David J. Olive

Survival Analysis

April 6, 2023

[redacted]

Preface

Many statistics departments offer a one semester undergraduate–graduate course in Reliability and Survival Analysis using texts such as Allison (2010), Collett (2014), and Hosmer et al. (2008). More advanced texts include Harrell (2015), Kalbfleisch and Prentice (2002), Klein and Moeschberger (2003), Lawless (2002), Miller (1981), and Smith (2002). Also see Kleinbaum and Klein (2012), Lee and Wang (2003), Leemis (1995), Meeker and Escobar (1998), and Tableman and Kim (2003).

The prerequisite for this text is a calculus based course in statistics at the level of Chihara and Hesterberg (2011), Hogg et al. (2015), Larsen and Marx (2017), Wackerly, Mendenhall and Scheaffer (2008) or Walpole et al. (2016). Linear algebra and knowledge of regression would be useful. See Olive (2017a) and Cook and Weisberg (1999).

Some highlights of this text follow.

- The response plot is useful for checking the model.

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

```
source("http://parker.ad.siu.edu/Olive/survpack.txt")
source("http://parker.ad.siu.edu/Olive/survdata.txt")
```

The *R* software is used in this text. See R Core Team (2023). Some packages used in the text include

Acknowledgements

Teaching Survival Analysis in Math 473 at Southern Illinois University was very useful.

Contents

Chapter 1
Univariate Survival Analysis

This chapter considers univariate survival analysis: there is a response variable but no predictors. In the analysis of "time to event" data, there are n individuals and the time until an event is recorded for each individual. Typical events are failure of a product or death of a person or reoccurrence of cancer after surgery, but other events such as first use of cigarettes or the time that baboons come down from trees (early in the morning) can also be modeled. The data is typically right skewed and censored data is often present.

Censoring occurs because of time and cost constraints. A product such as light bulbs may be tested for 1000 hours. Perhaps 30% fail in that time but the remaining 70% are still working. These are censored: they give partial information on the lifetime of the bulbs because it is known that about 70% last longer than 1000 hours. Handling censoring and time dependent covariates is what makes the analysis of time to event data different from other fields of statistics.

Reliability analysis is used in *engineering* to study the lifetime (time until failure) of manufactured products, while survival analysis is used in *actuarial sciences*, *statistics*, and *biostatistics* to study the lifetime (time until death) of humans, often after contracting a deadly disease. In the *social sciences*, the study of the time until the occurrence of an event is called the analysis of event time data or event history analysis. In *economics*, the study is called duration analysis or transition analysis. Hence reliability data = failure time data = lifetime data = survival data = event time data.

1.1 Functions Related to the Survival Function

In this text $\log(t) = \ln(t) = \log_e(t)$ while $\exp(t) = e^t$. One of the difficulties with survival analysis is that the response Y = survival time is usually not

observed, instead the censored response is observed. In this chapter the data will be right censored, and "right" will often be omitted. In the following definition, note that both $T \geq 0$ and $Y \geq 0$ are nonnegative.

Definition 1.1. Let $Y \geq 0$ be the time until an event occurs. Then Y is called the **survival time** or time until event. The survival time is **censored** if the event of interest has not been observed. Let Y_i be the ith survival time. Let Z_i be the time the ith observation (possibly an individual or machine) leaves the study for any reason other than the event of interest. Then Z_i is the time until the ith observation is censored. Then the **right censored survival time** T_i of the ith observation is $T_i = \min(Y_i, Z_i)$. Let $\delta_i = 0$ if T_i is (right) censored ($T_i = Z_i$) and let $\delta_i = 1$ if T_i is not censored ($T_i = Y_i$). Then the univariate survival analysis data is $(T_1, \delta_1), (T_2, \delta_2), ..., (T_n, \delta_n)$. Alternatively, the data is $T_1, T_2^*, T_3, ..., T_{n-1}^*, T_n$ where the * means that the case was (right) censored. Sometimes the asterisk * is replaced by a plus +, and Y_i, y_i or t_i can replace T_i.

In this chapter we will assume that the censoring mechanism is independent of the time to event: Y_i and Z_i are independent. Often censoring occurs because of cost and time constraints.

For example, in a study breast cancer patients who receive a lumpectomy, suppose the researchers want to keep track of 100 patients for five years after receiving a lumpectomy (tumor removal). The response is time until death after a lumpectomy. Patients who are lost to the study (move or eventually refuse to cooperate), and patients who are still alive after the study are censored. Perhaps 15% die, 5% move away and so leave the study, and 80% are still alive after 5 years. Then 85% of the cases are (right) censored. The actual study may take two years to recruit patients, follow each patient for 5 years, but end 5 years after the end of the two year recruitment period. So patients enter the study at different times, but the censored response is the time until death or censoring from the time the patient entered the study.

Definition 1.2. i) The **cumulative distribution function** (cdf) of Y is $F(t) = P(Y \leq t)$. Since $Y \geq 0$, $F(0) = 0, F(\infty) = 1$, and $F(t)$ is nondecreasing.

ii) The probability density function (**pdf**) of Y is $f(t) = F'(t)$.

iii) The **survival function** of Y is $S(t) = P(Y > t)$. $S(0) = 1, S(\infty) = 0$ and $S(t)$ is nonincreasing.

iv) The **hazard function** of Y is $h(t) = \dfrac{f(t)}{1 - F(t)}$ for $t > 0$ and $F(t) < 1$. Note that $h(t) \geq 0$ if $F(t) < 1$.

v) The **cumulative hazard function** of Y is $H(t) = \int_0^t h(u)du$ for $t > 0$. It is true that $H(0) = 0, H(\infty) = \infty$, and $H(t)$ is nondecreasing.

Assume $Y \geq 0$. Then $F(0) = 0$, $S(0) = 1$, and $H(0) = 1$. Note that $S(\infty) = 0$ implies that $H(\infty) = \infty$ where $\lim_{t\to\infty} H(t) = H(\infty)$. Memorize that $0 \leq F(t) \leq 1$, $0 \leq S(t) \leq 1$, $f(t) \geq 0$, $h(t) \geq 0$, and $H(t) \geq 0$.

Given one of $F(t), f(t), S(t), h(t)$ or $H(t)$, the following theorem shows how to find the other 4 quantities for $t > 0$. Each of these five quanities completely determines the distribution of the random variable. In reliability analysis, the reliability function $R(t) = S(t)$, and in economics, Mill's ratio $= 1/h(t)$. In actuarial sciences, $h(t)$ is the force of mortality.

Theorem 1.1.
A) $F(t) = \int_0^t f(u)du = 1 - S(t) = 1 - \exp[-H(t)] = 1 - \exp[-\int_0^t h(u)du]$.
B) $f(t) = F'(t) = -S'(t) = h(t)[1 - F(t)] = h(t)S(t) = h(t)\exp[-H(t)] = H'(t)\exp[-H(t)]$.
C) $S(t) = 1 - F(t) = 1 - \int_0^t f(u)du = \int_t^\infty f(u)du = \exp[-H(t)] = \exp[-\int_0^t h(u)du]$.
D)

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} = \frac{F'(t)}{1 - F(t)} = \frac{-S'(t)}{S(t)} = -\frac{d}{dt}\log[S(t)] = H'(t).$$

E) $H(t) = \int_0^t h(u)du = -\log[S(t)] = -\log[1 - F(t)]$.

Tips: i) If $F(t) = 1 - \exp[G(t)]$ for $t > 0$, then $H(t) = -G(t)$ and $S(t) = \exp[G(t)]$.
ii) For $S(t) > 0$, note that $S(t) = \exp[\log(S(t))] = \exp[-H(t)]$. Finding $\exp[\log(S(t))]$ and setting $H(t) = -\log[S(t)]$ is easier than integrating $h(t)$.

Next an interpretation for the hazard function is given. If $P(B) > 0$, then

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)} \quad \text{and} \quad P(A|B) = \frac{P(A)}{P(B)}$$

if $A \subseteq B$. Suppose the time until event is the time until death. Note that

$$P[t < Y < t + \Delta t | Y > t] = \frac{P[t < Y \leq t + \Delta t]}{P(Y > t)} = \frac{F(t + \Delta t) - F(t)}{1 - F(t)}.$$

So

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} P[t < Y \leq t + \Delta t | Y > t] = \lim_{\Delta t \to 0} \frac{\frac{F(t+\Delta t) - F(t)}{\Delta t}}{1 - F(t)}$$

$$= \frac{f(t)}{1 - F(t)} = h(t).$$

So for small Δt, it follows that $h(t)\Delta t \approx P[t < Y < t + \Delta t | Y > t] \approx$ P(person dies in interval $(t, t + \Delta t]$ given that the person has survived up to time t).

Larger $h(t)$ implies that the hazard of death is higher. The hazard function takes into account the *aging* of the observation (person or product).

For example, an 80 year old white male has about a 50% chance of living to 85 while a 100 year old white male has about a 50% chance of living to 101, although the percentage of white males living to 101 is tiny.

Example 1.1. Suppose $Y \sim EXP(\lambda)$ where $\lambda > 0$, then $h(t) = \lambda$ for $t > 0$, $f(t) = \lambda e^{-\lambda t}$ for $t > 0$, $F(t) = 1 - e^{-\lambda t}$ for $t > 0$, $S(t) = e^{-\lambda t}$ for $t > 0$, $H(t) = \lambda t$ for $t > 0$ and $E(Y) = 1/\lambda$. The **exponential distribution** can be a good model if failures are due to random shocks that follow a Poisson process (light bulbs, electrical components), but constant hazard means that a used product is as good as a new product: aging has no effect on the probability of failure of the product. The exponential distribution is the only distribution of a continuous random variable Y with a constant hazard function since $h(t)$ completely determines the distribution of the random variable Y. Derive $H(t), S(t), F(t)$, and $f(t)$ from the constant hazard function $h(t) = \lambda$ for $t > 0$ and some $\lambda > 0$.

Solution: $H(t) = \int_0^t h(u)du = \int_0^t \lambda du = \lambda t$ for $t > 0$.
$S(t) = e^{-H(t)} = e^{-\lambda t}$, for $t > 0$.
$F(t) = 1 - S(t) = 1 - e^{-\lambda t}$ for $t > 0$.
Finally, $f(t) = h(t)S(t) = \lambda e^{-\lambda t} = F'(t)$ for $t > 0$.

Suppose the observed survival times $T_1, ..., T_n$ are a censored data set from an exponential $EXP(\lambda)$ distribution. Let $T_i = Y_i^*$. Let $\delta_i = 0$ if the case is censored and let $\delta_i = 1$, otherwise. Let $r = \sum_{i=1}^n \delta_i =$ the number of uncensored cases. Then the maximum likelihood estimator (MLE) $\hat{\lambda} = r/\sum_{i=1}^n T_i$. So $\hat{\lambda} = r/\sum_{i=1}^n Y_i^*$. A 95% confidence interval (CI) for λ is $\hat{\lambda} \pm 1.96\hat{\lambda}/\sqrt{r}$. See Section 1.2.

Example 1.2. If $Y \sim$ Weibull(γ, λ) where $\gamma > 0$ and $\lambda > 0$, then $h(t) = \lambda\gamma t^{\gamma-1}$ for $t > 0$, $f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^\gamma)$ for $t > 0$, $F(t) = 1 - \exp(-\lambda t^\gamma)$ for $t > 0$, $S(t) = \exp(-\lambda t^\gamma)$ for $t > 0$, $H(t) = \lambda t^\gamma$ for $t > 0$. The Weibull($\lambda, \gamma = 1$) distribution is the EXP(λ) distribution. The hazard function can be increasing, decreasing or constant. Hence the **Weibull distribution** often fits reliability data well, and the Weibull distribution is an important distribution in reliability analysis. Derive $H(t), S(t), F(t)$, and $f(t)$ if $Y \sim$ Weibull(λ, γ).

Solution:

$$H(t) = \int_0^t h(u)du = \int_0^t \lambda\gamma u^{\gamma-1}du = \lambda\gamma \frac{u^\gamma}{\gamma}\bigg|_0^t = \lambda t^\gamma \quad \text{for} \quad \text{t} > 0.$$

$S(t) = \exp[-H(t)] = \exp[-\lambda t^\gamma]$, for $t > 0$.
$F(t) = 1 - S(t) = 1 - \exp[-\lambda t^\gamma]$ for $t > 0$.
Finally, $f(t) = h(t)S(t) = \lambda\gamma t^{\gamma-1}\exp[-\lambda t^\gamma]$ for $t > 0$.

Recall from the central limit theorem that the sample mean $\overline{X} = \sum_{i=1}^n X_i/n$ is approximately normal for many distributions. For many distributions, $\min(X_1, ..., X_n)$ is approximately Weibull. Suppose a product is made of m components with iid failure times X_{im}. Suppose the product fails as soon as one of the components fails, eg a chain of links fails when the weakest link fails. Then often the failure time $Y_i = \min(X_{im}, ..., X_{im})$ is approximately Weibull.

Notation: The set $\{t : f(t) > 0\}$ is the support of Y. Often the support of Y is $(0, \infty) = t > 0$, and the formulas will omit the $t > 0$.

Theorem 1.2. $E(Y) = \int_0^\infty y f(y)dy = \int_0^\infty t f(t)dt = \int_0^\infty S(t)dt$ if $\lim_{t\to\infty} tS(t) = 0$.

1.2 Estimating the Survival Function

Notation: Let the indicator variable $I_A(Y_i) = 1$ if $Y_i \in A$ and $I_A(Y_i) = 0$ otherwise. Often write $I_{(t,\infty)}(Y_i)$ as $I(Y_i > t)$.

Definition 1.3. If none of the survival times are censored, then the **empirical survival function** $\hat{S}_E(t) =$ (number of individual with survival times $> t$)/(number of individuals) $= a/n$. So

$$\hat{S}_E(t) = \frac{1}{n}\sum_{i=1}^n I(Y_i > t) = \hat{p}_t =$$

sample proportion of lifetimes $> t$.

Assume $Y_1, ..., Y_n$ are iid with $Y_i \geq 0$. Fix $t > 0$. Then $I(Y_i > t)$ are iid binomial(1,$p = P(Y_i > t)$). So $n\hat{S}_E(t) \sim$ binomial(n,$p = P(Y_i > t)$). Hence $E[n\hat{S}_E(t)] = nP(Y > t)$ and $V[n\hat{S}_E(t)] = nS(t)F(t)$. Thus $E[\hat{S}_E(t)] = S(t)$ and $V[\hat{S}_E(t)] = S(t)F(t)/n = [S(t)(1-S(t))]/n \leq 0.25/n$. Thus $SD[\hat{S}_E(t)] = \sqrt{V[\hat{S}_E(t)]} \leq 0.5/\sqrt{n}$. So need $n \approx 100$ for $SD[\hat{S}_E(t)] < 0.05$.

Let $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ be the observed ordered survival times (= lifetimes = death times). Let $t_0 = 0$ and let $0 < t_1 < t_2 < \cdots < t_m$ be the distinct survival times. Let $d_i =$ number of deaths at time t_i. If $m = n$ and $d_i = 1$ for $i = 1, ..., n$ then there are **no ties**. If $m < n$ and some $d_i \geq 2$, then there are **ties**.

Then $\hat{S}_E(t)$ is a step function with $\hat{S}_E(0) = 1$ and $\hat{S}_E(t) = \hat{S}_E(t_{i-1})$ for $t_{i-1} \leq t < t_i$. Note that $\sum_{i=1}^m d_i = n$. Know how to compute and plot $\hat{S}_E(t)$ given the $t_{(i)}$ or given the t_i and d_i. Use a table like the one below. Let

$a_0 = n$ and $a_i = \sum_{k=1}^{n} I(T_i > t_i) = \#$ of cases $t_{(j)} > t_i$ for $i = 1, ..., m$. Then $\hat{S}_E(t_i) = a_i/n = \sum_{k=1}^{n} I(T_i > t_i)/n = \hat{S}_E(t_{i-1}) - \dfrac{d_i}{n}$.

t_i	d_i	$\hat{S}_E(t_i) = \hat{S}_E(t_{i-1}) - \dfrac{d_i}{n} = a_i/n$
$t_0 = 0$		$\hat{S}_E(0) = 1 = \dfrac{n}{n} = \dfrac{a_0}{n}$
t_1	d_1	$\hat{S}_E(t_1) = \hat{S}_E(t_0) - \dfrac{d_1}{n} = \dfrac{a_0 - d_1}{n} = \dfrac{a_1}{n}$
t_2	d_2	$\hat{S}_E(t_2) = \hat{S}_E(t_1) - \dfrac{d_2}{n} = \dfrac{a_1 - d_2}{n} = \dfrac{a_2}{n}$
$\vdots$	$\vdots$	$\vdots$
t_j	d_j	$\hat{S}_E(t_j) = \hat{S}_E(t_{j-1}) - \dfrac{d_j}{n} = \dfrac{a_{j-1} - d_j}{n} = \dfrac{a_j}{n}$
$\vdots$	$\vdots$	$\vdots$
t_{m-1}	d_{m-1}	$\hat{S}_E(t_{m-1}) = \hat{S}_E(t_{m-2}) - \dfrac{d_{m-1}}{n} = \dfrac{a_{m-2} - d_{m-1}}{n} = \dfrac{a_{m-1}}{n}$
t_m	d_m	$\hat{S}_E(t_m) = 0 = \hat{S}_E(t_{m-1}) - \dfrac{d_m}{n} = \dfrac{a_{m-1} - d_m}{n} = \dfrac{a_m}{n}$

Let $\hat{S}(t)$ be the estimated survival function. Let $t(p)$ be the pth percentile of Y: $P(Y \leq t(p)) = F(t(p)) = p$ so $1 - p = S(t(p)) = P(Y > t(p))$. Then $\hat{t}(p)$, the estimated time when 100 p % have died, can be estimated from a graph of $\hat{S}(t)$ with "over" and "down" lines. a) Find $1-p$ on the vertical axis and draw a horizontal "over" line to $\hat{S}(t)$. Draw a vertical "down" line until it intersects the horizontal axis at $\hat{t}(p)$. Usually want $p = 0.5$ but sometimes $p = 0.25$ and $p = 0.75$ are used.

Example 1.3. Smith (2002, p. 68) gives steroid induced remission times for leukemia patients. The $t_{(j)}$, $t - i$ and d_i are given in the following table. The a_i and $\hat{S}_E(t)$ needed to be computed. Note that $a_i = \#$ of cases with $t_{(j)} > t_i$. For the following table, the 2nd column $t_{(j)}$ gives the 21 ordered survival times. The 3rd column t_i gives the distinct ordered survival times. Often just the number is given, so $t_1 = 1$ would be replaced by 1. The 4th column d_i tells how many events (remissions) occurred at time t_i and the last column computes $\hat{S}_E(t_i)$. A good check is that the 1st column entry divided by n is equal to $a_i/n = \hat{S}_E(t_i) =$ last column entry. A graph of the estimated survival function would be a step function with times 0, 1, ..., 23 on the horizontal axis and $\hat{S}_E(t)$ on the vertical axis. A convention is to draw vertical lines at the jumps (at the t_i). So the step function would be 1 on

(0,1), 19/21 on (1,2), ..., 1/21 on (22,23) and 0 for $t > 23$. The vertical lines connecting the steps are at $t = 1, 2, ..., 23$.

a_i	$t_{(j)}$	t_i	d_i	$\hat{S}_E(t_i) = \hat{S}_E(t_{i-1}) - \frac{d_i}{n}$
21		$t_0 = 0$		$\hat{S}_E(0) = 1 = 21/21$
	1			
19	1	$t_1 = 1$	2	$\hat{S}_E(1) = (21-2)/21 = 19/21$
	2			
17	2	$t_2 = 2$	2	$\hat{S}_E(2) = (19-2)/21 = 17/21$
16	3	$t_3 = 3$	1	$\hat{S}_E(3) = (17-1)/21 = 16/21$
	4			
14	4	$t_4 = 4$	2	$\hat{S}_E(4) = (16-2)/21 = 14/21$
	5			
12	5	$t_5 = 5$	2	$\hat{S}_E(5) = (14-2)/21 = 12/21$
	8			
	8			
	8			
8	8	$t_6 = 8$	4	$\hat{S}_E(8) = (12-4)/21 = 8/21$
	11			
6	11	$t_7 = 11$	2	$\hat{S}_E(11) = (8-2)/21 = 6/21$
	12			
4	12	$t_8 = 12$	2	$\hat{S}_E(12) = (6-2)/21 = 4/21$
3	15	$t_9 = 15$	1	$\hat{S}_E(15) = (4-1)/21 = 3/21$
2	17	$t_{10} = 17$	1	$\hat{S}_E(17) = (3-1)/21 = 2/21$
1	22	$t_{11} = 22$	1	$\hat{S}_E(22) = (2-1)/21 = 1/21$
0	23	$t_{12} = 23$	1	$\hat{S}_E(23) = (1-1)/21 = 0$ good check

Example 1.4. If $d_i = 1, 1, 1, 1$ and if $t_i = 1, 3, 5, 7$, then $a_1 = 3$, $a_2 = 2$ and $a_3 = 1$. Hence $\hat{S}_E(1) = 0.75, \hat{S}_E(3) = 0.5, \hat{S}_E(5) = 0.25$, and $\hat{S}_E(7) = 0$, and the estimated survival function is graphed as below.

```
^
S_E(t)
   ____
  |    |_____
  |          |______
  |                 |______
  |_________________________|_  t
     1   2   3   4   5   6   7
```

Let $t_1 \leq t < t_m$. Then the **classical large sample 95% CI** for $S(t_c)$ based on $\hat{S}_E(t)$ is

$$\hat{S}_E(t_c) \pm 1.96\sqrt{\frac{\hat{S}_E(t_c)[1-\hat{S}_E(t_c)]}{n}} = \hat{S}_E(t_c) \pm 1.96 SE[\hat{S}_E(t_c)] = [L, U].$$

Use $[\max(0, L), \min(1, U)]$.

Let $0 < t$. Let

$$\tilde{p}_{t_c} = \frac{n\hat{S}_E(t_c) + 2}{n + 4}.$$

Then the **plus four 95% CI** for $S(t_c)$ based on $\hat{S}_E(t)$ is

$$\tilde{p}_{t_c} \pm 1.96\sqrt{\frac{\tilde{p}_{t_c}[1 - \tilde{p}_{t_c}]}{n + 4}} = \tilde{p}_{t_c} \pm 1.96SE[\tilde{p}_{t_c}] = [L, U].$$

Use $[\max(0, L), \min(1, U)]$. For this CI, four imaginary T_i^* are added to the sample, with two of the $T_i^* > t_m > t_c$ and two $< t_1 < t_c$. See Agresti and Coull (1998).

The 95% large sample CI $\hat{S}_E(t_c) \pm 1.96SE[\tilde{p}_{t_c}]$ is also interesting.

Example 1.5. Let $n = 21$ and $\hat{S}_E(12) = 4/21$.
a) Find the 95% classical CI for $\hat{S}_E(12)$.
b) Find the 95% plus four CI for $\hat{S}_E(12)$.
Solution: a)

$$\frac{4}{21} + 1.96\sqrt{\frac{\frac{4}{21}(1 - \frac{4}{21})}{21}} = \frac{4}{21} \pm 0.16795 = [0.0225, 0.3584].$$

b)

$$\tilde{p}_{12} = \frac{21\frac{4}{21} + 2}{21 + 4} = \frac{6}{25}.$$

So the 95% CI is

$$\frac{6}{25} + 1.96\sqrt{\frac{\frac{6}{25}(1 - \frac{6}{25})}{25}} = \frac{6}{25} \pm 0.16742 = [0.0726, 0.4074].$$

Note that the CIs are not very short since $n = 21$ is small.

Let Y_i = time to event for ith person. $T_i = \min(Y_i, Z_i)$ where Z_i is the censoring time for the ith person (the time the ith person is lost to the study for any reason other than the time to event under study). The censored data is $y_1, y_2+, y_3, ..., y_{n-1}, y_n+$ where y_i means the time was uncensored and y_i+ means the time was censored. Let $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ be the ordered survival times (so if y_4+ is the smallest survival time, then $t_{(1)} = y_4+$). A status variable will be 1 if the time was uncensored and 0 if censored.

Let $[t_0, t_m) = I_1 \cup I_2 \cup \cdots \cup I_m = [t_0, t_1) \cup [t_1, t_2) \cdots \cup [t_{m-1}, t_m)$ where $t_0 \geq 0$ and $t_m = \infty$ is possible. It is possible that the 1st interval will have left endpoint $t_0 > 0$ and the last interval will have finite right endpoint $t_m < \infty$. Suppose that the following quantities are known: d_j = # deaths in I_j,
c_j = # of censored survival times in I_j, and
n_j = # at risk in I_j = # who were alive and not yet censored at the start of

I_j (at time t_{j-1}). Note that $n_1 = n$ and $n_j = n_{j-1} - d_{j-1} - c_{j-1}$ for $j > 1$. This equation shows how those at risk in the $(j-1)$th interval propagate to the jth interval.
Let $n'_j = n_j - \dfrac{c_j}{2}$ = average number at risk in I_j.

Definition 1.4. The **lifetable estimator** or actuarial method estimator of $S_Y(t)$ takes $\hat{S}_L(0) = 1$ and

$$\hat{S}_L(t_k) = \prod_{j=1}^{k} \frac{n'_j - d_j}{n'_j} = \prod_{j=1}^{k} \tilde{p}_j$$

for $k = 1, ..., m-1$. If $t_m = \infty$, $\hat{S}_L(t)$ is undefined for $t > t_{m-1}$. Suppose $t_m \neq \infty$. Then take $\hat{S}_L(t) = 0$ for $t \geq t_m$ if $c_m = 0$. If $c_m > 0$, then $\hat{S}_L(t)$ is undefined for $t \geq t_m$. (Some programs use $\hat{S}_L(t) = 0$ for $t \geq t_m$ if $t_m \neq \infty$.)

To graph $\hat{S}_L(t)$, use linear interpolation (connect the dots). If $n'_j = 0$, take $\tilde{p}_j = 0$. Note that

$$\hat{S}_L(t_k) = \hat{S}_L(t_{k-1}) \frac{n'_k - d_k}{n'_k} \quad \text{for} \quad \mathrm{k} = 1, ..., \mathrm{m} - 1.$$

The lifetable estimator is used to estimate $S_Y(t) = P(Y > t)$ when there is censoring. Also, the actual event or censoring times are unknown, but the number of event and censoring times in each interval I_j is known for $j = 1, ..., m$. Let $p_j = \mathrm{P(surviving\ through}\ I_j|\ \mathrm{alive\ at\ the\ start\ of}\ I_j) = P(Y > t_j | Y > t_{j-1}) = \dfrac{P(Y > t_j, Y > t_{j-1})}{P(Y > t_{j-1})} = \dfrac{S(t_j)}{S(t_{j-1})}$. Now $p_1 = S(t_1)/S(t_0) = S(t_1)$ since $S(0) = S(t_0) = 1$. Writing $S(t_k)$ as a telescoping product gives

$$S(t_k) = S(t_1) \frac{S(t_2)}{S(t_1)} \frac{S(t_3)}{S(t_2)} \cdots \frac{S(t_{k-1})}{S(t_{k-2})} \frac{S(t_k)}{S(t_{k-1})} = p_1 p_2 \cdots p_k = \prod_{j=1}^{k} p_j.$$

Let $\hat{p}_j = 1-$ (number dying in I_j)/(number with potential to die in I_j). Then $\tilde{p}_j = 1 - d_j/n'_j$ is the estimate of p_j used by the lifetable estimator, assuming that the censoring is roughly uniform over each interval.

Know how to get the lifetable estimator and $SE(\hat{S}_L(t_i))$ from output.

```
 (left output)                        (right output)
interval survival survival SE or interval survival survival SE
0    50   1.00        0            0    50  0.7594    0.0524
50  100   0.7594    0.0524        50   100  0.5889    0.0608
100 200   0.5889    0.0608       100   200  0.5253    0.0602
```

Since $\hat{S}_L(0) = 1$, $\hat{S}_L(t)$ is for the left endpoint for the "left output", and for the right endpoint for the "right output." For both cases, $\hat{S}_L(50) = 0.7594$ and $SE(\hat{S}_L(50)) = 0.0524$.

A 95% CI for $S_Y(t_i)$ based on the lifetable estimator is

$$\hat{S}_L(t_i) \pm 1.96\ SE[\hat{S}_L(t_i)].$$

Hence for the above output, a 95% CI for $S_Y(50)$ is $0.7594 \pm 1.96(0.0524) = [0.6567, 0.8621]$.

Know how to compute $\hat{S}_L(t)$ with a table like the one below. The first 4 entries need to be given but the last 3 columns may need to be filled in. On an exam you may be given a table with all but a few entries filled.

I_j, d_j, c_j, n_j	n'_j	$\frac{n'_j - d_j}{n'_j}$	$\hat{S}_L(t)$
$[t_0 = 0, t_1)$, d_1, c_1, n_1	$n_1 - \frac{c_1}{2}$	$\frac{n'_1 - d_1}{n'_1}$	$\hat{S}_L(t_0) = \hat{S}_L(0) = 1$
$[t_1, t_2)$, d_2, c_2, n_2	$n_2 - \frac{c_2}{2}$	$\frac{n'_2 - d_2}{n'_2}$	$\hat{S}_L(t_1) = \hat{S}_L(t_0)\frac{n'_1 - d_1}{n'_1}$
$[t_2, t_3)$, d_3, c_3, n_3	$n_3 - \frac{c_3}{2}$	$\frac{n'_3 - d_3}{n'_3}$	$\hat{S}_L(t_2) = \hat{S}_L(t_1)\frac{n'_2 - d_2}{n'_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[t_{k-1}, t_k)$, d_k, c_k, n_k	$n_k - \frac{c_k}{2}$	$\frac{n'_k - d_k}{n'_k}$	$\hat{S}_L(t_{k-1}) =$ $\hat{S}_L(t_{k-2})\frac{n'_{k-1} - d_{k-1}}{n'_{k-1}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[t_{m-2}, t_{m-1})$, d_{m-1}, c_{m-1}, n_{m-1}	$n_{m-1} - \frac{c_{m-1}}{2}$	$\frac{n'_{m-1} - d_{m-1}}{n'_{m-1}}$	$\hat{S}_L(t_{m-2}) =$ $\hat{S}_L(t_{m-3})\frac{n'_{m-2} - d_{m-2}}{n'_{m-2}}$
$[t_{m-1}, t_m = \infty)$, d_m, c_m, n_m	$n_m - \frac{c_m}{2}$	$\frac{n'_m - d_m}{n'_m}$	$\hat{S}_L(t_{m-1}) =$ $\hat{S}_L(t_{m-2})\frac{n'_{m-1} - d_{m-1}}{n'_{m-1}}$

Also get a 95% CI from output like that below. So the 95% CI for S(50) is [0.65666,0.86213] $\approx$ [0.6567,0.8621].

```
time survival SDF_LCL SDF_UCL
0     1.0     1.0     1.0
50    0.7594  0.65666 0.86213
```

Example 1.6. Allison (1995, p. 49-51) gives time until death after heart transplant for 68 patients. The 1st 5 columns are given, but the last 3 columns need to be computed. Use 4 digits in the computations.

