics and Computing*, 24, 51-63.

Lopuhaä, H.P. (1999), "Asymptotics of Reweighted Estimators of Multivariate Location and Scatter," *The Annals of Statistics,* 27, 1638-1665.

Lukacs, E. (1970), *Characteristic Functions*, 2nd ed., Hafnir, New York, NY.

Lukacs, E. (1975), *Stochastic Convergence*, Academic Press, New York, NY.

Machado, J.A.F., and Parente, P. (2005), "Bootstrap Estimation of Covariance Matrices Via the Percentile Method," *Econometrics Journal,* 8, 70–78.

MacKinnon, J.G., and White, H. (1985), "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics*, 29, 305-325.

Mallows, C. (1973), "Some Comments on C_p," *Technometrics,* 15, 661-676.

Mann, H. B., and Wald, A. (1943a), "On Stochastic Limit and Order Relationships," *The Annals of Mathematical Statistics*, 14, 217-226.

Mann, H.B., and Wald, A. (1943b), "On the Statistical Treatment of Linear Stochastic Difference Equations," *Econometrica*, 11, 173-220.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979), *Multivariate Analysis,* Academic Press, London, UK.

Maronna, R.A., Martin, R.D., and Yohai, V.J. (2006), *Robust Statistics: Theory and Methods*, 1st ed., Wiley, Hoboken, NJ.

Maronna, R.A., Martin, R.D., and Yohai, V.J. (2019), *Robust Statistics: Theory and Methods*, 2nd ed., Wiley, Hoboken, NJ.

Maronna, R.A., and Yohai, V.J. (2002), "Comment on 'Inconsistency of Resampling Algorithms for High Breakdown Regression and a New Algo-

rithm' by D.M. Hawkins and D.J. Olive," *Journal of the American Statistical Association,* 97, 154-155.

Maronna, R.A., and Yohai, V.J. (2015), "High-Sample Efficiency and Robustness Based on Distance-Constrained Maximum Likelihood," *Computational Statistics & Data Analysis*, 83, 262-274.

Marsden, J.E., and Hoffman, M.J. (1993), *Elementary Classical Analysis*, 2nd ed., W.H. Freeman, New York, NY.

Mašíček, L. (2004), "Optimality of the Least Weighted Squares Estimator," *Kybernetika*, 40, 715-734.

MathSoft (1999a), *S-Plus 2000 User's Guide,* Data Analysis Products Division, MathSoft, Seattle, WA.

MathSoft (1999b), *S-Plus 2000 Guide to Statistics,* Volume 2, Data Analysis Products Division, MathSoft, Seattle, WA.

McCullagh, P., and Nelder, J.A. (1989), *Generalized Linear Models*, 2nd ed., Chapman & Hall, London, UK.

McCulloch, R.E. (1988), "Information and the Likelihood Function in Exponential Families," *The American Statistician*, 42, 73-75.

McElroy, T.S., and Politis, D.N. (2020), *Time Series: a First Course with Bootstrap Starter*, CRC Press Taylor & Francis, Boca Raton, FL.

McKinney, D. (2021), "Some t–Type Confidence Intervals," MS paper, Southern Illinois University, at (https://opensiuc.lib.siu.edu/cgi/viewcontent.cgi?article=2403&context=gs_rp).

Meinshausen, N. (2007), "Relaxed Lasso," *Computational Statistics & Data Analysis*, 52, 374-393.

Moore, D.S. (2007), *The Basic Practice of Statistics,* 4th ed., W.H. Freeman, New York, NY.

Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., and Wu, A.Y. (2014), "On the Least Trimmed Squares Estimator," *Algorithmica*, 69, 148-183.

Myers, R.H., Montgomery, D.C., and Vining, G.G. (2002), *Generalized Linear Models with Applications in Engineering and the Sciences*, Wiley, New York, NY.

Navarro, J. (2014), "Can the Bounds in the Multivariate Chebyshev Inequality be Attained?" *Statistics & Probability Letters*, 91, 1-5.

Navarro, J. (2016), "A Very Simple Proof of the Multivariate Chebyshev's Inequality," *Communications in Statistics: Theory and Methods,* 45, 3458-3463.

