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. Regression is the study of the conditional distribution $Y|\boldsymbol{x}$ of the response variable Y given the vector of predictors $\boldsymbol{x} = (x_1, ..., x_p)^T$.

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**, Y is conditionally independent of $\boldsymbol{x}$ given the **sufficient predictor** $SP = h(\boldsymbol{x})$, written

$$Y \perp\!\!\!\perp \boldsymbol{x} | SP \quad \text{or} \quad \mathrm{Y} \perp\!\!\!\perp \boldsymbol{x} | \mathrm{h}(\boldsymbol{x}), \tag{6.1}$$

where the real valued function $h : \mathbb{R}^p \to \mathbb{R}$. The **estimated sufficient predictor** 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 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), where Y is independent of $\boldsymbol{x} = (x_1, ..., x_p)^T$ given the *additive predictor* $AP = \alpha + \sum_{j=1}^{p} S_j(x_j)$ for some (usually unknown) functions S_j. The *estimated additive predictor* EAP $=$ ESP $= \hat{\alpha} + \sum_{j=1}^{p} \hat{S}_j(x_j)$.

Notation. Often the index i will be suppressed. For example, the *multiple linear regression model*

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i \tag{6.2}$$

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, $Y|\boldsymbol{x} = \boldsymbol{x}^T \boldsymbol{\beta} + e$, but the conditioning on $\boldsymbol{x}$ will often be suppressed. Often the errors $e_1, ..., e_n$ are **iid** (independent and identically distributed) 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

Definition 6.5. Suppose that the response variable Y and at least one predictor variable x_i are quantitative. Then the **multiple linear regression (MLR) model** is

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i \tag{6.3}$$

for $i = 1, \ldots, n$. Here n is the *sample size* and the random variable e_i is the ith *error*. Suppressing the subscript i, the model is $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$.

In matrix notation, these n equations become

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{6.4}$$

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. Equivalently,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}. \tag{6.5}$$

Often the first column of $\boldsymbol{X}$ is $X_1 = \mathbf{1}$, 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 $\boldsymbol{x}_i$. If the e_i are **iid** (independent and identically distributed) with zero mean $E(e_i) = 0$ and variance $\text{VAR}(e_i) = V(e_i) = \sigma^2$, then 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 $r_i \equiv r_i(\boldsymbol{b}) = Y_i - \hat{Y}_i(\boldsymbol{b}) = Y_i - x_{i,1} b_1 - \cdots - x_{i,p} b_p$.

Most MLR methods attempt to find an estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ which minimizes some criterion function $Q(\boldsymbol{b})$ of the residuals.

6.1.1 Ordinary Least Squares

Definition 6.10. The full rank MLR model has $\text{rank}(\boldsymbol{X}) = p$.

Definition 6.11. The *ordinary least squares (OLS) estimator* $\hat{\boldsymbol{\beta}}_{OLS}$ minimizes

$$Q_{OLS}(\boldsymbol{b}) = \sum_{i=1}^{n} r_i^2(\boldsymbol{b}), \tag{6.6}$$

$$\text{and} \quad \hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}.$$

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.

The following theorem 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 result 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. See Sen and Singer (1993, p. 280) for the theorem, which implies that $\hat{\boldsymbol{\beta}} \approx N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T \boldsymbol{X})^{-1})$. Let $h_i = \boldsymbol{H}_{ii}$ where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T$. Note that in the following theorem, $\text{rank}(\boldsymbol{X}) = p$ since $\boldsymbol{X}^T \boldsymbol{X}$ is nonsingular.

Theorem 6.1, OLS CLT: Consider the MLR model $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ and assume that the zero mean errors are iid with $E(e_i) = 0$ and $\mathrm{V}(e_i) = \sigma^2$. Also assume that $\max_i(h_1, ..., h_n) \to 0$ in probability as $n \to \infty$ and

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \to \boldsymbol{W}^{-1}$$

as $n \to \infty$. Then the least squares (OLS) estimator $\hat{\boldsymbol{\beta}}$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{W}). \tag{6.7}$$

Equivalently,

$$(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{I}_p). \tag{6.8}$$

Definition 6.12. Let the r_i be the OLS residuals and let

$$\hat{\sigma}^2 = MSE = \frac{1}{n}\sum_{i=1}^{n} r_i^2. \tag{6.9}$$

Theorem 6.2 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 6.2. 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 $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{W}$, then $\hat{\boldsymbol{\Sigma}}_n = nMSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Hence

$$\hat{\boldsymbol{\beta}} \sim AN_p(\boldsymbol{\beta}, MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}), \quad \text{and}$$

$$rF_R = \frac{1}{MSE}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c})^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \xrightarrow{D} \chi_r^2 \tag{6.10}$$

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 6.1. 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

$$\sqrt{n}\max_{i=1,...,n}|r_i - e_i| \le (\max_{i=1,...,n}\|\boldsymbol{x}_i\|) \ \ \|\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\| = O_P(1)$$

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 6.13. 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 6.2. 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 6.3. 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 6.1 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 2.15.

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 6.5, 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 6.3: 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 6.4. 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 6.5, 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 $\boldsymbol{PP}_R = \boldsymbol{P}_R$. 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 6.4 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 6.4 c) using $\boldsymbol{e}$ and $\boldsymbol{e}_R$ instead of $\boldsymbol{Y}$, and Craig's Theorem 6.3 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 2.15, 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). □

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 $k=1$, $\boldsymbol{X}_R=\mathbf{1}$, $\boldsymbol{P}_R=\boldsymbol{P}_1$, 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: $\mathbf{1}$ is the 1st column of $\boldsymbol{X}$. $MS=SS/df$.

$SSTO=\boldsymbol{Y}^T(\boldsymbol{I}-\frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{Y}=\sum_{i=1}^n(Y_i-\overline{Y})^2$, $SSE=\sum_{i=1}^n r_i^2$, $SSR=\sum_{i=1}^n(\hat{Y}_i-\overline{Y})^2$, $SSTO=SSR+SSE$. SSTO is the SSE (residual sum of squares) for the location model $\boldsymbol{Y}=\mathbf{1}\beta_1+\boldsymbol{e}$ that contains a constant but no nontrivial predictors. The location model has projection matrix $\boldsymbol{P}_1=\mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T=\frac{1}{n}\mathbf{1}\mathbf{1}^T$. Hence $\boldsymbol{P}\boldsymbol{P}_1=\boldsymbol{P}_1$ and $\boldsymbol{P}\mathbf{1}=\boldsymbol{P}_1\mathbf{1}=\mathbf{1}$.

Source	df	SS	MS	F	p-value
Regression	p-1	$SSR=\boldsymbol{Y}^T(\boldsymbol{P}-\frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{Y}$	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 $\mathbf{0}$. Hence if $\boldsymbol{e}\sim N_n(\mathbf{0},\sigma^2\boldsymbol{I})$ 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 $\boldsymbol{Y}\sim N_n(\mathbf{1}\beta_1,\sigma^2\boldsymbol{I})$, 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 $rank(\boldsymbol{P}-\frac{1}{n}\mathbf{1}\mathbf{1}^T)=tr(\boldsymbol{P}-\frac{1}{n}\mathbf{1}\mathbf{1}^T)=p-1$. 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 $X^2>F_{1,n-p}(1-\delta)$.

Continue to assume the $n\times p$ matrix $\boldsymbol{X}$ has full rank p. There are two ways to compute $\hat{\boldsymbol{\beta}}$. Use $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, and use sample covariance matrices. The population OLS coefficients are defined below. Let $\boldsymbol{x}_i^T=(1,\boldsymbol{u}_i^T)$ where $\boldsymbol{u}_i$ is the vector of nontrivial predictors. Let $\frac{1}{n}\sum_{j=1}^n X_{jk}=\overline{X}_{ok}=\overline{u}_{ok}$ for $k=2,...,p$. The subscript "ok" means sum over the first subscript j. Let $\overline{\boldsymbol{u}}=(\overline{u}_{o,2},...,\overline{u}_{o,p})^T$ be the sample mean of the $\boldsymbol{u}_i$. Note that regressing on

$\boldsymbol{u}$ is equivalent to regressing on $\boldsymbol{x}$ if there is an intercept β_1 in the model. See Theorem 6.7 to show that $\hat{\boldsymbol{\beta}}$ estimates the population coefficients in the following definition.

Definition 6.14. Using the above notation, let $\boldsymbol{x}_i^T = (1, \boldsymbol{u}_i^T)$, and let $\boldsymbol{\beta}^T = (\beta_1, \boldsymbol{\beta}_2^T)$ where β_1 is the intercept and the slopes vector $\boldsymbol{\beta}_2 = (\beta_2, ..., \beta_p)^T$. Let the population covariance matrices

$$\text{Cov}(\boldsymbol{u}) = E[(\boldsymbol{u} - E(\boldsymbol{u}))(\boldsymbol{u} - E(\boldsymbol{u}))^T] = \boldsymbol{\Sigma}_{\boldsymbol{u}}, \text{ and}$$

$$\text{Cov}(\boldsymbol{u}, Y) = E[(\boldsymbol{u} - E(\boldsymbol{u}))(Y - E(Y))] = \boldsymbol{\Sigma}_{\boldsymbol{u}Y}.$$

Then the population coefficients from an OLS regression of Y on $\boldsymbol{x}$ (even if a linear model does not hold) are

$$\beta_1 = E(Y) - \boldsymbol{\beta}_2^T E(\boldsymbol{u}) \text{ and } \boldsymbol{\beta}_2 = \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{u}\text{Y}}.$$

Definition 6.15. Let the sample covariance matrices be

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T \text{ and } \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(Y_i - \overline{Y}).$$

Let the method of moments estimators be $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T$ and

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(Y_i - \overline{Y}) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{u}_i Y_i - \overline{\boldsymbol{u}}\, \overline{Y}.$$

The method of moment estimators are often called the maximum likelihood estimators, but are the MLE if the $(Y_i, \boldsymbol{u}_i^T)^T$ are iid from a multivariate normal distribution, a very strong assumption. In Theorem 6.6, note that $\boldsymbol{D} = \boldsymbol{X}_1^T \boldsymbol{X}_1 - n\overline{\boldsymbol{u}}\, \overline{\boldsymbol{u}}^T = (n-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}$.