I_j	t_j	d_j	c_j	n_j	$n'_j = n_j - c_j/2$	$\tilde{p}_j = \frac{n'_j - d_j}{n'_j}$	$\hat{S}_L(t_j) = \hat{S}_L(t_{j-1})\tilde{p}_j$
[0,50)	0	16	3	68	66.5	0.7594	$\hat{S}(0) = 1$
[50,100)	50	11	0	49	49	0.7755	$\hat{S}(50) = 0.7594$
[100,200)	100	14	2	38	37	0.8919	$\hat{S}(100) = 0.5889$
[200,400)	200	5	4	32	30	0.8333	$\hat{S}(0) = 0.5252$
[400,700)	400	2	6	23	20	0.90	$\hat{S}(400) = 0.4376$
[700,1000)	700	4	3	15	13.5	0.7037	$\hat{S}(700) = 0.7037$
[1000,1300)	1000	1	2	8	7	0.8571	$\hat{S}(1000) = 0.2771$
[1300,1600)	1300	1	3	5	3.5	0.7143	$\hat{S}(1300) = 0.2375$
[1600,∞)	1600	0	1	1	0.5	1.0	$\hat{S}(1600) = 0.1696$

Greenwood's formula is

$$SE[\hat{S}_L(t_j)] = \hat{S}_L(t_j)\sqrt{\sum_{i=1}^{j} \frac{1-\tilde{p}_i}{\tilde{p}_i n'_i}}$$

where $j = 1, ..., m-1$. The formula is best computed using software.

Now suppose the data is censored but the event or censoring times T_i are known with $Y_i^* = T_i = \min(Y_i, Z_i)$ where Y_i and Z_i are independent. Let $\delta_i = I(Y_i \le Z_i)$ so $\delta_i = 1$ if T_i is uncensored and $\delta_i = 0$ if T_i is censored. Let $t_{(1)} \le t_{(2)} \le \cdots \le t_{(n)}$ be the observed ordered survival times. Let $\gamma_j = 1$ if $t_{(j)}$ is uncensored and 0, otherwise. Let $t_0 = 0$ and let $0 < t_1 < t_2 < \cdots < t_m$ be the distinct survival times corresponding to the $t_{(j)}$ with $\gamma_j = 1$. Let $d_i =$ number of events (deaths) at time t_i. If $m = n$ and $d_i = 1$ for $i = 1, ..., n$ then there are **no ties**. If $m < n$ and some $d_i \ge 2$, then there are **ties**. Let $n_i = \sum_{j=1}^{n} I(t_{(j)} \ge t_i) = \#$ at risk at $t_i = \#$ alive and not yet censored just before t_i (and just after t_{i-1}).

Example 1.7. Suppose $n = 6$, $Y_i \sim EXP(1)$, $E(Y_i) = 1$, $Z_i \sim EXP(0.1)$, and $E(Z_i) = 10$. In the table below, Y_i and Z_i are not observed, $m = 5$, and the observed data is T_i and δ_i.

Y_i	0.2887	0.1796	1.1301	1.4165	0.2720	0.6667
Z_i	0.8967	1.6158	10.5266	1.0520	2.2329	4.2917
$T_i = Y_i^*$	0.2887	0.1796	1.1301	1.0520	0.2720	0.6667
δ_i	1	1	1	0	1	1
$t_{(j)}$	0.1796	0.2720	0.2887	0.6667	1.0522	1.1301
γ_j	1	1	1	1	0	1
t_i	0.1796	0.2720	0.2887	0.6667		1.1301

Consider intervals $I_1 = (0, t_1], I_2 = (t_1, t_2], ..., I_m = (t_{m-1}, t_m]$. Let n_k be the number at risk for interval I_k, $d_k =$ number of deaths in $I_k =$ number of deaths at t_k, and

$$\hat{p}_k = 1 - \frac{d_k}{n_k} = 1 - \frac{\text{number dying in I}_\text{k}}{\text{number with potential to die in I}_\text{k}} \approx \frac{S(t_k)}{S(t_{k-1})} \approx$$

P(survive in interval $(t_{k-1}, t_k)|$ alive at start of I_k). Then

$$\hat{S}_K(t_i) = \prod_{k=1}^{i} \hat{p}_k.$$

Note that individuals who die or are censored at time t_k are "at risk at t_k."

Definition 1.5. The **Kaplan Meier estimator = product limit estimator** of $S_Y(t_i) = P(Y > t_i)$ is $\hat{S}_K(0) = 1$ and

$$\hat{S}_K(t_i) = \prod_{k=1}^{i} (1 - \frac{d_k}{n_k}) = \hat{S}_K(t_{i-1})(1 - \frac{d_i}{n_i}).$$

$\hat{S}_K(t)$ is a step function with $\hat{S}_K(t) = \hat{S}_K(t_{i-1})$ for $t_{i-1} \leq t < t_i$ and $i = 1, ..., m$. If $t_{(n)}$ is uncensored then $t_m = t_{(n)}$ and $\hat{S}_K(t) = 0$ for $t > t_m$. If $t_{(n)}$ is censored, then $\hat{S}_K(t) = \hat{S}_K(t_m)$ for $t_m \leq t \leq t_{(n)}$, but $\hat{S}_K(t)$ is undefined for $t > t_{(n)}$.

Know how to compute and plot $\hat{S}_k(t_i)$ given the $t_{(j)}$ and γ_j or given the t_i, n_i and d_i. Use a table like the one below. Let $n_0 = n$. If f_{i-1} = number of events (deaths) and number censored in time interval $[t_{i-1}, t_i)$, then $n_i = n_{i-1} - f_{i-1}$ = number of $t_{(j)} \geq t_i$.

t_i	n_i	d_i	$\hat{S}_K(t)$
$t_0 = 0$			$\hat{S}_K(0) = 1$
t_1	n_1	d_1	$\hat{S}_K(t_1) = \hat{S}_K(t_0)[1 - \frac{d_1}{n_1}]$
t_2	n_2	d_2	$\hat{S}_K(t_2) = \hat{S}_K(t_1)[1 - \frac{d_2}{n_2}]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_j	n_j	d_j	$\hat{S}_K(t_j) = \hat{S}_K(t_{j-1})[1 - \frac{d_j}{n_j}]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_{m-1}	n_{m-1}	d_{m-1}	$\hat{S}_K(t_{m-1}) = \hat{S}_K(t_{m-2})[1 - \frac{d_{m-1}}{n_{m-1}}]$
t_m	n_m	d_m	$\hat{S}_K(t_m) = 0 = \hat{S}_K(t_{m-1})[1 - \frac{d_m}{n_m}]$

Example 1.8. Modifying Smith (2002, p. 113) slightly, suppose that the ordered censored survival times in days until repair of $n = 13$ street lights is 36, 38, 38, 38+, 78 112, 112, 114+, 162+, 189, 198, 237, 489+.

f_j	$t_{(j)}$	γ_j	t_i	n_i	d_i	$\hat{S}(t)$
						$\hat{S}(0) = 1$
1	36	1	36	13	1	$\hat{S}(36) = 0.9231$
3	38	1	38	12	2	$\hat{S}(38) = 0.7692$
	38	1				
	38	0				
1	78	1	78	9	1	$\hat{S}(78) = 0.6837$
4	112	1	112	8	2	$\hat{S}(112) = 0.5128$
	112	1				
	114	0				
	162	0				
1	189	1	189	4	1	$\hat{S}(189) = 0.3846$
1	198	1	198	3	1	$\hat{S}(198) = 0.2564$
1	237	1	237	2	1	$\hat{S}(36) = 0.1282$
	489	0				

Know how to find a 95% CI for $S_Y(t_i)$ based on $\hat{S}_K(t_i)$ using output: the 95% CI is $\hat{S}_K(t_i) \pm 1.96\ SE[\hat{S}_K(t_i)]$. The R output below gives $t_i, n_i, d_i, \hat{S}_K(t_i)$, $SE(\hat{S}_K(t_i))$ and the 95% CI for $S_Y(36)$ is $[0.7782, 1]$.

```
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   36     13       1    0.923  0.0739       0.7782        1.000
```

In general, a 95% CI for $S_Y(t_i)$ is $\hat{S}(t_i) \pm 1.96\ SE[\hat{S}(t_i)]$. If the lower endpoint of the CI is negative, round it up to 0. If the upper endpoint of the CI is greater than 1, round it down to 1. **Do not use impossible values of** $S_Y(t)$. Note that $\hat{S}_K(36) \pm 1.96 SE[\hat{S}_K(36)] = 0.923 \pm 1.96(0.0793) = 0.923 \pm 0.1448 = [0.7782, 1.0678]$. So use [0.7782, 1] since $S(y) \in [0, 1]$.

Let $P(Y \leq t(p)) = F_Y(t(p)) = p$ for $0 < p < 1$. Be able to get $t(p)$ and 95% CIs for $t(p)$ from SAS output for $p = 0.25, 0.5, 0.75$. For the output below, the CI for $t(0.75)$ is not given. The 95% CI for $t(0.50) \approx 210$ is [63,1296]. The 95% CI for $t(0.25) \approx 63$ is [18,195].

```
Quartile estimates
Percent point estimate lower upper
75          .           220.0  .
50        210.00        63.00 1296.00
25         63.00        18.00 195.00
```

R plots the KM survival estimator along with the pointwise 95% CIs for $S_Y(t)$. If we guess a distribution for Y, say $Y \sim W$, with a formula for $S_W(t)$, then the guessed $S_W(t_i)$ can be added to the plot. If roughly 95% of the $S_W(t_i)$ fall within the bands, then $Y \sim W$ may be reasonable. For example, if $W \sim EXP(1)$, use $S_W(t) = \exp(-t)$. If $W \sim EXP(\lambda)$, then $S_W(t) = \exp(-\lambda t)$. Recall that $E(W) = 1/\lambda$.

If $\lim_{t\to\infty} tS_Y(t) \to 0$, then $E(Y) = \int_0^\infty t f_Y(t)dt = \int_0^\infty S_Y(t)dt$. Hence an estimate of the mean $\hat{E}(Y)$ can be obtained from the area under $\hat{S}(t)$.

Greenwood's formula is

$$SE[\hat{S}_K(t_j)] = \hat{S}_K(t_j)\sqrt{\sum_{i=1}^{j} \frac{d_j}{n_j(n_j - d_j)}}$$

where $j = 1, ..., m-1$. The formula is best computed using software.

Definition 1.6. The **Nelson Aelen estimator** of $S_Y(t)$ is

$$\hat{S}_N(t_i) = \prod_{k=1}^{i} \exp\left(\frac{-d_k}{n_k}\right) = \exp\left(-\sum_{k=1}^{i} \frac{d_k}{n_k}\right) = \hat{S}_N(t_{i-1}) \exp\left(-\frac{d_i}{n_i}\right)$$

where $\hat{S}_N(0) = 1$ and t_i, d_i, and n_i are the same as for the Kaplan Meier estimator.

1.3 Estimating the (Cumulative) Hazard Function

Two important estimators of the cumulative hazard function use $\hat{H}(t_i) = -\log(\hat{S}(t_i))$.

Definition 1.7. The Kaplan Meier estimator of $H_Y(t)$ is $\hat{H}_K(0) = 0$,

$$\hat{H}_K(t_i) = -\log(\hat{S}_k(t_i) = -\sum_{k=1}^{i} \log\left(1 - \frac{d_k}{n_k}\right) = \hat{H}_K(t_{i-1}) - \log\left(1 - \frac{d_i}{n_i}\right).$$

Definition 1.8. The Nelson Aelen estimator of $H_Y(t)$ is $\hat{H}_N(0) = 0$,

$$\hat{H}_N(t_i) = \sum_{k=1}^{i} \frac{d_k}{n_k} = \hat{H}_K(t_{i-1}) + \frac{d_i}{n_i}.$$

Note that $\hat{S}_N(t_i) = \exp(-\hat{H}_n(t_i))$ and $\hat{H}_N(t_i) = -\log(\hat{S}_N(t_i))$.

A 95% CI for $H_Y(t_i)$ is $\hat{H}(t_i) \pm 1.96 SE[\hat{H}(t_i)] = [L, U]$. Use $[\max(0, L), U]$. Also,

$$SE[\hat{H}_N(t_i)] = \sqrt{\sum_{k=1}^{i} \frac{d_k}{n_k^2}} = \sqrt{SE[\hat{H}_N(t_{i-1})] + \frac{d_i}{n_i^2}}.$$

For the hazard function with $t_0 = 0$,

$$\hat{h}_K(t_i) = \hat{h}_N(t_i) = \frac{d_i}{n_i(t_{i+1} - t_i)}$$

for $i = 1, ..., m-1$.

Example 1.9.

t_i	n_i	d_i	$\hat{h}_K(t_i)$
10	18	1	$\frac{1}{18(19-10)} = 0.00617$
19	15	1	$\frac{1}{15(30-19)} = 0.00606$
30	13	1	

For the life table estimator with interval $I_j = [t_{j-1}, t_j)$, d_j, and n'_j,

$$\hat{h}_L(t) = \frac{d_j}{(n'_j - \frac{d_j}{2})(t_j - t_{j-1})}$$

$t_{j-1} \leq t < t_j$ with $\hat{h}_L(t)$ undefined for the last interval $[t_{m-1}, t_m)$. Sometime $t^* = (t_{j-1} + t_j)/2$ is used.

Example 1.10.

I_j	t^*	d_j	n'_j	$\hat{h}_L(t^*)$
$[0, 50)$	25	16	66.5	$\dfrac{16}{(66.5 - 16/2)(50 - 0)} = 0.00547$
$[50, 100)$	75	11	49	$\dfrac{11}{(49 - 11/2)(100 - 50)} = 0.005058$

Example 1.11. The data is from Klein and Moeschberger (2002, pp. 2, 86). There were 21 children with acute leukemia in complete or partial remission induced by the drug Prednisone, and the children were given the drug over a six month period. Note that $t_0 = 0$, $t_{(j)}$ = time until relapse, and $n_j = \sum_j t_{(j)} \geq t_i$. See the following table for computations. Using that table, a 95% CI for $H_Y(13)$ is $\hat{H}_N(13) \pm 1.96 SE[\hat{H}_n(13)] = 0.3517 \pm 1.96\sqrt{0.0217} = 0.3517 \pm 0.2888 = [0.0630, 0.6404]$.

$t_{(j)}$	γ_j	t_i	n_i	d_i	$\hat{H}_N(t_i)$	$(SE[\hat{H}_N(t_i)])^2$
		$t_0 = 0$			0	
6	1	6	21	3	$0 + 3/21 = 0.1428$	$0 + 3/21^2 = 0.0068$
6	1					
6	1					
6	0					
7	1	1	17	1	$0.1428 + 1/17 = 0.2017$	$0.0068 + 1/17^2 = 0.0103$
9	0					
10	1	10	15	1	$0.2017 + 1/15 = 0.2683$	$0.0103 + 1/15^2 = 0.0147$
10	0					
11	0					
13	1	13	12	1	$0.2683 + 1/12 = 0.3517$	$0.0147 + 1/12^2 = 0.0217$
16	1	16	11	1	$0.3517 + 1/11 = 0.4426$	$0.0217 + 1/11^2 = 0.0299$
17	0					
19	0					
20	0					
22	1	22	7	1	$0.4426 + 1/7 = 0.5854$	$0.0299 + 1/7^2 = 0.2243$
23	1	23	6	1	$0.5854 + 1/6 = 0.7521$	$0.2244 + 1/6^2 = 0.2795$
25	0					
32	0					
32	0					
34	0					
35	0					

1.4 Maximum Likelihood Estimation

Definition 1.9. Let $f(\boldsymbol{y}|\boldsymbol{\theta})$ be the pdf of a sample $\boldsymbol{Y}$ with parameter space Θ. If $\boldsymbol{Y} = \boldsymbol{y}$ is observed, then the **likelihood function** is $L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|\boldsymbol{y}) =$

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

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

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

$$L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|y_1, ..., y_n) = \prod_{i=1}^{n} g(y_i|\boldsymbol{\theta}). \tag{1.1}$$

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

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

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

Theorem 1.3: Invariance Principle. If $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, then $\tau(\hat{\boldsymbol{\theta}})$ is the MLE of $\tau(\boldsymbol{\theta})$ where τ is a function with domain Θ.

Really just need $\Theta \in dom(\tau)$ so $tau(\hat{\theta})$ is well defined: can't have $\log(-7.89)$ or $\sqrt{-1.57}$.

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

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

The one parameter case can often be solved by hand with the following technique. To show that $\hat{\theta}$ is the MLE of θ is equivalent to showing that $\hat{\theta}$ is the global maximizer of $\log L(\theta)$ on Θ where Θ is an interval with endpoints a and b, not necessarily finite. Suppose that $\log L(\theta)$ is continuous on Θ. Show that $\log L(\theta)$ is differentiable on (a, b). Then show that $\hat{\theta}$ is the unique

solution to the equation $\frac{d}{d\theta}\log L(\theta) = 0$ and that the 2nd derivative evaluated at $\hat{\theta}$ is negative: $\left.\dfrac{d^2}{d\theta^2}\log L(\theta)\right|_{\hat{\theta}} < 0$. See Remark 1.1V below.

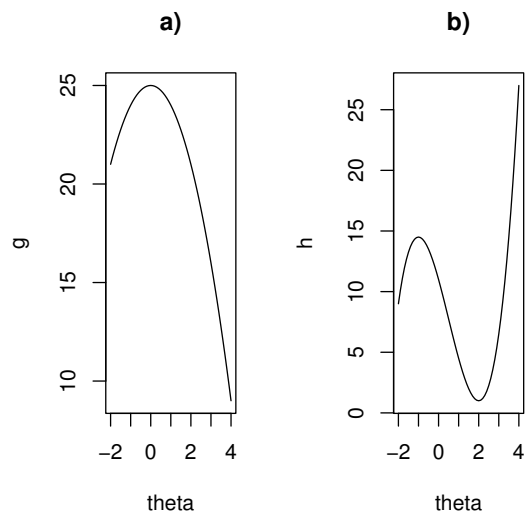

Fig. 1.1 The local max in a) is a global max, but the local max at $\theta = -1$ in b) is not the global max.

Remark 1.1. From calculus, recall the following facts. I) If the function g is continuous on an interval $[a, b]$ then both the max and min of g exist. Suppose that g is continuous on an interval $[a, b]$ and differentiable on (a, b). Solve $g'(\theta) \equiv 0$ and find the places where $g'(\theta)$ does not exist. These values are the **critical points**. Evaluate g at a, b, and the critical points. One of these values will be the min and one the max.

II) Assume g is continuous. Then g has a local max at the critical point θ_o if g is increasing for $\theta < \theta_o$ in a neighborhood of θ_o and if g is decreasing for $\theta > \theta_o$ in a neighborhood of θ_o (and θ_o is a global max if you can remove the phrase "in a neighborhood of θ_o"). The first derivative test is often used: if g is continuous at θ_o and if there exists some $\delta > 0$ such that $g'(\theta) > 0$ for all θ in $(\theta_o - \delta, \theta_o)$ and $g'(\theta) < 0$ for all θ in $(\theta_o, \theta_o + \delta)$, then g has a local max at θ_o.

III) If g is strictly concave ($\dfrac{d^2}{d\theta^2}g(\theta) < 0$ for all $\theta \in \Theta$), then any local max of g is a global max.

IV) Suppose $g'(\theta_o) = 0$. The 2nd derivative test states that if $\frac{d^2}{d\theta^2} g(\theta_o) < 0$, then g has a local max at θ_o.

V) If $g(\theta)$ is a continuous function on an interval with endpoints $a < b$ (not necessarily finite), differentiable on (a, b) and if the **critical point is unique**, then the critical point is a **global maximum** if it is a local maximum. To see this claim, note that if the critical point is not the global max then there would be a local minimum and the critical point would not be unique. Let $a = -2$ and $b = 4$. In Figure 1.1 a), the critical point for $g(\theta) = -\theta^2 + 25$ is at $\theta = 0$, is unique, and is both a local and global maximum. In Figure 1.1 b), $h(\theta) = \theta^3 - 1.5\theta^2 - 6\theta + 11$, the critical point $\theta = -1$ is not unique and is a local max but not a global max.

VI) If g is strictly convex ($\frac{d^2}{d\theta^2} g(\theta) > 0$ for all $\theta \in \Theta$), then any local min of g is a global min. If $g'(\theta_o) = 0$, then the 2nd derivative test states that if $\frac{d^2}{d\theta^2} g(\theta_o) > 0$, then θ_o is a local min.

VII) If $g(\theta)$ is a continuous function on an interval with endpoints $a < b$ (not necessarily finite), differentiable on (a, b) and if the **critical point is unique**, then the critical point is a **global minimum** if it is a local minimum. To see this claim, note that if the critical point is not the global min then there would be a local maximum and the critical point would not be unique.

Tips: a) $\exp(a) = e^a$ and $\log(y) = \ln(y) = \log_e(y)$ is the **natural logarithm**.
b) $\log(a^b) = b \log(a)$ and $\log(e^b) = b$.
c) $\log(\prod_{i=1}^n a_i) = \sum_{i=1}^n \log(a_i)$.
d) $\log L(\theta) = \log(\prod_{i=1}^n f(y_i|\theta)) = \sum_{i=1}^n \log(f(y_i|\theta))$.
e) If t is a differentiable function and $t(\theta) \neq 0$, then $\frac{d}{d\theta} \log(|t(\theta)|) = \frac{t'(\theta)}{t(\theta)}$ where $t'(\theta) = \frac{d}{d\theta} t(\theta)$. In particular, $\frac{d}{d\theta} \log(\theta) = 1/\theta$.
f) Any additive term that does not depend on θ is treated as a constant with respect to θ and hence has derivative 0 with respect to θ.

With censoring and truncation, the likelihood function changes. Often $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|y_1, ..., y_n) = \prod_{i=1}^n L(\boldsymbol{\theta}|y_i)$. Note that $1 - F(w) = S(w)$.
a) For iid individual data, $L(\boldsymbol{\theta}|y_i) = f(y_i)$ if Y has pdf $f(y)$.
b) For iid individual data, $L(\boldsymbol{\theta}|y_i) = p(x_i)$ if Y has pmf $p(y)$.
c) If it is only known that y_i is in some interval $(c_{j-1}, c_j]$, then $L(\boldsymbol{\theta}|y_i) = P(y_i \in (c_{j-1}, c_j]) = F(c_j) - F(c_{j-1})$.
The endpoints can be open or closed if Y is from a continuous distribution.
d) If y_i is right censored at u_i, then the interval is $[u_i, \infty)$, and $L(\boldsymbol{\theta}|y_i) = 1 - F(u_i)$.

e) For grouped data from the table below, $L(\boldsymbol{\theta}) = \prod_{j=1}^{m} [F(c_j) - F(c_{j-1})]^{n_j}$.

interval	number
$(c_0, c_1]$	n_1
$(c_1, c_2]$	n_2
$(c_2, c_3]$	n_3
$\vdots$	$\vdots$
$(c_{m-2}, c_{m-1}]$	n_{m-1}
$(c_{m-1}, c_m]$	n_m

f) If y_i is left truncated at d_i, then $L(\boldsymbol{\theta}|y_i) = \dfrac{f(y_i)}{1 - F(d_i)}$.

g) If y_i is left truncated at d_i and right censored at u_i, then $L(\boldsymbol{\theta}|y_i) = \dfrac{1 - F(u_i)}{1 - F(d_i)}$.

h) If the data are left truncated at d with $n - k$ uncensored cases y_i and k cases right censored at u, then $L(\boldsymbol{\theta}) = \dfrac{[\prod_{i=1}^{n-k} f(y_i)][1 - F(u)]^k}{[1 - F(d)]^n}$.

i) (**Rare**, the interval is $(0, d]$.): If y_i is censored below at d, $L(\boldsymbol{\theta}|y_i) = F(d)$.

j) (**Rare**): If y_i is truncated above at u, $L(\boldsymbol{\theta}|x_i) = \dfrac{f(y_i)}{F(u)}$.

Note that left truncated = truncated below = truncated, and right censored = censored above = censored are often used.

1.5 Simulations for KM Confidence Intervals

Section 1.2 described confidence intervals for the Kaplan Meier estimator. We will describe another CI, and two more CIs are easy to compute with R. Then we will simulate the four CIs.

The Agresti and Coull (1998) plus four 95% CI adds two successes (deaths) and two failures (survives) to the data set from a binomial distribution, and then computes the classical binomial 95% CI from the modified data set. For $t \in [t_1, t_m]$, Olive (2010, problem 16.45) modifies this procedure by adding two artificial deaths just before time t_1 and two artificial censored observations after the largest death time t_m. Then the classical 95% CI for the Kaplan Meier estimator is computed from the modified data set.

Hence

$$\tilde{S}_K(t_i) = \left(1 - \frac{1}{n+4}\right)\left(1 - \frac{1}{n+3}\right)\prod_{k=1}^{i}\left(1 - \frac{d_k}{n_k + 2}\right)$$

for $i = 1, ..., m$ where the first two terms are due to the two artificial deaths at the just before t_1 and $n_k + 2$ is used in the product due to the two artificial cases censored at time t_m. Also $[SE(\tilde{S}_K(t_i)]^2 =$

$$[\tilde{S}_K(t_i)]^2 \left(\sum_{k=1}^{i} \frac{d_k}{(n_k+2)(n_k+2-d_k)} + \frac{1}{(n+4)(n+4-1)} + \frac{1}{(n+3)(n+3-1)} \right)$$

for $i = 1, ..., m-1$.

If the CI is initially [L,U], then the CI $[\max(0, L), \min(1, U)]$ is used. In addition to the classical Kaplan Meier CI, there is a log CI that uses $log(\hat{S})$ and a log–log CI that uses $\log(-\log(\hat{S}))$ that are easy to compute with software.

Simulations were done in *R*. The function *kmsim2* simulates the classical, log, log–log, and plus four CIs for the Kaplan Meier estimator and is in the collection of *R* functions *survpack*. See Yang (2016) for a bigger simulation. The plus four CI worked well for $S(t_{(1)})$ and $S(t_{(n)})$.

The program *kmsim2* computes censored data $T = \min(Y, Z)$ where $Y \sim EXP(1)$. Then a 95% CI is made for $S_Y(t_{(j)})$ for each of the n $t_{(j)}$. This is done for runs=5000 data sets and the program computes the proportion of times the CI contains $S_Y(t_{(j)}) = \exp(-t_{(j)})$. The average scaled CI lengths (the average of $\sqrt{n}$ CI length) are also computed. The ccov is the proportion for the classical $\hat{S} \pm 1.96SE(\hat{S})$ interval while p4cov is for the plus 4 CI. The lcov is based on a CI that uses $\log(\hat{S})$ and llcov is based on a CI that uses $\log(-\log(\hat{S}))$. The three classical CIs are not made if the last case is censored so NA is given. The plus four CI seems to be good at $t_{(1)}$ and $t_{(n)}$. With 5000 runs, coverage between 0.94 and 0.96 would not give much evidence that the coverage is different from the nominal covarage of 0.95.

```
library(survival)
kmsim2(n=10,runs=5000)
$ccov
 [1] 0.8852 0.9604 0.9736 0.9720 0.9666 0.9544 0.9380
     0.9062 0.8404  NA

$lcov
 [1] 0.8772 0.9470 0.9564 0.9618 0.9632 0.9670 0.9702
     0.9800 0.9828  NA

$llcov
 [1] 0.7694 0.8886 0.9130 0.9222 0.9242 0.9230 0.9258
     0.9246 0.9208  NA

$p4cov
 [1] 0.9978 0.9082 0.9090 0.9132 0.9200 0.9236 0.9330
     0.9410 0.9550 0.9734
```

```
$clen
 [1] 0.8213907 1.3221304 1.7054981 1.8938355 1.9760212
     1.9803150 1.9032412 1.5986898 1.0969514       NA

$llen
 [1] 0.7698268 1.2214843 1.5940815 1.9111395 2.0769800
     2.1522128 2.1692379 2.1519330 2.2099754        NA

$lllen
 [1] 1.471560 1.679038 1.776042 1.826047 1.832765
     1.791306 1.692973 1.526673 1.264845       NA

$p4len
 [1] 1.327469 1.471418 1.569004 1.632534 1.665829
     1.669772 1.641687 1.578521 1.470567 1.189487
```

The above output is for $n = 10$ with 5000 runs. The table below summarizes the CI coverages and scaled lengths for t_1, t_3, t_{n-2}, and t_{n-1}.

Table 1.1 Simulated CI Coverages and Scaled Lengths

n	t_i	cov/len	clas	log	loglog	plus4
10	t_1	cov	0.885	0.877	0.769	0.998
		len	0.821	0.770	1.471	1.327
10	t_3	cov	0.974	0.956	0.913	0.909
		len	1.705	1.594	1.776	1.569
10	t_{n-2}	cov	0.906	0.980	0.925	0.941
		len	1.599	2.512	1.527	1.579
10	t_{n-1}	cov	0.840	0.983	0.921	0.955
		len	1.097	2.210	1.265	1.470

1.6 Summary

Let $Y \geq 0$ be a nonnegative random variable.

Then the **cumulative distribution function** (cdf) $F(t) = P(Y \leq t)$. Since $Y \geq 0$, $F(0) = 0, F(\infty) = 1$, and $F(t)$ is nondecreasing.

The probability density function (**pdf**) $f(t) = F'(t)$.

The **survival function** $S(t) = P(Y > t)$. $S(0) = 1, S(\infty) = 0$ and $S(t)$ is nonincreasing.

The **hazard function** $h(t) = \dfrac{f(t)}{1 - F(t)}$ for $t > 0$ and $F(t) < 1$. Note that $h(t) \geq 0$ if $F(t) < 1$.

The **cumulative hazard function** $H(t) = \int_0^t h(u)du$ for $t > 0$. It is true that $H(0) = 0, H(\infty) = \infty$, and $H(t)$ is nondecreasing.

1) Given one of $F(t), f(t), S(t), h(t)$ or $H(t)$, be able to find the other 4 quantities for $t > 0$.

A) $F(t) = \int_0^t f(u)du = 1 - S(t) = 1 - \exp[-H(t)] = 1 - \exp[-\int_0^t h(u)du]$.