Nelder, J.A., and Wedderburn, R.W.M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society, A*, 135, 370-384.

Nishi, R. (1984), "Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression," *The Annals of Statistics,* 12, 758-765.

Olive, D.J. (2001), "High Breakdown Analogs of the Trimmed Mean," *Statistics & Probability Letters,* 51, 87-92.

Olive, D.J. (2004a), "A Resistant Estimator of Multivariate Location and Dispersion," *Computational Statistics & Data Analysis*, 46, 99-102.

Olive, D.J. (2004b), "Visualizing 1D Regression," in *Theory and Applications of Recent Robust Methods*, eds. Hubert, M., Pison, G., Struyf, A., and Van Aelst, S., Birkhäuser, Basel, Switzerland, 221-233.

Olive, D.J. (2005), "Two Simple Resistant Regression Estimators," *Computational Statistics & Data Analysis*, 49, 809-819.

Olive, D.J. (2007), "Prediction Intervals for Regression Models," *Computational Statistics & Data Analysis,* 51, 3115-3122.

Olive, D.J. (2008), *Applied Robust Statistics,* Unpublished Online Text, see (http://parker.ad.siu.edu/Olive/ol-bookp.htm).

Olive, D.J. (2013a), "Plots for Generalized Additive Models," *Communications in Statistics: Theory and Methods*, 42, 2610-2628.

Olive, D.J. (2013b), "Asymptotically Optimal Regression Prediction Intervals and Prediction Regions for Multivariate Data," *International Journal of Statistics and Probability*, 2, 90-100.

Olive, D.J. (2014), *Statistical Theory and Inference*, Springer, New York, NY.

Olive, D.J. (2017a), *Linear Regression*, Springer, New York, NY.

Olive, D.J. (2017b), *Robust Multivariate Analysis*, Springer, New York, NY.

Olive, D.J. (2018), "Applications of Hyperellipsoidal Prediction Regions," *Statistical Papers,* 59, 913-931.

Olive, D.J. (2025a), *Prediction and Statistical Learning*, online course notes, see (http://parker.ad.siu.edu/Olive/slearnbk.htm).

Olive, D.J. (2025b), *Theory for Linear Models*, online course notes, (http://parker.ad.siu.edu/Olive/linmodbk.htm).

Olive, D.J. (2025c), *Robust Statistics*, online course notes, (http://parker.ad.siu.edu/Olive/robbook.htm).

Olive, D.J. (2025d), *Survival Analysis*, online course notes, see (http://parker.ad.siu.edu/Olive/survbk.htm).

Olive, D.J. (2025e), *Probability and Measure*, online course notes, see (http://parker.ad.siu.edu/Olive/probbook.pdf).

Olive, D.J., Alshammari, A., Pathiranage, K.G., and Hettige, L.A.W. (2025), "Testing with the One Component Partial Least Squares and the Marginal Maximum Likelihood Estimators," is at (http://parker.ad.siu.edu/Olive/pphdwls.pdf).

Olive, D.J., and Johana Lemonge, S. (2025), "OLS Testing with Predictors Scaled to Have Unit Sample Variance." See (http://parker.ad.siu.edu/Olive/ppsols.pdf).

Olive, D.J., and Hawkins, D.M. (2005), "Variable Selection for 1D Regression Models," *Technometrics*, 47, 43-50.

Olive, D.J., and Hawkins, D.M. (2010), "Robust Multivariate Location and Dispersion," preprint, see (http://parker.ad.siu.edu/Olive/pphbmld.pdf).

Olive, D.J., and Hawkins, D.M. (2011), "Practical High Breakdown Regression," preprint at (http://parker.ad.siu.edu/Olive/pphbreg.pdf).

Olive, D.J., Pelawa Watagoda, L.C.R., and Rupasinghe Arachchige Don, H.S. (2015), "Visualizing and Testing the Multivariate Linear Regression Model," *International Journal of Statistics and Probability*, 4, 126-137.

Olive, D.J. and Quaye, P. (2024), "Testing Poisson Regression and Related Models with the One Component Partial Least Squares Estimator," is at (http://parker.ad.siu.edu/Olive/pphdpois.pdf).