Theorem 6.6: Seber and Lee (2003, p. 106). Let $\boldsymbol{X} = (\boldsymbol{1} \;\; \boldsymbol{X}_1)$. Then $\boldsymbol{X}^T\boldsymbol{Y} = \begin{pmatrix} n\overline{Y} \\ \boldsymbol{X}_1^T\boldsymbol{Y} \end{pmatrix} = \begin{pmatrix} n\overline{Y} \\ \sum_{i=1}^{n} \boldsymbol{u}_i Y_i \end{pmatrix}$, $\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} n & n\overline{\boldsymbol{u}}^T \\ n\overline{\boldsymbol{u}} & \boldsymbol{X}_1^T\boldsymbol{X}_1 \end{pmatrix}$,

$$\text{and } (\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} \frac{1}{n} + \overline{\boldsymbol{u}}^T \boldsymbol{D}^{-1} \overline{\boldsymbol{u}} & -\overline{\boldsymbol{u}}^T \boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{u}} & \boldsymbol{D}^{-1} \end{pmatrix}$$

where the $(p-1) \times (p-1)$ matrix $\boldsymbol{D}^{-1} = [(n-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}]^{-1} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}/(n-1)$.

Theorem 6.7: Second way to compute $\hat{\boldsymbol{\beta}}$: Assume a constant β_1 is in the model so the first column of $\boldsymbol{X}$ is $\boldsymbol{x}_1 = \boldsymbol{1}$, an $n \times 1$ vector of ones.

a) If $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}$ exists, then $\hat{\beta}_1 = \overline{Y} - \hat{\boldsymbol{\beta}}_2^T \overline{\boldsymbol{u}}$ and

$$\hat{\boldsymbol{\beta}}_2 = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y}.$$

b) Suppose that $(Y_i, \boldsymbol{u}_i^T)^T$ are iid random vectors such that σ_Y^2, $\boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}$, and $\boldsymbol{\Sigma}_{\boldsymbol{u}Y}$ exist. Then $\hat{\beta}_1 \xrightarrow{P} \beta_1$ and

$$\hat{\boldsymbol{\beta}}_2 \xrightarrow{P} \boldsymbol{\beta}_2 \text{ as } \mathrm{n} \to \infty.$$

Proof. Note that

$$\boldsymbol{Y}^T\boldsymbol{X}_1 = (Y_1 \cdots Y_n)\begin{bmatrix} \boldsymbol{u}_1^T \\ \vdots \\ \boldsymbol{u}_n^T \end{bmatrix} = \sum_{i=1}^{n} Y_i \boldsymbol{u}_i^T$$

and

$$\boldsymbol{X}_1^T\boldsymbol{Y} = [\boldsymbol{u}_1 \cdots \boldsymbol{u}_n]\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \sum_{i=1}^{n} \boldsymbol{u}_i Y_i.$$

So

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{n} + \overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{u}} & -\overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{u}} & \boldsymbol{D}^{-1} \end{bmatrix}\begin{bmatrix} \boldsymbol{1}^T \\ \boldsymbol{X}_1^T \end{bmatrix}\boldsymbol{Y} =$$

$$\begin{bmatrix} \frac{1}{n} + \overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{u}} & -\overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{u}} & \boldsymbol{D}^{-1} \end{bmatrix}\begin{bmatrix} n\overline{Y} \\ \boldsymbol{X}_1^T\boldsymbol{Y} \end{bmatrix}.$$

Thus $\hat{\boldsymbol{\beta}}_2 = -n\boldsymbol{D}^{-1}\overline{\boldsymbol{u}}\,\overline{Y} + \boldsymbol{D}^{-1}\boldsymbol{X}_1^T\boldsymbol{Y} = \boldsymbol{D}^{-1}(\boldsymbol{X}_1^T\boldsymbol{Y} - n\overline{\boldsymbol{u}}\,\overline{Y}) =$

$$\boldsymbol{D}^{-1}\left[\sum_{i=1}^{n} \boldsymbol{u}_i Y_i - n\overline{\boldsymbol{u}}\,\overline{Y}\right] = \frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}}{n-1} n\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y}. \text{ Then}$$

$\hat{\beta}_1 = \overline{Y} + n\overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{u}}\,\overline{Y} - \overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1}\boldsymbol{X}_1^T\boldsymbol{Y} = \overline{Y} + [n\overline{Y}\overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1} - \boldsymbol{Y}^T\boldsymbol{X}_1\boldsymbol{D}^{-1}]\overline{\boldsymbol{u}}$ $= \overline{Y} - \hat{\boldsymbol{\beta}}_2^T\boldsymbol{u}$. The convergence in probability results hold since sample means and sample covariance matrices are consistent estimators of the population means and population covariance matrices. □

It is important to note that the convergence in probability results are for iid $(Y_i, \boldsymbol{u}_i^T)^T$ with second moments and nonsingular $\boldsymbol{\Sigma}_{\boldsymbol{u}}$: a linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ does not need to hold. Also, $\boldsymbol{X}$ is a random matrix, and the least squares regression is conditional on $\boldsymbol{X}$. 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 the LS CLT. The population results of Definition 6.14 can be shown when

$$\begin{bmatrix} Y \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \sim N_p \left[\begin{pmatrix} E(Y) \\ E(\boldsymbol{u}) \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \boldsymbol{\Sigma}_{Y\boldsymbol{u}} \\ \boldsymbol{\Sigma}_{\boldsymbol{u}Y} & \boldsymbol{\Sigma}_{\boldsymbol{uu}} \end{pmatrix} \right].$$

See Remark 1.6.

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 $\text{Cov}(\boldsymbol{e}) = \sigma^2 \boldsymbol{I}$. i) The least squares estimator $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$: $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$. ii) $\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}$.

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) $\text{Cov}(\hat{\boldsymbol{\beta}}) = \text{Cov}[(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}] = \text{Cov}(\boldsymbol{A}\boldsymbol{Y}) = \boldsymbol{A}\text{Cov}(\boldsymbol{Y})\boldsymbol{A}^T =$

$$\sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{I} \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{X})^{-1} = \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}. \ \square$$

6.1.2 L_1

Definition 6.16. 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

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{L_1} - \boldsymbol{\beta}) \xrightarrow{D} N_p \left(\boldsymbol{0}, \frac{1}{4[f(0)]^2} \boldsymbol{W} \right) \tag{6.11}$$

when $\boldsymbol{X}^T \boldsymbol{X}/n \to \boldsymbol{W}^{-1}$.

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 $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, V(\hat{\boldsymbol{\beta}}, F)\ \boldsymbol{W})$ 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 $S \subseteq I$, $\boldsymbol{Y} = \boldsymbol{X}_I\boldsymbol{\beta}_I + \boldsymbol{e}$. See Section 6.7 for more on this notation. 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

$$\frac{\boldsymbol{X}_I^T\boldsymbol{X}}{n} \to \boldsymbol{W}_I^{-1}.$$

Then by the OLS CLT, $\sqrt{n}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}_I)$ where $\boldsymbol{V}_I = \sigma^2\boldsymbol{W}_I$. See Section 6.8 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 $\boldsymbol{Y}^* \sim N_n(\boldsymbol{X}\hat{\boldsymbol{\beta}}, \hat{\sigma}_n^2\boldsymbol{I}) \sim N_n(\boldsymbol{H}\boldsymbol{Y}, \hat{\sigma}_n^2\boldsymbol{I})$ 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 Theorem 6.2 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 $E(\hat{\boldsymbol{\beta}}_I^*) = (\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{X}_I^T\boldsymbol{H}\boldsymbol{Y} = \hat{\boldsymbol{\beta}}_I$ because $\boldsymbol{H}\boldsymbol{X}_I = \boldsymbol{X}_I$, and $\text{Cov}(\hat{\boldsymbol{\beta}}_I^*) = \hat{\sigma}_n^2(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}$. Hence

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_I^* - \hat{\boldsymbol{\beta}}_I) \sim N_{a_I}(\boldsymbol{0}, n\hat{\sigma}_n^2(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}_I)$$

as $n \to \infty$ if $S \subseteq I$. In particular, for the full model $I = F$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \sim N_p(\boldsymbol{0}, n\hat{\sigma}_n^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V})$$

as $n \to \infty$.