B) $f(t) = F'(t) = -S'(t) = h(t)[1 - F(t)] = h(t)S(t) = h(t)\exp[-H(t)] = H'(t)\exp[-H(t)]$.

C) $S(t) = 1 - F(t) = 1 - \int_0^t f(u)du = \int_t^\infty f(u)du = \exp[-H(t)] = \exp[-\int_0^t h(u)du]$.

D)

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} = \frac{F'(t)}{1 - F(t)} = \frac{-S'(t)}{S(t)} = -\frac{d}{dt}\log[S(t)] = H'(t).$$

E) $H(t) = \int_0^t h(u)du = -\log[S(t)] = -\log[1 - F(t)]$.

Tip: if $F(t) = 1 - \exp[G(t)]$ for $t > 0$, then $H(t) = -G(t)$ and $S(t) = \exp[G(t)]$.

Tip: For $S(t) > 0$, note that $S(t) = \exp[\log(S(t))] = \exp[-H(t)]$. Finding $\exp[\log(S(t))]$ and setting $H(t) = -\log[S(t)]$is easier than integrating $h(t)$.

Know that if $Y \sim EXP(\lambda)$ where $\lambda > 0$, then $h(t) = \lambda$ for $t > 0$, $f(t) = \lambda e^{-\lambda t}$ for $t > 0$, $F(t) = 1 - e^{-\lambda t}$ for $t > 0$, $S(t) = e^{-\lambda t}$ for $t > 0$, $H(t) = \lambda t$ for $t > 0$ and $E(T) = 1/\lambda$. The **exponential distribution** can be a good model if failures are due to random shocks that follow a Poisson process, but constant hazard means that a used product is as good as a new product.

2) Suppose the observed survival times $T_1, ..., T_n$ are a censored data set from an exponential (λ) distribution. Let $T_i = Y_i^*$. Let $\delta_i = 0$ if the case is censored and let $\delta_i = 1$, otherwise. Let $r = \sum_{i=1}^n \delta_i$ = the number of uncensored cases. Then the MLE $\hat{\lambda} = r/\sum_{i=1}^n T_i$. So $\hat{\lambda} = r/\sum_{i=1}^n Y_i^*$. A 95% CI for λ is $\hat{\lambda} \pm 1.96\hat{\lambda}/\sqrt{r}$.

Know that if $Y \sim \text{Weibull}(\lambda, \gamma)$ where $\lambda > 0$ and $\gamma > 0$, then $h(t) = \lambda\gamma t^{\gamma-1}$ for $t > 0$, $f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^\gamma)$ for $t > 0$, $F(t) = 1 - \exp(-\lambda t^\gamma)$ for $t > 0$, $S(t) = \exp(-\lambda t^\gamma)$ for $t > 0$, $H(t) = \lambda t^\gamma$ for $t > 0$. The Weibull$(\lambda, \gamma = 1)$ distribution is the EXP(λ) distribution. The hazard function can be increasing, decreasing or constant. Hence the **Weibull distribution** often fits reliability data well, and the Weibull distribution is the most important distribution in reliability analysis.

3) Let $\hat{S}(t)$ be the estimated survival function. Let $t(p)$ be the pth percentile of Y: $P(Y \leq t(p)) = F(t(p)) = p$ so $1 - p = S(t(p)) = P(Y > t(p))$. Then $\hat{t}(p)$, the estimated time when 100 p % have died, can be estimated from a graph of $\hat{S}(t)$ with "over" and "down" lines. a) Find $1-p$ on the vertical axis and draw a horizontal "over" line to $\hat{S}(t)$. Draw a vertical "down" line until it intersects the horizontal axis at $\hat{t}(p)$. Usually want $p = 0.5$ but sometimes $p = 0.25$ and $p = 0.75$ are used.

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

If none of the survival times are censored, then the **empirical survival function** = (number of individual with survival times $> t$)/(number of individuals) $= a/n =$

$\hat{S}_E(t) = \dfrac{1}{n}\sum_{i=1}^{n} I(T_i > t) = \hat{p}_t =$ sample proportion of lifetimes $> t$.

Let $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ be the observed ordered survival times (= lifetimes = death times). Let $t_0 = 0$ and let $0 < t_1 < t_2 < \cdots < t_m$ be the distinct survival times. Let d_i = number of deaths at time t_i. If $m = n$ and $d_i = 1$ for $i = 1, ..., n$ then there are **no ties**. If $m < n$ and some $d_i \geq 2$, then there are **ties**.

$\hat{S}_E(t)$ is a step function with $\hat{S}_E(0) = 1$ and $\hat{S}_E(t) = \hat{S}_E(t_{i-1})$ for $t_{i-1} \leq t < t_i$. Note that $\sum_{i=1}^{m} d_i = n$.

4) Know how to compute and plot $\hat{S}_E(t)$ given the $t_{(i)}$ or given the t_i and d_i. Use a table like the one below. Let $a_0 = n$ and $a_i = \sum_{k=1}^{n} I(T_i > t_i) = \#$ of cases $t_{(j)} > t_i$ for $i = 1, ..., m$. Then $\hat{S}_E(t_i) = a_i/n = \sum_{k=1}^{n} I(T_i > t_i)/n = \hat{S}_E(t_{i-1}) - \frac{d_i}{n}$.

t_i	d_i	$\hat{S}_E(t_i) = \hat{S}_E(t_{i-1}) - \frac{d_i}{n}$
$t_0 = 0$		$\hat{S}_E(0) = 1 = \frac{n}{n} = \frac{a_0}{n}$
t_1	d_1	$\hat{S}_E(t_1) = \hat{S}_E(t_0) - \frac{d_1}{n} = \frac{a_0 - d_1}{n} = \frac{a_1}{n}$
t_2	d_2	$\hat{S}_E(t_2) = \hat{S}_E(t_1) - \frac{d_2}{n} = \frac{a_1 - d_2}{n} = \frac{a_2}{n}$
$\vdots$	$\vdots$	$\vdots$
t_j	d_j	$\hat{S}_E(t_j) = \hat{S}_E(t_{j-1}) - \frac{d_j}{n} = \frac{a_{j-1} - d_j}{n} = \frac{a_j}{n}$
$\vdots$	$\vdots$	$\vdots$
t_{m-1}	d_{m-1}	$\hat{S}_E(t_{m-1}) = \hat{S}_E(t_{m-2}) - \frac{d_{m-1}}{n} = \frac{a_{m-2} - d_{m-1}}{n} = \frac{a_{m-1}}{n}$
t_m	d_m	$\hat{S}_E(t_m) = 0 = \hat{S}_E(t_{m-1}) - \frac{d_m}{n} = \frac{a_{m-1} - d_m}{n} = \frac{a_m}{n}$

5) Let $t_1 \leq t < t_m$. Then the **classical large sample 95% CI** for $S(t_c)$ based on $\hat{S}_E(t)$ is

$$\hat{S}_E(t_c) \pm 1.96\sqrt{\frac{\hat{S}_E(t_c)[1 - \hat{S}_E(t_c)]}{n}} = \hat{S}_E(t_c) \pm 1.96 SE[\hat{S}_E(t_c)].$$

6) Let $0 < t$. Let

$$\tilde{p}_{t_c} = \frac{n\hat{S}_E(t_c) + 2}{n + 4}.$$

Then the **plus four 95% CI** for $S(t_c)$ based on $\hat{S}_E(t)$ is

$$\tilde{p}_{t_c} \pm 1.96\sqrt{\frac{\tilde{p}_{t_c}[1 - \tilde{p}_{t_c}]}{n + 4}} = \tilde{p}_{t_c} \pm 1.96 SE[\tilde{p}_{t_c}].$$

Let Y_i = time to event for ith person. $T_i = \min(Y_i, Z_i)$ where Z_i is the censoring time for the ith person (the time the ith person is lost to the study for any reason other than the time to event under study). The censored data is $y_1, y_2+, y_3, ..., y_{n-1}, y_n+$ where y_i means the time was uncensored and y_i+ means the time was censored. $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ are the ordered survival times (so if y_4+ is the smallest survival time, then $t_{(1)} = y_4+$). A status variable will be 1 if the time was uncensored and 0 if censored.

Let $[0, \infty) = I_1 \cup I_2 \cup \cdots \cup I_m = [t_0, t_1) \cup [t_1, t_2) \cdots \cup [t_{m-1}, t_m)$ where $t_o = 0$ and $t_m = \infty$. It is possible that the 1st interval will have left endpoint > 0 ($t_0 > 0$) and the last interval will have finite right endpoint ($t_m < \infty$). Suppose that the following quantities are known: $d_j = \#$ deaths in I_j,
$c_j = \#$ of censored survival times in I_j,

$n_j = \#$ at risk in $I_j = \#$ who were alive and not yet censored at the start of I_j (at time t_{j-1}).
Let $n'_j = n_j - \frac{c_j}{2}$ = average number at risk in I_j.

7) The **lifetable estimator** or actuarial method estimator of $S_Y(t)$ takes $\hat{S}_L(0) = 1$ and

$$\hat{S}_L(t_k) = \prod_{j=1}^{k} \frac{n'_j - d_j}{n'_j} = \prod_{j=1}^{k} \tilde{p}_j$$

for $k = 1, ..., m-1$. If $t_m = \infty$, $\hat{S}_L(t)$ is undefined for $t > t_{m-1}$. Suppose $t_m \neq \infty$. Then take $\hat{S}_L(t) = 0$ for $t \geq t_m$ if $c_m = 0$. If $c_m > 0$, then $\hat{S}_L(t)$ is undefined for $t \geq t_m$. **To graph** $\hat{S}_L(t)$, use linear interpolation (connect the dots). If $n'_j = 0$, take $\tilde{p}_j = 0$. Note that

$$\hat{S}_L(t_k) = \hat{S}_L(t_{k-1}) \frac{n'_k - d_k}{n'_k} \quad \text{for} \quad \mathrm{k} = 1, ..., \mathrm{m} - 1.$$

8) Know how to get the lifetable estimator and $SE(\hat{S}_L(t_i))$ from output.

```
(left output)                          (right output)
interval survival survival SE or interval survival survival SE
0    50   1.00        0           0    50   0.7594   0.0524
50  100   0.7594    0.0524        50   100  0.5889   0.0608
100 200   0.5889    0.0608       100   200  0.5253   0.0602
```

Since $\hat{S}_L(0) = 1$, $\hat{S}_L(t)$ is for the left endpoint for the "left output," and for the right endpoint for the "right output." For both cases, $\hat{S}_L(50) = 0.7594$ and $SE(\hat{S}_L(50)) = 0.0524$.

9) A 95% CI for $S_Y(t_i)$ based on the lifetable estimator is

$$\hat{S}_L(t_i) \pm 1.96\ SE[\hat{S}_L(t_i)].$$

10) Know how to compute $\hat{S}_L(t)$ with a table like the one below. The first 4 columns need to be given but the last 3 columns may need to be filled in. On an exam you may be given a table with all but a few entries filled.

I_j, d_j, c_j, n_j	n'_j	$\frac{n'_j-d_j}{n'_j}$	$\hat{S}_L(t)$
$[t_0=0,t_1)$, d_1, c_1, n_1	$n_1-\frac{c_1}{2}$	$\frac{n'_1-d_1}{n'_1}$	$\hat{S}_L(t_o)=\hat{S}_L(0)=1$
$[t_1,t_2)$, d_2, c_2, n_2	$n_2-\frac{c_2}{2}$	$\frac{n'_2-d_2}{n'_2}$	$\hat{S}_L(t_1)=\hat{S}_L(t_0)\frac{n'_1-d_1}{n'_1}$
$[t_2,t_3)$, d_3, c_3, n_3	$n_3-\frac{c_3}{2}$	$\frac{n'_3-d_3}{n'_3}$	$\hat{S}_L(t_2)=\hat{S}_L(t_1)\frac{n'_2-d_2}{n'_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[t_{k-1},t_k)$, d_k, c_k, n_k	$n_k-\frac{c_k}{2}$	$\frac{n'_k-d_k}{n'_k}$	$\hat{S}_L(t_{k-1})=$ $\hat{S}_L(t_{k-2})\frac{n'_{k-1}-d_{k-1}}{n'_{k-1}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[t_{m-2},t_{m-1})$, d_{m-1}, c_{m-1}, n_{m-1}	$n_{m-1}-\frac{c_{m-1}}{2}$	$\frac{n'_{m-1}-d_{m-1}}{n'_{m-1}}$	$\hat{S}_L(t_{m-2})=$ $\hat{S}_L(t_{m-3})\frac{n'_{m-2}-d_{m-2}}{n'_{m-2}}$
$[t_{m-1},t_m=\infty)$, d_m, c_m, n_m	$n_m-\frac{c_m}{2}$	$\frac{n'_m-d_m}{n'_m}$	$\hat{S}_L(t_{m-1})=$ $\hat{S}_L(t_{m-2})\frac{n'_{m-1}-d_{m-1}}{n'_{m-1}}$

11) Also get a 95% CI from output like that below. So the 95% CI for S(50) is (0.65666,0.86213).

```
time survival SDF_LCL SDF_UCL
0     1.0     1.0     1.0
50    0.7594  0.65666 0.86213
```

Let $Y_i^* = T_i = \min(Y_i, Z_i)$ where Y_i and Z_i are independent. Let $\delta_i = I(Y_i \le Z_i)$ so $\delta_i = 1$ if T_i is uncensored and $\delta_i = 0$ if T_i is censored. Let $t_{(1)} \le t_{(2)} \le \cdots \le t_{(n)}$ be the observed ordered survival times. Let $\gamma_j = 1$ if $t_{(j)}$ is uncensored and 0, otherwise. Let $t_0 = 0$ and let $0 < t_1 < t_2 < \cdots < t_m$ be the distinct survival times corresponding to the $t_{(j)}$ with $\gamma_j = 1$. Let $d_i =$ number of deaths at time t_i. If $m = n$ and $d_i = 1$ for $i = 1, ..., n$ then there are **no ties**. If $m < n$ and some $d_i \ge 2$, then there are **ties**.

12) Let $n_i = \sum_{j=1}^n I(t_{(j)} \ge t_i) = \#$ at risk at $t_i = \#$ alive and not yet censored just before t_i. Let $d_i = \#$ of events (deaths) at t_i. The **Kaplan Meier estimator = product limit estimator** of $S_Y(t_i) = P(Y > t_i)$ is $\hat{S}_K(0) = 1$ and $\hat{S}_K(t_i) = \prod_{k=1}^i (1 - \frac{d_k}{n_k}) = \hat{S}_K(t_{i-1})(1 - \frac{d_i}{n_i})$. $\hat{S}_K(t)$ is a step function with $\hat{S}_K(t) = \hat{S}_K(t_{i-1})$ for $t_{i-1} \le t < t_i$ and $i = 1, ..., m$. If $t_{(n)}$ is uncensored then $t_m = t_{(n)}$ and $\hat{S}_K(t) = 0$ for $t > t_m$. If $t_{(n)}$ is censored, then $\hat{S}_K(t) = \hat{S}_K(t_m)$ for $t_m \le t \le t_{(n)}$, but $\hat{S}_K(t)$ is undefined for $t > t_{(n)}$.

13) Know how to compute and plot $\hat{S}_k(t_i)$ given the $t_{(j)}$ and γ_j or given the t_i, n_i and d_i. Use a table like the one below.

t_i	n_i	d_i	$\hat{S}_K(t)$
$t_0 = 0$			$\hat{S}_K(0) = 1$
t_1	n_1	d_1	$\hat{S}_K(t_1) = \hat{S}_K(t_0)[1 - \frac{d_1}{n_1}]$
t_2	n_2	d_2	$\hat{S}_K(t_2) = \hat{S}_K(t_1)[1 - \frac{d_2}{n_2}]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_j	n_j	d_j	$\hat{S}_K(t_j) = \hat{S}_K(t_{j-1})[1 - \frac{d_j}{n_j}]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_{m-1}	n_{m-1}	d_{m-1}	$\hat{S}_K(t_{m-1}) = \hat{S}_K(t_{m-2})[1 - \frac{d_{m-1}}{n_{m-1}}]$
t_m	n_m	d_m	$\hat{S}_K(t_m) = 0 = \hat{S}_K(t_{m-1})[1 - \frac{d_m}{n_m}]$

14) Know how to find a 95% CI for $S_Y(t_i)$ based on $\hat{S}_K(t_i)$ using output: the 95% CI is $\hat{S}_K(t_i) \pm 1.96\ SE[\hat{S}_K(t_i)]$. The R output below gives $t_i, n_i, d_i, \hat{S}_K(t_i), SE(\hat{S}_K(t_i))$ and the 95% CI for $S_Y(36)$ is $(0.7782, 1)$.

```
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   36     13       1    0.923  0.0739       0.7782        1.000
```

15) In general, a 95% CI for $S_Y(t_i)$ is $\hat{S}(t_i) \pm 1.96\ SE[\hat{S}(t_i)]$. If the lower endpoint of the CI is negative, round it up to 0. If the upper endpoint of the CI is greater than 1, round it down to 1. **Do not use impossible values of** $S_Y(t)$.

16) Let $P(Y \leq t(p)) = p$ for $0 < p < 1$. Be able to get $t(p)$ and 95% CIs for $t(p)$ from SAS output for $p = 0.25, 0.5, 0.75$. For the output below, the CI for $t(0.75)$ is not given. The 95% CI for $t(0.50) \approx 210$ is (63,1296). The 95% CI for $t(0.25) \approx 63$ is (18,195).

```
Quartile estimates
Percent point estimate lower upper
75          .           220.0  .
50        210.00        63.00 1296.00
25         63.00        18.00 195.00
```

17) R plots the KM survival estimator along with the pointwise 95% CIs for $S_Y(t)$. If we guess a distribution for Y, say $Y \sim W$, with a formula for $S_W(t)$, then the guessed $S_W(t_i)$ can be added to the plot. If roughly 95% of the $S_W(t_i)$ fall within the bands, then $Y \sim W$ may be reasonable. For example, if $W \sim EXP(1)$, use $S_W(t) = \exp(-t)$. If $W \sim EXP(\lambda)$, then $S_W(t) = \exp(-\lambda t)$. Recall that $E(W) = 1/\lambda$.

18) If $\lim_{t\to\infty} tS_Y(t) \to 0$, then $E(Y) = \int_0^\infty t f_Y(t)dt = \int_0^\infty S_Y(t)dt$. Hence an estimate of the mean $\hat{E}(Y)$ can be obtained from the area under $\hat{S}(t)$.

19) Let $\boldsymbol{Y} = (Y_1, ..., Y_n)$. If $\boldsymbol{y} = (y_1, ..., y_n)$ is the data then the **likelihood function** $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\boldsymbol{y})$. For each sample point $\boldsymbol{y} = (y_1, ..., y_n)$, let $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ be a parameter value at which $L(\boldsymbol{\theta}|\boldsymbol{y})$ attains its maximum as a function of $\boldsymbol{\theta}$ with $\boldsymbol{y}$ held fixed. Then a maximum likelihood estimator (**MLE**) of the parameter $\boldsymbol{\theta}$ based on the sample $\boldsymbol{Y}$ is $\hat{\boldsymbol{\theta}}(\boldsymbol{Y})$. Note: it is crucial to observe that the likelihood function is a function of $\boldsymbol{\theta}$ (and that $y_1, ..., y_n$ act as fixed constants). Often $\boldsymbol{\theta} = \theta$ is a scalar.

20) If the MLE $\hat{\boldsymbol{\theta}}$ exists, then $\hat{\boldsymbol{\theta}} \in \Theta$. If the MLE $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, ..., \hat{\theta}_k)$, then the MLE of θ_i is $\hat{\theta}_i$, the MLE of (θ_1, θ_5) is $(\hat{\theta}_1, \hat{\theta}_5)$, etc.

21) **Invariance Principle:** If $\hat{\boldsymbol{\theta}}$ is the MLE of θ, then $\tau(\hat{\boldsymbol{\theta}})$ is the MLE of $\tau(\boldsymbol{\theta})$. Here τ is a function of $\boldsymbol{\theta}$ with domain Θ.

22) For **individual data**, $Y_1, ..., Y_n$ are iid, usually with pdf $f(y)$ or pmf $p(y)$. Let $\boldsymbol{y} = (y_1, ..., y_n)$ be the observed data. Then the **likelihood function** $L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^n g(y_i)$ where $g(y)$ is $f(y)$ or $p(y)$. The **log likelihood function** $\log(L(\boldsymbol{\theta})) = \sum_{i=1}^n \log(g(x_i))$. Usually use 22) to find the MLE.

23) For this class, assume that the maximum likelihood estimator (MLE) is a solution to $\dfrac{\partial}{\partial\theta_i} \log L(\boldsymbol{\theta}) \stackrel{set}{=} 0$ for $i = 1, ..., k$ where usually $k = 1$ or 2. (We will not use second derivatives to show that the MLE was the global max.)

Tips: a) $\exp(a) = e^a$. b) $\log(a^b) = b\log(a)$ and $\log(e^b) = b$. c) $\log(\prod_{i=1}^n a_i) = \sum_{i=1}^n \log(a_i)$.
d) Often $\log[L(\theta)] = \log(\prod_{i=1}^n f(x_i|\theta)) = \sum_{i=1}^n \log(f(x_i|\theta))$. e) If t is a differentiable function and $t(\theta) \neq 0$, then $\frac{d}{d\theta} \ln(|t(\theta)|) = \frac{t'(\theta)}{t(\theta)}$ where $t'(\theta) = \frac{d}{d\theta} t(\theta)$. In particular, $\frac{d}{d\theta} \ln(\theta) = 1/\theta$. f) Anything that does not depend on θ is treated as a constant with respect to θ and hence has derivative 0 with respect to θ.

24) For small n, if given $\boldsymbol{y}$ it can be easier to plug in the y_i to find the MLE. Sometimes you will solve for the MLE as a statistic, then plug $\boldsymbol{x}$ into the statistic.

25) Let $g(\boldsymbol{x}|\boldsymbol{\theta})$ be the pmf or pdf of a sample $\boldsymbol{Y}$. If $\boldsymbol{Y} = \boldsymbol{y}$ is observed, then **the likelihood function** $L(\boldsymbol{\theta}) = g(\boldsymbol{y}|\boldsymbol{\theta})$.

26) Often $L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|y_1, ..., y_n) = \prod_{i=1}^n L(\boldsymbol{\theta}|y_i)$. Note that $1 - F(w) = S(w)$.

a) For iid individual data, $L(\boldsymbol{\theta}|y_i) = f(y_i)$ if Y has pdf $f(y)$.

b) For iid individual data, $L(\boldsymbol{\theta}|x_i) = p(y_i)$ if Y has pmf $p(y)$.

c) If it is only known that y_i is in some interval $(c_{j-1}, c_j]$, then
$L(\boldsymbol{\theta}|y_i) = P(y_i \in (c_{j-1}, c_j]) = F(c_j) - F(c_{j-1})$.
The endpoints can be open or closed if Y is from a continuous distribution.

d) If y_i is right censored at u_i, then the interval is $[u_i, \infty)$, and $L(\boldsymbol{\theta}|y_i) = 1 - F(u_i)$.

e) For grouped data from the table below, $L(\boldsymbol{\theta}) = \prod_{j=1}^{m} [F(c_j) - F(c_{j-1})]^{n_j}$.

interval	number
$(c_0, c_1]$	n_1
$(c_1, c_2]$	n_2
$(c_2, c_3]$	n_3
$\vdots$	$\vdots$
$(c_{m-2}, c_{m-1}]$	n_{m-1}
$(c_{m-1}, c_m]$	n_m

f) If y_i is left truncated at d_i, then $L(\boldsymbol{\theta}|y_i) = \dfrac{f(y_i)}{1 - F(d_i)}$.

g) If y_i is left truncated at d_i and right censored at u_i, then $L(\boldsymbol{\theta}|y_i) = \dfrac{1 - F(u_i)}{1 - F(d_i)}$.

h) If the data are left truncated at d with $n - k$ uncensored cases y_i and k cases right censored at u, then $L(\boldsymbol{\theta}) = \dfrac{[\prod_{i=1}^{n-k} f(y_i)][1 - F(u)]^k}{[1 - F(d)]^n}$.

i) (**Rare**, the interval is $(0, d]$.): If y_i is censored below at d, $L(\boldsymbol{\theta}|y_i) = F(d)$.

j) (**Rare**): If y_i is truncated above at u, $L(\boldsymbol{\theta}|y_i) = \dfrac{f(y_i)}{F(u)}$.

Note that left truncated = truncated below = truncated, and right censored = censored above = censored are often used.

1.7 Complements

For some of the MLE rules, see Klugman et al. (2008) and Kellison and London (2011). Olive (2014, pp. 145-147) gives a correct proof of the invariance principle (most "proofs" in the literature are not valid).

Important papers include Aalen (1978), Kaplan and Meier (1958), Nelson (1969, 1972). For Greenwood's formula, see Kaplan and Meier (1958). Advanced works use theory from counting processes and martingales.

1.8 Problems

Problems with an asterisk * are especially important.

1.1. Suppose $H(t) = \dfrac{\lambda}{\theta}[e^{\theta t} - 1]$ for $t > 0$ where $\lambda > 0$ and $\theta > 0$. Find $a)h(t)$, b)$S(t)$, c) $F(t)$ and d) $f(t)$ for $t > 0$.

1.2. Suppose $T \sim \text{EXP}(\lambda)$. Show $P(T > t + s|T > s) = P(T > t)$ for any $t > 0$ and $s > 0$. This property is known as the memoryless property and implies that the future survival of the product does not depend on the past if the lifetime T of the product is exponential.

1.3. Suppose $F(t) = 1 - \exp[-at - (bt)^2]$ where $a > 0$, $b > 0$ and $t > 0$. Find $a)S(t)$, b)$f(t)$, c) $h(t)$ and d) $H(t)$ for $t > 0$.

1.4. Suppose $F(t) = 1 - \exp[-at - (ct)^3]$ where $a > 0$, $c > 0$ and $t > 0$. Find the following quantities for $t > 0$.

a) $S(t)$

b)$f(t)$

c) $h(t)$

d) $H(t)$

1.5. Suppose $H(t) = \alpha + \beta t^2$ for $t > 0$ where $\alpha > 0$ and $\beta > 0$.

a) Find $h(t)$.

b) Find $S(t)$.

c) Find $F(t)$.

1.6. Suppose

$$F(t) = 1 - \exp\left(\frac{-t^2}{2\sigma^2}\right)$$

where $\sigma > 0$ and $t > 0$. Find the following quantities for $t > 0$.

a) $S(t)$

b)$f(t)$

c) $h(t)$

d) $H(t)$

1.7. Eleven death times from Collett (2003b, p. 16) are given below. The patients had malignant bone tumours.

```
11  13  13  13  13  13  14  14  15  15  17
```

a) Following Example 1.3, make a table with headers $t_{(j)}, t_i, d_i, \hat{S}_E(t) = \sum(T_i > t)/n$.

b) Plot $\hat{S}_E(t)$.

c) Find the 95% classical CI for $S(13)$ based on $\hat{S}_E(t)$.

d) Find the 95% plus four CI for $S(13)$ based on $\hat{S}_E(t)$.

1.8. Find the 95% classical CI for $S_Y(32)$ if $n = 9$ and $\hat{S}_E(32) = 4/9$.

1.9. Find the 95% plus four CI for $S_Y(32)$ if $n = 9$ and $\hat{S}_E(32) = 4/9$.

1.10. Find the 95% plus four CI for $S_Y(32)$ if $n = 9$ and $\hat{S}_E(32) = 6/9$.

1.11. Find the 95% classical CI for $S_Y(32)$ if $n = 9$ and $\hat{S}_E(32) = 6/9$.

1.12. Survival times for nine electrical components are given below.
8, 8, 23, 32, 32, 46, 57, 88, 109
Compute the empirical survival function $\hat{S}_E(t_i)$ by filling in the table below. Then plot the function.

$t_{(j)}$	t_i	d_i	$\hat{S}_E(t)$
	$t_0 = 0$		$\hat{S}_E(0) = 1 = \frac{9}{9}$
8			
8	8	2	$\hat{S}_E(8) =$
23			$\hat{S}_E(23) =$
32			
32			$\hat{S}_E(32) =$
46			$\hat{S}_E(46) =$
57			$\hat{S}_E(57) =$
88			$\hat{S}_E(88) =$
109			$\hat{S}_E(109) =$

1.13. The Klein and Moeschberger (1997, p. 141-142) data set consists of information from 927 1st born children to mothers who chose to breast feed their child. The event was time in weeks until weaned (instead of death). Complete the following table used to produce the lifetable estimator (on a separate sheet of paper).

I_j	d_j	c_j	n_j	n'_j	$\frac{n'_j - d_j}{n'_j}$	$\hat{S}_L(t)$
$[0,2)$	77	2	927	926	0.9168	1.0000
$[2,3)$	71	3	848	846.5	0.9161	0.9168
$[3,5)$	119	6	774	771	0.8457	0.8399
$[5,7)$	75	9	649	644.5	0.8836	0.7103
$[7,11)$	109	7	565	561.5	0.8059	0.6276
$[11,17)$	148	5	449	446.5	0.6685	0.5058
$[17,25)$	107	3	296			0.3381
$[25,37)$	74	0	186			
$[37,53)$	85	0	112			
$[53,\infty)$	27	0	27			

```
time n.risk n.event survival std.err lower 95% CI upper 95% CI
   9     11       1    0.909  0.0867       0.7392        1.000
  13     10       1    0.818  0.1163       0.5903        1.000
  18      8       1    0.716  0.1397       0.4422        0.990
  23      7       1    0.614  0.1526       0.3145        0.913
  31      5       1    0.491  0.1642       0.1691        0.813
  34      4       1    0.368  0.1627       0.0494        0.687
  48      2       1    0.184  0.1535       0.0000        0.485
```

1.14. The length of times of remission (time until relapse) in acute myelogeneous leukemia under maintenance chemotherapy for 11 patients is $9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+$. See Miller (1981, p. 49). From the output above what is the 95% CI for $S_Y(34)$?

1.15. The Lindsey (2004, p. 280) data set is for survival times for 110 women with stage 1 cervical cancer studied over a 10 year period. Use the life table estimator to compute the estimated survival function $\hat{S}_L(t_i)$ by filling in the table below. Then plot the function.