Olive, D.J., Rathnayake, R.C., and Haile, M.G. (2022), "Prediction Intervals for GLMs, GAMs, and Some Survival Regression Models," *Communications in Statistics: Theory and Methods*, 51, 8012-8026.

Olive, D.J., and Zhang, L. (2025), "One Component Partial Least Squares, High Dimensional Regression, Data Splitting, and the Multitude of Models," *Communications in Statistics: Theory and Methods*, 54, 130-145.

Pankratz, A. (1983), *Forecasting with Univariate Box-Jenkins Models*, Wiley, New York, NY.

Park, Y., Kim, D., and Kim, S. (2012), "Robust Regression Using Data Partitioning and M-Estimation," *Communications in Statistics: Simulation and Computation*, 8, 1282-1300.

Pelawa Watagoda, L.C.R., and Olive, D.J. (2021a), "Bootstrapping Multiple Linear Regression after Variable Selection," *Statistical Papers,* 62, 681-700.

Pelawa Watagoda, L.C.R., and Olive, D.J. (2021b), "Comparing Six Shrinkage Estimators with Large Sample Theory and Asymptotically Optimal Prediction Intervals," *Statistical Papers,* 62, 2407-2431.

Perlman, M.D. (1972), "Maximum Likelihood–an Introduction," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability,* 1, 263-281.

Pesch, C. (1999), "Computation of the Minimum Covariance Determinant Estimator," in *Classification in the Information Age, Proceedings of the 22nd Annual GfKl Conference, Dresden 1998,* eds. Gaul, W., and Locarek-Junge, H., Springer, Berlin, 225–232.

Petrov, V.V. (1995), *Limit Theorems of Probability Theory: Sequences of Independent Random Variables*, Clarendon Press, Oxford, UK.

Pewsey, A. (2002), "Large-Sample Inference for the Half-Normal Distribution," *Communications in Statistics: Theory and Methods,* 31, 1045-1054.

Pfanzagl, J. (1993), "Sequences of Optimal Unbiased Estimators Need Not Be Asymptotically Optimal," *Scandinavian Journal of Statistics,* 20, 73-76.

Polansky, A.M. (2011), *Introduction to Statistical Limit Theory*, CRC Press, Boca Raton, FL.

Politis, D.N, and Romano, J.P. (1994), "Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions," *The Annals of Statistics*, 22, 2031-2050.

Pollard, D. (1991), "Asymptotics for Least Absolute Deviation Regression Estimators," *Econometric Theory*, 7, 186-199.

Pollard, D. (1984), *Convergence of Stochastic Processes*, Springer, Berlin.

Portnoy, S. (1977), "Asymptotic Efficiency of Minimum Variance Unbiased Estimators," *The Annals of Statistics,* 5, 522-529.

Pötscher, B.M. (1990), "Estimation of Autoregressive Moving-Average Order Given an Infinite Number of Models and Approximation of Spectral Densities," *Journal of Time Series Analysis*, 11, 165-179.

Pötscher, B. (1991), "Effects of Model Selection on Inference," *Econometric Theory*, 7, 163-185.

Pratt, J.W. (1959), "On a General Concept of "in Probability"," *The Annals of Mathematical Statistics,* 30, 549-558.

Proschan, M.A., and Shaw, P.A. (2016), *Essentials of Probability Theory for Statisticians*, Chapman & Hall/CRC Press, Boca Raton, FL.

Qi, X., Luo, R., Carroll, R.J., and Zhao, H. (2015), "Sparse Regression by Projection and Sparse Discriminant Analysis," *Journal of Computational and Graphical Statistics,* 24, 416-438.

R Core Team (2016), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

Rajapaksha, K.W.G.D.H., and Olive, D.J. (2024), "Wald Type Tests with the Wrong Dispersion Matrix," *Communications in Statistics: Theory and Methods*, 53, 2236-2251.

Rao, C.R. (1965, 1973), *Linear Statistical Inference and Its Applications,* 1st and 2nd ed., Wiley, New York, NY.