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_i^W of $\boldsymbol{r}^W$ are iid from the empirical distribution of the OLS full model residuals r_i. Hence

$$E(r_i^W) = \frac{1}{n}\sum_{i=1}^{n} r_i = 0, \quad V(r_i^W) = \sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} r_i^2 = \frac{n-p}{n}MSE,$$

$$E(\boldsymbol{r}^W) = \boldsymbol{0}, \text{ and } \text{Cov}(\boldsymbol{Y}^*) = \text{Cov}(\boldsymbol{r}^{\text{W}}) = \sigma_{\text{n}}^2\boldsymbol{I}_{\text{n}}.$$

Let $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS}$. Then $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}^*$ with $\text{Cov}(\hat{\boldsymbol{\beta}}^*) = \sigma_n^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \frac{n-p}{n}MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, 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 dif-

ference from the usual OLS MLR model is that $\sigma_n^2 \xrightarrow{P} \sigma^2$ 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 6.1 says

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \lim_{n\to\infty} n\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{OLS})) \sim N_p(\mathbf{0}, \sigma^2\boldsymbol{W})$$

where $n(\boldsymbol{X}^T\boldsymbol{X})^{-1} \to \boldsymbol{W}$. Since $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{r}^W$ follows a standard linear model, it may not be surprising that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}_{OLS}) \xrightarrow{D} N_p(\mathbf{0}, \lim_{n\to\infty} n\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}^*)) \sim N_p(\mathbf{0}, \sigma^2\boldsymbol{W}). \tag{6.12}$$

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, $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_{n,N} - \hat{\boldsymbol{\beta}}_{OLS}) \xrightarrow{D} N_p(\mathbf{0}, \sigma_N^2\boldsymbol{W})$ as $n \to \infty$, and $N_p(\mathbf{0}, \sigma_N^2\boldsymbol{W}) \xrightarrow{D} N_p(\mathbf{0}, \sigma^2\boldsymbol{W})$ as $N \to \infty$. Hence Theorem 5.1 is satisfied, and Equation (6.12) holds. See Freedman (1981) for an alternative proof.

For the above residual bootstrap, $\hat{\boldsymbol{\beta}}^*_{I_j} = (\boldsymbol{X}^T_{I_j}\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}^T_{I_j}\boldsymbol{Y}^* = \boldsymbol{D}_j\boldsymbol{Y}^*$ with $\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*_{I_j}) = \sigma_n^2(\boldsymbol{X}^T_{I_j}\boldsymbol{X}_{I_j})^{-1}$ and $E(\hat{\boldsymbol{\beta}}^*_{I_j}) = (\boldsymbol{X}^T_{I_j}\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}^T_{I_j}E(\boldsymbol{Y}^*) = (\boldsymbol{X}^T_{I_j}\boldsymbol{X}_{I_j})^{-1}\boldsymbol{X}^T_{I_j}\boldsymbol{H}\boldsymbol{Y} = \hat{\boldsymbol{\beta}}_{I_j}$ 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 $\hat{\boldsymbol{\beta}}^*_I - \hat{\boldsymbol{\beta}}_I \xrightarrow{P} \mathbf{0}$ 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.

Remark 6.4. Note that both the residual bootstrap and parametric bootstrap for OLS are robust to the unknown error distribution of the iid e_i. For the residual bootstrap with $S \subseteq I$ where I is not the full model, we conjecture that $\sqrt{n}(\hat{\boldsymbol{\beta}}^*_I - \hat{\boldsymbol{\beta}}_I) \xrightarrow{D} N_{a_I}(\mathbf{0}, \boldsymbol{V}_I)$ 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 $\boldsymbol{\beta}_E = \mathbf{0}$ where E 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_i^W depends on n and $\hat{\boldsymbol{\beta}}_E \neq \mathbf{0}$ although $\sqrt{n}\hat{\boldsymbol{\beta}}_E = O_P(1)$.

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,

$$\boldsymbol{Y}^* = \boldsymbol{X}_I^* \hat{\boldsymbol{\beta}}_{I,OLS} + \boldsymbol{r}_I^W.$$

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 $S \subseteq I_j$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_I^* - \hat{\boldsymbol{\beta}}_I) \xrightarrow{D} N_{a_I}(\boldsymbol{0}, \boldsymbol{V}_I)$$

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

$$\text{Cov}(\boldsymbol{w}^*) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_i - \overline{\boldsymbol{w}})(\boldsymbol{w}_i - \overline{\boldsymbol{w}})^T = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \begin{bmatrix} \tilde{S}_Y^2 & \tilde{\boldsymbol{\Sigma}}_{Y\boldsymbol{u}} \\ \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} & \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}} \end{bmatrix}.$$

Note that $\hat{\boldsymbol{\beta}}$ is a constant with respect to the bootstrap distribution. Assume all inverse matrices exist. Then by Theorem 6.7,

$$\hat{\boldsymbol{\beta}}^* = \begin{bmatrix} \hat{\beta}_1^* \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}^* \end{bmatrix} = \begin{bmatrix} \overline{Y}^* - \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}^{*T} \overline{\boldsymbol{u}}^* \\ \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1^*} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y}^* \end{bmatrix} \xrightarrow{P} \begin{bmatrix} \overline{Y} - \hat{\boldsymbol{\beta}}_{\boldsymbol{u}}^T \overline{\boldsymbol{u}} \\ \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{u}} \end{bmatrix} = \hat{\boldsymbol{\beta}}$$

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

Remark 6.5. For many MLR estimators, a method is needed so that everyone who uses the same units of measurements for the predictors and Y gets the same $(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{\beta}})$. Let the nontrivial predictors $\boldsymbol{u}_i^T = (x_{i,2}, ..., x_{i,p}) = (u_{i,1}, ..., u_{i,p-1})$. Then $\boldsymbol{x}_i = (1, \boldsymbol{u}_i^T)^T$. Let the $n \times (p-1)$ matrix of standardized nontrivial predictors $\boldsymbol{W}_g = (W_{ij})$ when the predictors are standardized such that $\sum_{i=1}^n W_{ij} = 0$ and $\sum_{i=1}^n W_{ij}^2 = n - g$ for $j = 1, ..., p-1$. Hence

$$W_{ij} = \frac{x_{i,j+1} - \overline{x}_{j+1}}{\hat{\sigma}_{j+1}} \text{ where } \hat{\sigma}_{\mathrm{j+1}}^2 = \frac{1}{\mathrm{n-g}} \sum_{\mathrm{i=1}}^{\mathrm{n}} (\mathrm{x_{i,j+1}} - \overline{\mathrm{x}}_{\mathrm{j+1}})^2$$

is for the $(j+1)$th variable x_{j+1}. Let $\boldsymbol{w}_i^T = (w_{i,1}, ..., w_{i,p-1})$ be the standardized vector of nontrivial predictors for the ith case. Since the standardized data are also centered, $\overline{\boldsymbol{w}} = \boldsymbol{0}$. Then the sample covariance matrix of the $\boldsymbol{w}_i$ is the sample correlation matrix of the $\boldsymbol{u}_i$:

$$\hat{\boldsymbol{\rho}}_{\boldsymbol{u}} = \boldsymbol{R}_{\boldsymbol{u}} = (r_{ij}) = \frac{\boldsymbol{W}_g^T \boldsymbol{W}_g}{n - g}$$

where r_{ij} is the sample correlation of $u_i = x_{i+1}$ and $u_j = x_{j+1}$. Thus the sample correlation matrix $\boldsymbol{R}_{\boldsymbol{u}}$ 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 $n \times (p-1)$ matrix $\boldsymbol{W}$ does not include a vector $\boldsymbol{1}$ of ones. Then regression through the origin is used for the model

$$\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e} \tag{6.13}$$

where $\boldsymbol{Z} = (Z_1, ..., Z_n)^T$ and $\boldsymbol{\eta} = (\eta_1, ..., \eta_{p-1})^T$. The vector of fitted values $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$. Then $\hat{Y}_i = \hat{Z}_i + \overline{Y}$. The software obtains $\hat{\boldsymbol{\beta}}$ from $\hat{\boldsymbol{\eta}}$.

Definition 6.17. 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* $\hat{\boldsymbol{Z}}_{OLS} = \boldsymbol{W}\hat{\boldsymbol{\eta}}_{OLS} = \boldsymbol{H}\boldsymbol{Z}$ where $\boldsymbol{H} = \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T$. The vector of residuals $\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{Z}, \boldsymbol{W}) = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Z}$.

Assume that the sample correlation matrix

$$\boldsymbol{R}_{\boldsymbol{u}} = \frac{\boldsymbol{W}^T\boldsymbol{W}}{n} \xrightarrow{P} \boldsymbol{V}^{-1}. \tag{6.14}$$

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

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V}). \tag{6.15}$$

Definition 6.18. 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

Consider the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Ridge regression uses the centered response $Z_i = Y_i - \overline{Y}$ and standardized nontrivial predictors in the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. Then $\hat{Y}_i = \hat{Z}_i + \overline{Y}$. The software obtains $\hat{\boldsymbol{\beta}}$ from $\hat{\boldsymbol{\eta}}$. See Remark 6.3.

Definition 6.19. Consider fitting the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ using $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. Let $\lambda \geq 0$ be a constant. The *ridge regression estimator* $\hat{\boldsymbol{\eta}}_R$ minimizes the *ridge regression criterion*

$$Q_R(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1}\eta_i^2 \tag{6.16}$$

over all vectors $\boldsymbol{\eta} \in \mathbb{R}^{p-1}$ where $\lambda_{1,n} \geq 0$ and $a > 0$ are known constants with $a = 1, 2, n$, and $2n$ common. Then

$$\hat{\boldsymbol{\eta}}_R = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{Z}. \tag{6.17}$$

The residual sum of squares $RSS(\boldsymbol{\eta}) = (\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator $\hat{\boldsymbol{\eta}}_{OLS}$. The ridge regression vector of fitted values is $\hat{\boldsymbol{Z}} = \hat{\boldsymbol{Z}}_R = \boldsymbol{W}\hat{\boldsymbol{\eta}}_R$, and the ridge regression vector of residuals $\boldsymbol{r}_R = \boldsymbol{r}(\hat{\boldsymbol{\eta}}_R) = \boldsymbol{Z} - \hat{\boldsymbol{Z}}_R$. The estimator is said to be *regularized* if $\lambda_{1,n} > 0$. Obtain $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{\beta}}_R$ using $\hat{\boldsymbol{\eta}}_R$, $\hat{\boldsymbol{Z}}$, and $\overline{\boldsymbol{Y}}$.