I_j	d_j	c_j	n_j	n'_j	$\frac{n'_j - d_j}{n'_j}$	$\hat{S}_L(t)$
$[0,1)$	5	5	110	107.5	0.9535	1.0000
$[1,2)$	7	7	100	96.5	0.9275	0.9535
$[2,3)$	7	7	86	82.5	0.9152	0.8843
$[3,4)$	3	8	72	68	0.9559	0.8093
$[4,5)$	0	7	61	57.5	1.0	0.7736
$[5,6)$	2	10	54	49	0.9591	0.7736
$[6,7)$	3	6	42	39	0.9230	0.7420
$[7,8)$	0	5	33			
$[8,9)$	0	4	28			
$[9,10)$	1	8	24			
$[10,\infty)$	15	0	15			

1.16. Survival times for 13 women with tumors from breast cancer that were negatively stained with HPA are given below.
23, 47, 69, 70+, 71+, 100+, 101+, 148, 181, 198+, 208+, 212+, 224+
See Collett (2003b, p. 6). Compute the Kaplan Meier survival function $\hat{S}_K(t_i)$ by filling in the table below. Then plot the function.

$t_{(j)}$	γ_j	t_i	n_i	d_i	$\hat{S}_K(t)$
		$t_0 = 0$			$\hat{S}_K(0) = 1$
23	1	23	13	1	$\hat{S}_K(23) =$
47	1	47			$\hat{S}_K(47) =$
69	1	69			$\hat{S}_K(69) =$
70	0				
71	0				
100	0				
101	0				
148	1	148			$\hat{S}_K(148) =$
181	1	181			$\hat{S}_K(181) =$
198	0				
208	0				
212	0				
224	0				

1.17. The Lindsey (2004, p. 280) data is for survival times for 234 women with stage 2 cervical cancer studied over a 10 year period. Use the life table estimator to compute the estimated survival function $\hat{S}_L(t_i)$ by filling in the table below. Show what you multiply to find $\hat{S}_L(t_i)$. Then plot the function.

I_j	d_j	c_j	n_j	n'_j	$\frac{n'_j - d_j}{n'_j}$	$\hat{S}_L(t)$
$[0,1)$	24	3	234	232.5	0.8968	1.0000
$[1,2)$	27	11	207	201.5	0.8660	0.8968
$[2,3)$	31	9	169	164.5	0.8116	0.7766
$[3,4)$	17	7	129	125.5	0.8645	0.6302
$[4,5)$	7	13	105	98.5	0.9289	0.5448
$[5,6)$	6	6	85	82	0.9268	0.5061
$[6,7)$	5	6	73	70	0.9286	0.4691
$[7,8)$	3	10	62			
$[8,9)$	2	13	49			
$[9,10)$	4	6	34			
$[10,\infty)$	24	0	24			

1.18. Times (in weeks) until relapse below are for 12 patients with acute myelogeneous leukemia who reached a state of remission after chemotherapy. See Miller (1981, p. 49).
5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45
Compute the Kaplan Meier survival function $\hat{S}_K(t_i)$ by filling in the table below. Show what you multiply to find $\hat{S}_k(t_i)$. Then plot the function.

$t_{(j)}$	γ_j	t_i	n_i	d_i	$\hat{S}_K(t)$
		$t_0 = 0$			$\hat{S}_K(0) = 1$
5	1	5	12	2	$\hat{S}_K(5) =$
5	1				
8	1	8			$\hat{S}_K(8) =$
8	1				
12	1	12			$\hat{S}_K(12) =$
16	0				
23	1	23			$\hat{S}_K(23) =$
27	1	27			$\hat{S}_K(27) =$
30	1	30			$\hat{S}_K(30) =$
33	1	33			$\hat{S}_K(33) =$
43	1	43			$\hat{S}_K(43) =$
45	1	45			$\hat{S}_K(45) =$

1.19. Suppose the random variable Y has probability density function (pdf) $f(y)$ where $f(y) = 0$ for $y < 0$ and the expected value $E(Y)$ exists. One way to get a new pdf $g(y)$ is to use

$$g(y) = \frac{yf(y)}{E(Y)}.$$

See Cox (1962, p. 65). Show $\int_0^\infty g(y)dy = 1$.

SAS Problems

SAS is a statistical software package that will be used in this course. You will need a disk. There are SAS manuals and books at the library, but they are not needed in this course. To use SAS on windows (PC), use the following steps.

i) Click the lower left icon to see programs in the icons Window. You can click on the desktop icon to escape. If your computer does not have SAS, go to another computer. If you click on something and can't get out of the information window, there is a Windows key that looks like 4 rectangles and is on the lower left of the keyboard near the Ctrl key. This Windows key can get you back to icons Windows.

ii) Use the homework link or (http://parker.ad.siu.edu/Olive/survhw.txt) to copy and paste the program for Problem 1.20 into *SAS*. Highlight the program for problem 1.20. Hit Ctrl–c. Click the lower left icon to see programs. Double click the `SAS 9.4` icon. The editor window is the lower window. Click on that window,then hit Ctrl–v to paste in the program. Then run>submit. Output will appear in a few minutes.

(You can copy and paste the program from (http://parker.ad.siu.edu/Olive/M473hw.txt) or (http://parker.ad.siu.edu/Olive/regsas.txt) problem 16.36. The *ls* stands for linesize so *l* is a lowercase *L*, not the number one.)

If you were not successful, look at the *log window* for hints on errors. A single typo can cause failure. Reopen your file in *Word* or *Notepad* and make corrections. Occasionally you can not find your error. Then find your instructor or wait a few hours and reenter the program. *Word* seems to make better looking tables, and copying from *Notepad* to *Word* can completely ruin the table.

iii) To copy and paste relevant output into *Word,* click on the output window and use the top menu commands "Edit>Select All" and then the menu commands "Edit>Copy".

(In *Notepad* use the commands "Edit>Paste". Then use the mouse to highlight the relevant output (**the table and statistics for the table**). Then use the commands "Edit>Copy".)

Finally, in *Word*, use the commands "Edit>Paste".

iv) This point explains the SAS commands. The semicolon ";" is used to end SAS commands and the "options ls = 70;" command makes the output readable. (An "*" can be used to insert comments into the SAS program. Try putting an * before the options command and see what it does to the output.) The next step is to get the data into SAS. The command "data heart;" gives the name "heart" to the data set. The command "input time status number;" says the first entry is the censored variable time, the 2nd variable status (0 if censored 1 if uncensored) and the third variable number (= number of deaths or number of cases censored, depending on status). The command "cards;" means that the data is entered below. Then the data in entered and the isolated semicolon indicates that the last case has been entered. The next 4 lines make perform the lifetable estimates for S(t) and

the corresponding confidence intervals. Also plots of the estimated survival and hazard functions are given. The command "run;" tells SAS to execute the program.

You may want to save your SAS output as the file **hw1d20.doc** It may be easier to save output from each problem as a *Word* document, but you may get an extra page printed whenever you use the printer.

1.20. The following problem gets the lifetable estimator using SAS. The data is on 68 patients that received heart transplants at about the time when getting a heart transplant was new. The following problem gets the lifetable estimator using SAS. See Allison (1995, p. 49-50).

a) Do i) through iii) above, and look at iv).

b) From the 1st page of output, *Number Failed* $= d_i$, *Number Censored* $= c_i$, *Effective Sample Size* $= n'_i$, *Survival* $= \hat{S}_L(t_{i-1}) =$ estimated survival for the left endpoint of the interval and *Survival Standard Error* $= SE[\hat{S}_L(t_{i-1})]$.

What is $SE[\hat{S}_L(200)]$?

c) From the 2nd page of output, *SDF_LCL SDF_UCL* gives a 95% CI for $S(t_{i-1})$.

What is the 95% CI for $S(200)$ using output?

d) Compute the 95% CI for $S(200)$ using the formula and $SE[\hat{S}_L(200)]$.

e) The SAS program (with plots(s,h)) plots both the survival and the hazard function (scroll down!). From the 2nd page of output, plot MIDPOINT vs HAZARD (so the first point is (25,0.0055)) **by hand**. Connect the dots to make an estimated hazard function. Notice that the estimated hazard function decreases sharply to about 200 days after surgery and then is fairly stable.

1.21. This problem examines the Allison (1995, p. 31-34) myelomatosis data (a cancer causing tumors in the bone marrow) with SAS using the Kaplan Meier product limit estimator. Obtain the SAS program for this problem from (http://parker.ad.siu.edu/Olive/survhw.txt). Obtain the output from the program in the same manner as i) through iv) above Problem 1.20.

a) The output should be roughly 3 pages and a graph. Include this output in *Word*.

b) From the summary statistics of the first page of output, about when do 50% of the patients die?

c) From the first page of output (perhaps), what is the 95% CI for the time when 50% of the patients die?

d) From the 3rd page of output (perhaps), what is the 95% CI for $S_Y(13)$. This is the log log transformed CI, so will differ from the CI in e).

e) Make the CI using $\hat{S}_K(13)$ and $SE(\hat{S}_K(13))$ obtained from the 1st page of output (perhaps). If the interval is (L, U), use $[\max(0, L), \min(U, 1)]$ as the final interval.

f) From the plot of $\hat{S}_K(t)$ for the KM estimator, briefly explain survival for days 0–250 and for days 250–2250.

1.22. This Miller (1981, p. 49-50) data set is on remission times in weeks for leukemia patients. Twenty patients received treatment A and 20 received treatment B. The predictor *group* was 0 for A and 1 for B.

a) Obtain the SAS program for this problem from (http://parker.ad.siu.edu/Olive/survhw.txt). Obtain the output from the program in the same manner as i) through iv) above Problem 1.20.

b) Do a 4 step test for $Ho : \beta = 0$.

c) Do a 4 step PLRT for $Ho : \boldsymbol{\beta} = \mathbf{0}$ (for $\beta = 0$). (The PLRT is better than the Wald test in b).)

R Problems

R is the free version of *Splus.* Click on the *Rgui* icon to get into *R.* Then typing *q()* gets you out of *R.*

Use the command *source("G:/survdata.txt")* **to download the data. See Preface or Section 5.1.** For the following problems, the *R* command can be copied and pasted from (http://parker.ad.siu.edu/Olive/survhw.txt) into *R*.

1.23. Miller (1981, p. 49) gives the length of times of remission (time until relapse) in acute myelogeneous leukemia under maintenance chemotherapy for 11 patients is
$9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+$.

a) Following Example 1.3, make a table with headers $t_{(j)}$, γ_j, t_i, n_i, d_i and $\hat{S}_K(t_i)$. Then compute the Kaplan Meier estimator. (You can check it with the *R* output obtained in b).)

b) Get into *R*. Copy and paste commands for this problem from (http://parker.ad.siu.edu/Olive/survhw.txt) into *R*. Hit `Enter` and a plot should appear. Copy and paste the *R* output with header (time ... upper 95% CI) into *Word*. Following the *R* handout, click on the plot and hold down the *Ctrl* and *c* buttons simultaneously. Then in the *Word* Edit menu, select "paste."

Include this output with the homework. The center step function is the Kaplan Meier estimator $\hat{S}_K(t)$ while the lower and upper limits correspond to the confidence interval for $S_Y(t)$.

c) Write down the 95% CI for $S_Y(23)$ and then verify the CI by computing $\hat{S}_K(23) \pm 1.96SE(\hat{S}_K(23))$.

1.24. Copy and paste commands for parts a) and b) for this problem from (http://parker.ad.siu.edu/Olive/survhw.txt) into *R*.

The commands make the KM estimator for censored data $T = \min(Y, Z)$ where $Y \sim EXP(1)$. The KM estimator attempts to estimate $S_Y(t) = \exp(-t)$. The points in the plot are $S_Y(t_{(j)}) = \exp(-t_{(j)})$, and the points

should be within the confidence intervals roughly 95% of the time (actually, if you make many plots the points should be in the intervals about 95% of the time, but for a given plot you could get a "bad data set" and then the rather more than 5% of the points are outside of the intervals).

a) Copy and paste the commands for a) and hit `Enter`. Then copy and paste the plot into *Word*.

b) Copy and paste the commands for b) and hit `Enter`. Then copy and paste the plot into *Word*.

c) As the sample size increases from $n = 20$ to $n = 200$, the CIs should become more narrow. Can you see this in the two plots? Are about 95% of the plotted points within the CIs?

1.25. Go to (http://parker.ad.siu.edu/Olive/survhw.txt) and copy and paste the source command near the top of the file into *R*.

Type the command `kmsim2(n=10)`, hit `Enter` and include the output in *Word*.

This program computes censored data $T = \min(Y, Z)$ where $Y \sim EXP(1)$. Then a 95% CI is made for $S_Y(t_{(j)})$ for each of the $n = 10$ $t_{(j)}$. This is done for 100 data sets and the program counts how many times the CI contains $S_Y(t_{(j)}) = \exp(-t_{(j)})$. The scaled lengths are also computed. The ccov is the count for the classical $\hat{S} \pm 1.96SE(\hat{S})$ interval while p4cov is for the plus 4 CI. The lcov is based on a CI that uses $\log(\hat{S})$ and llcov is based on a CI that uses $\log(-\log(\hat{S}))$. The 1st 3 CIs are not made if the last case is censored so NA is given. The plus 4 CI seems to be good at $t_{(1)}$ and $t_{(n)}$.

Problems from Quizzes and Exams

1.40. Suppose

$$F(t) = 1 - \frac{1}{1+t}$$

where $t > 0$. Find the following quantities for $t > 0$.

a) $S(t)$
b) $f(t)$
c) $h(t)$
d) $H(t)$

1.41. A survival function for treatment A is plotted below.

```
1.0|________
.9 |        |
.8 |        |
.7 |        |______________
.6 |                       |___________
.5 |                                   |
.4 |                                   |___
.3 |                                       |
.2 |                                       |
.1 |                                       |______________________
0  |_____________________________________________________________|___
       |   |   |   |   |   |   |   |   |   |   |   |   |   |
       10  20  30  40  50  60  70  80  90  100 110 120 130 140
```

a) Estimate when 50% of the patients from treatment A have died. Show the over and down lines.

b) Suppose treatment B had lower hazard than treatment A for $0 < t < 140$. Would you expect the survivor function for treatment B to be lower or higher that that for treatment A in the above plot?

Chapter 2
Cox Proportional Hazards Regression

This chapter give the first 1D regression model for survival analysis. The survival 1D regression models differ from the multiple linear regression, experimental design models, and generalized linear models in that the conditional mean function is no longer of primary interest. Instead, the conditional survival function and the conditional hazard functions are of interest. For survival regression, the ith case will often be $(T_i = Y_i^*, \delta_i, \boldsymbol{x}_i^T)^T$ for $i = 1, ..., n$ where $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$ is a $p \times 1$ vector of predictors. Predictors are also called independent variables, risk factors, or explanatory variables.

Definition 2.1. 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 2.2. *Regression* investigates how the response variable Y changes with the value of a $p \times 1$ vector $\boldsymbol{x}$ of predictors. Often this *conditional distribution* $Y|\boldsymbol{x}$ is described by a *1D regression model*, where Y is conditionally independent of $\boldsymbol{x}$ given the *sufficient predictor* $SP = h(\boldsymbol{x})$, written

$$Y \perp\!\!\!\perp \boldsymbol{x}|SP \quad \text{or} \quad \mathrm{Y} \perp\!\!\!\perp \boldsymbol{x}|\mathrm{h}(\boldsymbol{x}), \tag{2.1}$$

where the real valued function $h : \mathbb{R}^p \to \mathbb{R}$. The *estimated sufficient predictor* $\mathrm{ESP} = \hat{h}(\boldsymbol{x})$. An important special case is a model with a linear predictor $h(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{\beta}$ where $\mathrm{ESP} = \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$. This class of models includes several important survival regression models.

One of the simplest examples of a regression model has $\boldsymbol{x} = (x_1) = x_1 = x$ where $x = 1$ for a new treatment and $x = 0$ for a standard treatment or for a placebo = sham treatment. Then $\hat{S}(t|x = 1)$ and $\hat{S}(t|x = 0)$ are of interest.

Suppose $S(t|\boldsymbol{x}_j)$ is of interest. If there was enough data at $\boldsymbol{x}_j$, say $Y_1^*(\boldsymbol{x}_j), ..., Y_m^*(\boldsymbol{x}_j)$, then you could make, for example, the Kaplan Meier es-

timator for various values of $\boldsymbol{x}_j$ and plot the survival curves, e.g, $\hat{S}_k(t|\boldsymbol{x}_1), ...,$ $\hat{S}_k(t|\boldsymbol{x}_J)$.

Often there is only one censored survival time $Y_i^*|\boldsymbol{x}_i$ for each vector of predictors $\boldsymbol{x}_i$. The training data set is $(Y_i^*, \delta_i, \boldsymbol{x}_i^T)^T$ for $i = 1, ..., n$. Often interest is in estimating the conditional hazard function $h_i(t) = h(t|\boldsymbol{x}_i) = h_{Y_i|\boldsymbol{x}_i}(t) = h_{Y_i|\boldsymbol{\beta}^T\boldsymbol{x}_i}(t)$.

2.1 Proportional Hazards Regression

Definition 2.3. The **Cox proportional hazards regression (PH) model** is

$$h_i(t) = h_{Y_i|\boldsymbol{x}_i}(t) = h_{Y_i|\boldsymbol{\beta}^T\boldsymbol{x}_i}(t) = \exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)h_0(t)$$

where $h_0(t)$ is the **unknown baseline function** and $\exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)$ is the **hazard ratio**. The sufficient predictor $\mathbf{SP} = \boldsymbol{\beta}^T\boldsymbol{x}_i = \sum_{j=1}^p \beta_j x_{ij}$.

The Cox PH model (= Cox PH regression model = Cox regression model = Cox proportional hazards regression model) is a 1D regression model since the conditional distribution $Y|\boldsymbol{x}$ is completely determined by the hazard function, and the hazard function only depends on $\boldsymbol{x}$ through $\boldsymbol{\beta}^T\boldsymbol{x}$. Inference for the PH model uses computer output that is used almost exactly as the output for generalized linear models such as the logistic and Poisson regression models. The Cox PH model is semiparametric: the conditional distribution $Y|\boldsymbol{x}$ depends on the sufficient predictor $\boldsymbol{\beta}^T\boldsymbol{x}$, but the parametric form of the hazard function $h_{Y|\boldsymbol{x}}(t)$ is not specified. The Cox PH model is the most widely used survival regression model in survival analysis. For the Cox PH model, often we will use $\boldsymbol{\beta} = \boldsymbol{\beta}_C$.

Regression models are used to study the conditional distribution $Y|\boldsymbol{x}$ given the $p \times 1$ vector of nontrivial predictors $\boldsymbol{x}$. In survival regression, Y is the time until an event such as death. Many of the most important survival regression models are 1D regression models with $SP = \boldsymbol{\beta}^T\boldsymbol{x}$: the nonnegative response variable Y is independent of $\boldsymbol{x}$ given $\boldsymbol{\beta}^T\boldsymbol{x}$, written $Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x}$. Let the sufficient predictor $SP = \boldsymbol{\beta}^T\boldsymbol{x}$, and the estimated sufficient predictor $ESP = \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$. The ESP is sometimes called the estimated risk score. The sufficient predictor is also called a linear component or linear predictor.

The conditional distribution $Y|\boldsymbol{x}$ is completely determined by the probability density function $f_{\boldsymbol{x}}(t)$, the distribution function $F_{\boldsymbol{x}}(t)$, the survival function

$$S_{\boldsymbol{x}}(t) \equiv S_{Y|SP}(t) = P(Y > t|SP = \boldsymbol{\beta}^T\boldsymbol{x}),$$

the cumulative hazard function $H_{\boldsymbol{x}}(t) = -\log(S_{\boldsymbol{x}}(t))$ for $t > 0$, or the hazard function $h_{\boldsymbol{x}}(t) = \frac{d}{dt}H_{\boldsymbol{x}}(t) = f_{\boldsymbol{x}}(t)/S_{\boldsymbol{x}}(t)$ for $t > 0$. High hazard implies low survival times while low hazard implies long survival times.

Survival data is usually right censored so Y is not observed. Instead, the survival time $T_i = \min(Y_i, Z_i)$ where $Y_i \perp\!\!\!\perp Z_i$ and Z_i is the censoring time. Also $\delta_i = 0$ if $T_i = Z_i$ is censored and $\delta_i = 1$ if $T_i = Y_i$ is uncensored. Hence the data is $(T_i, \delta_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$.

The *Cox proportional hazards* regression model (Cox 1972) is a semiparametric model with $SP = \boldsymbol{\beta}_C^T \boldsymbol{x}$ and

$$h_{\boldsymbol{x}}(t) \equiv h_{Y|SP}(t) = \exp(\boldsymbol{\beta}_C^T \boldsymbol{x}) h_0(t) = \exp(SP) h_0(t)$$

where the baseline hazard function $h_0(t)$ is left unspecified. Note that

$$\frac{h_{Y|SP}(t)}{h_0(t)} = e^{SP}, \quad \text{and} \quad SP = \log\left(\frac{h_{Y|SP}(t)}{h_0(t)}\right).$$

The survival function is

$$S_{\boldsymbol{x}}(t) \equiv S_{Y|SP}(t) = [S_0(t)]^{\exp(\boldsymbol{\beta}_C^T \boldsymbol{x})} = [S_0(t)]^{\exp(SP)}. \tag{2.2}$$

If $\boldsymbol{x} = \mathbf{0}$ is within the range of the predictors, then the baseline survival and hazard functions correspond to the survival and hazard functions of $\boldsymbol{x} = \mathbf{0}$. First $\boldsymbol{\beta}_C$ is estimated by the maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}_C$, then estimators $\hat{h}_0(t)$ and $\hat{S}_0(t)$ can be found (see Breslow 1974), and

$$\hat{S}_{\boldsymbol{x}}(t) = [\hat{S}_0(t)]^{\exp(\hat{\boldsymbol{\beta}}_C^T \boldsymbol{x})} = [\hat{S}_0(t)]^{\exp(ESP)}, \tag{2.3}$$

$\hat{h}_{\boldsymbol{x}}(t) = \exp(\boldsymbol{x}^T \hat{\boldsymbol{\beta}}) \hat{h}_0(t)$, and $\hat{H}_{\boldsymbol{x}}(t) = \exp(\boldsymbol{x}^T \hat{\boldsymbol{\beta}}) \hat{H}_0(t)$.

Let $h_i(t) = h_{\boldsymbol{x}}(t) = e^{\boldsymbol{x}^T \boldsymbol{\beta}} h_0(t) = \exp(x_1\beta_1 + \cdots + x_i\beta_i + \cdots x_p\beta_p) h_0(t)$. Suppose x_i changes by r units while the other x_j are held fixed. Then $SP(x_i + r) = x_1\beta_1 + \cdots + (x_i + r)\beta_i + \cdots x_p\beta_p = SP + r\beta_i$, and

$$h_{i|x_i+r}(t) = \exp(r\beta_i) \exp(\boldsymbol{x}^T \boldsymbol{\beta}) h_0(t) = \exp(r\beta_i) h_i(t).$$

Then the hazard ratio

$$\frac{h_{i|x_i+r}(t)}{h_0(t)} = \exp(r\beta_i) \frac{h_i(t)}{h_0(t)}$$

changes by a factor of $\exp(r\beta_i)$. The log hazard ratio

$$\log\left(\frac{h_{i|x_i+r}(t)}{h_0(t)}\right) = r\beta_i + \log\left(\frac{h_i(t)}{h_0(t)}\right) = r\beta_i + \boldsymbol{x}^T \boldsymbol{\beta}.$$

Thus β_i is the change in the log hazard ratio when $\boldsymbol{x}_i$ is changed by $r = 1$ unit with all other x_j held fixed.

2.2 Visualizing the Cox PH Regression Model

Grambsch and Therneau (1994) give a useful graphical check for whether the PH model is a reasonable approximation for the data. Suppose the ith case had an uncensored survival time t_i. Let the scaled Schoenfeld residual for the ith observation and jth variable x_j be $r^*_{P,j}(t_i)$. For each variable, plot the t_i versus the $r^*_{P,j}(t_i) + \hat{\beta}_j$ and add the loess curve. If the loess curve is approximately horizontal for each of the p plots, then the proportional hazards assumption is reasonable. Alternatively, fit a line to each plot and test that each of the p slopes is equal to 0. The R function `cox.zph` makes both the plots and tests. See MathSoft (1999b, p. 267, 275). Hosmer and Lemeshow (1999, p. 211) suggest also testing whether the interactions $x_i \log(t)$ are significant for $i = 1, ..., p$.

Definition 2.4. The **slice survival plot** divides the ESP into J groups of roughly the same size. For each group j with n_j cases, the model estimated survival function $\hat{S}_j(t)$ is computed using the $\boldsymbol{x}$ corresponding to the "median ESP" of the group (the kth order statistic of the ESP in group j, where $k = 1 + floor[(n_j - 1)/2]$). Let $\hat{S}_{KMj}(t)$ be the Kaplan Meier estimator computed from the survival times (T_i, δ_i) in the jth group. For each group, $\hat{S}_j(t)$ is plotted and $\hat{S}_{KMj}(t_i)$ is plotted as circles at the uncensored event times t_i. The survival regression model is reasonable if the circles "track $\hat{S}_j$ well" in each of the J plots.

If the slice widths go to zero, but the number of cases per slice increases to ∞ as $n \to \infty$, then the Kaplan Meier estimator and the model estimator converge to $S_{Y|SP}(t)$ if the model holds. Simulations suggest that the two survival functions are "close" for moderate n and nine slices. For small n and skewed predictors, some slices may be too wide in that the model is correct but $\hat{S}_{KMj}(t)$ is not a good approximation of $S_{Y|SP}(t)$ where SP corresponds to the $\boldsymbol{x}$ used to compute $\hat{S}_j(t)$.

For the Cox model, if pointwise confidence interval (CI) bands are added to the plot, then $\hat{S}_{KMj}$ "tracks $\hat{S}_j$ well" if most of the plotted circles do not fall very far outside the pointwise CI bands since these pointwise bands are not as wide as simultaneous bands. Collett (2003, p. 241-243) places several observed Kaplan Meier curves with fitted curves on the same plot.

Survival regression is the study of the conditional survival $S_{Y|SP}(t)$, and the slice survival plot is a useful tool for visualizing $S_{Y|SP}(t)$ in the background of the data. Suppose the jth slice is narrow so that $ESP \approx w_j$. If the model is reasonable, $ESP \approx SP$, and the number of uncensored cases in the jth slice is not too small, then $S_{Y|SP=w_j}(t) \approx \hat{S}_j(t) \approx \hat{S}_{KMj}(t)$. (These quantities approximate $[\hat{S}_0(t)]^{\exp(w_j)}$ for the Cox model.) Thus the nonparametric Kaplan Meier estimator is used to check the model estimator $\hat{S}_j(t)$ in each slice.

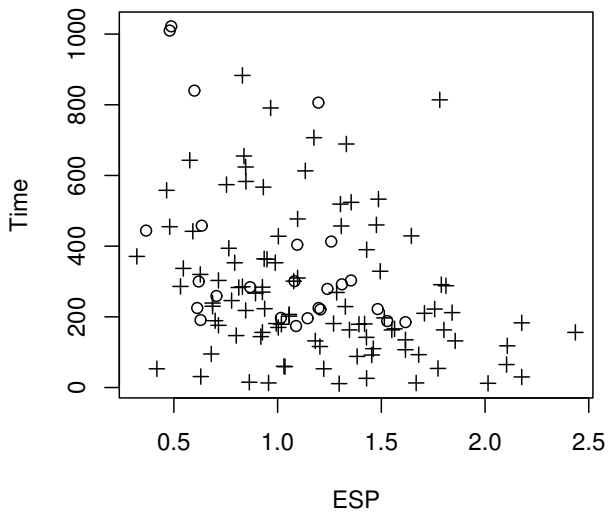

Fig. 2.1 Censored Response Plot for R Lung Cancer Data

The slice survival plot tailored to the Cox model is closely related to the May and Hosmer (1998) test. Also, van Houwelingen et al. (2006) use similar ideas, but place the J Kaplan Meier curves on one plot and the J Cox survival curves on another plot. For a 1D regression model, the ESP is a scalar while $\boldsymbol{x}$ is a $p \times 1$ vector. Using the ESP instead of $\boldsymbol{x}$ in plots is an important dimension reduction technique (and is similar to using a scalar valued minimal sufficient statistic instead of the p-dimensional sufficient statistic $\boldsymbol{x}$.) Inferior plots have been suggested by several authors with $\boldsymbol{x}$ divided into J groups instead of the ESP. For example, see Miller (1981, p. 168). Hosmer and Lemeshow (1999, p. 141–145) suggests making plots based on the quartiles of the ith predictor x_i, and note that a problem with Cox survival curves (2.3) is that they may use inappropriate extrapolation. Using the ESP results in narrow slices with many cases, and adding Kaplan Meier curves shows if there is extrapolation. The main use of the next plot is to check for cases with unusual survival times. Hazard increases and survival decreases as ESP increases if ESP $\approx$ SP.

Definition 2.5. A **censored response plot** is a plot of the ESP versus T with plotting symbol 0 for censored cases and + for uncensored cases. Slices in this plot correspond to the slices used in the slice survival plot.

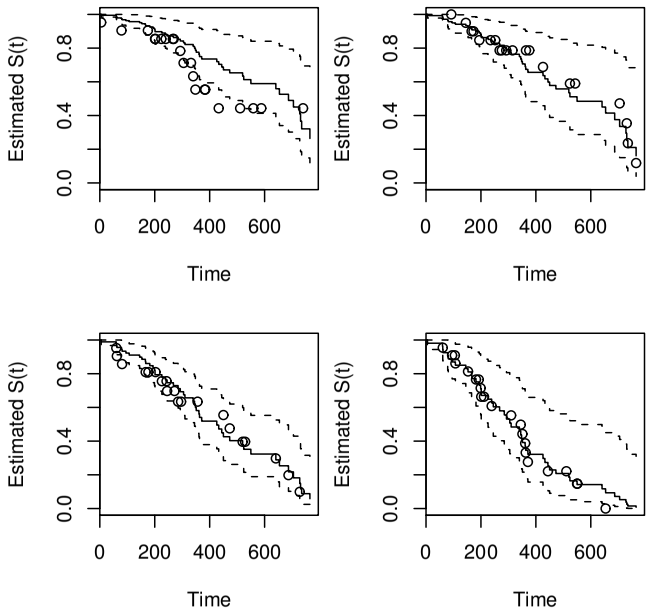

Fig. 2.2 Slice Survival Plots for R Lung Cancer Data

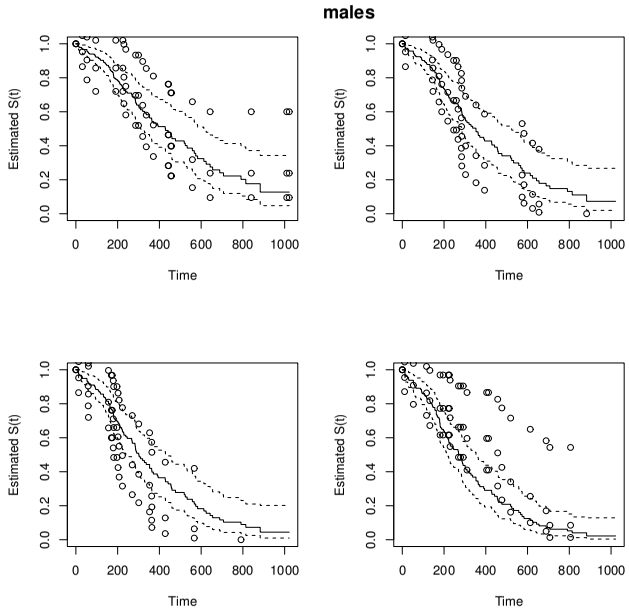

Fig. 2.3 Slice Survival Plots for R Lung Cancer Data-Males

Suppose the ESP is a good estimator of the SP. Consider a narrow vertical slice taken in the censored response plot about $ESP = w$. The points in the slice are a censored sample with $S_{Y|SP}(t) \approx S_{Y|w}(t)$. For proportional hazards models, $h_{Y|SP}(t) \approx \exp(ESP)h_0(t)$, and the hazard increases while the survival decreases as the ESP increases.