Rathnayake, R.C., and Olive, D.J. (2023), "Bootstrapping Some GLM and Survival Regression Variable Selection Estimators," *Communications in Statistics: Theory and Methods*, 52, 2625-2645.

Rejchel, W. (2016), "Lasso with Convex Loss: Model Selection Consistency and Estimation," *Communications in Statistics: Theory and Methods,* 45, 1989-2004.

Ren, J.-J. (1991), "On Hadamard Differentiability of Extended Statistical Functional," *Journal of Multivariate Analysis*, 39, 30-43.

Ren, J.-J., and Sen, P.K. (1995), "Hadamard Differentiability on $D[0,1]^p$," *Journal of Multivariate Analysis*, 55, 14-28.

Reyen, S.S., Miller, J.J., and Wegman, E.J. (2009), "Separating a Mixture of Two Normals with Proportional Covariances," *Metrika*, 70, 297-314.

Rinaldo, A., Wasserman, L., and G'Sell, M. (2019), "Bootstrapping and Sample Splitting for High-Dimensional, Assumption-Lean Inference," *The Annals of Statistics*, 47, 3438-3469.

Ro, K., Zou, C., Wang, W., and Yin, G. (2015), "Outlier Detection for High–Dimensional Data," *Biometrika*, 102, 589-599.

Rocke, D.M., and Woodruff, D.L. (1996), "Identification of Outliers in Multivariate Data," *Journal of the American Statistical Association,* 91, 1047-1061.

Rohatgi, V.K. (1976), *An Introduction to Probability Theory and Mathematical Statistics,* Wiley, New York, NY.

Rohatgi, V.K. (1984), *Statistical Inference,* Wiley, New York, NY.

Romano, J.P., and Wolf, M. (2017), "Resurrecting Weighted Least Squares," *Journal of Econometrics*, 197, 1-19.

Ross, S.M. (2014), *Introduction to Probability Models*, 11th ed., Academic Press, San Diego, CA.

Rousseeuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association,* 79, 871-880.

Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection,* Wiley, New York, NY.

Rousseeuw, P.J., and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics,* 41, 212-223.

Rupasinghe Arachchige Don, H.S. (2018), "A Relationship Between the One-Way MANOVA Test Statistic and the Hotelling Lawley Trace Test Statistic," *International Journal of Statistics and Probability*, 7, 124-131.

Rupasinghe Arachchige Don, H.S., and Olive, D.J. (2019), "Bootstrapping Analogs of the One Way MANOVA Test," *Communications in Statistics: Theory and Methods*, 48, 5546-5558.

Rupasinghe Arachchige Don, H.S., and Pelawa Watagoda, L.C.R. (2018), "Bootstrapping Analogs of the Two Sample Hotelling's T^2 Test," *Communications in Statistics: Theory and Methods*, 47, 2172-2182.

Schervish, M.J. (1995), *Theory of Statistics,* Springer, New York, NY.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics,* 6, 461-464.

Searle, S.R. (1982), *Matrix Algebra Useful for Statistics*, Wiley, New York, NY.

Seber, G.A.F., and Lee, A.J. (2003), *Linear Regression Analysis,* 2nd ed., Wiley, New York, NY.

Sen, P.K., and Singer, J.M. (1993), *Large Sample Methods in Statistics: an Introduction with Applications,* Chapman & Hall, New York, NY.

Sen, P.K., Singer, J.M., and Pedrosa De Lima, A.C. (2010), *From Finite Sample to Asymptotic Methods in Statistics,* Cambridge University Press, New York, NY.

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics,* Wiley, New York, NY.

Severini, T.A. (1998), "Some Properties of Inferences in Misspecified Linear Models," *Statistics & Probability Letters*, 40, 149-153.

Severini, T.A. (2005), *Elements of Distribution Theory*, Cambridge University Press, New York, NY.

Shao, J. (1993), "Linear Model Selection by Cross-Validation," *Journal of the American Statistical Association*, 88, 486-494.

Shao, J., and Tu, D.S. (1995), *The Jackknife and the Bootstrap*, Springer, New York, NY.

Shibata, R. (1976), "Selection of the Order of an Autoregressive Model by Akaike's Information Criterion," *Biometrika*, 63, 117-126.