Using a vector of parameters $\boldsymbol{\eta}$ and a dummy vector $\boldsymbol{\eta}$ in Q_R is common for minimizing a criterion $Q(\boldsymbol{\eta})$, often with estimating equations. We could also write

$$Q_R(\boldsymbol{b}) = \frac{1}{a}\boldsymbol{r}(\boldsymbol{b})^T\boldsymbol{r}(\boldsymbol{b}) + \frac{\lambda_{1,n}}{a}\boldsymbol{b}^T\boldsymbol{b}$$

where the minimization is over all vectors $\boldsymbol{b} \in \mathbb{R}^{p-1}$. Note that $\sum_{i=1}^{p-1}\eta_i^2 = \boldsymbol{\eta}^T\boldsymbol{\eta} = \|\boldsymbol{\eta}\|_2^2$. The literature often uses $\lambda_a = \lambda = \lambda_{1,n}/a$.

Note that $\lambda_{1,n}\boldsymbol{b}^T\boldsymbol{b} = \lambda_{1,n}\sum_{i=1}^{p-1} b_i^2$. Each coefficient b_i is penalized equally by $\lambda_{1,n}$. Hence using standardized nontrivial predictors makes sense so that if η_i is large in magnitude, then the standardized variable w_i is important.

Remark 6.6. i) If $\lambda_{1,n} = 0$, the ridge regression estimator becomes the OLS full model estimator: $\hat{\boldsymbol{\eta}}_R = \hat{\boldsymbol{\eta}}_{OLS}$.

ii) If $\lambda_{1,n} > 0$, then $\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1}$ is nonsingular. Hence $\hat{\boldsymbol{\eta}}_R$ exists even if $\boldsymbol{X}$ and $\boldsymbol{W}$ are singular or ill conditioned, or if $p > n$.

iii) Following Hastie et al. (2009, p. 96), let the augmented matrix $\boldsymbol{W}_A$ and the augmented response vector $\boldsymbol{Z}_A$ be defined by

$$\boldsymbol{W}_A = \begin{pmatrix} \boldsymbol{W} \\ \sqrt{\lambda_{1,n}}\ \boldsymbol{I}_{p-1} \end{pmatrix}, \text{ and } \boldsymbol{Z}_{\mathrm{A}} = \begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{0} \end{pmatrix},$$

where $\boldsymbol{0}$ is the $(p-1)\times 1$ zero vector. For $\lambda_{1,n} > 0$, the OLS estimator from regressing $\boldsymbol{Z}_A$ on $\boldsymbol{W}_A$ is

$$\hat{\boldsymbol{\eta}}_A = (\boldsymbol{W}_A^T\boldsymbol{W}_A)^{-1}\boldsymbol{W}_A^T\boldsymbol{Z}_A = \hat{\boldsymbol{\eta}}_R$$

since $\boldsymbol{W}_A^T\boldsymbol{Z}_A = \boldsymbol{W}^T\boldsymbol{Z}$ and

$$\boldsymbol{W}_A^T\boldsymbol{W}_A = \begin{pmatrix} \boldsymbol{W}^T & \sqrt{\lambda_{1,n}}\ \boldsymbol{I}_{p-1} \end{pmatrix}\begin{pmatrix} \boldsymbol{W} \\ \sqrt{\lambda_{1,n}}\ \boldsymbol{I}_{p-1} \end{pmatrix} = \boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\ \boldsymbol{I}_{p-1}.$$

Remark 6.6 iv) is interesting. Note that for $\lambda_{1,n} > 0$, the $(n+p-1)\times(p-1)$ matrix $\boldsymbol{W}_A$ has full rank $p-1$. The augmented OLS model consists of adding $p-1$ pseudo-cases $(\boldsymbol{w}_{n+1}^T, Z_{n+1})^T, ..., (\boldsymbol{w}_{n+p-1}^T, Z_{n+p-1})^T$ where $Z_j = 0$ and $\boldsymbol{w}_j = (0, ..., \sqrt{\lambda_{1,n}}, 0, ..., 0)^T$ for $j = n+1, ..., n+p-1$ where the nonzero entry is in the kth position if $j = n + k$. For centered response and standardized nontrivial predictors, the population OLS regression fit runs through the origin $(\boldsymbol{w}^T, Z)^T = (\boldsymbol{0}^T, 0)^T$. Hence for $\lambda_{1,n} = 0$, the augmented OLS model adds $p-1$ typical cases at the origin. If $\lambda_{1,n}$ is not large, then the pseudo-data can still be regarded as typical cases. If $\lambda_{1,n}$ is large, the pseudo-data act as w–outliers (outliers in the standardized predictor variables), and the OLS slopes go to zero as $\lambda_{1,n}$ gets large, making $\hat{\boldsymbol{Z}} \approx \boldsymbol{0}$ so $\hat{\boldsymbol{Y}} \approx \overline{\boldsymbol{Y}}$.

To prove Remark 6.6 ii), let $(\psi, \boldsymbol{g})$ be an eigenvalue eigenvector pair of $\boldsymbol{W}^T\boldsymbol{W} = n\boldsymbol{R}_{\boldsymbol{u}}$. Then $[\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1}]\boldsymbol{g} = (\psi + \lambda_{1,n})\boldsymbol{g}$, and $(\psi + \lambda_{1,n}, \boldsymbol{g})$ is an eigenvalue eigenvector pair of $\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1} > 0$ provided $\lambda_{1,n} > 0$.

The following identity from Gunst and Mason (1980, p. 342) is useful for ridge regression inference: $\hat{\boldsymbol{\eta}}_R = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$

$$= (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$$

$$= (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{W}\hat{\boldsymbol{\eta}}_{OLS} = \boldsymbol{A}_n\hat{\boldsymbol{\eta}}_{OLS} =$$

$$[\boldsymbol{I}_{p-1} - \lambda_{1,n}(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}]\hat{\boldsymbol{\eta}}_{OLS} = \boldsymbol{B}_n\hat{\boldsymbol{\eta}}_{OLS} =$$

$$\hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda_{1n}}{n} n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\hat{\boldsymbol{\eta}}_{OLS}$$

since $\boldsymbol{A}_n - \boldsymbol{B}_n = \boldsymbol{0}$. See Problem 6.3. Assume Equation (6.14) holds. If $\lambda_{1,n}/n \to 0$ then

$$\frac{\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1}}{n} \xrightarrow{P} \boldsymbol{V}^{-1}, \text{ and } \mathrm{n}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W} + \lambda_{1,\mathrm{n}}\boldsymbol{I}_{\mathrm{p-1}})^{-1} \xrightarrow{\mathrm{P}} \boldsymbol{V}.$$

Note that

$$\boldsymbol{A}_n = \boldsymbol{A}_{n,\lambda} = \left(\frac{\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1}}{n}\right)^{-1} \frac{\boldsymbol{W}^T\boldsymbol{W}}{n} \xrightarrow{P} \boldsymbol{V}\ \boldsymbol{V}^{-1} = \boldsymbol{I}_{p-1}$$

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 (6.15) hold for the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$.

a) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$, then

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_R - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V}).$$

b) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$ then

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_R - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(-\tau\boldsymbol{V}\boldsymbol{\eta}, \sigma^2\boldsymbol{V}).$$

Proof: If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$, then by the above Gunst and Mason (1980) identity,

$$\hat{\boldsymbol{\eta}}_R = [\boldsymbol{I}_{p-1} - \hat{\lambda}_{1,n}(\boldsymbol{W}^T\boldsymbol{W} + \hat{\lambda}_{1,n}\boldsymbol{I}_{p-1})^{-1}]\hat{\boldsymbol{\eta}}_{OLS}.$$

Hence

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_R - \boldsymbol{\eta}) = \sqrt{n}(\hat{\boldsymbol{\eta}}_R - \hat{\boldsymbol{\eta}}_{OLS} + \hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) =$$

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) - \sqrt{n}\frac{\hat{\lambda}_{1,n}}{n} n(\boldsymbol{W}^T\boldsymbol{W} + \hat{\lambda}_{1,n}\boldsymbol{I}_{p-1})^{-1}\hat{\boldsymbol{\eta}}_{OLS}$$

$$\xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V}) - \tau\boldsymbol{V}\boldsymbol{\eta} \sim N_{p-1}(-\tau\boldsymbol{V}\boldsymbol{\eta}, \sigma^2\boldsymbol{V}). \ \square$$

For p fixed, Knight and Fu (2000) note i) that $\hat{\boldsymbol{\eta}}_R$ 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 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 $\boldsymbol{\eta}$ 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

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_R - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(-\tau \boldsymbol{V}\boldsymbol{\eta}, \sigma^2 \boldsymbol{V}).$$

Hence the bias can be considerable if $\tau \neq 0$. If $\tau = 0$, then OLS and ridge regression have the same limiting distribution.

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 (6.26) 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.