Example 2.1. *R* and *Splus* contain a data set *lung* where the response variable Y is the time until death for patients with lung cancer. See MathSoft (1999b, p. 268). Consider the data set for males with predictors *ph.ecog* = Ecog performance score 0-4, *ph.karno* = a competitor to ph.ecog, *pat.karno* = patient's assessment of their karno score and *wt.loss* = weight loss in last 6 months. Figure 2.1 shows the censored response plot. Notice that the survival times decrease rapidly as the ESP increases and that there is one time that is unusually large for ESP ≈ 1.8. If the Cox regression model is a good approximation to the data, then the response variables corresponding to the cases in a narrow vertical strip centered at $ESP = w$ are approximately a censored sample from a distribution with hazard function $h_{\boldsymbol{x}}(t) \approx \exp(w)h_0(t)$. Figure 2.2 shows the slice survival plots. The ESP was divided into 4 groups and the ESP increases from the upper left, upper right, lower left and lower right corners of the plot where $\hat{S}(400) \approx (0.70, 0.60, 0.55, 0.30)$. The circles corresponding to the Kaplan Meier estimator are "close" to the Cox survival curves in that the circles do not fall very far outside the pointwise CI bands.

Figure 2.3 shows the slice survival plots for males. See Problem 2.19. Some versions of *R* add the poinwise confidence interval bands for the Kaplan Meier estimator to the plot. Then there are three curves of circles. The center curve of circles is the Kaplan Meier estimator while the two outer curves of cirlces are the pointwise CI bands. The pointwise CI bands for the PH survival curve are narrower than those for the Kaplan Meier estimator since the PH survival curve is based on all n cases while the Kaplan Meier estimator is based on the n_i cases corresponding to the ith slice. The center curve does not fall very far outside the Cox PH pointwise survival bands, although the lower right plot looks the worst.

Example 2.2. *R* contains a data set *nwtco* where the response variable Y is the time until relapse with $n = 4028$. The model used predictors *histol* = tumor histology from central lab, *instit* = tumor histology from local institution, *age* in months, and *stage* of disease from 1 to 4 (treated as a continuous variable). In Figure 2.4, the Grambsch and Therneau (1994) plots suggest that the Cox model is not valid since not all of the loess curves are flat, and the global test has p-value $\approx 5.66 \times 10^{-11}$. The slice survival plot in Figure 2.5 shows that the Cox survival estimators and Kaplan Meier estimators are nearly identical in the six slices, suggesting that the Cox model is a reasonable approximation to the data. The greatest contributors to lack of fit seem to be the predictors age and stage corresponding to the bottom two plots of Figure 2.4, and survival for small ESP corresponding to the upper left plot in Figure 2.5.

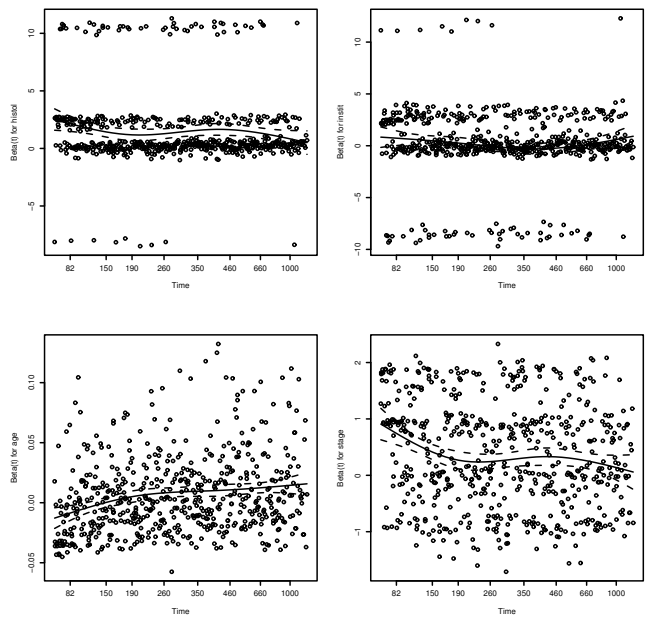

Fig. 2.4 Grambsch and Therneau Plots for NWTCO Data

Residuals are quantities calculated for each individual or case, and the residual behavior is roughly known with the fitted model is satisfactory. Let $T_i = t_i$ be the observed death or censoring time of individual i.

Definition 2.6. a) The **Cox Snell residual** $r_{ci} = \exp(\boldsymbol{x}^T\hat{\boldsymbol{\beta}})\hat{H}_0(t_i) = \hat{H}_{\boldsymbol{x}}(t_i)$ for $i = 1, ..., n$.

b) Let $\gamma_i = 1$ if t_i is uncensored and $\gamma_i = 0$ if t_i is censored. Then the **Martingale residual** $r_{mi} = \gamma_i - r_{ci}$.

The Martingale residual has mean 0 for uncensored cases and $r_{mi} < 0$ if $\gamma_i = 0$ if case i is censored. Also, $-\infty < r_{mi} \leq 1$. It can be shown that $-\log(S(Y)) \sim EXP(1)$. So if $\hat{S}(t)$ is a good approximation to $S(t)$, then $-\log(\hat{S}_{\boldsymbol{x}_i}(t_i)) = \hat{H}_{\boldsymbol{x}_i}(t_i) = r_{ci}$ should behave like n observations from a censored $EXP(1)$ distribution.

2.3 Testing

For regression models, we want to test i) whether the predictors $\boldsymbol{x}$ are needed in the model: $H_0 : \boldsymbol{\beta} = \mathbf{0}$ versus $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$, ii) whether a reduced model that that does not use predictors $x_{i_1}, ..., x_{ik}$ is good: $H_0 : (\beta_{i1}, ..., \beta_{ik})^T = \mathbf{0}$

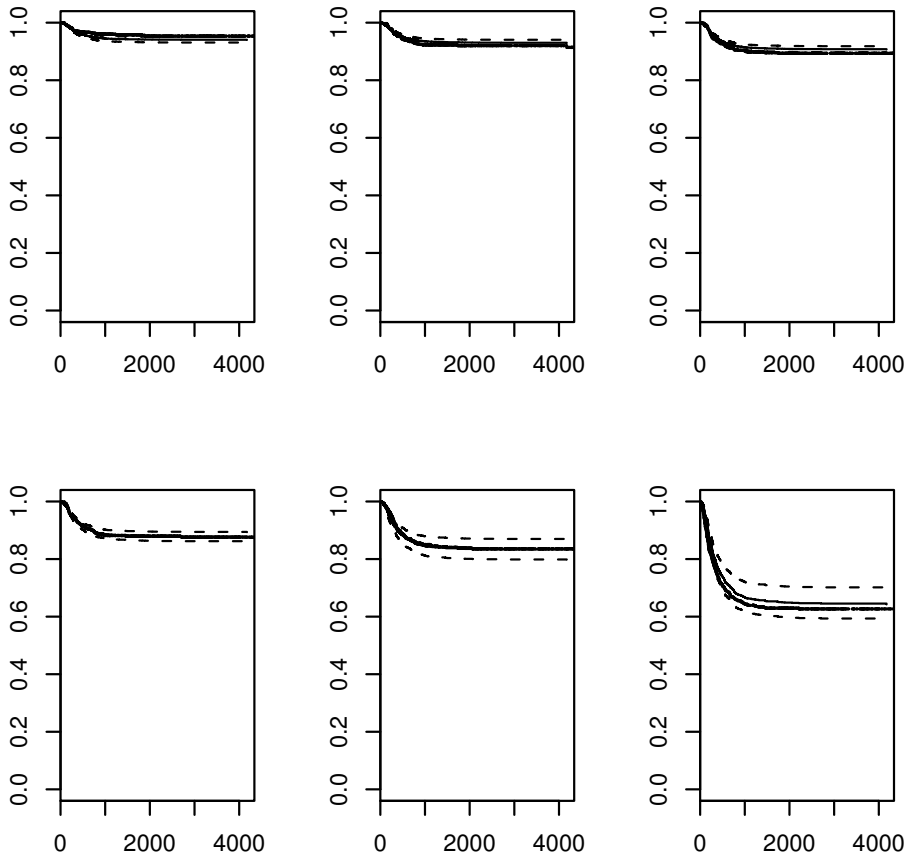

Fig. 2.5 Slice Survival Plot for NWTCO Data: Horizontal Axis is the Estimated Survival Function S(t)

versus $H_1 : (\beta_{i1}, ..., \beta_{ik})^T \neq \mathbf{0}$, and iii) whether predictor x_i is needed in the model given that the other predictors are needed in the model $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0$. Note that tests i) and iii) are special cases of test ii). We also want confidence intervals for β_i. We also want to find $ESP = \hat{\boldsymbol{\beta}}_C^T \boldsymbol{x}_i$ and $\hat{h}_i(t) = e^{ESP} \hat{h}_0(t)$ given $\boldsymbol{x}_i$. Often the hypothesis $H_1 = H_A$.

Computer output will be needed, and shown below is output in symbols from *SAS* and *R*. The estimated coefficient is $\hat{\beta}_j$. The Wald chi square $= X_{0,j}^2$ while p and "pr > chisqu" are both p-values. Sometimes "Std. Err." replaces "SE." Note that $z_{0,j}^2 = X_{0,j}$ where $z_{0,j} \approx N(0,1)$, a standard normal random variable.

variable	Est.	SE	Est./SE	or $(Est/SE)^2$	pvalue for
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{0,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	$X^2_{0,1} = z^2_{0,1}$	$H_0 : \beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$z_{0,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	$X^2_{0,p} = z^2_{0,p}$	$H_0 : \beta_p = 0$

```
SAS                                Wald          pr >
variable  df Estimate  SE     chi square    chisqu
age       1  0.1615    0.0499 10.4652       0.0012
ecog.ps   1  0.0187    0.5991 0.00097       0.9800

R         coef  exp(coef) se(coef)    z      p
age     0.1615      1.18   0.0499 3.2350 0.0012
ecog.ps 0.0187      1.02   0.5991 0.0312 0.9800

Likelihood ratio test=14.3  on 2 df, p=0.000787  n= 26
```

The estimated sufficient predictor **ESP** $= \hat{\boldsymbol{\beta}}'\boldsymbol{x}_j = \sum_{i=1}^p \hat{\beta}_i x_{ij}$. Given $\hat{\boldsymbol{\beta}}$ from output and given $\boldsymbol{x}$, be able to find ESP and $\hat{h}_i(t) = \exp(ESP)\hat{h}_0(t) = \exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x})\hat{h}_o(t)$ where $\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x})$ is the **estimated hazard ratio**.

For tests, the p–value is an important quantity. The hypothesis H_0 is rejected if the p–value $< \delta$. A p–value between 0.07 and 1.0 provides little evidence that H_0 should be rejected, a p–value between 0.01 and 0.07 provides moderate evidence and a p–value less than 0.01 provides strong statistical evidence that H_0 should be rejected. Statistical evidence is not necessarily practical evidence, and reporting the p–value along with a statement of the strength of the evidence is more informative than stating that the p–value is less than some chosen value such as $\delta = 0.05$. Nevertheless, as a **homework convention**, use $\delta = 0.05$ if δ is not given.

The Wald confidence interval (CI) for β_j can also be obtained from the output: the large sample 95% CI for β_j is

$$\hat{\beta}_j \pm 1.96\ se(\hat{\beta}_j).$$

Investigators often test whether a predictor x_j is needed in the model given that the other $p-1$ predictors $x_1, ..., x_{i-1}, x_{i+1}, ..., x_p$ are in the model with a **4 step Wald test of hypotheses**:

i) State the hypotheses $H_0 : \beta_j = 0 \quad H_1 : \beta_j \neq 0$.

ii) Find the test statistic $z_{0,j} = \hat{\beta}_j/se(\hat{\beta}_j)$ or $X^2_{0,j} = z^2_{0,j}$ or obtain it from output.

iii) The p–value $= 2P(Z < -|z_{0j}|) = P(\chi^2_1 > X^2_{0,j})$. Find the p–value from output or use the standard normal table.

iv) State whether you reject Ho or fail to reject Ho and give a nontechnical sentence restating your conclusion in terms of the story problem.

If Ho is rejected, then conclude that x_j is needed in the PH survival model given that the other $p-1$ predictors are in the model. If you fail to reject Ho, then conclude that the values of x_j do not (significantly) affect the PH survival model given that the other $p-1$ predictors are in the model. (Or state that there is not enough evidence to conclude that the values of x_j affect the PH survival model.)

Typically the "p-value" is actually an estimated p-value called **pval**. When a normal table is used, if $-|z_{0,j}| < -3.9$, then take pval $= 0$.

Remark 2.1. Suppose the test fails to reject Ho. Then x_j could be a very useful PH survival predictor, but may not be needed if other predictors are added to the model (often due to correlation with other predictors). Also, x_j could be needed in the survival model, but survival does not depend on the observed values of x_j. This result can be extremely important if x_j is treatment. For example, suppose that there are two different, but equally effective treatments, where $x_j = 1$ for treatment 1 and $x_j = 0$ for treatment 2. Then the test may fail to reject Ho for Ho: $\beta_j = 0$, but not giving either treatment may greatly reduce survival. If treatment 2 is a placebo = sham treatment, then failing to reject Ho suggests that treatment 1 in not effective. It is also possible that the sample size is not large enough to determine whether the values of x_j affect the survival model.

Example 2.3. Allison (1995, p. 120) considers one of the first heat transplant studies with Y = days from acceptance until death, x_1 = trans = 1 if the patient received a heart transplant with $x_1 = 0$, otherwise, x_2 = surg = 1 if the transplant was before the date of acceptance with $x_2 = 0$, otherwise, and x_3 = ageaccept = age at date of acceptance. Using the following output, a) find ESP $\hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ if $\boldsymbol{x} = (1, 0, 64)^T$, b) find $\hat{h}_i(t)$, c) find a 95% Wald CI for β_2, d) perform a 4 step test of hypotheses for $\beta_2 = 0$ without using output to find the test statistic and p–value, e) perform the 4 step test of hypotheses of $\beta_3 = 0$ using output.

variable	df	estimate	SE	Wald chisquare $X^2_{0,j} = z^2_{0,j}$	pr > chisq pval	risk ratio
trans	1	−1.70814	0.2786	37.59	0.0001	0.181
surg	1	−0.42140	0.3710	1.29	0.2560	0.656
ageaccpt	1	0.05861	0.0151	15.16	0.0001	1.060

Solution: a) ESP $= \hat{\boldsymbol{\beta}}^T \boldsymbol{x} = -1.70814(1) - 0.170814(0) + 0.05861(64) = 2.0429$

b) $\hat{h}_i(t) = e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{x}} \hat{h}_0(t) = e^{2.0429} \hat{h}_0(t) = 7.7129\ \hat{h}_0(t)$

c) $\hat{\beta}_2 \pm 1.96SE(\hat{\beta}_2) = -0.4214 \pm 1.96(0.3710) = -0.4214 \pm 0.72716 = [-1.1486, 0.3058]$

Note that the 95% CI gives reasonable values for β_2 and includes 0. thus x_2 may not be important given that x_1 and x_2 are in the model.

d) i) $H_0 : \beta_2 = 0$, $H_1 : \beta_2 \neq 0$

ii)

$$z_{0,2} = \frac{-0.4214}{0.3710} = -1.136$$

iii) Using a normal table and rounding $z_{0,2}$ to 2 digits, $pval = 2P(Z < -|z_{0,2}|) = 2P(Z < -1.14) = 2(0.1271) = 0.2542$. From the t-table near the back of Chapter 5, line Z and the last line "two tail" gives $0.1 < pval < 1$.

iv) Since pval $> \delta = 0.05$, fail to reject H_0. Hence the values of surg do not affect the survival model given that trans and ageaccpt are in the model.

e) i) $H_0 : \beta_3 = 0$, $H_1 : \beta_3 \neq 0$

ii) $X^2_{0,3} = 15.16$

iii) $pval = 0.001 < \delta = 0.05$

iv) Since $pval < \delta$, reject H_0. Hence ageaccpt is needed in the survival model given that trans and surg are in the model.

For a PH, often 3 models are of interest: the **full model** that uses all p of the predictors $\boldsymbol{x}^T = (\boldsymbol{x}_R^T, \boldsymbol{x}_O^T)$, the **reduced model** that uses the r predictors $\boldsymbol{x}_R$, and the **null model** that uses none of the predictors. The null model has $h_i(t) \equiv h_0(t)$ regardless of the value of $\boldsymbol{x}_i$.

The *partial likelihood ratio test* (**PLRT**) is used to test whether $\boldsymbol{\beta} = \mathbf{0}$. If this is the case, then the predictors are not needed in the PH model (so survival times $Y \perp\!\!\!\perp \boldsymbol{x}$). If $H_0 : \boldsymbol{\beta} = \mathbf{0}$ is not rejected, then the Kaplan Meier estimator should be used. If H_0 is rejected, use the PH model.

Know that the 4 step **PLRT** is

i) $H_0 : \boldsymbol{\beta} = \mathbf{0} \quad H_1 : \boldsymbol{\beta} \neq \mathbf{0}$

ii) test statistic $X^2(N|F) = [-2\log L(none)] - [-2\log L(full)]$ is often obtained from output.

iii) The p–value $= P(\chi^2_p > X^2(N|F))$ where χ^2_p has a chi–square distribution with p degrees of freedom. The p–value is often obtained from output.

iv) Reject H_0 if the p–value $< \delta$ and conclude that there is a PH survival relationship between Y and the predictors $\boldsymbol{x}$. If p–value $\geq \delta$, then fail to reject H_0 and conclude that the values of the predictors $\boldsymbol{x}$ do not (significantly) affect the PH survival model. (Or state that there is not enough evidence to conclude that the values of $\boldsymbol{x}$ affect the PH survival model.)

Remark 2.2. Suppose the test fails to reject Ho. Then it is possible that predictors are not useful for predicting survival. It is also possible that that the predictors are very useful for increasing survival, but survival does not depend on the observed values of the predictors. For example, there could be two or more equally effective treatments, but if no treatment was given, then survival would decrease greatly. It is also possible that the sample size is not

large enough to determine whether the values of $\boldsymbol{x}$ affect the survival model.

Output in symbols is often given in three ways.

variables in model	$-2\log\hat{L}$
none	$-2\log\hat{L}(none)$
$\vdots$	$\vdots$
$x_1, ..., x_p$	$-2\log\hat{L}(full)$

or

Model test	Fit chisq	Statistics DF	pr > chisq
likelihood ratio	$X^2(N\|F)$	p	pval $= P(\chi^2_p > X^2(N\|F))$

or

Testing criterion	Global without covariates	Null Hypotheses: with covariates	BETA = 0 model chisq
likelihood ratio $-2\log L$	$-2\log\hat{L}(none)$	$-2\log\hat{L}(full)$	$X^2(N\|F)$

R output for the PLRT uses a line like
Likelihood ratio test=14.3 on 2 df, p=0.000787.
Some *SAS* output for the PLRT is shown next.

```
Model Fit Statistics or
SAS Testing Global Null Hypotheses: BETA = 0
           without     with
criterion covariates covariates model Chi-square with
-2 LOG L  596.651    551.1888  45.463 3 DF (p=0.0001)
```

$x_1, ..., x_p$	$-2\log\hat{L}(full)$
none	$-2\log\hat{L}(none)$

Example 2.4.

$x_1, ..., x_5$	$-2\log L = 162.479$
none	$-2\log L = 177.667$

or *R* output: likelihood ratio test = 15.188 on 5 df p = 0.00959
or

```
SAS Testing Global Null Hypotheses: BETA = 0
  Test              chisq    DF    pr > chisq
likelihood ratio    15.188   5     0.00959
```

Using the above output, shown in 3 different formats, do a 4 step test for $\boldsymbol{\beta} = \mathbf{0}$.

Solution: i) $H_0 : \boldsymbol{\beta} = \mathbf{0}$ $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$

ii) $X^2(N|F) = 15.188 = 177.667 - 162.479$

iii) $pval = 0.00959$

iv) Reject H_0, there is a PH survival relationship between survival times Y and the predictors $x_1, ..., x_5$.

Example 2.5. Suppose there are treatments A and B for leukemia patients in remission. Let $x = 0$ for treatment A and $x = 1$ for treatment B. Then $\boldsymbol{\beta} = \beta$ is a scalar since $p = 1$. Do a 4 step test for $\beta = 0$ i $n = 40$ and the output is R likelihood ration test = 1.32 on 1 df, p=0.025.

Solution: i) $H_0 : \beta = 0 \;\; H_1 : \beta \neq 0$

ii) $X^2(N|F) = 1.32$

iii) pval = 0.25

iv) Fail to reject H_0: the values of x do not affect the PH model for relapse times (so no difference between treatments A and B for survival (relapse) times).

Let the **full model** be

$$SP = SP(F) = \beta_1 x_1 + \cdots + \beta_p x_p = \boldsymbol{\beta}^T \boldsymbol{x} = \boldsymbol{\beta}_R^T \boldsymbol{x}_R + \boldsymbol{\beta}_O^T \boldsymbol{x}_O.$$

let the **reduced model**

$$SP = SP(R) = \beta_{R1} x_{R1} + \cdots + \beta_{Rr} x_{Rr} = \boldsymbol{\beta}_R^T \boldsymbol{x}_R$$

where the reduced model uses r of the predictors used by the full model and $\boldsymbol{x}_O$ denotes the vector of $p - r$ predictors that are in the full model but not the reduced model.

Assume that the full model is useful. Then we want to test H_0: the reduced model is good (can be used instead of the full model, so the values of $\boldsymbol{x}_O$ do not affect the survival model given $\boldsymbol{x}_R$ is in the model) versus H_A: use the full model (the full model is significantly better than the reduced model). Fit the full model and the reduced model to get $X^2(N|F)$ and $X^2(N|R)$ where $X^2(N|F)$ is used in the PLRT to test whether $\boldsymbol{\beta} = \mathbf{0}$ and $X^2(N|R)$ is used in the PLRT to test whether $\boldsymbol{\beta}_R = \mathbf{0}$ (treating the reduced model as the model in the PLRT).

Shown below in symbols is output for the full model and output for the reduced model. The output shown on can be used to perform the change in PLR test. For simplicity, the reduced model used in the output is $\boldsymbol{x}_R = (x_1, ..., x_r)^T$.

Notice that $X^2(R|F) \equiv X^2(N|F) - X^2(N|R) =$

$$[-2 \log L(none)] - [-2 \log L(full)] - ([-2 \log L(none)] - [-2 \log L(red)]) =$$

$$[-2 \log L(red)] - [-2 \log L(full)] = -2 \log \left(\frac{L(red)}{L(full)} \right).$$

variable	Est.	SE	Est/SE	or $(Est/SE)^2$	pvalue for
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{0,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	$X^2_{0,1} = z^2_{0,1}$	$H_0 : \beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$z_{0,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	$X^2_{0,p} = z^2_{0,p}$	Ho $H_0 : \beta_p = 0$

R: Likelihood ratio test $= X^2(N|F)$ on p df

```
SAS: Testing Global Null Hypotheses: BETA = 0
Test          Chi-Square     DF          Pr > Chisq
```

Likelihood ratio $X^2(N|F)$ p pval for $H_0 : \boldsymbol{\beta} = \mathbf{0}$

variable	Est.	SE	Est/SE	or $(Est/SE)^2$	pvalue for
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{0,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	$X^2_{0,1} = z^2_{0,1}$	$H_0 : \beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	$\hat{\beta}_r$	$se(\hat{\beta}_r)$	$z_{0,r} = \hat{\beta}_r/se(\hat{\beta}_r)$	$X^2_{0,r} = z^2_{0,r}$	$H_0 : \beta_r = 0$

R: Likelihood ratio test $= X^2(N|R)$ on r df

```
SAS: Testing Global Null Hypotheses: BETA = 0
Test          Chi-Square     DF          Pr > Chisq
```

Likelihood ratio $X^2(N|R)$ r pval for Ho: $\boldsymbol{\beta}_R = \mathbf{0}$

Know that the 4 step **change in PLR test** is

i) H_0: the reduced model is good H_1: use the full model

ii) test statistic $X^2(R|F) = X^2(N|F) - X^2(N|R) = [-2\log L(red)] - [-2\log L(full)]$.

iii) The p–value $= P(\chi^2_{p-r} > X^2(R|F))$ where χ^2_{p-r} has a chi–square distribution with $p - r$ degrees of freedom.

iv) Reject H_0 if the p–value $< \delta$ and conclude that the full model should be used. If p–value $\geq \delta$, then fail to reject H_0 and conclude that the reduced model is good (the values of $\boldsymbol{x}_O$ do not (significantly) affect the survival model, or there is not enough evidence to conclude that the values of $\boldsymbol{x}_O$ affect the survival model).

Remark 2.3. Suppose the test fails to reject Ho. Then it is possible that predictors $\boldsymbol{x}_O$ are not useful for predicting survival. It is also possible that that the predictors $\boldsymbol{x}_O$ are very useful for increasing survival, but survival does not depend on the observed values of the predictors. For example, there could be two or more equally effective treatments, but if no treatment was given, then survival would decrease greatly. It is also possible that the sample

size is not large enough to determine whether the values of $\boldsymbol{x}_O$ affect the survival model.

If the reduced model leaves out a single variable x_i, then the change in PLR test becomes $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0$. This change in partial likelihood ratio test is a competitor of the Wald test. The change in PLRT is usually better than the Wald test if the sample size n is not large, but the Wald test is often easier for software to produce. For large n the test statistics from the two tests tend to be very similar (asymptotically equivalent tests).

Example 2.6. Data is from Smith (2002, pp. 179-180). Aids patients received low dose or high dose of a drug or a placebo. Let $v1 = 1$ for low dose and $v1 = 0$, else. Let $v2 = 1$ for high dose and $v2 = 0$, else. The time until a blood test was positive was measured, and the bood test was taken each day for a month. Note that $(v1, v2) = (0, 0)$ means a placebo was given to the patient. Let the full model output be as below.

```
R    coef       se coef    z                      p
SAS parameter   standard        chisquare   Pr > chisq
    estimate    error
v1  -1.51       0.528     -2.86  8.1796       0.0043
v2  -1.03       0.455     -2.26  5.1076       0.0240

R: likelihood ratio test = 8.99 on 2 df, p = 0.0111

SAS  Test          chisq    df  Pr > chisq
 Likelihood ratio  8.99     2    0.0111
```

Let the reduced model have v1 alone with the following output.

```
R: likelihood ratio test = 3.88 on 1 df, p =

SAS  Test          chisq    df  Pr > chisq
 Likelihood ratio  3.88     1
```

Test whether the reduced model is good.
Solution: i) H_0 : the reduced model is good H_1 : use the full model
ii) $X^2(R|F) = X^2(N|F) - X^2(N|F) = 8.99 - 3.88 = 5.11$
iii) $pval = P(\chi^2_{2-1} > 5.11)$ with $0.01 < pval < 0.025$ using a χ^2 table as below

```
df | 0.025  0.01
-----------------
1  | 5.02   6.63
```

iv) Reject H_0, use the full model.

Example 2.7. Data is from Collett (2003, p. 79). Test whether the reduced model is good using the following output.

```
model    variables in model   -2 log L
reduced  A2, A3, N             165.508
full     A2, A3, N, A2N, A3N   162.479
```

Solution: i) H_0: the reduced model is good H_1: use the full model
ii) $X^2(R|F) = X^2(N|F) - X^2(N|F) = 165.508 - 162.479 = 3.029$
iii) The $df = 5 - 3 = 2 =$ number of terms left out of full model. Hence $pval = P(\chi^2_2 > 3.029)$ with $0.1 < pval < 0.25$ using a χ^2 table as below

```
df | 0.25   0.1
-----------------
2  | 2.77   4.61
```

iv) Fail to reject H_0, the reduced model is good.

If the reduced model is good, then the **EE plot** of $ESP(R) = \hat{\boldsymbol{\beta}}_R^T \boldsymbol{x}_{Ri}$ versus $ESP = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ should have plotted points that cluster tightly about the identity line with unit slope and zero intercept.