Shorack, G.R., and Wellner, J.A. (1986), *Empirical Processes With Applications to Statistics,* Wiley, New York, NY.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011), "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent," *Journal of Statistical Software*, 39, 1-13.

Simonoff, J.S. (2003), *Analyzing Categorical Data*, Springer, New York, NY.

Slawski, M., zu Castell, W., and Tutz, G. (2010), "Feature Selection Guided by Structural Information," *Annals of Applied Statistics*, 4, 1056-1080.

Staudte, R.G., and Sheather, S.J. (1990), *Robust Estimation and Testing,* Wiley, New York, NY.

Steinberger, L., and Leeb, H. (2023), "Conditional Predictive Inference for Stable Algorithms," *The Annals of Statistics*, 51, 290-311.

Stewart, G.M. (1969), "On the Continuity of the Generalized Inverse," *SIAM Journal on Applied Mathematics*, 17, 33-45.

Stigler, S.M. (1973), "The Asymptotic Distribution of the Trimmed Mean," *The Annals of Mathematical Statistics,* 1, 472-477.

Su, W., Bogdan, M., and Candés, E. (2017), "False Discoveries Occur Early on the Lasso Path," *The Annals of Statistics*, 45, 2133-2150.

Su, Z., and Cook, R.D. (2012), "Inner Envelopes: Efficient Estimation in Multivariate Linear Regression," *Biometrika*, 99, 687-702.

Sun, T., and Zhang, C.-H. (2012), "Scaled Sparse Linear Regression," *Biometrika*, 99, 879-898.

Swift, M.B. (2009), "Comparison of Confidence Intervals for a Poisson Mean–Further Considerations," *Communications in Statistics: Theory and Methods,* 38, 748-759.

Tardiff, R.M. (1981), "L'Hospital's Rule and the Central Limit Theorem," *The American Statistician*, 35, 43.

Tarr, G., Müller, S., and Weber, N.C. (2016), "Robust Estimation of Precision Matrices Under Cellwise Contamination," *Computational Statistics & Data Analysis,* 93, 404-420.

Tay, J.K., Narasimhan, B. and Hastie, T. (2023), "Elastic Net Regularization Paths for All Generalized Linear Models," *Journal of Statistical Software*, 106, 1-31.

Thomopoulos, N.T. (2018), *Probability Distributions*, Springer, New York, NY.

Tian, Q., Nordman, D.J., and Meeker, W.Q. (2022), "Methods to Compute Prediction Intervals: a Review and New Results," *Statistical Science,* 37, 580-597.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, B*, 58, 267-288.

Tibshirani, R, (1997), "The Lasso Method for Variable Selection in the Cox Model," *Statistics in Medicine*, 16, 385-395.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R.J. (2012), "Strong Rules for Discarding Predictors in Lasso-Type Problems," *Journal of the Royal Statistical Society, B,* 74, 245–266.

Tibshirani, R.J. (2013), "The Lasso Problem and Uniqueness," *Electronic Journal of Statistics*, 7, 1456-1490.

Tibshirani, R.J., Rinaldo, A., Tibshirani, R., and Wasserman, L. (2018), "Uniform Asymptotic Inference and the Bootstrap after Model Selection," *The Annals of Statistics*, 46, 1255-1287.

Tremearne, A.J.N. (1911), "Notes on Some Nigerian Tribal Marks," *Journal of the Royal Anthropological Institute of Great Britain and Ireland,* 41, 162-178.

van der Vaart, A.W. (1998), *Asymptotic Statistics*, Cambridge University Press, Cambridge, UK.

Ver Hoef, J.M. (2012),"Who Invented the Delta Method?" *The American Statistician*, 66, 124-127.

Wald, A. (1949), "Note on the Consistency of the Maximum Likelihood Estimate," *The Annals of Mathematical Statistics*, 20, 595-601.

Wade, W.R. (2000), *Introduction to Analysis*, 2nd ed., Prentice Hall, Upper Saddle River, NJ.

Welch, B.L. (1937), "The Significance of the Difference Between Two Means When the Population Variances are Unequal," *Biometrika*, 29, 350-362.