6.3.2 Lasso

Consider the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Lasso uses the centered response $Z_i = Y_i - \overline{Y}$ and standardized nontrivial predictors in the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$ as described in Remark 6.5. Then $\hat{Y}_i = \hat{Z}_i + \overline{Y}$. The residuals $\boldsymbol{r} = \boldsymbol{r}(\hat{\boldsymbol{\beta}}_L) = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$. Recall that $\overline{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1}$.

Definition 6.20. Consider fitting the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ using $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. The *lasso estimator* $\hat{\boldsymbol{\eta}}_L$ minimizes the *lasso criterion*

$$Q_L(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1}|\eta_i| \tag{6.18}$$

over all vectors $\boldsymbol{\eta} \in \mathbb{R}^{p-1}$ 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 $RSS(\boldsymbol{\eta}) = (\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator

$\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$ if $\boldsymbol{W}$ has full rank $p-1$. The lasso vector of fitted values is $\hat{\boldsymbol{Z}} = \hat{\boldsymbol{Z}}_L = \boldsymbol{W}\hat{\boldsymbol{\eta}}_L$, and the lasso vector of residuals $\boldsymbol{r}(\hat{\boldsymbol{\eta}}_L) = \boldsymbol{Z} - \hat{\boldsymbol{Z}}_L$. The estimator is said to be *regularized* if $\lambda_{1,n} > 0$. Obtain $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{\beta}}_L$ using $\hat{\boldsymbol{\eta}}_L$, $\hat{\boldsymbol{Z}}$, and $\overline{\boldsymbol{Y}}$.

Using a vector of parameters $\boldsymbol{\eta}$ and a dummy vector $\boldsymbol{\eta}$ in Q_L is common for minimizing a criterion $Q(\boldsymbol{\eta})$, often with estimating equations. We could also write

$$Q_L(\boldsymbol{b}) = \frac{1}{a}\boldsymbol{r}(\boldsymbol{b})^T\boldsymbol{r}(\boldsymbol{b}) + \frac{\lambda_{1,n}}{a}\sum_{j=1}^{p-1}|b_j|, \tag{6.19}$$

where the minimization is over all vectors $\boldsymbol{b} \in \mathbb{R}^{p-1}$. The literature often uses $\lambda_a = \lambda = \lambda_{1,n}/a$.

For fixed $\lambda_{1,n}$, the lasso optimization problem is convex. Hence fast algorithms exist. As $\lambda_{1,n}$ increases, some of the $\hat{\eta}_i = 0$. If $\lambda_{1,n}$ is large enough, then $\hat{\boldsymbol{\eta}}_L = \boldsymbol{0}$ and $\hat{Y}_i = \overline{Y}$ for $i = 1, ..., n$. If none of the elements $\hat{\eta}_i$ of $\hat{\boldsymbol{\eta}}_L$ are zero, then $\hat{\boldsymbol{\eta}}_L$ can be found, in principle, by setting the partial derivatives of $Q_L(\boldsymbol{\eta})$ to 0. Potential minimizers also occur at values of $\boldsymbol{\eta}$ where not all of the partial derivatives exist. An analogy is finding the minimizer of a real valued function of one variable $h(x)$. Possible values for the minimizer include values of x_c satisfying $h'(x_c) = 0$, and values x_c where the derivative does not exist. Typically some of the elements $\hat{\eta}_i$ of $\hat{\boldsymbol{\eta}}_L$ that minimizes $Q_L(\boldsymbol{\eta})$ are zero, and differentiating does not work.

The following identity from Efron and Hastie (2016, p. 308), for example, is useful for inference for the lasso estimator $\hat{\boldsymbol{\eta}}_L$:

$$\frac{-1}{n}\boldsymbol{W}^T(\boldsymbol{Z} - \boldsymbol{W}\hat{\boldsymbol{\eta}}_L) + \frac{\lambda_{1,n}}{2n}\boldsymbol{s}_n = \boldsymbol{0} \quad \text{or} \quad -\boldsymbol{W}^{\mathrm{T}}(\boldsymbol{Z} - \boldsymbol{W}\hat{\boldsymbol{\eta}}_{\mathrm{L}}) + \frac{\lambda_{1,\mathrm{n}}}{2}\boldsymbol{s}_{\mathrm{n}} = \boldsymbol{0}$$

where $s_{in} \in [-1, 1]$ and $s_{in} = \text{sign}(\hat{\eta}_{i,L})$ if $\hat{\eta}_{i,L} \neq 0$. Here $\text{sign}(\eta_i) = 1$ if $\eta_i > 1$ and $\text{sign}(\eta_i) = -1$ if $\eta_i < 1$. Note that $\boldsymbol{s}_n = \boldsymbol{s}_{n,\hat{\boldsymbol{\eta}}_L}$ depends on $\hat{\boldsymbol{\eta}}_L$. Thus $\hat{\boldsymbol{\eta}}_L$

$$= (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z} - \frac{\lambda_{1,n}}{2n}\, n(\boldsymbol{W}^T\boldsymbol{W})^{-1}\, \boldsymbol{s}_n = \hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda_{1,n}}{2n}\, n(\boldsymbol{W}^T\boldsymbol{W})^{-1}\, \boldsymbol{s}_n.$$

If none of the elements of $\boldsymbol{\eta}$ are zero, and if $\hat{\boldsymbol{\eta}}_L$ is a consistent estimator of $\boldsymbol{\eta}$, then $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$. 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 $\hat{\boldsymbol{\eta}}_{\lambda_M} = \boldsymbol{0}$. Hence $\hat{\boldsymbol{\eta}}_{\lambda_i} \neq \boldsymbol{0}$ 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 $\sqrt{n}(\hat{\boldsymbol{\eta}}_L - \hat{\boldsymbol{\eta}}_{OLS}) = o_p(1)$.

Theorem 6.11, Lasso CLT. Assume p is fixed and that the conditions of the OLS CLT Theorem Equation (6.15) hold for the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$.

a) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$, then

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_L - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2 \boldsymbol{V}).$$

b) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$ and $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$, then

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_L - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}\left(\frac{-\tau}{2}\boldsymbol{V}\boldsymbol{s}, \sigma^2 \boldsymbol{V}\right).$$

Proof. If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$ and $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$, then

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_L - \boldsymbol{\eta}) = \sqrt{n}(\hat{\boldsymbol{\eta}}_L - \hat{\boldsymbol{\eta}}_{OLS} + \hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) =$$

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) - \sqrt{n}\frac{\lambda_{1,n}}{2n} n(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{s}_n \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V}) - \frac{\tau}{2}\boldsymbol{V}\boldsymbol{s}$$

$$\sim N_{p-1}\left(\frac{-\tau}{2}\boldsymbol{V}\boldsymbol{s}, \sigma^2\boldsymbol{V}\right)$$

since under the LS CLT, $n(\boldsymbol{W}^T\boldsymbol{W})^{-1} \xrightarrow{P} \boldsymbol{V}$.

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 $\hat{\boldsymbol{\eta}}_L$ 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 $\boldsymbol{\eta}$ 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 $n \to 0$.

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 6.21. The *elastic net* estimator $\hat{\boldsymbol{\beta}}_{EN}$ minimizes the criterion

$$Q_{EN}(\boldsymbol{\beta}) = \frac{1}{2} RSS(\boldsymbol{\beta}) + \lambda_{1,n} \left[\frac{1}{2}(1-\alpha)\|\boldsymbol{\beta}_S\|_2^2 + \alpha \|\boldsymbol{\beta}_S\|_1 \right], \text{ or} \tag{6.20}$$

$$Q_2(\boldsymbol{\beta}) = RSS(\boldsymbol{\beta}) + \lambda_1 \|\boldsymbol{\beta}_S\|_2^2 + \lambda_2 \|\boldsymbol{\beta}_S\|_1 \tag{6.21}$$

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.

As with lasso, it is often convenient to use the centered response $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$ where $\overline{\boldsymbol{Y}} = \overline{Y}\mathbf{1}$, and the $n \times (p-1)$ matrix of standardized nontrivial predictors $\boldsymbol{W}$. Then regression through the origin is used for the model

$$\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e} \tag{6.22}$$

where the vector of fitted values $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$.