In R, there is a useful shortcut for doing the change in PLR test. In the code below let "fit" be for the full model and "fitR" be for the reduced model. The anova command gives the following output in symbols. Values left blank are not needed for the test.

	loglik	chisq	df	$P(> \|chi\|)$
1				
2		$X^2(R\|F)$	for test	pval for test

Then for the output below, $X^2(R|F) = 2.0469 = 8.08 - 6.03$ up to rounding, the df = 1, and the pval = 0.01525. So fail to reject H_0 and conclude that the reduced model is good.

```
fit <- coxph(Surv(time,status)~x1*x2 + x3, data = dat)
fitR <- coxph(Surv(time,status)~x1 + x2 + x3, data = dat)

full     coef  exp(coef)  SE(coef)  Z     P
x1       4.236            2.326     1.79  0.073
x2       2.674            2.556     1.05  0.296
x3       0.473            0.592     0.80  0.424
x1:x2   -1.936            1.421    -1.38  0.167
LRT = 8.08 on 4 df  p = 0.0888

reduced coef  exp(coef)  SE(coef)  Z     P
x1       1.347            0.680     1.98  0.048
x2      -0.749            0.595    -1.26  0.208
x3       0.453            0.590     0.77  0.443
LRT = 6.03 on 3 df  p = 0.011
```

```
anova(fitR, fit, test = "Chisq")
    loglik   chisq   df  P( > |chi|)
1   -31.970
2   -30.494  2.0469   1   0.1525
```

Remark 2.4. This remark summarizes Remarks 2.1, 2.2, and 2.3. For testing, $\boldsymbol{\beta} = \mathbf{0}$ means changing values of $\boldsymbol{x}$, within the observed range of $\boldsymbol{x}$ or of $\hat{\boldsymbol{\beta}}^T \boldsymbol{x}$, does not affect survival. For example, suppose $p = 1$ and $x = 1$ for treatment 1 and $x = 0$ for treatment 0. If treatments 1 and 0 are both very and equally effective, then $h_1(t) = h_0(t) = e^{\beta x} h_0(t)$ with $\beta = 0$. For this example, x is important for survival times Y, in that survival could be poor if neither treatment were given, but the value of x, 0 or 1, did not affect the value of Y. Hence $\boldsymbol{\beta} = \mathbf{0}$ could imply that the survival relationship between $\boldsymbol{x}$ and Y is the same for all observed values of $\hat{\boldsymbol{\beta}}^T \boldsymbol{x}$. Hence concluding $\boldsymbol{\beta} = \mathbf{0}$ does not necessarily mean that the predictors $\boldsymbol{x}$ are not important for survival times. Similarly, $\beta_i = 0$ means changing values of x_i, within the observed range of x_i, does not affect the survival times. If $\boldsymbol{\beta} = (\boldsymbol{\beta}_R^T, \boldsymbol{\beta}_O^T)^T$, then $\boldsymbol{\beta}_O = \mathbf{0}$ means changing the values of $\boldsymbol{x}_O$, within the observed values of $\boldsymbol{x}_O$ or $\hat{\boldsymbol{\beta}}_O^T \boldsymbol{x}_O$, does not affect the survival times. Then the reduced model is good in that you get the "same survival model" regardless of the $\boldsymbol{x}_O$ values. So "no survival relationship" between Y and $\boldsymbol{x}$ or $\boldsymbol{x}_O$ or x_i means within the observed range of $\boldsymbol{x}$, or $\boldsymbol{x}_O$, or x_i. This remark for testing applies to the other models in Chapters 2 and 3.

A **factor** A is a qualitative variable that takes on K categories called levels. Suppose A has a categories $c_1, ..., c_K$. Then the factor is incorporated into the PH model by using $a - 1$ indicator variables $x_{jA} = 1$ if $A = c_j$ and $x_{Aj} = 0$ otherwise, where the 1st indicator variable is omitted, eg, use $x_{2A}, ..., x_{aA}$. Each indicator has 1 degree of freedom. Hence the degrees of freedom of the $K - 1$ indicator variables associated with the factor is $K - 1$.

Example 2.8. Let factor A have levels squamous, adeno, and small cell with respective indicator variables x_{1A}, x_{2A}, and X_{3A}. Then $(x_{2A}, x_{3A}) = (1, 0)$ corresponds to adeno, $(x_{2A}, x_{3A}) = (0, 0)$ corresponds to squamous, and $(x_{2A}, x_{3A}) = (0, 1)$ corresponds to small cell.

The x_j corresponding to variates (quantitative variables that take on numerical values) or to indicator variables from a factor are called **main effects**.

An **interaction** is a product of two or more main effects, but for a factor include products for the $K - 1$ indicator variables of the factor. Hence an interaction between a variate x_1 and a factor A with indicator variables $x_{2A}, ..., x_{KA}$ is incorporated into the model with $x_1 x_{2A}, ..., x_1 x_{KA}$. An interaction between factor A and factor B with indicators $x_{2B}, ..., x_{bB}$ is incorpo-

rated into the model with the $(K-1)(b-1)$ pairs
$x_{2A}x_{2B}, ..., x_{2A}x_{bB}$
$\vdots$
$x_{KA}x_{KB}, ..., x_{KA}x_{bB}$.

If an interaction is in the full or reduced model, also include the corresponding main effects in the model. For example, if x_1x_3 is in the model, also include the main effects x_1 and x_2. In Example 2.7, A2N and A3N are interactions. Sometimes an interaction is denoted by $x_{12} = x_1x_2$ and $x_{123} = x_1x_2x_3$.

Suppose x_1 is quantitative and x_2 is qualitative with 2 levels and $x_2 = 1$ for level c_2 and $x_2 = 0$ for level c_1. Then a first order model with interaction is $SP = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$. This model yields two unrelated lines in the sufficient predictor depending on the value of x_2: $SP = \beta_2 + (\beta_1 + \beta_3)x_1$ if $x_2 = 1$ and $SP = \beta_1 x_1$ if $x_2 = 0$. If $\beta_3 = 0$, then there are two parallel lines: $SP = \beta_2 + \beta_1 x_1$ if $x_2 = 1$ and $SP = \beta_1 x_1$ if $x_2 = 0$. If $\beta_2 = \beta_3 = 0$, then the two lines are coincident: $SP = \beta_1 x_1$ for both values of x_2. If $\beta_2 = 0$, then the two lines both have the intercept at the origin: $SP = (\beta_1 + \beta_3)x_1$ if $x_2 = 1$ and $SP = \beta_1 x_1$ if $x_2 = 0$. In general, as factors have more levels and interactions have more terms, e.g. $x_1x_2x_3x_4$, the interpretation of the model rapidly becomes very complex.

A **scatterplot** is a plot of x_i versus x_j. A **scatterplot matrix** is an array of scatterplots. It is used to examine the marginal relationships of the predictors. Variables with outliers, missing values or strong nonlinearities may be so bad that they should not be included in the full model.

Suppose that all values of the variable x are positive. The **log rule** says add $\log(x)$ to the full model if $\max(x_i)/\min(x_i) > 10$.

2.4 Variable Selection

Variable selection, also called subset selection, is the search for a subset of predictor variables that can be deleted with little loss of information if n/p is large. Consider the 1D regression model where $Y \perp\!\!\!\perp \boldsymbol{x}|SP$ where $SP = \boldsymbol{x}^T\boldsymbol{\beta}$. See Definition 2.2. A *model for variable selection* can be described by

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S \tag{2.4}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ is a $p \times 1$ vector of predictors, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \mathbf{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model.

Since S is unknown, candidate subsets will be examined. Let $\boldsymbol{x}_I$ be the vector of a terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Then

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

Suppose that S is a subset of I and that model (2.4) holds. Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T\mathbf{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \mathbf{0}$ and the sample correlation $\text{corr}(\boldsymbol{x}_i^T\boldsymbol{\beta}, \boldsymbol{x}_{I,i}^T\boldsymbol{\beta}_I) = 1.0$ for the population model if $S \subseteq I$. The estimated sufficient predictor (ESP) is $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}$, and *a submodel I is worth considering if the correlation* $\text{corr}(ESP, ESP(I)) \geq 0.95$.

Definition 2.7. The model $Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta}$ that uses all of the predictors is called the *full model.* A model $Y \perp\!\!\!\perp \boldsymbol{x}_I|\boldsymbol{x}_I^T\boldsymbol{\beta}_I$ that uses a subset $\boldsymbol{x}_I$ of the predictors is called a *submodel.* The **full model is always a submodel**. The full model has *sufficient predictor* $SP = \boldsymbol{x}^T\boldsymbol{\beta}$ and the submodel has $SP = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$. Underfitting occurs if submodel I does not contain S. Fitting unnecessary predictors is sometimes called *fitting noise* or *overfitting.*

Definition 2.8. An **EE plot** for variable selection is a plot of $ESP(I)$ versus ESP where $ESP(I) = \hat{\boldsymbol{\beta}}_I^T\boldsymbol{x}_I$ and $ESP = \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$.

Forward selection or backward elimination with the Akaike (1973) AIC criterion or Schwarz (1978) BIC criterion are often used for variable selection. The relaxed lasso or relaxed elastic net estimator fits the regression method, such as a Cox (1972) proportional hazards regression, to the predictors than had nonzero lasso or elastic net coefficients. Underfitting occurs if submodel I does not contain S: a PH model may not hold for submodel I even if the PH model does hold for the full model.

Variable selection is closely related to the change in PLR test for a reduced model. You are seeking a subset I of the variables to keep in the model. The $AIC(I)$ statistic is used as an aid in backward elimination and forward selection. The full model and the model I_{min} with the smallest AIC (among models considered) are always of interest. Create a full model. The full model has a -2 log(L) at least as small as that of any submodel.

Backward elimination starts with the full model with p variables and the predictor that optimizes some criterion is deleted. Then there are $p-1$ variables left and the predictor that optimizes some criterion is deleted. This process continues for models with $p-2, p-3, ..., 3$ and 2 predictors.

Forward selection starts with the model with 0 variables and the predictor that optimizes some criterion is added. Then there is p variable in the model and the predictor that optimizes some criterion is added. This process continues for models with $2, 3, ..., p-2$ and $p-1$ predictors. Both

forward selection and backward elimination result in a sequence of p models $\{x_1^*\}, \{x_1^*, x_2^*\}, ..., \{x_1^*, x_2^*, ..., x_{p-1}^*\}, \{x_1^*, x_2^*, ..., x_p^*\}$ = full model.

Consider models I with a_I predictors. Often the criterion is the minimum value of $-2\log(L(\hat{\boldsymbol{\beta}}_I))$ or the minimum AIC(I) $= -2\log(L(\hat{\boldsymbol{\beta}}_I)) + 2a_I$. For forward selection and backward elimination, these two criterion generate the same sequence of models if each variable has 1 degree of freedom (no factors with more than 2 levels since a factor with $K \geq 2$ levels uses $K-1$ indicator variables with df $= K-1$). To see this, let model I_i have i predictors $\{x_1^*, ..., x_i^*\}$ with $a_{I_i} = i$. Forward selection moves from I_{i-1} to I_i while backward elimination moves from I_{i+1} to I_i, but all models I being considered for I_i have i predictors with $a_{I_i} = i$ a constant.

Heuristically, backward elimination tries to delete the variable that will increase the -2 log(L) the least. An increase in -2 log(L) greater than 4 (if the predictor has 1 degree of freedom) may be troubling in that a good predictor may have been deleted. In practice, the backward elimination program may delete the variable such that the submodel I with i predictors has 1) the smallest AIC(I), 2) the smallest $-2\log(L(\hat{\boldsymbol{\beta}}_I))$ or 3) the biggest p–value (preferably from a change in PLR test but possibly from a Wald test) in the test Ho $\beta_j = 0$ versus H_A $\beta_j \neq 0$ where the current model with $i+1$ variables is treated as the full model.

Heuristically, forward selection tries to add the variable that will decrease the -2 log(L) the most. An decrease in -2 log(L) less than 4 (if the predictor has 1 degree of freedom) may be troubling in that a bad predictor may have been added. In practice, the forward selection program may add the variable such that the submodel I with i predictors has 1) the smallest AIC(I), 2) the smallest $-2\log(L(\hat{\boldsymbol{\beta}}_I))$ or 3) the smallest p–value (preferably from a change in PLR test but possibly from a Wald test) in the test Ho $\beta_i = 0$ versus $H_A \beta_i \neq 0$ where the current model with $i-1$ terms plus the predictor x_j is treated as the full model (for all variables x_j not yet in the model).

Rule of thumb: a) If an interaction (e.g. $x_3x_7x_9$) is in the submodel, then the main effects (x_3, x_7, and x_9) should be in the submodel.

b) If $x_{i+1}, x_{i+2}, ..., x_{i+K-1}$ are the $K-1$ indictor variables corresponding to factor A, submodel I should either contain none or all of the $K-1$ indicator variables.

Given a list of submodels along with the number of predictors and AIC, be able to find the "initial submodel to examine" I_I. Let I_{min} be the minimum AIC model. Then I_I is the submodel with the fewest predictors such that $AIC(I_I) \leq AIC(I_{min}) + 2$. It is possible that $I_I = I_{min} = I_{full}$. Also look at submodels I with fewer predictors than I_I such that $AIC(I) \leq AIC(I_{min}) + 7$.

Submodels I with more predictors than I_I should not be used.

Submodels I with $AIC(I) > AIC(I_{min}) + 7$ should not be used.

Assume $n > 5p$, that the full PH model is reasonable and all predictors are equally important. The following rules of thumb for a good PH submodel I are in roughly decreasing order of importance.
i) Do not use more predictors than I_I.
ii) The slice survival plots for I looks like the slice survival plot for the full model.
iii) corr(ESP,ESP(I))$\geq$ 0.95.
iv) The plotted points in the EE plot of ESP(I) vs. ESP cluster tightly about the identity line.
v) Want p-value ≥ 0.01 for the change in PLR test that uses I as the reduced model. (So for variable selection use $\delta = 0.01$ instead of $\delta = 0.05$.)
vi) Want the number of predictors $a_I \leq n/10$.
vii) Want $-2\log(L(\hat{\boldsymbol{\beta}}_I)) \geq -2\log(L(\hat{\boldsymbol{\beta}}_{full}))$ but close.
viii) Want $AIC(I) \leq AIC(I_{min}) + 7$.
ix) Want hardly any predictors with p-values > 0.05.
x) Want few predictors with p-values between 0.01 and 0.05.

But for factors with $K - 1$ indicators, modify ix) and x) so that the indicator with the smallest p-value is examined.

Suppose that the full model is good and is stored in M1. Let M2, M3, M4, and M5 be candidate submodels found after forward selection, backward elimination, etc. Typically one of the submodels is the min(AIC) model. Given a list of properties of each submodel, be able to pick out "good submodels."
Tips: i) submodels with more predictors then I_I have too many predictors.
ii) The initial submodel to look at is I_I which has $AIC(I_I) \leq AIC(I_{min}) + 2$.
iii) Submodels I with $AIC(I) > AIC(I_{min}) + 7$ are not good submodels.
iv) Submodels I with a pvalue < 0.01 for the change in PLR test have too few predictors.
v) The full model I_{full} may be the best starting submodel if $I_{full} = I_I$ and M2–M5 satisfy iii). Similarly, the min(AIC) model I_{min} may be the best starting submodel if $I_{min} = I_I$ and models with fewer predictors satisfy iii).
vi) Submodels I with fewer predictors than I_I and $AIC(I) \leq AIC(I_{min}) + 7$ are worth considering. For fixed a, take the candidate that minimizes AIC.

Example 2.9. Given a list of variables with their AIC, be able to find I_I, I_{min}, and candidate submodels. The list below comes from Collett (2003, p. 86). For this list, $I_{min} = I_I = \{size, index\}$ since the model I with the fewest predictors $a_I \leq 2 = a_{I_{min}}$ and smallest $AIC(I) \leq AIC(I_{min}) + 2 = 29.533$ is $I_I = I_{min}$. A candidate submodel is $I = \{size\}$ since $AIC(I) = 31.042 \leq AIC(I_{min}) + 7 = 34.533$ and $a_I = 1 < a_{I_{min}}$. This model also has the smallest AIC for models with $a = 1$. Note that there are four models with $a = 1$, six with $a = 2$, four with $a = 3$ and one with $a = 1$. For each value of a, the model with the lowest $-2logL$ is also the one with the lowest AIC. Note that adding predictors does not increase $-2\log L$.

```
variables           -2 log L    AIC= -2 log L + 2a
none                 36.349     36.349
age                  36.269     38.269
shb                  36.196     38.196
size                 29.042     31.042 candidate
index                29.127     31.127
age,shb              36.151     40.151
age,size             28.854     32.854
age,index            28.760     32.760
shb,size             29.019     33.019
shb,index            27.981     31.981
size,index           23.533     27.533 Imin= I_I
age,shb,size         28.852     34.853
age,shb,index        27.893     33.893
age,size,index       23.269     29.269
shb,size,index       23.508     29.508
age,shb,size,index   23.231     31.231
```

Example 2.10. Given summaries on several models, be able to pick out the "best starting model" I_I. In the table below, M1 is the full model and M3 is the minimum AIC model I_{min}. M2 and M2 have more predictors than the minimum AIC model and the AIC for M4 is to large to be the starting model. So use M3 as the starting model.

If M4 has $-2logL = 27.042$, $AIC = 29.042$ and p-value $= 0.283$, then M4 would be the starting value. Any model p-value < 0.01 in the last row has a p-value that is too small.

	M1	M2	M3	M4
# of predictors	4	3	2	1
# with $0.01 \leq$ p-value ≤ 0.05	1	2	1	0
# with p-value > 0.05	2	1	0	0
$-2\log(L)$	23.231	23.269	23.533	29.042
$AIC(I)$	31.231	29.269	27.533	31.042
p-value for change in PLR test	1.0	0.8454	0.8598	0.12

If there are important predictors such as treatment that must be in the submodel, either force the variable selection procedures to contain the important predictors or do variable selection on the less important predictors and then add the important predictors to the submodel.

Suppose the PH model contains $x_1, ..., x_p$. Leave out x_j, find the martingale residuals $r_{m(j)}$, plot x_j vs $r_{m(j)}$ and add the lowess or loess curve. If the curve is linear then x_j has the correct functional form. If the curve looks like $t(x_j)$ (e.g. $(x_j)^2$), then replace x_j by $t(x_j)$, find the martingale residuals, plot $t(x_j)$ vs the residuals and check that the loess curve is linear.

Warning: A common mistake is to act as if the variable selection model I_{min} as the reduced model and to use inference for the reduced model. This type of inference is not valid: the pvalue for the change in PLRT that used $x_{I_{min}}$ as the reduced model is too high and the pvalues for $H_0 : \beta_i = 0$ are too small if x_i is a variable in I_{min}. A reduced model needs to be chosen before looking at the data. The variable selection model fits the data a bit to well since many submodels are examined. Chapter 5 will explain how to do inference after variable selection.

Lasso also does variable selection. Below is *R* code for backward elimination, forward selection, and lasso for the Lawless (1982, p. 286) `alung` data.

```
source("http://parker.ad.siu.edu/Olive/survdata.txt")
library(MASS)
library(survival)

alung<-as.data.frame(alung)
zc <- coxph(Surv(alung[,1],alung[,2])~perf+age+ttoent+
size+type+ttype+trt,data=alung)
outb<-stepAIC(zc) #default is backward

fit1 <- coxph(Surv(time,status) ~ ., data=alung)
fit2 <- coxph(Surv(time,status) ~ 1, data=alung)
#fit1 <- coxph(Surv(alung[,1],alung[,2]) .,data=alung)
#fails because it uses time and status as predictors
outb<-stepAIC(fit1,direction="backward")
Start:  AIC=189.22
Surv(time, status) ~ perf + age + ttoent + size + type + ttype +
    trt

         Df    AIC
- type    1 187.22
- ttoent  1 187.22
- age     1 187.63
- size    1 187.78
- trt     1 188.21
<none>      189.22
- ttype   1 190.28
- perf    1 206.73

Step:  AIC=187.22
Surv(time, status) ~ perf + age + ttoent + size + ttype + trt

         Df    AIC
- ttoent  1 185.22
```

```
- age     1 185.63
- size    1 185.87
- trt     1 186.22
<none>      187.22
- ttype   1 188.71
- perf    1 204.93

Step:  AIC=185.22
Surv(time, status) ~ perf + age + size + ttype + trt

        Df    AIC
- age    1 183.63
- size   1 184.00
- trt    1 184.29
<none>     185.22
- ttype  1 186.79
- perf   1 205.16

Step:  AIC=183.63
Surv(time, status) ~ perf + size + ttype + trt

        Df    AIC
- trt    1 182.41
- size   1 182.53
<none>     183.63
- ttype  1 184.92
- perf   1 203.18

Step:  AIC=182.41
Surv(time, status) ~ perf + size + ttype

        Df    AIC
- size   1 181.52
<none>     182.41
- ttype  1 183.27
- perf   1 203.14

Step:  AIC=181.52
Surv(time, status) ~ perf + ttype

        Df    AIC
<none>     181.52
- ttype  1 183.12
- perf   1 203.16
```

```
#Imin has perf and ttype

outf<-stepAIC(fit2,direction="forward",scope=
list(upper=fit1,lower=fit2))
Start:  AIC=204.69
Surv(time, status) ~ 1

         Df    AIC
+ perf    1 183.12
+ ttype   1 203.16
+ size    1 203.58
<none>      204.69
+ type    1 205.09
+ ttoent  1 205.30
+ trt     1 205.45
+ age     1 206.50

Step:  AIC=183.12
Surv(time, status) ~ perf

         Df    AIC
+ ttype   1 181.52
<none>      183.12
+ size    1 183.27
+ trt     1 184.59
+ ttoent  1 185.09
+ age     1 185.10
+ type    1 185.12

Step:  AIC=181.52
Surv(time, status) ~ perf + ttype

         Df    AIC
<none>      181.52
+ size    1 182.41
+ trt     1 182.53
+ type    1 183.28
+ age     1 183.41
+ ttoent  1 183.52


#The following code also works.
fit1 <- zc  #full model
fit2 <- coxph(Surv(alung[,1],alung[,2])~ 1,data=alung) #null model
#fit2 <- coxph(Surv(alung[,1],alung[,2])~ NULL,data=alung) #null model
outb<-stepAIC(fit1,direction="backward")
```

```
outf<-stepAIC(fit2,direction="forward",scope=list(upper=fit1,lower=fit2))
outboth<-stepAIC(fit2,direction="both",scope=list(upper=fit1,lower=fit2))

library(glmnet)
y <- as.matrix(alung[,1:2])
x <- as.matrix(alung[,3:9])
outlasso<-cv.glmnet(x,y,family="cox")
lam <- outlasso$lambda.min
betahat <- as.vector(predict(outlasso,
type="coefficients",s=lam))
betahat
-0.04331  0.0  0.0 -0.09863  0.0  0.43485 0.0
#perf, size, ttype have nonzero lasso coefficients
```

2.5 Stratified Proportional Hazards Regression

Definition 2.9. The **stratified proportional hazards regression (SPH) model** is

$$h_{\boldsymbol{x},j}(t) = h_{Y_i|\boldsymbol{x},j}(t) = h_{Y_i|\boldsymbol{\beta}'\boldsymbol{x}_i}(t) = \exp(\boldsymbol{\beta}'\boldsymbol{x}_i)h_{0,j}(t)$$

where $h_{0,j}(t)$ is the **unknown baseline function** for the jth stratum, $j = 1, ..., J$ where $J \geq 2$.

A SPH model is not a PH model, but a PH model is fit to each of the J strata. The same $\boldsymbol{\beta}$ is used for each group = stratum, but the baseline hazard functions differ. Stratification can be useful if there are clusters of cases such that the observations within the clusters are not independent. A common example is the variable *study sites* and the stratification should be on site. For example, the sites could be hospitals where the hospitals are fixed by the design of the study, rather than being a random sample of sites (hospitals). Sometimes stratification is done on a categorical variable such as gender. Sometimes stratification is done on a continuous varaible by grouping the variable and using the groups as strata. For example, use low, medium and high incomes as the strata for the variable income.

Inference is done almost exactly as done for the PH model. Except the conclusion is changed slightly: replace "PH" by "SPH".

Let A be a categorical variable with the J levels corresponding to the J groups for the SPH model. This categorical variable is not included as a predictor variable for the SPH model. A Cox PH regression model would use $J - 1$ indicator variables as predictor variable for a categorical variable included in the Cox PH regression.

Since J Cox PH regression models are fit for SPH , one for each group, check each Cox PH model with graphs. Another useful method is to divide

the ESP $\hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ into k groups where $4 \leq k \leq 9$. Choose an $\boldsymbol{x}_i$ from near the center of each group. Then plot t versus $\hat{S}_{\boldsymbol{x}_i,j}(t)$ for $j = 1, ..., J$ on the same graph for $\boldsymbol{x}_i$. Make such graphs for $\boldsymbol{x}_1, ..., \boldsymbol{x}_k$.

2.6 Generalized Cox Regression

In the Cox PH regression model, the predictors x_j are not allowed to depend on time.

Definition 2.10. In the *generalized Cox regression (GCR)* model, the predictors $x_j(t)$ do depend on time for at least one j. These predictors are called *time dependent variables*. Let $\boldsymbol{x}_i(t) = (x_{i1}(t), ..., x_{ip}(t))^T$. If x_j is not a time dependent variable, then interpret $x_j(t) \equiv x_j(0) = x_j$. Then $x_{ij}(t) \equiv x_{ij}(0)$. Then the generalized Cox regression model has

$$h_{Y|\boldsymbol{\beta}^T \boldsymbol{x}_i(t)} = h_i(t) = h_{\boldsymbol{x}(t)}(t) = \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i(t)) h_0(t).$$

The GCR model is not a PH model, but $h_0(t)$ is still the baseline function. Note that $\boldsymbol{\beta}$ does not depend on t. If subjects can have $\boldsymbol{x}_i(t) \equiv \boldsymbol{x}_i(0) = 0$ $\forall t > 0$, so that the subject's predictor variables are 0 at the time of the origin and remain at 0 regardless of the time $t > 0$, then $h_0(t)$ is the hazard function for such subjects.

Note that

$$\frac{h_i(t)}{h_0(t)} = \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i(t))$$

depends on time. Also $h_i(t) \neq c\ h_0(t)$ for some constant c that does not depend on time. These results again show that the GCR model is not a PH model.

Often patients are monitored for the duration of the study, and some variables are recorded on a regular basis. Some examples are size of tumor, PSA levels for prostate cancel, white blood cell count, and weight. If $x_j(t)$ is the value of x_j measured at time t, the time t is the study time, not the calendar time. Hence if subject 1 began on May 1 and subject 2 on July 1, and both are measured weekly, then the time in days will be 7, 14, 21, ...

There are two types of time dependent variables. An *internal time dependent variable* is suject specific and requires the subject to be under periodic observation. An *external time dependent variable* does not require the subject to be under direct observation, and often only needs one initial measurement. For example, if the patient's birthdate is known, then the patient's age can be computed at any time after the patient enters the study.

Example 2.11.

presence of side effect	internal
$x_j * \log(time)$ interaction	external
age measured yearly	external
environmental variables such as pollen count	internal
serum cholesterol level measured monthly	internal
white blood cell count measured monthly	internal

Know: Inference is almost the same as that for the Cox PH regression model, but in the conclusions, replace "PH" by "GCR."

Data management and computing the GCR model is much more difficult than that for the Cox PH model. For the GCR model, $x_j(t)$ needs to be known for "all individuals" who are in the risk set at time t_i for $i = 1, ..., m$ if there are m distinct death times, or there are missing values.

One type of time dependent covariate that is easy to work with is an interaction like $x_j * time$ or $x_j * \log(time)$. As an application, suppose a Cox PH model is fit with predictor variables $x_1, ..., x_p$. To test the Cox PH assumption, add the variables $x_1 * \log(time), ..., x_p * \log(time)$, and fit a GCR model. Want the pvalues for the interactions to be larger than 0.05. This procedure uses multiple testing. So if $p = 20$, $\beta_{p+i} = 0$ is the coefficient for $x_i * \log(time)$ for $i = 1, ..., 20$, then about 1 in 20 will have pvalue < 0.05.

2.7 Summary

1) The **Cox proportional hazards regression (PH) model** is

$$h_i(t) = h_{Y_i|\boldsymbol{x}_i}(t) = h_{Y_i|\boldsymbol{\beta}^T\boldsymbol{x}_i}(t) = \exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)h_0(t)$$

where $h_0(t)$ is the **unknown baseline function** and $\exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)$ is the **hazard ratio**.

For now, assume that the PH model is appropriate, although this assumption should be checked before performing inference.

2) The sufficient predictor $\mathbf{SP} = \boldsymbol{\beta}^T\boldsymbol{x}_j = \sum_{i=1}^p \beta_i x_{ij}$.

variable	Est.	SE	Est./SE	or $(Est/SE)^2$	pvalue for
x_1	$\hat\beta_1$	$se(\hat\beta_1)$	$z_{o,1} = \hat\beta_1/se(\hat\beta_1)$	$X^2_{o,1} = z^2_{o,1}$	Ho $\beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat\beta_p$	$se(\hat\beta_p)$	$z_{o,p} = \hat\beta_p/se(\hat\beta_p)$	$X^2_{o,p} = z^2_{o,p}$	Ho $\beta_p = 0$

```
SAS                                  Wald          pr >
variable  df Estimate  SE             chi square    chisqu
```

```
age       1  0.1615    0.0499 10.4652       0.0012
ecog.ps   1  0.0187    0.5991 0.00097       0.9800

R         coef  exp(coef) se(coef)    z      p
age     0.1615      1.18   0.0499 3.2350 0.0012
ecog.ps 0.0187      1.02   0.5991 0.0312 0.9800

Likelihood ratio test=14.3  on 2 df, p=0.000787  n= 26
```

Shown above is output in symbols from and *SAS* and *R*. The estimated coefficient is $\hat{\beta}_j$. The Wald chi square $= X^2_{o,j}$ while p and "pr > chisqu" are both p-values.

3) The estimated sufficient predictor $\mathbf{ESP} = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_j = \sum_{i=1}^p \hat{\beta}_i x_{ij}$. Given $\hat{\boldsymbol{\beta}}$ from output and given $\boldsymbol{x}$, be able to find ESP and $\hat{h}_i(t) = \exp(ESP)\hat{h}_0(t) = \exp(\hat{\boldsymbol{\beta}}^T \boldsymbol{x})\hat{h}_o(t)$ where $\exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x})$ is the **estimated hazard ratio**.

For tests, the p–value is an important quantity. Recall that H_o is rejected if the p–value $< \delta$. A p–value between 0.07 and 1.0 provides little evidence that H_o should be rejected, a p–value between 0.01 and 0.07 provides moderate evidence and a p–value less than 0.01 provides strong statistical evidence that H_o should be rejected. Statistical evidence is not necessarily practical evidence, and reporting the p–value along with a statement of the strength of the evidence is more informative than stating that the p–value is less than some chosen value such as $\delta = 0.05$. Nevertheless, as a **homework convention**, use $\delta = 0.05$ if δ is not given.

4) The Wald confidence interval (CI) for β_j can also be obtained from the output: the large sample 95% CI for β_j is

$$\hat{\beta}_j \pm 1.96\ se(\hat{\beta}_j).$$

5) Investigators also sometimes test whether a predictor x_j is needed in the model given that the other $k-1$ nontrivial predictors are in the model with a **4 step Wald test of hypotheses**:
i) State the hypotheses Ho: $\beta_j = 0$ Ha: $\beta_j \neq 0$.
ii) Find the test statistic $z_{o,j} = \hat{\beta}_j / se(\hat{\beta}_j)$ or $X^2_{o,j} = z^2_{o,j}$ or obtain it from output.
iii) The p–value $= 2P(Z < -|z_{oj}|) = P(\chi^2_1 > X^2_{o,j})$. Find the p–value from output or use the standard normal table.
iv) State whether you reject Ho or fail to reject Ho and give a nontechnical sentence restating your conclusion in terms of the story problem.

If Ho is rejected, then conclude that x_j is needed in the PH survival model given that the other $p-1$ predictors are in the model. If you fail to reject Ho, then conclude the values of x_j do not affect the PH survival model given that the other $p-1$ predictors are in the model. (Or state that there is not enough

evidence to conclude that the values of x_j affect the PH survival model.) Note that x_j could be a very useful PH survival predictor, but the observed values of x_j may not affect survival or x_j may not be needed if other predictors are added to the model.