Welch, B.L. (1947), "The Generalization of Student's Problem When Several Different Population Variances Are Involved," *Biometrika*, 34, 28-35.

Welch, B.L. (1951), "On the Comparison of Several Mean Values: an Alternative Approach," *Biometrika*, 38, 330-336.

Welagedara, W.A.D.M., Haile, M.G., and Olive, D.J. (2024), "ARIMA Model Selection and Prediction Intervals," is at (http://parker.ad.siu.edu/Olive/pptspi.pdf).

Welagedara, W.A.D.M., and Olive, D.J. (2024), "Calibrating and Visualizing Some Bootstrap Confidence Regions," *Axions*, 13, 659.

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817-838.

White, H. (1984), *Asymptotic Theory for Econometricians,* Academic Press, San Diego, CA.

Wieczorek, J., and Lei, J. (2022), "Model-Selection Properties of Forward Selection and Sequential Cross-Validation for High-Dimensional Regression," *Canadian Journal of Statistics*, 50, 454-470.

Wisseman, S.U., Hopke, P.K., and Schindler-Kaudelka, E. (1987), "Multielemental and Multivariate Analysis of Italian Terra Sigillata in the World Heritage Museum, University of Illinois at Urbana-Champaign," *Archeomaterials*, 1, 101-107.

Wold, H. (1975), "Soft Modelling by Latent Variables: the Non-Linear Partial Least Squares (NIPALS) Approach," *Journal of Applied Probability*, 12, 117-142.

Wood, S.N. (2017), *Generalized Additive Models: an Introduction with R*, 2nd ed., Chapman & Hall/CRC, Boca Rotan, FL.

Woodruff, D.L., and Rocke, D.M. (1994), "Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimators," *Journal of the American Statistical Association,* 89, 888-896.

Woodroofe, M. (1975), *Probability With Applications*, McGraw-Hill, New York, NY.

Wu, C.F.J. (1990), "On the Asymptotic Properties of the Jackknife Histogram," *The Annals of Statistics*, 18, 1438-1452.

Yang, Y. (2003), "Regression with Multiple Candidate Models: Selecting or Mixing?" *Statistica Sinica*, 13, 783-809.

Yao, Q. and Brockwell, P.J. (2006), "Gaussian Maximum Likelihood Estimation for ARMA Models I: Time Series," *Journal of Time Series Analysis*, 27, 857-875.

Yee, T. (2015), *Vector Generalized Linear and Additive Models*, Springer, New York, NY.

Yuen, K.K. (1974), "The Two-Sample Trimmed t for Unequal Population Variances," *Biometrika,* 61, 165-170.

Zehna, P.W. (1966), "Invariance of Maximum Likelihood Estimators," *The Annals of Mathematical Statistics,* 37, 744.

Zhang, J. (2020), "Consistency of MLE, LSE and M-Estimation under Mild Conditions," *Statistical Papers*, 61, 189-199.

Zhang, J., Olive, D.J., and Ye, P. (2012), "Robust Covariance Matrix Estimation with Canonical Correlation Analysis," *International Journal of Statistics and Probability*, 1, 119-136.

Zhang, J.-T., and Liu, X. (2013), "A Modified Bartlett Test for Heteroscedastic One-Way MANOVA," *Metrika,* 76, 135–152.

Zhang, P. (1992), "Inference after Variable Selection in Linear Regression Models," *Biometrika*, 79, 741-746.

Zhang, X., and Cheng, G. (2017), "Simultaneous Inference for High-Dimensional Linear Models," *Journal of the American Statistical Association*, 112, 757-768.

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research* 7, 2541-2563.

Zhou, L., Cook, R.D., and Zou, H. (2024), "Enveloped Huber Regression," *Journal of the American Statistical Association*, 119, 2722-2732.

Zhou, M. (2001), "Understanding the Cox Regression Models with Time–Change Covariates," *The American Statistician*, 55, 153-155.

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection Via the Elastic Net," *Journal of the Royal Statistical Society Series, B,* 67, 301-320.

Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A., and Smith, G.M. (2009), *Mixed Effects Models and Extensions in Ecology with R,* Springer, New York, NY.

Index