Ridge regression can be computed using OLS on augmented matrices. Similarly, the elastic net can be computed using lasso on augmented matrices. Let the elastic net estimator $\hat{\boldsymbol{\eta}}_{EN}$ minimize

$$Q_{EN}(\boldsymbol{\eta}) = RSS_W(\boldsymbol{\eta}) + \lambda_1 \|\boldsymbol{\eta}\|_2^2 + \lambda_2 \|\boldsymbol{\eta}\|_1 \tag{6.23}$$

where $\lambda_1 = (1-\alpha)\lambda_{1,n}$ and $\lambda_2 = 2\alpha\lambda_{1,n}$. Let the $(n+p-1) \times (p-1)$ augmented matrix $\boldsymbol{W}_A$ and the $(n+p-1) \times 1$ augmented response vector $\boldsymbol{Z}_A$ be defined by

$$\boldsymbol{W}_A = \begin{pmatrix} \boldsymbol{W} \\ \sqrt{\lambda_1}\, \boldsymbol{I}_{p-1} \end{pmatrix}, \text{ and } \boldsymbol{Z}_A = \begin{pmatrix} \boldsymbol{Z} \\ \mathbf{0} \end{pmatrix},$$

where $\mathbf{0}$ is the $(p-1) \times 1$ zero vector. Let $RSS_A(\boldsymbol{\eta}) = \|\boldsymbol{Z}_A - \boldsymbol{W}_A\boldsymbol{\eta}\|_2^2$. Then $\hat{\boldsymbol{\eta}}_{EN}$ can be obtained from the lasso of $\boldsymbol{Z}_A$ on $\boldsymbol{W}_A$: that is, $\hat{\boldsymbol{\eta}}_{EN}$ minimizes

$$Q_L(\boldsymbol{\eta}) = RSS_A(\boldsymbol{\eta}) + \lambda_2 \|\boldsymbol{\eta}\|_1 = Q_{EN}(\boldsymbol{\eta}). \tag{6.24}$$

Proof: We need to show that $Q_L(\boldsymbol{\eta}) = Q_{EN}(\boldsymbol{\eta})$. Note that $\boldsymbol{Z}_A^T \boldsymbol{Z}_A = \boldsymbol{Z}^T \boldsymbol{Z}$,

$$\boldsymbol{W}_A\, \boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{W}\boldsymbol{\eta} \\ \sqrt{\lambda_1}\, \boldsymbol{\eta} \end{pmatrix},$$

and $\boldsymbol{Z}_A^T \boldsymbol{W}_A\, \boldsymbol{\eta} = \boldsymbol{Z}^T \boldsymbol{W}\boldsymbol{\eta}$. Then

$$RSS_A(\boldsymbol{\eta}) = \|\boldsymbol{Z}_A - \boldsymbol{W}_A\boldsymbol{\eta}\|_2^2 = (\boldsymbol{Z}_A - \boldsymbol{W}_A\boldsymbol{\eta})^T (\boldsymbol{Z}_A - \boldsymbol{W}_A\boldsymbol{\eta}) =$$

$$\boldsymbol{Z}_A^T \boldsymbol{Z}_A - \boldsymbol{Z}_A^T \boldsymbol{W}_A\boldsymbol{\eta} - \boldsymbol{\eta}^T \boldsymbol{W}_A^T \boldsymbol{Z}_A + \boldsymbol{\eta}^T \boldsymbol{W}_A^T \boldsymbol{W}_A\boldsymbol{\eta} =$$

$$\boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{\eta} - \boldsymbol{\eta}^T\boldsymbol{W}^T\boldsymbol{Z} + \begin{pmatrix}\boldsymbol{\eta}^T\boldsymbol{W}^T & \sqrt{\lambda_1}\ \boldsymbol{\eta}^T\end{pmatrix}\begin{pmatrix}\boldsymbol{W}\boldsymbol{\eta} \\ \sqrt{\lambda_1}\ \boldsymbol{\eta}\end{pmatrix}.$$

Thus

$$Q_L(\boldsymbol{\eta}) = \boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{\eta} - \boldsymbol{\eta}^T\boldsymbol{W}^T\boldsymbol{Z} + \boldsymbol{\eta}^T\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{\eta} + \lambda_1\boldsymbol{\eta}^T\boldsymbol{\eta} + \lambda_2\|\boldsymbol{\eta}\|_1 =$$

$$RSS(\boldsymbol{\eta}) + \lambda_1\|\boldsymbol{\eta}\|_2^2 + \lambda_2\|\boldsymbol{\eta}\|_1 = Q_{EN}(\boldsymbol{\eta}). \ \square$$

Remark 6.7. i) You could compute the elastic net estimator using a grid of 100 $\lambda_{1,n}$ values and a grid of $J \geq 10$ α values, which would take about $J \geq 10$ times as long to compute as lasso. The above equivalent lasso problem (6.24) still needs a grid of $\lambda_1 = (1-\alpha)\lambda_{1,n}$ and $\lambda_2 = 2\alpha\lambda_{1,n}$ values. Often $J = 11$, 21, 51, or 101. The elastic net estimator tends to be computed with fast methods for optimizing convex problems, such as coordinate descent. ii) Like lasso and ridge regression, the elastic net estimator is asymptotically equivalent to the OLS full model if p is fixed and $\hat{\lambda}_{1,n} = o_P(\sqrt{n})$, but behaves worse than the OLS full model otherwise. See Theorem 6.6. iii) For prediction intervals, let d be the number of nonzero coefficients from the equivalent augmented lasso problem (5.23). Alternatively, use d_2 with $d \approx d_2 = tr[\boldsymbol{W}_{AS}(\boldsymbol{W}_{AS}^T\boldsymbol{W}_{AS} + \lambda_{2,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}_{AS}^T]$ where $\boldsymbol{W}_{AS}$ corresponds to the active set (not the augmented matrix). See Tibshirani and Taylor (2012, p. 1214). Again $\lambda_{2,n}$ may not be the λ_2 given by the software. iv) The number of nonzero lasso components (not including the constant) is at most $\min(n, p-1)$. Elastic net tends to do variable selection, but the number of nonzero components can equal $p-1$ (make the elastic net equal to ridge regression). Note that the number of nonzero components in the augmented lasso problem (6.24) is at most $\min(n+p-1, p-1) = p-1$. However, when transforming back to $\boldsymbol{X}$ and $\boldsymbol{\beta}$, a constant is added in with at most p nonzero components. vi) The elastic net can be computed with `glmnet`, and there is an *R* package `elasticnet`. vii) For fixed $\alpha > 0$, we could get λ_M for elastic net from the equivalent lasso problem. For ridge regression, we could use the λ_M for an α near 0.

Since lasso uses at most $\min(n, p-1)$ nontrivial predictors, elastic net and ridge regression can perform better than lasso if the true number of active nontrivial predictors $a_S > \min(n, p-1)$. For example, suppose $n = 1000$, $p = 5000$, and $a_S = 1500$.

Following Jia and Yu (2010), by standard Karush-Kuhn-Tucker (KKT) conditions for convex optimality for Equation (6.23), $\hat{\boldsymbol{\eta}}_{EN}$ is optimal if

$$2\boldsymbol{W}^T\boldsymbol{W}\hat{\boldsymbol{\eta}}_{EN} - 2\boldsymbol{W}^T\boldsymbol{Z} + 2\lambda_1\hat{\boldsymbol{\eta}}_{EN} + \lambda_2\boldsymbol{s}_n = 0, \quad \text{or}$$

$$(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})\hat{\boldsymbol{\eta}}_{EN} = \boldsymbol{W}^T\boldsymbol{Z} - \frac{\lambda_2}{2}\boldsymbol{s}_n, \quad \text{or}$$

$$\hat{\boldsymbol{\eta}}_{EN} = \hat{\boldsymbol{\eta}}_R - n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})^{-1}\frac{\lambda_2}{2n}\boldsymbol{s}_n. \tag{6.25}$$

Hence

$$\hat{\boldsymbol{\eta}}_{EN} = \hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda_1}{n}\, n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1 \boldsymbol{I}_{p-1})^{-1}\, \hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda_2}{2n}\, n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1 \boldsymbol{I}_{p-1})^{-1}\, \boldsymbol{s}_n$$

$$= \hat{\boldsymbol{\eta}}_{OLS} - n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1 \boldsymbol{I}_{p-1})^{-1}\, [\frac{\lambda_1}{n}\hat{\boldsymbol{\eta}}_{OLS} + \frac{\lambda_2}{2n}\boldsymbol{s}_n].$$

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,

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{EN} - \boldsymbol{\eta}) = \sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) - n(\boldsymbol{W}^T\boldsymbol{W} + \hat{\lambda}_1 \boldsymbol{I}_{p-1})^{-1}\, [\frac{\hat{\lambda}_1}{\sqrt{n}}\hat{\boldsymbol{\eta}}_{OLS} + \frac{\hat{\lambda}_2}{2\sqrt{n}}\boldsymbol{s}_n].$$

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 (6.15) hold for the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$.

a) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$, then

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{EN} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2 \boldsymbol{V}).$$

b) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$, $\hat{\alpha} \xrightarrow{P} \psi \in [0,1]$, and $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$, then

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{EN} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}\left(-\boldsymbol{V}[(1-\psi)\tau\boldsymbol{\eta} + \psi\tau\boldsymbol{s}], \sigma^2\boldsymbol{V}\right).$$

Proof. By the above remarks and the RR CLT Theorem 6.4,

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{EN} - \boldsymbol{\eta}) = \sqrt{n}(\hat{\boldsymbol{\eta}}_{EN} - \hat{\boldsymbol{\eta}}_R + \hat{\boldsymbol{\eta}}_R - \boldsymbol{\eta}) = \sqrt{n}(\hat{\boldsymbol{\eta}}_R - \boldsymbol{\eta}) + \sqrt{n}(\hat{\boldsymbol{\eta}}_{EN} - \hat{\boldsymbol{\eta}}_R)$$

$$\xrightarrow{D} N_{p-1}\left(-(1-\psi)\tau\boldsymbol{V}\boldsymbol{\eta}, \sigma^2\boldsymbol{V}\right) \quad - \quad \frac{2\psi\tau}{2}\boldsymbol{V}\boldsymbol{s}$$

$$\sim N_{p-1}\left(-\boldsymbol{V}[(1-\psi)\tau\boldsymbol{\eta} + \psi\tau\boldsymbol{s}], \sigma^2\boldsymbol{V}\right).$$

The mean of the normal distribution is $\boldsymbol{0}$ under a) since $\hat{\alpha}$ and $\boldsymbol{s}_n$ are bounded. □

6.3.4 Ridge Type Regression Estimators

See Jin and Olive (2022).

6.4 Weighted Least Squares

See Rajapaksha and Olive (2022).