For a PH, often 3 models are of interest: the **full model** that uses all p of the predictors $\boldsymbol{x}^T = (\boldsymbol{x}_R^T, \boldsymbol{x}_O^T)$, the **reduced model** that uses the r predictors $\boldsymbol{x}_R$, and the **null model** that uses none of the predictors.

The partial likelihood ratio test (**PLRT**) is used to test whether $\boldsymbol{\beta} = \mathbf{0}$. If this is the case, then the predictors are not needed in the PH model (so survival times $Y \perp\!\!\!\perp \boldsymbol{x}$). If $H_o : \boldsymbol{\beta} = \mathbf{0}$ is not rejected, then the Kaplan Meier estimator should be used. If H_o is rejected, use the PH model.

6) The 4 step **PLRT** is

i) $H_o : \boldsymbol{\beta} = \mathbf{0} \quad H_A : \boldsymbol{\beta} \neq \mathbf{0}$

ii) test statistic $X^2(N|F) = [-2 \log L(none)] - [-2 \log L(full)]$ is often obtained from output.

iii) The p–value $= P(\chi_p^2 > X^2(N|F))$ where χ_p^2 has a chi–square distribution with p degrees of freedom. The p–value is often obtained from output.

iv) Reject H_o if the p–value $< \delta$ and conclude that there is a PH survival relationship between Y and the predictors $\boldsymbol{x}$. If p–value $\geq \delta$, then fail to reject H_0 and conclude that the values of the predictors $\boldsymbol{x}$ do not (significantly) affect the PH survival model. (Or state that there is not enough evidence to conclude that the values of $\boldsymbol{x}$ affect the PH survival model.)

Some *SAS* output for the PLRT is shown next. *R* output is above 20).

```
SAS Testing Global Null Hypotheses: BETA = 0
           without     with
criterion  covariates covariates  model Chi-square
-2 LOG L   596.651     551.1888   45.463 with 3 DF (p=0.0001)
```

Let the **full model** be

$$SP = \beta_1 x_1 + \cdots + \beta_p x_p = \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_R^T \boldsymbol{x}_R + \boldsymbol{\beta}_O^T \boldsymbol{x}_O.$$

let the **reduced model**

$$SP = \beta_{R1} x_{R1} + \cdots + \beta_{Rr} x_{Rr} = \boldsymbol{\beta}_R^T \boldsymbol{x}_R$$

where the reduced model uses r of the predictors used by the full model and $\boldsymbol{x}_O$ denotes the vector of $p - r$ predictors that are in the full model but not the reduced model.

Assume that the full model is useful. Then we want to test H_o: the reduced model is good (can be used instead of the full model, so $\boldsymbol{x}_O$ is not needed in the model given $\boldsymbol{x}_R$ is in the model) versus H_A: use the full model (the full model is significantly better than the reduced model). Fit the full model and the reduced model to get $X^2(N|F)$ and $X^2(N|R)$ where $X^2(N|F)$ is used in

the PLRT to test whether $\boldsymbol{\beta} = \mathbf{0}$ and $X^2(N|R)$ is used in the PLRT to test whether $\boldsymbol{\beta}_R = \mathbf{0}$ (treating the reduced model as the model in the PLRT).

variable	Est.	SE	Est/SE	or $(Est/SE)^2$	pvalue for
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	$X^2_{o,1} = z^2_{o,1}$	Ho $\beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$z_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	$X^2_{o,p} = z^2_{o,p}$	Ho $\beta_p = 0$

R: Likelihood ratio test $= X^2(N|F)$ on p df

```
SAS: Testing Global Null Hypotheses: BETA = 0
Test                Chi-Square       DF      Pr > Chisq
```

Likelihood ratio $X^2(N|F)$ p pval for Ho: $\boldsymbol{\beta} = \mathbf{0}$

variable	Est.	SE	Est/SE	or $(Est/SE)^2$	pvalue for
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	$X^2_{o,1} = z^2_{o,1}$	Ho: $\beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	$\hat{\beta}_r$	$se(\hat{\beta}_r)$	$z_{o,r} = \hat{\beta}_r/se(\hat{\beta}_r)$	$X^2_{o,r} = z^2_{o,r}$	Ho: $\beta_r = 0$

R: Likelihood ratio test $= X^2(N|R)$ on r df

```
SAS: Testing Global Null Hypotheses: BETA = 0
Test                Chi-Square       DF      Pr > Chisq
```

Likelihood ratio $X^2(N|R)$ r pval for Ho: $\boldsymbol{\beta}_R = \mathbf{0}$

The output shown above in symbols, can be used to perform the change in PLR test. For simplicity, the reduced model used in the output is $\boldsymbol{x}_R = (x_1, ..., x_r)^T$.

Notice that $X^2(R|F) \equiv X^2(N|F) - X^2(N|R) =$

$$[-2\log L(none)] - [-2\log L(full)] - ([-2\log L(none)] - [-2\log L(red)]) =$$

$$[-2\log L(red)] - [-2\log L(full)] = -2\log\left(\frac{L(red)}{L(full)}\right).$$

7) The 4 step **change in PLR test** is

i) H_o: the reduced model is good H_A: use the full model

ii) test statistic $X^2(R|F) = X^2(N|F) - X^2(N|R) = [-2\log L(red)] - [-2\log L(full)]$.

iii) The p–value $= P(\chi^2_{p-r} > X^2(R|F))$ where χ^2_{p-r} has a chi–square distribution with $p-r$ degrees of freedom.

iv) Reject H_o if the p–value $< \delta$ and conclude that the full model should be used. If p–value $\geq \delta$, then fail to reject H_0 and conclude that the reduced model is good (the values of $\boldsymbol{x}_O$ do not affect the survival model, or there is not enough evidence to conclude that the values of $\boldsymbol{x}_O$ affect the survival model).

If the reduced model leaves out a single variable x_i, then the change in PLR test becomes $H_o : \beta_i = 0$ versus $H_A : \beta_i \neq 0$. This change in partial likelihood ratio test is a competitor of the Wald test. The change in PLRT is usually better than the Wald test if the sample size n is not large, but the Wald test is currently easier for software to produce. For large n the test statistics from the two tests tend to be very similar (asymptotically equivalent tests).

8) If the reduced model is good, then the **EE plot** of $ESP(R) = \hat{\boldsymbol{\beta}}_R^T \boldsymbol{x}_{Ri}$ versus $ESP = \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ should be highly correlated with the identity line with unit slope and zero intercept.

A **factor** A is a variable that takes on a categories called levels. Suppose A has a categories $c_1, ..., c_a$. Then the factor is incorporated into the PH model by using $a-1$ indicator variables $x_{jA} = 1$ if $A = c_j$ and $x_{Aj} = 0$ otherwise, where the 1st indicator variable is omitted, eg, use $x_{2A}, ..., x_{aA}$. Each indicator has 1 degree of freedom. Hence the degrees of freedom of the $a-1$ indicator variables associated with the factor is $a-1$.

The x_j corresponding to variates (variables that take on numerical values) or to indicator variables from a factor are called **main effects**.

An **interaction** is a product of two or more main effects, but for a factor include products for all indicator variables of the factor.

If an interaction is in the model, also include the corresponding main effects. For example, if $x_1 x_3$ is in the model, also include the main effects x_1 and x_2.

A **scatterplot** is a plot of x_i vs. x_j. A **scatterplot matrix** is an array of scatterplots. It is used to examine the marginal relationships of the predictors. Variables with outliers, missing values or strong nonlinearities may be so bad that they should not be included in the full model.

9) Suppose that all values of the variable x are positive. The **log rule** says add $\log(x)$ to the full model if $\max(x_i)/\min(x_i) > 10$.

Variable selection is closely related to the change in PLR test for a reduced model. You are seeking a subset I of the variables to keep in the model. The $AIC(I)$ statistic is used as an aid in backward elimination and forward selection. The full model and the model with the smallest AIC are

always of interest. Create a full model. The full model has a -2 log(L) at least as small as that of any submodel. The full model is a submodel.

Backward elimination starts with the full model with p variables and the predictor that optimizes some criterion is deleted. Then there are $p-1$ variables left and the predictor that optimizes some criterion is deleted. This process continues for models with $p-2, p-3, ..., 3$ and 2 predictors.

Forward selection starts with the model with 0 variables and the predictor that optimizes some criterion is added. Then there is p variable in the model and the predictor that optimizes some criterion is added. This process continues for models with $2, 3, ..., p-2$ and $p-1$ predictors. Both forward selection and backward elimination result in a sequence of p models $\{x_1^*\}, \{x_1^*, x_2^*\}, ..., \{x_1^*, x_2^*, ..., x_{p-1}^*\}, \{x_1^*, x_2^*, ..., x_p^*\}$ = full model.

Consider models I with r_I predictors. Often the criterion is the minimum value of $-2\log(L(\hat{\boldsymbol{\beta}}_I))$ or the minimum AIC(I) $= -2\log(L(\hat{\boldsymbol{\beta}}_I)) + 2r_I$.

Heuristically, backward elimination tries to delete the variable that will increase the -2 log(L) the least. An increase in -2 log(L) greater than 4 (if the predictor has 1 degree of freedom) may be troubling in that a good predictor may have been deleted. In practice, the backward elimination program may delete the variable such that the submodel I with k predictors has 1) the smallest AIC(I), 2) the smallest $-2\log(L(\hat{\boldsymbol{\beta}}_I))$ or 3) the biggest p–value (preferably from a change in PLR test but possibly from a Wald test) in the test Ho $\beta_i = 0$ versus H_A $\beta_i \neq 0$ where the current model with $k+1$ variables is treated as the full model.

Heuristically, forward selection tries to add the variable that will decrease the -2 log(L) the most. An decrease in -2 log(L) less than 4 (if the predictor has 1 degree of freedom) may be troubling in that a bad predictor may have been added. In practice, the forward selection program may add the variable such that the submodel I with k predictors has 1) the smallest AIC(I), 2) the smallest $-2\log(L(\hat{\boldsymbol{\beta}}_I))$ or 3) the smallest p–value (preferably from a change in PLR test but possibly from a Wald test) in the test Ho $\beta_i = 0$ versus $H_A \beta_i \neq 0$ where the current model with $k-1$ terms plus the predictor x_i is treated as the full model (for all variables x_i not yet in the model).

10) If an interaction (e.g. $x_3x_7x_9$) is in the submodel, then the main effects (x_3, x_7, and x_9) should be in the submodel.

11) If $x_{i+1}, x_{i+2}, ..., x_{i+a-1}$ are the $a-1$ indictor variables corresponding to factor A, submodel I should either contain none or all of the $a-1$ indictor variables.

12) Given a list of submodels along with the number of predictors and AIC, be able to find the "initial submodel to examine" I_I. Let I_{min} be the

minimum AIC model. Then I_I is the submodel with the fewest predictors such that $AIC(I_I) \leq AIC(I_{min}) + 2$. It is possible that $I_I = I_{min} = I_{full}$. Also look at submodels I with fewer predictors than I_I such that $AIC(I) \leq AIC(I_{min}) + 7$.

13) Submodels I with more predictors than I_I should not be used.

14) Submodels I with $AIC(I) > AIC(I_{min}) + 7$ should not be used.

15) Let the survival times $T_i = min(Y_i, Z_i)$, and let $\gamma_i = 1$ if $T_i = Y_i$ (uncensored) and $\gamma_i = 0$ if $T_i = Z_i$ (censored). For PH models, an **censored response plot** is a plot of the ESP vs T with plotting symbol 0 for censored cases and + for uncensored cases. If the ESP is a good estimator of the SP and $h_{SP}(t) = \exp(SP)h_0(t)$, then the hazard increases and survival decreases as the ESP increases.

16) The **slice survival plot** divides the ESP into J groups of roughly the same size. For each group j, $\hat{S}_{PHj}(t)$ is computed using the $\boldsymbol{x}$ corresponding to the "median ESP" of the group. The Kaplan Meier estimator $\hat{S}_{KMj}(t)$ is computed from the survival times in the jth group. For each group, $\hat{S}_{PHj}(t)$ is plotted and $\hat{S}_{KMj}(t_i)$ as circles at the deaths t_i. The proportional hazards assumption is reasonable if the circles track the curve well in each of the J plots. If pointwise CI bands are added to the plot, then $\hat{S}_{KMj}$ tracks $\hat{S}_{PHj}$ well if most of the plotted circles do not fall very far outside the pointwise CI bands.

17) Assume $n > 5p$, that the full PH model is reasonable and all predictors are equally important. The following rules of thumb for a good PH submodel I are in roughly decreasing order of importance.
i) Do not use more predictors than I_I.
ii) The slice survival plots for I looks like the slice survival plot for the full model.
iii) corr(ESP,ESP(I))$\geq$ 0.95.
iv) The plotted points in the EE plot of ESP(I) vs ESP cluster tightly about the identity line.
v) Want pvalue ≥ 0.01 for the change in PLR test that uses I as the reduced model. (So for variable selection use $\delta = 0.01$ instead of $\delta = 0.05$.)
vi) Want the number of predictors $r_I \leq n/10$.
vii) Want $-2\log(L(\hat{\boldsymbol{\beta}}_I)) \geq -2\log(L(\hat{\boldsymbol{\beta}}_{full}))$ but close.
viii) Want $AIC(I) \leq AIC(I_{min}) + 7$.
ix) Want hardly any predictors with pvalues > 0.05.
x) Want few predictors with pvalues between 0.01 and 0.05.

But for factors with $a-1$ indicators, modify ix) and x) so that the indicator with the smallest pvalue is examined.

18) Suppose that the full model is good and is stored in M1. Let M2, M3, M4, and M5 be candidate submodels found after forward selection, backward elimination, etc. Typically one of the submodels is the min(AIC) model. Given a list of properties of each submodel, be able to pick out "good submodels."
Tips: i) submodels with more predictors then I_I have too many predictors.
ii) The initial submodel to look at is I_I which has $AIC(I_I) \leq AIC(I_{min}) + 2$.
iii) Submodels I with $AIC(I) > AIC(I_{min}) + 7$ are not good submodels.
iv) Submodels I with a pvalue < 0.01 for the change in PLR test have too few predictors.
v) The full model I_{full} may be the best starting submodel if $I_{full} = I_I$ and M2–M5 satisfy iii). Similarly, the min(AIC) model I_{min} may be the best starting submodel if $I_{min} = I_I$ and models with fewer predictors satisfy iii).
vi) Submodels I with fewer predictors than I_I and $AIC(I) \leq AIC(I_{min}) + 7$ are worth considering.

19) If there are important predictors such as treatment that must be in the submodel, either force the variable selection procedures to contain the important predictors or do variable selection on the less important predictors and then add the important predictors to the submodel.

20) Suppose the PH model contains $x_1, ..., x_p$. Leave out x_j, find the martingale residuals $r_{m(j)}$, plot x_j vs $r_{m(j)}$ and add the lowess or loess curve. If the curve is linear then x_j has the correct functional form. If the curve looks like $t(x_j)$ (eg $(x_j)^2$), then replace x_j by $t(x_j)$, find the martingale residuals, plot $t(x_j)$ vs the residuals and check that the loess curve is linear.

21) Let the scaled Schoenfeld residual for the jth variable x_j be $r^*_{pj} + \hat{\beta}_j$. Plot the death times t_i vs the scaled residuals and add the loess curve. If the loess curve is approximately horizontal for each of the p plots, then the PH assumption is reasonable. Alternatively, fit a line to each plot and test that each of the p slopes is equal to 0. The *R* function `cox.zph` makes both the plots and tests.

22) The **stratified proportional hazards regression (SPH) model** is

$$h_{\boldsymbol{x},j}(t) = h_{Y_i|\boldsymbol{x},j}(t) = h_{Y_i|\boldsymbol{\beta}'\boldsymbol{x}_i}(t) = \exp(\boldsymbol{\beta}'\boldsymbol{x}_i)h_{0,j}(t)$$

where $h_{0,j}(t)$ is the **unknown baseline function** for the jth stratum, $j = 1, ..., J$ where $J \geq 2$.

A SPH model is not a PH model, but a PH model is fit to each of the J strata. The same $\boldsymbol{\beta}$ is used for each group = stratum, but the baseline hazard functions differ. Stratification can be useful if there are clusters of cases such that the observations within the clusters are not independent. A common example is the variable *study sites* and the stratification should be on site. Sometimes stratification is done on a categorical variable such as gender.

23) Inference is done exactly as for the PH model. See points 3), 4), 5), 6), and 7). Except the conclusion is changed slightly: in 5) and 6) replace "PH" by "SPH".

2.8 Complements

Sometimes the Cox PH regression model does not fit the data set, but there is a categorical variable A with J levels such that a Cox PH regression model fits each group corresponding to the levels of A. Then each group has a $\boldsymbol{\beta}_j$ for $j = 1, ..., J$. For example, men and women could follow a different Cox PH regression model. The stratified proportional hazards regression model is a special case where $\boldsymbol{\beta}_g \equiv \boldsymbol{\beta}$ for $j = 1, ..., J$, but the baseline hazard functions $h_{0j}(t)$ differ.

For multiple linear regression, the ANOVA F test is like the PLRT and the partial F test is like the change in PLR test.

Oakes (2000) notes that the proportional hazards model is not preserved when variables are added or deleted from the model, eg by variable selection. Any 1D regression model can be invalidated by adding or deleting variables with nonzero coefficients. Variable selection is a search for variables $\boldsymbol{x}_O$ where $\boldsymbol{x} = (\boldsymbol{x}_I^T, \boldsymbol{x}_O^T)^T$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_I^T, \boldsymbol{\beta}_O^T)$. If variable selection is successful to a useful approximation, so that $\boldsymbol{\beta}_O = \mathbf{0}$, then the 1D regression model and proportional hazards is preserved.

From the CRAN website, e.g. (https://cran.r-project.org/), click on *packages*, then *survival*, then *survival.pdf* to obtain the *R* reference manual on the *survival* package. Much of this material is also in MathSoft (1999b, Ch. 8–13).

For SAS, see the SAS Institute (1999). The chapters on PHREG, LIFEREG and LIFETEST procedures are useful. These chapters can be found online at (www.google.com) with a search of the keywords *SAS/STAT User's Guide*.

The most used survival regression models satisfy $Y \perp\!\!\!\perp \boldsymbol{x}|SP$, and the slice survival plot is useful for visualizing $S_{Y|SP}(t)$ in the background of the data. Simultaneous or pointwise CI bands are needed to determine whether the nonparametric Kaplan Meier estimator is close to the model estimator. If the two estimators are close for each slice, then the graph suggests that the model is giving a useful approximation to $S_{Y|SP}(t)$ for the observed data if the number of uncensored cases is large compared to the number of predictors p. The plots are also useful for teaching survival regression to students and for explaining the models to consulting clients.

The slice survival, censored response, LCR, and EE plots are due to Olive (2011). Emphasis was on proportional hazards models since pointwise CI bands are available for the Cox proportional hazards model. Thus the slice survival plot can be made for the Cox model, and then the estimated sur-

vival function from a parametric proportional hazards model can be added as crosses for each slice if points in the EE plot cluster tightly about the identity line. Stratified proportional hazards models can be checked by making one slice survival plot per stratum. EE plots can be made for parametric models if software for a semiparametric analog is available. For some parametric survival models, see Chapter 3, Bennett (1983), Yang and Prentice (1999), Wei (1992), and Zeng and Lin (2007).

The censored response plot and LCR plot can be regarded as special cases of the model checking plots of Cook and Weisberg (1997) applied to censored data.

If pointwise bands are not available for the parametric or semiparametric model, but the number of cases in each slice is large, then simultaneous or pointwise CI bands for the Kaplan Meier estimator could be added for each slice.

Plots were made in *R* and the function `coxph` produces the survival curves for Cox regression. The collection of *R* functions *survpack* available from (http://parker.ad.siu.edu/Olive/survpack.txt) contains functions for reproducing simulations and some of the plots. The functions `vlung2`, `vovar`, and `vnwtco` were used to produce Figures 2.1, 2.2, and 2.5. The function `bphsim3` shows that the Kaplan Meier estimator was close to the Cox survival curves for 2 groups (a single binary predictor) when censoring was light and $n = 10$.

Zhou (2001) shows how to simulate Cox proportional hazards regression data. Simulated Weibull proportional hazards regression data was made following Zhou (2001) but with three iid N(0,1) covariates. The function `phsim5` showed that for 9 groups and $p = 3$, the Kaplan Meier and Cox curves were close (with respect to the pointwise CI bands) for $n \geq 80$. The function `wphsim` showed a similar result for Kaplan Meier curves (circles), and the function `wregsim2` shows that for $n \geq 30$, the plotted points in an EE plot cluster tightly about the identity line with correlation greater than 0.99 with high probability.

2.9 Problems

Problems with an asterisk * are especially important.

2.1. Suppose that a proportional hazards model holds so that $h_{\boldsymbol{x}}(t) = \exp(\boldsymbol{\beta}^T \boldsymbol{x}) h_0(t)$ where $h_0(t)$ is the baseline hazard function. Let $f_0(t)$, $S_0(t)$ $F_0(t)$ and $H_0(t)$ denote the baseline pdf, survival function, distribution function and cumulative hazard function.

a) Show

$$H_{\boldsymbol{x}}(t) = \exp(\boldsymbol{\beta}^T \boldsymbol{x}) H_0(t).$$

b) Show

$$S_{\boldsymbol{x}}(t) = [S_0(t)]^{\exp(\boldsymbol{\beta}^T \boldsymbol{x})}.$$

c) Show

$$f_{\boldsymbol{x}}(t) = f_0(t) \exp(\boldsymbol{\beta}^T \boldsymbol{x})[S_0(t)]^{\exp(\boldsymbol{\beta}^T \boldsymbol{x}) - 1}.$$

2.2. Suppose that $h_0(t) = 1$ for $t > 0$. This corresponds to the exponential proportional hazards model $h_{\boldsymbol{x}}(t) = \exp(\boldsymbol{\beta}^T \boldsymbol{x}) h_0(t) = \exp(\boldsymbol{\beta}^T \boldsymbol{x})$.

a) Find $H_0(t)$.

b) Find $H_{\boldsymbol{x}}(t)$.

```
Data for 2.3
Variables in model         -2 log L
none                         36.349
size                         29.042
size, index                  23.533
size, index, treatment       22.572
```

2.3. The Collett (2003b, p. 86) data studies the time until death from prostate cancer from the date the patient was randomized to a treatment. The variable *treatment* was a 0 for a placebo and a 1 for DES (a drug). The variable *size* was tumor size, and *index* the Gleason index. Let the full model contain *size*, *index* and *treatment*. Use the table above.

a) If the reduced model uses *size* and *index*, test whether the reduced model is good.

b) If the reduced model uses *size*, test whether the reduced model is good.

```
 data for 2.4
full model     coef    exp(coef)  se(coef)   z      p
age           0.00318     1.003   0.0111  0.285 0.78
sex          -1.48314     0.227   0.3582 -4.140 0.000035
diseaseGN     0.08796     1.092   0.4064  0.216 0.83
diseaseAN     0.35079     1.420   0.3997  0.878 0.38
diseasePKD   -1.43111     0.239   0.6311 -2.268 0.023

Likelihood ratio test=17.6  on 5 df, p=0.00342  n= 76

reduced model  coef    exp(coef) se(coef)    z      p
age           0.00203     1.002  0.00925  0.220 0.8300
sex          -0.82931     0.436  0.29895 -2.774 0.0055

Likelihood ratio test=7.12  on 2 df, p=0.0285  n= 76
```

2.4. The *R* `kidney` data is on the recurrence times Y to infection, at the point of insertion of the catheter, for kidney patients. Predictors are *age*, *sex* (M=1,F=2), and the factor *disease* (0=GN, 1=AN, 2=PKD, 3=Other).

a) For the reduced model, test $\boldsymbol{\beta} = \mathbf{0}$.

b) For the reduced model, test $\boldsymbol{\beta} = \mathbf{0}$ **using** $\delta = 0.01$.

c) Test whether the reduced model is good.

```
 Output for 2.5
             coef exp(coef) se(coef)      z       p
rxLev     -0.0423     0.959   0.1103 -0.384 0.70000
rxLev+5FU -0.3787     0.685   0.1189 -3.186 0.00140
extent     0.4930     1.637   0.1117  4.412 0.00001
node4      0.9154     2.498   0.0968

Likelihood ratio test=122  on 4 df, p=0  n= 929
```

2.5. The *R* `colon` data from one of the first successful trials of adjuvant chemotherapy for colon cancer. Levamisole is a low-toxicity compound, 5-FU is a moderately toxic chemotherapy agent. The treatment was nothing, levamisole, or levamisole and 5-FU. Y is time until death. The 4 predictors are $x_1 = 1$ if treatment was levamisole, $x_2 = 1$ if the treatment was levamisole and 5-FU, *extent* of local spread (treated as a variate with 1=submucosa, 2=muscle, 3=serosa, 4=contiguous structures), and *node4* $= 1$ for more than 4 positive lymph nodes.

a) Find the ESP and $\hat{h}_i(t)$ if $\boldsymbol{x} = (0, 1, 2, 1)$.

b) Find a 95% CI for β_1.

c) Do a 4 step test for $Ho : \beta_1 = 0$.

d) Do a 4 step test for $Ho : \beta_4 = 0$.

```
 Output for 2.6.
full model            coef  exp(coef) se(coef)    z     p
trt                 0.295      1.343  0.20755  1.4194 0.16
celltypesmallcell  0.862      2.367  0.27528  3.1297 0.017
celltypeadeno      1.20       3.307  0.30092  3.9747 0.000
celltypelarge      0.401      1.494  0.28269  1.4196 0.16
karno             -0.0328     0.968  0.00551 -5.9580 0.000
diagtime           0.000081   1.000  0.00914  0.0089 0.99
age               -0.00871    0.991  0.00930 -0.9361 0.35
prior              0.00716    1.007  0.02323  0.3082 0.76

Likelihood ratio test=62.1  on 8 df, p=1.8e-10  n= 137

reduced model        coef exp(coef) se(coef)     z     p
trt                0.2617      1.30  0.20092  1.30 0.19
celltypesmallcell  0.8250      2.28  0.26891  3.07 0.022
celltypeadeno      1.1540      3.17  0.29504  3.91 0.0009
celltypelarge      0.3946      1.48  0.28224  1.40 0.16
karno             -0.0313      0.97  0.00517 -6.05 0.000
```

```
Likelihood ratio test=61.1  on 5 df, p=7.3e-12  n= 137
```

2.6. The *R* `veteran` lung cancer data has Y = survival time. The predictors are *trt* (1=standard, 2=test), the factor *celltype* (1=squamous, 2=small-cell, 3=adeno, 4=large), *karno* = Karnofsky performance score (100=good), *diagtime* = months from diagnosis to randomization, *age* in years, and *prior* = prior therapy (0=no, 1=yes).

a) For the full model, test Ho $\boldsymbol{\beta} = \mathbf{0}$.

b) Test whether the reduced model is good.

```
Full model   Output for 2.7
 variable    coef  std._err.  z     pval
      age   -0.029 0.008     -3.53  0.000
  bectota    0.008 0.005      1.68  0.094
  ndrugtx    0.028 0.008      3.42  0.001
  herco_2    0.065 0.150      0.44  0.663
  herco_3   -0.094 0.166     -0.57  0.572
  herco_4    0.028 0.160      0.18  0.861
   ivhx_2    0.174 0.139      1.26  0.208
   ivhx_3    0.281 0.147      1.91  0.056
     race   -0.203 0.117     -1.74  0.082
    treat   -0.240 0.094     -2.54  0.011
     site   -0.102 0.109     -0.94  0.348

Likelihood ratio test = 24.436 on 11 df, p = 0.011

Reduced model
 variable    coef  std._err.  z     pval
      age   -0.026 0.008     -3.25  0.001
  bectota    0.008 0.005      1.70  0.090
  ndrugtx    0.029 0.008      3.54  0.000
   ivhx_3    0.256 0.106      2.41  0.016
     race   -0.224 0.115     -1.95  0.051
    treat   -0.232 0.093     -2.48  0.013
     site   -0.087 0.108     -0.80  0.422

Likelihood ratio test = 21.038 on 7 df, p = 0.004
```

2.7. The Hosmer and Lemeshow (1999, p. 165 - 170) data studies time until illegal drug use relapse. Variables were *age*, *becktota*, *ndrugtx*, $herco_2$ = 1 if heroin user and 0 else, $herco_3 = 1$ if cocaine user and 0 else, $herco_4 = 1$ if used neither heroin nor cocaine and 0 else, $ivhx_2 = 1$ if previous but not recent IV drug use and 0 else, $ivhx_3 = 1$ if recent IV drug use and 0 else, *race* = 1 for white and 0 else, *treat* = 1 for short treatment and 0 for long and *site*.

Using the output for the full and reduced model above, test whether the reduced model is good.

```
output for 2.8     variables                                AIC
trt sex race pburn bhd bbut btor bupleg blowleg bresp 439.470
trt sex race pburn bhd bbut btor bupleg blowleg       437.479
trt sex race pburn     bbut btor bupleg blowleg       435.540
trt sex race pburn     bbut      bupleg blowleg       433.677
trt sex race           bbut      bupleg blowleg       431.952
trt sex race           bbut      bupleg               430.281
trt sex race           bbut                           429.617
trt sex race                                          428.708
trt     race                                          429.704
        race                                          431.795
```

2.8. Data from Klein and Moeschberger (1997, p. 7) is on severely burned patients. The response variable is *time* until infection. Predictors include *treatment* (0-routine bathing 1-Body cleansing), *sex* (0=male 1=female), *race* (0=nonwhite 1=white), *pburn* = percent of body burned. The remaining variables are burn cite indicators. For example, *bhd* is head (1 yes 0 no). Results from backward elimination are shown.

a) What is the minimum AIC submodel I_{min}?

b) What is the submodel I_I?

c) Are there any other good candidate submodels? Explain briefly.

	M1	M2	M3	M4
# of predictors	10	3	2	1
# with $0.01 \leq$ p-value ≤ 0.05	2	2	1	1
# with p-value > 0.05	8	1	0	0
$-2\log(L)$	419.470	422.708	425.704	429.795
$AIC(I)$	439.470	428.708	429.704	431.795
p-value for change in PLR test	1.0	0.862	0.304	0.325

2.9. Data from Klein and Moeschberger (1997, p. 7) is on severely burned patients. The above table gives summary statistics for 4 PH regression models considered as final submodels after performing variable selection. Assume that the PH assumptions hold for all 4 models. The full model was M1, and M2 was the minimum AIC model found. Which submodel is the initial model to examine I_I? Explain briefly why each of the other 3 submodels should not be used as the starting submodel.