6.5 GLMs and Related Regression Models

Definition 6.22. A **parametric 1D regression model** is $Y|\boldsymbol{x} \sim D(h(\boldsymbol{x}), \boldsymbol{\gamma})$ or $Y_i|\boldsymbol{x}_i \sim D(h(\boldsymbol{x}_i), \boldsymbol{\gamma})$, 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 6.23. 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 $Y \sim \text{binomial}\left(1, \rho = \dfrac{\mathrm{e}^{\mathrm{SP}}}{1 + \mathrm{e}^{\mathrm{SP}}}\right)$. 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 $Y_i \sim \text{binomial}\left(\mathrm{m_i}, \rho = \dfrac{\mathrm{e}^{\mathrm{SP}}}{1 + \mathrm{e}^{\mathrm{SP}}}\right)$. 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 6.24. 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 6.25. 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 6.8. 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 $Y_i|\boldsymbol{x}_i \sim D(\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{\gamma})$, 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 $Y_i|\boldsymbol{x}_i \sim D(\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{\gamma})$, 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 6.13. 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 $n \to \infty$.

6.6 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 fail-*

ure 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 6.10.

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 6.13 applies.

Theorem 6.14. 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 $n \to \infty$.

6.7 Bootstrapping Some Regression Models

6.7.1 Parametric Bootstrap

For the parametric regression model $Y_i|\boldsymbol{x}_i \sim D(\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{\gamma})$ of Definition 6.22, 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

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}(\boldsymbol{\beta})) \tag{6.26}$$

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.

6.7.2 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).

6.8 Variable Selection

Consider 1D regression models where the response variable Y is independent of the $p \times 1$ vector of predictors $\boldsymbol{x}$ given $\boldsymbol{x}^T\boldsymbol{\beta}$, written $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T\boldsymbol{\beta}$. 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.

Some shrinkage methods do variable selection: the regression method, such as a GLM, uses the predictors that had nonzero shrinkage estimator coefficients. These methods include least angle regression, lasso, relaxed lasso, and elastic net. Least angle regression variable selection is the LARS-OLS hy-

brid 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), Simon et al. (2011), 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

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S \tag{6.27}$$

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. 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). Suppose that S is a subset of I and that model (6.27) holds. Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_I^T\boldsymbol{\beta}_I + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I.$$

Thus $\boldsymbol{\beta}_O = \boldsymbol{0}$ if $S \subseteq I$. The model using $\boldsymbol{x}^T\boldsymbol{\beta}$ is the full model.

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 $\boldsymbol{x}_{I_2} = (x_1, x_2)^T$.

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 $\hat{\boldsymbol{\beta}}_{VS}$. 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 variable selection estimator $\hat{\boldsymbol{\beta}}_{VS}$, but the I_k are randomly selected. Review Section 1.8 for mixture distributions.

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.

6.8.1 Large Sample Theory for Variable Selection Estimators

The Theorems 6.15 and 6.16 in this section are due to Rathnayake and Olive (2021), 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 (6.27) holds, and that if $S \subseteq I_j$ where the dimension of I_j is a_j, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{V}_j)$ where $\boldsymbol{V}_j$ is the covariance matrix of the asymptotic multivariate normal distribution. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}_{j,0}) \tag{6.28}$$

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 6.15 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 shrinkage 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 6.15c) 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 6.16. The assumption on $\boldsymbol{u}_{jn}$ in Theorem 6.15

is reasonable by (6.28) 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 give an argument for the Mallows (1973) C_p criterion when each submodel contains a constant. Let submodel I have $k \leq p$ predictors including a constant. Then

$$C_p(I) = \frac{SSE(I)}{MSE} + 2k - n$$

where MSE is for the full model,and $C_p(I) \geq -p$. Assume the full model is one of the submodels considered with $C_p(full) = p$, 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 $n \to \infty$. Suppose S is not a subset of I. Under the model $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$, $corr(r, r_I)$ will not converge to 1 as $n \to \infty$, and for large enough n, $[corr(r, r_I)]^2 \leq \gamma < 1$. Thus $C_p(I) \to \infty$ as $n \to \infty$. Hence $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$ if the zero mean iid errors have constant variance σ^2.

Theorem 6.15 a) proves that $\boldsymbol{u}$ is a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j, $E(\boldsymbol{u}) = \boldsymbol{0}$, and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma u} = \sum_j \pi_j \boldsymbol{V}_{j,0}$. 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.

Theorem 6.15. *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*

$$\boldsymbol{u}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \tag{6.29}$$

where the cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$.

b) Let $\boldsymbol{A}$ *be a* $g \times p$ *full rank matrix with* $1 \leq g \leq p$. *Then*

$$\boldsymbol{v}_n = \boldsymbol{A}\boldsymbol{u}_n = \sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} = \boldsymbol{v} \tag{6.30}$$

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.

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 $F_{\boldsymbol{u}_n}(\boldsymbol{t}) = \sum_k \pi_{kn} F_{\boldsymbol{u}_{kn}}(\boldsymbol{t}) \rightarrow F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$ at continuity points of the $F_{\boldsymbol{u}_j}(\boldsymbol{t})$ as $n \rightarrow \infty$.

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, $\boldsymbol{A}\boldsymbol{u}_j \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T) = N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. Then $\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. This special case occurs for $\hat{\boldsymbol{\beta}}_{S,MIX}$ if $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V})$ 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 6.15 can be used to justify prediction intervals after variable selection. See Olive, Rathnayake, and Haile (2021). Theorem 6.15d) is useful for *variable selection consistency* and the *oracle property* where $\pi_d = \pi_S = 1$ if $P(I_{min} = S) \rightarrow 1$ as $n \rightarrow \infty$. See Claeskens and Hjort (2008, pp. 101-114) and Fan and Li (2001) for references. A necessary condition for $P(I_{min} = S) \rightarrow 1$ is that S is one of the models considered with probability going to one. This condition holds under strong regularity conditions for fast methods. See

Wieczorek and Lei (2021) for forward selection and Hastie, Tibshirani, and Wainwright (2015, pp. 295-302) for lasso, where the predictors need a "near orthogonality" condition.

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 (10) where $E(\boldsymbol{v}) = \boldsymbol{0}$, and $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \sum_j \pi_j \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T$. By Theorem 5.3, if we had iid data $T_1, ..., T_B$, then R_c would be a large sample confidence region for $\boldsymbol{\theta}$. If $\sqrt{n}(T_n^* - T_n) \xrightarrow{D} \boldsymbol{v}$, then we could use the bootstrap sample and confidence regions (5.31) to (5.32). This condition holds only under strong regularity conditions such as $\pi_d = 1$ or $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{B}\boldsymbol{\beta}_S$ if $\boldsymbol{V}$ was diagonal. Section 6.8.2 explains why the bootstrap confidence regions may still be useful.

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 a generalization of Theorem 1.3 that defines $P(A|B_k)P(B_k) = 0$ if $P(B_k) = 0$,

$$F_{\boldsymbol{w}_n}(\boldsymbol{t}) = P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \le \boldsymbol{t}] =$$

$$\sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \le \boldsymbol{t}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})]P(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}) =$$

$$\sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta}) \le \boldsymbol{t}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})]\pi_{kn}$$

$$= \sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}^C_{I_k,0} - \boldsymbol{\beta}) \le \boldsymbol{t}]\pi_{kn} = \sum_{k=1}^{J} F_{\boldsymbol{w}_{kn}}(\boldsymbol{t})\pi_{kn}. \quad \square$$

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. Here $\boldsymbol{w}_j$ is a multivariate truncated normal distribution (where no truncation is possible) that is symmetric about $\boldsymbol{0}$. 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 (6.31) is the same as that for (6.29). Theorem 6.16 proves that $\boldsymbol{w}$ is a mixture distribution of the $\boldsymbol{w}_j$ with probabilities π_j.

Theorem 6.16. *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*

$$\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w} \tag{6.31}$$

where the cdf of $\boldsymbol{w}$ is $F_{\boldsymbol{w}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{w}_j}(\boldsymbol{t})$.

6.8.2 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 6.8.1, 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 $H_0 : \boldsymbol{A\beta} = \boldsymbol{B\beta}_E = \boldsymbol{0}$, 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 $\mathrm{Cov}(T^*)$.

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 (6.28),

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}_{j,0}) \text{ and } \sqrt{\mathrm{n}}(\hat{\boldsymbol{\beta}}^*_{\mathrm{I_j,0}} - \hat{\boldsymbol{\beta}}_{\mathrm{I_j,0}}) \xrightarrow{\mathrm{D}} \mathrm{N_p}(\boldsymbol{0}, \boldsymbol{V}_{\mathrm{j,0}}). \quad (6.32)$$

If Equation (6.32) 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 $(T_n - \boldsymbol{\theta}_0)^T [\boldsymbol{S}_T^*]^{-1} (T_n - \boldsymbol{\theta}_0) > D^2_{(U_B,T)}$. 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 $D^2_{(U_B,T)}$ 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 (6.32) 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 (6.28) 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. Equation (6.32) tends to hold for the nonparametric bootstrap by Section 6.7.2 and also for the the parametric bootstrap for OLS MLR by Section 6.2.1.

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.

6.9 Data Splitting

See Zhang and Olive (2022).

6.10 Prediction Intervals

See Olive, Rathnayake, and Haile (2021).

6.11 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 6.27. The **response variables** are the variables that you want to predict. The **predictor variables** are the variables used to predict the response variables.

Definition 6.28. 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 $\boldsymbol{X} = \mathbf{1}$ and $p = 1$.

The data matrix $\boldsymbol{W} = [\boldsymbol{X} \;\; \boldsymbol{Z}]$ except usually the first column $\mathbf{1}$ of $\boldsymbol{X}$ is omitted for software. 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 $n \times p$ design matrix of predictor variables is

$$\boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}$$

where $\boldsymbol{v}_1 = \mathbf{1}$.

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

$$\boldsymbol{E} = \begin{bmatrix} \epsilon_{1,1} & \epsilon_{1,2} & \dots & \epsilon_{1,m} \\ \epsilon_{2,1} & \epsilon_{2,2} & \dots & \epsilon_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n,1} & \epsilon_{n,2} & \dots & \epsilon_{n,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \dots & \boldsymbol{e}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{\epsilon}_1^T \\ \vdots \\ \boldsymbol{\epsilon}_n^T \end{bmatrix}.$$

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 $E(\boldsymbol{e}_j) = \mathbf{0}$ 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.

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 6.1. 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 6.29. 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* $\hat{\boldsymbol{E}} = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = \boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}} =$

$$\begin{bmatrix} \hat{\boldsymbol{\epsilon}}_1^T \\ \hat{\boldsymbol{\epsilon}}_2^T \\ \vdots \\ \hat{\boldsymbol{\epsilon}}_n^T \end{bmatrix} = \begin{bmatrix} \boldsymbol{r}_1 \ \boldsymbol{r}_2 \ \ldots \ \boldsymbol{r}_m \end{bmatrix} = \begin{bmatrix} \hat{\epsilon}_{1,1} & \hat{\epsilon}_{1,2} & \ldots & \hat{\epsilon}_{1,m} \\ \hat{\epsilon}_{2,1} & \hat{\epsilon}_{2,2} & \ldots & \hat{\epsilon}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\epsilon}_{n,1} & \hat{\epsilon}_{n,2} & \ldots & \hat{\epsilon}_{n,m} \end{bmatrix}.$$

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

$$\hat{\boldsymbol{E}} = [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}]\boldsymbol{Z}.$$

6.11.1 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 6.30. 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 6.31. 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 6.17. *The Hotelling-Lawley trace statistic*

$$U(\boldsymbol{L}) = \frac{1}{n-p}[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]. \quad (6.33)$$

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 (6.33) holds. □

Some notation is useful to show (6.33) 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 $\max_{1 \le i \le n} h_i \xrightarrow{P} 0$ 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 6.18: 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 6.19. If assumption D1 holds and if H_0 is true, then $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$.

Proof. By Theorem 6.18, $\sqrt{n}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W})$. 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 6.17,

$$T = (n-p)U(\boldsymbol{L}) = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}.$$

□

Some more details on the above results may be useful. Consider testing a linear hypothesis $H_0 : \boldsymbol{LB} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{LB} \neq \boldsymbol{0}$ where $\boldsymbol{L}$ is a full rank $r \times p$ matrix. For now assume the error distribution is multivariate normal $N_m(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon})$. Then

$$vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \\ \vdots \\ \hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m \end{pmatrix} \sim N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon} \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

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 $\boldsymbol{D} = \boldsymbol{\Sigma_\epsilon} \otimes \boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T = \boldsymbol{ACA}^T =$

$$\begin{bmatrix} \sigma_{11}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \sigma_{12}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \cdots & \sigma_{1m}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T \\ \sigma_{21}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \sigma_{22}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \cdots & \sigma_{2m}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{m1}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \sigma_{m2}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T & \cdots & \sigma_{mm}\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T \end{bmatrix}.$$

Under H_0, $vec(\boldsymbol{LB}) = \boldsymbol{A}\ \ vec(\boldsymbol{B}) = \boldsymbol{0}$, and

$$vec(\boldsymbol{L}\hat{\boldsymbol{B}}) = \begin{pmatrix} \boldsymbol{L}\hat{\boldsymbol{\beta}}_1 \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_2 \\ \vdots \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_m \end{pmatrix} \sim N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T).$$

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

$$T = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}. \tag{6.34}$$

A large sample level δ test will reject H_0 if $pval \leq \delta$ where

$$pval = P\left(\frac{T}{rm} < F_{rm,n-mp}\right). \tag{6.35}$$

Since least squares estimators are asymptotically normal, if the $\boldsymbol{\epsilon}_i$ are iid for a large class of distributions,

$$\sqrt{n}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) = \sqrt{n}\ \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \\ \vdots \\ \hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m \end{pmatrix} \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W})$$

where

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \xrightarrow{P} \boldsymbol{W}^{-1}.$$

Then under H_0,

$$\sqrt{n}\ \ vec(\boldsymbol{L}\hat{\boldsymbol{B}}) = \sqrt{n}\ \begin{pmatrix} \boldsymbol{L}\hat{\boldsymbol{\beta}}_1 \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_2 \\ \vdots \\ \boldsymbol{L}\hat{\boldsymbol{\beta}}_m \end{pmatrix} \xrightarrow{D} N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T),$$

and

$$n\ \ [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}.$$

Hence (6.34) holds, and (6.35) gives a large sample level δ test if the least squares estimators are asymptotically normal.

Kakizawa (2009) showed, under stronger assumptions than Theorem 6.19, 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 non-central $\chi^2_{rm}(\omega^2)$ 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 6.17 and 6.19 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 6.16.

By Theorem 6.19, 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),

$$V(\boldsymbol{L}) \leq -\log(\Lambda(\boldsymbol{L})) \leq U(\boldsymbol{L}).$$

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

$$\frac{gt-2u}{rm} \ \frac{1-\Lambda^{1/t}}{\Lambda^{1/t}} \approx F(rm, gt-2u) \ \text{ or } \ (\mathrm{n}-\mathrm{p})\mathrm{t} \ \frac{1-\Lambda^{1/\mathrm{t}}}{\Lambda^{1/\mathrm{t}}} \approx \chi^2_{\mathrm{rm}}.$$

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 $N_m(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon})$, 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 $H_0 : \boldsymbol{LB} = \boldsymbol{0}$ and $H_1 : \boldsymbol{LB} \neq \boldsymbol{0}$.
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 $\boldsymbol{LB} \neq \boldsymbol{0}$. If pval $> \delta$, fail to reject H_0 and conclude that $\boldsymbol{LB} = \boldsymbol{0}$ or that there is not enough evidence to conclude that $\boldsymbol{LB} \neq \boldsymbol{0}$.

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 $d_j = (\boldsymbol{X}^T \boldsymbol{X})_{jj}^{-1}$, 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 6.2
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 6.2. 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.

6.11.2 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 6.32. 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 6.21.

Theorem 6.20. 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 $P(B_n) \to 1$ as $n \to \infty$. 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 6.21 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 er-

ror 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 6.21.

Theorem 6.21. 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

$$\{\boldsymbol{z} : (\boldsymbol{z} - \hat{\boldsymbol{y}}_f)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\boldsymbol{z} - \hat{\boldsymbol{y}}_f) \leq D_{(U_n)}^2\} =$$

$$\{\boldsymbol{z} : D_{\boldsymbol{z}}^2(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{(U_n)}^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{(U_n)}\}. \tag{6.36}$$

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 (8.1) 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 (6.36) is asymptotically optimal.

Proof. a) Suppose $(\boldsymbol{x}_f, \boldsymbol{y}_f) = (\boldsymbol{x}_i, \boldsymbol{y}_i)$. Then

$$D_{\boldsymbol{y}_i}^2(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) = (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i) = \hat{\boldsymbol{\epsilon}}_i^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \hat{\boldsymbol{\epsilon}}_i = D_{\hat{\boldsymbol{\epsilon}}_i}^2(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}).$$

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 6.20 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

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \leq D^2_{(U_n)}\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r) \leq D_{(U_n)}\}. \tag{6.37}$$

Theorem 6.22 will show that this prediction region (6.37) 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 6.29 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 6.18 holds, and the $\boldsymbol{\epsilon}_i$ have a nonsingular covariance matrix, then (6.37) is a large sample $100(1-\delta)\%$ prediction region for $\boldsymbol{y}_f$.

Theorem 6.22. 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 (6.37) is the nonparametric prediction region (4.11) applied to the $\hat{\boldsymbol{z}}_i$.

Proof. Multivariate linear regression with least squares satisfies Theorem 6.21 by Su and Cook (2012). (See Theorem 6.18.) 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 6.22 uses $(T, \boldsymbol{C}) = (\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ in (6.36), and has simple geometry. Let R_r be the nonparametric prediction region (6.37) applied to the residuals $\hat{\boldsymbol{\epsilon}}_i$ with $\hat{\boldsymbol{y}}_f = \boldsymbol{0}$. Then R_r is a hyperellipsoid with center $\boldsymbol{0}$, 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 $EC_m(\boldsymbol{0}, \boldsymbol{\Sigma}, g)$ 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 (6.37) for the data since the prediction region (6.37) centered at $\hat{\boldsymbol{y}}_i$ contains $\boldsymbol{y}_i$ iff R_r, the prediction region centered at $\boldsymbol{0}$, 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.

6.12 Summary

6.13 Complements

Multiple Linear Regression

For linear model theory based on large sample theory, see Olive (2022b). 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 $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{OLS}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \to \boldsymbol{0}$ 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 (2022) 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 (2021), 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.

6.14 Problems

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 $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$ 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 $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$, $\hat{\boldsymbol{Z}} = \boldsymbol{W}\hat{\boldsymbol{\eta}}$, $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$, 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?