2.10. Suppose that the survival times are plotted versus the scaled Schoenfeld residuals for variable x_1. Sketch the loess curve if the PH assumption is reasonable.

SAS Problems

2.11. Data is from SAS Institute (1999) and is from a study on multiple myeloma (bone cancer) in which researchers treated 65 patients with alkylating agents. The variable *Time* is the survival time in months from diagnosis. The predictor variables are *LogBUN* (blood urea nitrogen), *HGB* (hemoglobin at diagnosis), *Platelet* (platelets at diagnosis: 0=abnormal, 1=normal), *Age* at diagnosis in years, *LogWBC*, *Frac* (fractures at diagnosis: 0=none, 1=present), *LogPBM* (log percentage of plasma cells in bone marrow), *Protein* (proteinuria at diagnosis), and *SCalc* (serum calcium at diagnosis).

a) Obtain the SAS program for this problem from (http://parker.ad.siu.edu/Olive/survhw.txt).

b) First backward elimination is considered. From the SAS output window, copy and paste the output for the full model that uses all 9 variables into *Word*. That is, scroll to the top of the output and copy and paste the following output.

```
Step  0. The model contains the following variables:

LogBUN  HGB  Platelet  Age  LogWBC  Frac  LogPBM  Protein  SCalc
.
.
.
SCalc 1 0.12595  0.10340 1.4837 0.2232 1.134
```

c) At step 7 of backward elimination, the final model considered uses LogBUN and HGB. Copy and paste the output for this model (similar to the output for b) into *Word*.

d) Backward elimination will consider 8 models. Write down the variables used for each model as well as the AIC. The first two models are shown below.

```
variables                                                    AIC
LogBUN HGB Platelet Age LogWBC Frac LogPBM Protein SCalc  310.588
LogBUN HGB          Age LogWBC Frac LogPBM Protein SCalc  308.827
```

e) Repeat d) for the 4 models considered by forward selection.

f) Repeat d) for the 4 models considered by stepwise selection.

g) For all subsets selection, complete the following table.

```
variables       chisq
      2                    LogBUN HGB
      9                    full
```

h) Perform a change in PLR test if the full model uses 9 variables and the reduced model uses LogBUN and HGB. (Use the output from b) and c).)

i) Are there any other good candidate models?

SAS forward selection, backward elimination, and stepwise selection produces too much output. Only submit some of the produced output. The AIC line in the With Covariates column is important.

2.12. Data is from Allison (1995, p. 270). The response variable *week* is time in weeks until arrest after release from prison (right censored if week = 52). The 7 variables are *Fin* (1 for those who received financial aid, 0 else), *Age* at time of release, *Race* (1 if black, 0 else), *Wexp*(1 if inmate had full time work experience prior to conviction, 0 else), *Mar* (1 if married at time of release, 0 else), *Paro* (1 if released on parole, 0 else), *Prio* (the number of prior convictions).

a) Obtain the SAS program for this problem from (http://parker.ad.siu.edu/Olive/survhw.txt). To execute the program, use the top menu commands "Run>Submit". An output window will appear if successful. **Warning: if you do not have the recid.txt file on e drive, then you need to change** the *infile* command in the SAS code to the drive that you are using, e.g. change *infile "e:redic.txt";* to *infile "f:recid.txt";* if you are using the f drive.

b) Obtain the SAS program for this problem from (http://parker.ad.siu.edu/Olive/survhw.txt). To execute the program, use the top menu commands "Run>Submit". An output window will appear if successful. **Warning: if you do not have the recid.txt file on e drive, then you need to change** the *infile* command in the SAS code to the drive that you are using, eg change *infile "e:redic.txt";* to *infile "f:recid.txt";* if you are using the f drive.

c) First backward elimination is considered. Scroll to the top of the copy and paste the 1st 2 pages of output for the full model into *Word*.

d) Backward elimination will consider 5 models. Write down the variables used for each model as well as the AIC. The first two models are shown below.

```
variables                                        AIC
fin  age  race  wexp  mar  paro  prio      1332.241
fin  age  race  wexp  mar        prio      1330.429
```

e) Repeat d) for the 4 models considered by forward selection.

f) Repeat d) for the 5 models considered by stepwise selection.

g) For all subsets selection, complete the following table (get the 2 chisq entries).

```
variables       chisq
      3                    fin age prio
      7                    full
```

2.13. This problem considers the ovarian data from Collett (2003b, p. 344-346).

a) Obtain the SAS program for this problem from (http://parker.ad.siu.edu/Olive/survhw.txt). Print the output.

b) Find the ESP if *age* = 40 and *treat 1* = 1. (Comment: treatment takes on 2 levels so only one indicator is needed. SAS output includes a 2nd indicator *treat 2* but its coefficient is $\hat{\beta}_3 = 0$ and hence can be ignored. In general if the category takes on J levels, SAS will give nonzero output for the first $J - 1$ levels and a line of 0s for the Jth level. This means level J was omitted and the line of 0s should be ignored.)

c) Give a 95% CI for β_1 corresponding to age from output and the CI using the formula.

d) Give a 95% CI for β_2 corresponding to treat 1 from output and the CI using the formula.

e) If the model statement in the SAS program is changed to
model survtime*status(0)=;
then the null model is fit and the SAS output says
Log Likelihood −29.76723997.

Test $\boldsymbol{\beta} = \mathbf{0}$ with the LR test.
(Hint: The full model log likelihood log(L) = −20.56313339. Want −2 log(L) for both the full and null models for the LR test.)

f) Suppose the reduced model does not include *treat*. Then SAS output says Log Likelihood −21.7830. Test whether the reduced model is good.
(Hint: The log likelihood for the full model is log(L) = −20.56313339. Want −2 log(L) for the full and reduced models for the change in LR test.)

2.14. Copy and paste commands for this problem from (http://parker.ad.siu.edu/Olive/survhw.txt) for this problem into *SAS*. The myelomatosis data is from Allison (1995, p. 31, 158-161, 269). The 25 patients have tumours in the bone marrow. The patients were randomly assigned 2 drug treatments *treat*. The variable *renal* is 1 if renal (kidney) functioning is normal and 0 otherwise.

A stratified proportional hazards (SPH) model makes sense if the effect of *Renal* varies with time since randomization (if there is a time–Renal interaction). In this situation the PH model would be inappropriate since time–variable interactions are not allowed in the PH model. Notice that the results in a) and b) below are different. The analysis does need to control for the variable *Renal* to obtain good estimates of the treatment effect, but both the SPH model in a) and the PH model in c) may be adequate

a) The SAS program produces output for 3 models. The first model is a SPH model with stratification on *Renal*. Perform a Wald test on β_1 corresponding to *treat*. (In the output, $\hat{\beta}_1 = 1.463986$.)

b) The 2nd model is a PH model with the predictor *treat*. Perform a Wald test on β_1 corresponding to *treat*. (In the output, $\hat{\beta}_1 = 0.56103$.)

c) The 3rd model is a PH model with the predictors *treat* and *Renal*. Perform a Wald test on β_1 corresponding to *treat*. (In the output, $\hat{\beta}_1 = 1.22191$.)

R Problems

2.15. This data is from a study on ovarian cancer. There were 26 patients. The variable *futime* was the time until death or censoring in days, the variable *fustat* was 1 for death and 0 for censored, *age* is age and *ecog.ps* is a measure of status ranging from 0 (fully functional) to 4 (completely disabled). Level 4 subjects are usually considered too ill to enter a study such as this one.

a) Copy and paste commands for this problem from (http://parker.ad.siu.edu/Olive/survhw.txt) into *R*. Hit `Enter` and a plot should appear. Copy and paste the *R* output into *Word*. The output is similar to that of Problem 2.16 but also contains the variable ecog.ps.

Click on the plot and hold down the *Ctrl* and *c* buttons simultaneously. Then in the *Word* Edit menu, select "paste." The plot is the Cox regression estimated survival function at the average age (56.17) and average ecog.ps (1.462).

b) Now copy and paste the command for b) and place the plot in *Word* as described in a). This plot is for the Cox regression estimated survival function at the (age,ecog.ps) = (66,4). Is survival better for (56.17,1.462) or (66,4)?

c) Find the ESP and $\hat{h}_i(t)$ if $\boldsymbol{x} = (56.17, 1.462)$.

d) Find the ESP and $\hat{h}_i(t)$ if $\boldsymbol{x} = (66, 4)$.

e) Find a 95% CI for β_1.

f) Find a 95% CI for β_2.

g) Do a 4 step test for $H_0 : \beta_1 = 0$.

h) Do a 4 step test for $H_0 : \beta_2 = 0$.

i) Do a 4 step PLRT for $H_0 : \boldsymbol{\beta} = \mathbf{0}$.

```
     coef  exp(coef) se(coef)    z      p
age 0.162      1.18   0.0497

Likelihood ratio test=14.3  output for 2.16
```

2.16. Use the output above which is for the same data as in 2.15 but only the predictor *age* is used.

a) Find a 95% CI for β.

b) Do a 4 step test for $Ho : \beta = 0$.

c) Do a 4 step PLRT for $Ho : \boldsymbol{\beta} = \mathbf{0}$ (for $\beta = 0$). (The PLRT is better than the Wald test in b).)

2.17. The *R* `lung` cancer data has the *time* until death or censoring and *status* = 0 for censored and 1 for uncensored. Then the covariates are *age*, *sex* = 1 for M and 2 for F, *ph.ecog* = Ecog performance score 0-4, *ph.karno*

= a competitor to ph.ecog, *pat.karno* = patient's assessment of their karno score, *meal.cal* = calories consumed at meals excluding beverages and snacks and *wt.loss* = weight loss in last 6 months. A stratified proportional hazards model with stratification on *sex* will be used.

a) Copy and paste commands for this problem from (http://parker.ad.siu.edu/Olive/survhw.txt) into *R*.

Type *zfull*, then *zred1* then *zred2*. Copy and paste the resulting output into *Word*. The full model uses *age*, *ph.ecog*, *ph.karno*, *pat.karno* and *wt.loss*.

b) Test whether the reduced model that omits *age* can be used.

c) Test whether the reduced model that omits *age* and *ph.karno* can be used.

2.18. Go to (http://parker.ad.siu.edu/Olive/survhw.txt) and copy and paste the source command source("http://parker.ad.siu.edu/Olive/survpack.txt") near the top of the file into *R*. This problem will use the program `bphgfit` to check the PH model with the Kaplan Meier KM estimator.

a) Copy and paste commands from (http://parker.ad.siu.edu/Olive/survhw.txt) for this problem into *R*. Copy and paste the output into *Word*. (You may need to press Enter to get the plot.)

b) Click on the plot and hold down the *Ctrl* and *c* buttons simultaneously. Then in the *Word* Edit menu, select "paste."

c) The data is remission time in weeks for leukemia patients receiving treatments A (x = 0) or B (x = 1). See Smith (2002, p. 174). The indicator variable x (leuk[,3]) is the single covariate. Do a PLRT to test whether $\beta = 0$. Is there a difference in the effectiveness of the 2 treatments?

d) The solid lines in the plot correspond to the estimated PH survival function for each treatment group. The plotted points correspond to the estimated Kaplan Meier estimator for each group. If the PH model is good, then the plotted points should track the solid lines fairly well. Is the PH model good? (When $\beta = 0$, the PH model for this data is $h_0(t) = h_1(t)$, but the PH model could fail, e.g. if the survival function for treatment A is higher than that of treatment B until time t_A and then the survival function for treatment B is higher: the survival functions cross at exactly one point $t_A > 0$.)

With some versions of *R*, there are three curves of circles. The center curve is the Kaplan Meier estimator while the two outer bands are pointwise CI bands.

2.19. An extension of the PH model is the stratified PH model where $h_{\boldsymbol{x},j} = \exp(\boldsymbol{\beta}^T \boldsymbol{x}) h_{0,j}(t)$ for $j = 1, ..., K$ where $K \geq 2$ is the number of strata (groups). Testing is done in exactly the same manner as for the PH model, and the same $\boldsymbol{\beta}$ is used for each strata, only the baseline function changes. The regression in Problem 2.17 used gender, male and female, as strata. If the model was good, then a PH model should hold for males and a

PH model should hold for females. For the lung cancer data, females had a higher survival curve than males for $\boldsymbol{x}$ set to the average values.

A censored response plot (ESSP) is a plot of the ESP $= \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ versus T, the survival times, where the symbol "0" means the time was censored and "+" uncensored. If the PH model holds, the variability of the plotted points should decrease rapidly as ESP increases.

a) Copy and paste commands from (http://parker.ad.siu.edu/Olive/survhw.txt) for this problem into *R*. Click on the plot and hold down the *Ctrl* and *c* buttons simultaneously. Then in the *Word* Edit menu, select "paste."

b) Repeat a) except use the commands for 2.19b.

How does the variability in the plot for a narrow vertical strip at ESP $=$ 0.5 compare to the variability for a narrow vertical strip at ESP $= -1.5$?

c) Copy and paste the commands for this part into *R*, and include the resulting plot in *Word*.

d) Copy and paste the commands for this part into *R*, and include the resulting plot in *Word*.

```
vlung2(2)
title("females")
```

e) The plots in c) and d) divide the ESP into 4 slices. The estimated PH survival function is evaluated at the last point in the first 3 slices and at the first point in the 4th slice. Pointwise confidence intervals are also included (dashed upper and lower lines). The plotted circles correspond to the Kaplan Meier estimator for the points in each slice. The 1st slice is in the NW corner, the 2nd slice in the NE, the 3rd slice in the SW and the 4th slice in the SE. Confidence bands that would include an entire reasonable survival function would be much wider. Hence if the plotted circles are not very far outside the pointwise CI bands, then the PH model is reasonable.

Is the PH model reasonable for males? Is the PH model reasonable for females?

With some versions of *R*, there are three curves of circles. The center curve is the Kaplan Meier estimator while the two outer bands are pointwise CI bands.

2.20. The lung cancer data is the same as that described in 2.17, but the PH model is stratified on *sex* with variables *ph.ecog*, *ph.karno*, *pat.karno* and *wt.loss*.

a) Copy and paste commands for this problem from (http://parker.ad.siu.edu/Olive/survhw.txt) into *R*. Click on the left window and hit *Enter*. Then 4 plots should appear. Include the plot in *Word*.

b) The plots are of x_j versus the martingale residuals when x_j is omitted. The loess curve should be roughly linear (or at least not taking on some simple shape such as a quadratic) if x_j is the correct functional form. If the

loess curve looks like $t(x_j)$ for some simple t (eg $t(x_j) = x_j^2$), then $t(x_j)$ should be used instead of x_j. Are the loess curves in the 4 plots roughly linear?

c) Copy and paste commands for this problem from (http://parker.ad.siu.edu/Olive/survhw.txt) into R. Click on the left window and hit *Enter*. Then 4 plots should appear. Include the plot in *Word*. Also include the output from *cox.zph(lungfit2)* in *Word*.

d) The plots are of survival times vs scaled Schoenfeld residuals for each of the 4 variables. The loess curves should be approximately horizontal (0 slope) lines if the PH assumption is reasonable. Alternatively, the pvalue for Ho slope = 0 from *cox.zph* should be greater than 0.05 for each of the 4 variables. Is the PH assumption is reasonable? Explain briefly.

2.21. Copy and paste the R commands for this problem into R. This problem shows how to do backward elimination for the PH model in R using the Leemis (1995, p. 249-250) and Lawless (1982, p. 286) lung survival data. List the AIC for the model chosen in each step. Some entries are below.

```
                                      model  AIC
perf, age, ttoent, size, type, ttype, trt  189.22  full model
perf, age, ttoent, size,       ttype, trt  187.22
.
.
.
perf,                          ttype       181.52
perf                                       183.12
```

2.22. Copy and paste the R command

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

from near the top of (http://parker.ad.siu.edu/Olive/survhw.txt) into R. **(Do not give any plots for this problem.)**

a) In R, type "library(survival)" if necessary. Then type "phsim(k=1)". Hit the up arrow to repeat this command several times. Repeat for "phsim(k=0.5)" and "" to make ET plots. The simulated data follows a PH Weibull regression model with $h_0(t) = kt^{k-1}$. For $k = 1$ the data follows a PH exponential regression model. Did the survival times decrease rapidly as ESP increases?

b) The function `phsim2` slices the ESP into 9 groups and computes the Kaplan Meier estimator for each group. If the PH model is reasonable and n is large enough, the 9 plots should have approximately the same shape. Type "phsim2(n=100,k=1)", then "phsim2(n=200,k=1)" and keep increasing n by 100 until the nine plots look similar (assuming survival decreases from 1 to 0, and ignoring the labels on the horizontal axis and the + signs that correspond to censored times). We will say that the plots look similar if $n = 800$. What value of n did you get?

c) The function `bphsim3` makes the slice survival plots when the single covariate is an indicator for 2 groups. The PH assumption is reasonable if the plotted circles corresponding to the Kaplan Meier estimator track the solid line corresponding to the PH estimated survival function. Type "bphsim3(n=10,k=1)" and repeat several times (use the up arrow). Do the plotted circle track the solid line fairly well?

d) The function `phsim5` is similar but the ESP takes on many values and is divided into 9 groups. Type "phsim5(n=50,k=1)", then "phsim5(n=60,k=1)" and keep increasing n by 10 until the circles track the solid lines well. We will say that the circles track the solid lines well if they are within or not very far outside the pointwise CI bands. What value of n do you get?

2.23. This problem produces output for the Stanford Heart Transplant data, but R is used instead of SAS. Obtain the R program for Problem 2.23 from (http://parker.ad.siu.edu/Olive/survhw.txt). The time dependent variable $x_1(t) = \textit{transplant} = 1$ if the patient has had a transplant by time t and is 0 otherwise. The variable $x_2 = \textit{surgery} = 1$ if the patient has had previous heart surgery and is 0 otherwise. The variable $x_3 = \textit{age}$ is the patient's age at time of acceptance into the program. The R program fits a generalized Cox regression (GCR) model. The SAS and R heart data sets seem to differ slightly and do not give the exact same answers.

a) Print the output. (Put into *Word*.)

b) Test $\beta_1 = 0$.

c) Test $\boldsymbol{\beta} = \mathbf{0}$.

Problems from Quizzes and Exams

```
Output for Problem 2.24   full model, n = 26
          coef exp(coef) se(coef)     z     p
age      0.121      1.13   0.0484 2.500 0.012
resid.ds 0.792      2.21   0.8078 0.980 0.330
ecog.ps  0.087      1.09   0.6592 0.132 0.890
Likelihood ratio test= 13.7  on 3 df,   p=0.00333

     coef exp(coef) se(coef)   z      p  reduced model
age 0.137      1.15   0.0474 2.9 0.0038
Likelihood ratio test= 12.7  on 1 df,   p=0.000368
```

2.24. The R `ovarian` data gives survival times for patients with ovarian cancer. Predictors are *age* in years, *resid.ds* (residual disease present 1=no,2=yes), and *ecog.ps* (ECOG performance status: 1 is better than 2). A stratified proportional hazards model is fit where the stratification variable *rx* is the treatment group.

a) Test whether $\beta_3 = 0$.

b) Test whether $\boldsymbol{\beta} = \mathbf{0}$ for the full model.

c) Test whether the reduced model is good.

```
Output for Problem 2.25
               Value Std. Error     z        p
(Intercept)  5.32632    0.66298  8.03 9.44e-16
age         -0.00891    0.00711 -1.25 0.210
sex          0.37019    0.12796  2.89 0.00382
ph.karno     0.00926    0.00446  2.08 0.0379
Log(scale)  -0.28085    0.06171 -4.55 5.33e-06

Scale= 0.755

Weibull distribution
Loglik(model)= -1138.7   Loglik(intercept only)= -1147.5
Chisq= 17.59 on 3 degrees of freedom, p= 0.00053
n=227 (1 observation deleted due to missingness)
```

2.25. A Weibull regression model was fit to the *R* `lung` data resulting in the above output.

a) Test whether $\boldsymbol{\beta} = \mathbf{0}$.

b) Test whether $\beta_1 = 0$.

c) Test whether $\beta_2 = 0$.

d) Sketch the Weibull EE plot if the Weibull model is good.

Chapter 3
Parametric Survival Regression

Definition 3.1. In a *1D regression model*, the response variable Y is conditionally independent of the $p \times 1$ vector of predictors $\boldsymbol{x}$ given the sufficient predictor $SP = h(\boldsymbol{x})$, written

$$Y \perp\!\!\!\perp \boldsymbol{x} | SP \quad \text{or} \quad \mathrm{Y} \perp\!\!\!\perp \boldsymbol{x} | \mathrm{h}(\boldsymbol{x}), \tag{3.1}$$

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 $h(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\beta}$ where $ESP = \boldsymbol{x}^T \hat{\boldsymbol{\beta}}$.

An important class of parametric 1D regression models has $Y|\boldsymbol{x} \sim D(\boldsymbol{x}^T \boldsymbol{\beta}, \boldsymbol{\gamma})$ where D is a parametric distribution that depends on the $p \times 1$ vector of predictors $\boldsymbol{x}$ only through $SP = \boldsymbol{x}^T \boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ is a $q \times 1$ vector of parameters. Several important survival regression models, including Weibull regression and accelerated failure time models, have this form, and will be covered in this chapter. Weibull regression and Exponential regression are parametric proportional hazards regression models.

3.1 Univariate Parametric Models

Assume that $Y_1, ..., Y_n$ are iid from a parametric distribution such as the Weibull or Exponential distribution. Let $T_1, ..., T_n$ be the observed right censored data. Often the parameters of the parametric distribution can be estimated by maximum likelihood.

Example 3.1. Suppose the observed survival times $T_1, ..., T_n$ are a censored data set from an Exponential (λ) distribution. Let $T_i = Y_i^*$. Let $\delta_i = 0$ if the case is censored and let $\delta_i = 1$, otherwise. Let $r = \sum_{i=1}^n \delta_i =$ the num-

ber of uncensored cases. Then the MLE $\hat{\lambda} = r/\sum_{i=1}^n T_i$. So $\hat{\lambda} = r/\sum_{i=1}^n Y_i^*$. A 95% CI for λ is $\hat{\lambda} \pm 1.96\hat{\lambda}/\sqrt{r}$.

Remark 3.1. It can be shown that a better CI than the one given in Example 3.1 is

$$\left[\frac{\hat{\lambda}\ \ \chi^2(2r, 1-\delta/2)}{2r}, \frac{\hat{\lambda}\ \ \chi^2(2r, \delta/2)}{2r}\right]$$

where $P[X \leq \chi^2(k, \delta)] = \delta$ if $X \sim \chi^2_k$ has a chi-square distribution with k degrees of freedom.

3.2 Weibull and Exponential Regression

Definition 3.2. For **parametric proportional hazards** regression models, the baseline function is parametric and the parameters are estimated via maximum likelihood. Then as a 1D regression model, $SP = \boldsymbol{\beta}_P^T \boldsymbol{x}$, and

$$h_{Y|SP}(t) \equiv h_{\boldsymbol{x}}(t) = \exp(\boldsymbol{\beta}_P^T \boldsymbol{x}) h_{0,P}(t) = \exp(SP) h_{0,P}(t)$$

where the parametric baseline function $h_{0,P}$ depends on k unknown parameters but does not depend on the predictors $\boldsymbol{x}$. The survival function is

$$S_{\boldsymbol{x}}(t) \equiv S_{Y|SP}(t) = [S_{0,P}(t)]^{\exp(\boldsymbol{\beta}_P^T \boldsymbol{x})} = [S_{0,P}(t)]^{\exp(SP)}, \tag{3.2}$$

and

$$\hat{S}_{\boldsymbol{x}}(t) = [\hat{S}_{0,P}(t)]^{\exp(\hat{\boldsymbol{\beta}}_P^T \boldsymbol{x})} = [\hat{S}_{0,P}(t)]^{\exp(ESP)}. \tag{3.3}$$

The following univariate results will be useful for Exponential and Weibull regression. If Y has a Weibull distribution, $Y \sim W(\gamma, \lambda)$, then $S_Y(t) = \exp(-\lambda t^\gamma)$ where t, λ and γ are positive. If $\gamma = 1$, then Y has an Exponential distribution, $Y \sim EXP(\lambda)$ where $E(Y) = 1/\lambda$. See Examples 1.1 and 1.2. Now V has a smallest extreme value distribution, $V \sim SEV(\theta, \sigma)$, if

$$S_V(t) = P(V > t) = \exp\left(-\exp\left(\frac{t-\theta}{\sigma}\right)\right)$$

where $\sigma > 0$ while t and θ are real. If $Z \sim SEV(0,1)$, then $V = \theta + \sigma Z \sim SEV(\theta, \sigma)$ since the SEV distribution is a location scale family. Also, $V = \log(Y) \sim SEV(\theta = -\sigma \log(\lambda), \sigma = 1/\gamma)$, and $Y = e^V \sim W(\gamma = 1/\sigma, \lambda = e^{-\theta/\sigma})$.

If Y_i follows a Weibull regression model, then $\log(Y_i)$ follows an accelerated failure time (AFT) model: $\log(Y_i) = \alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}_i + \sigma e_i$ where the e_i are iid $SEV(0,1)$, and $\log(Y)|\boldsymbol{x} \sim SEV(\alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}, \sigma)$. See Section 3.3.

Definition 3.3. The **Weibull proportional hazards regression (WPH) model** or **Weibull regression model** is a parametric proportional hazards model with $Y|\boldsymbol{x} \sim W(\gamma = 1/\sigma, \lambda_{\boldsymbol{x}})$ where

$$\lambda_{\boldsymbol{x}} = \exp\left[-\left(\frac{\alpha}{\sigma} + \frac{\boldsymbol{\beta}_A^T \boldsymbol{x}}{\sigma}\right)\right] = \lambda_0 \exp(\boldsymbol{\beta}_P^T \boldsymbol{x})$$

with $\lambda_0 = \exp(-\alpha/\sigma)$ and $\boldsymbol{\beta}_P = -\boldsymbol{\beta}_A/\sigma$. Thus for $t > 0$, $P(Y > t|\boldsymbol{x}) =$

$$S_{\boldsymbol{x}}(t) = \exp(-\lambda_{\boldsymbol{x}} t^\gamma) = \exp(-\lambda_0 \exp(\boldsymbol{\beta}_P^T \boldsymbol{x}) t^\gamma) = [\exp(-\lambda_0 t^\gamma)]^{\exp(\boldsymbol{\beta}_P^T \boldsymbol{x})} =$$

$$[S_{0,P}(t)]^{\exp(\boldsymbol{\beta}_P^T \boldsymbol{x})}.$$

As a 1D regression model, $Y|SP \sim W(\gamma, \lambda_0 \exp(SP))$. Also,

$$h_i(t) = h_{Y_i|\boldsymbol{x}_i}(t) = h_{Y_i|\boldsymbol{\beta}_P^T \boldsymbol{x}_i}(t) = \exp(\boldsymbol{\beta}_P^T \boldsymbol{x}_i) h_0(t)$$

where $h_0(t) = h_0(t|\boldsymbol{\theta}) = \lambda_0 \gamma t^{\gamma - 1}$ is the Weibull **baseline function**. **Exponential regression** is the special case of Weibull regression where $\sigma = 1$. Hence $Y|\boldsymbol{x} \sim W(1, \lambda_{\boldsymbol{x}}) \sim EXP(\lambda_{\boldsymbol{x}})$.

Since Weibull regression and Exponential regression are proportional hazards regression models, the plots from Chapter 2 can be used to check the models. The Weibull proportional hazard model is valid iff the Weibull accelerated failure time (AFT) model is valid. Similarly, the Exponential PH model is valid iff the Exponential AFT model is valid. Hence the following two plots are useful.

Definition 3.4. Let $T_i = \min(Y_i, Z_i)$ be the censored survival times, and let $\log(T_i) = \hat{\alpha} + \hat{\boldsymbol{\beta}}_A^T \boldsymbol{x}_i + r_i$. For accelerated failure time models, a **log censored response (LCR) plot** is a plot of $\hat{\alpha} + \hat{\boldsymbol{\beta}}_A^T \boldsymbol{x}_i$ versus $\log(T_i)$ with plotting symbol 0 for censored cases and + for uncensored cases. The identity line with unit slope and zero intercept is added to the plot, and the vertical deviations from the identity line $= r_i$. Collett (2003b, p. 231) defines a standardized residual $r_{Si} = r_i/\hat{\sigma}$.

The least squares line based on the +'s could be added to the plot and should have slope not too far from 1, especially if $\gamma \geq 1$ for the Weibull AFT. The plotted points should be linear with roughly constant variance. The censoring and long left tails of the smallest extreme value distribution make judging linearity and detecting outliers from the left tail difficult. Try

to ignore the bottom of the plot where there are few cases when assessing linearity.

Definition 3.5. For parametric proportional hazards models, an **EE plot** is a plot of the parametric ESP $\hat{\boldsymbol{\beta}}_P^T \boldsymbol{x}$ versus the Cox semiparametric ESP $\hat{\boldsymbol{\beta}}_C^T \boldsymbol{x}$.

If the parametric proportional hazards model is good, then the plotted points in the EE plot should track the identity line with unit slope and zero intercept. As $n \to \infty$, the correlation of the plotted points goes to 1 in probability for any finite interval, e.g., from the 1st percentile to the 99th percentile of $\hat{\boldsymbol{\beta}}_C^T \boldsymbol{x}$. Lack of fit is suggested if the plotted points do not cluster tightly about the identity line.

Software typically fits Exponential and Weibull regression models as accelerated failure time models: $\log(Y_i) = \alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}_i + \sigma e_i$. For the Exponential regression model, $\sigma = 1$ and $\boldsymbol{\beta}_C = -\boldsymbol{\beta}_A$, and the Exponential EE plot is a plot of

$$ESPE = -\hat{\boldsymbol{\beta}}_A^T \boldsymbol{x} \text{ versus } \mathrm{ESPC} = \hat{\boldsymbol{\beta}}_{\mathrm{C}}^{\mathrm{T}} \boldsymbol{x}.$$

For the Weibull regression model, $\boldsymbol{\beta}_C = -\boldsymbol{\beta}_A/\sigma$, and the Weibull EE plot is a plot of

$$ESPW = \frac{-1}{\hat{\sigma}} \hat{\boldsymbol{\beta}}_A^T \boldsymbol{x} \text{ versus } \mathrm{ESPC} = \hat{\boldsymbol{\beta}}_{\mathrm{C}}^{\mathrm{T}} \boldsymbol{x}.$$

Suppose the plotted points cluster tightly about the identity line in the EE plot with $corr(\hat{\boldsymbol{\beta}}_C^T \boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_P^T \boldsymbol{x}_i) > 0.99$. Thus $\hat{\boldsymbol{\beta}}_C^T \boldsymbol{x} \approx \hat{\boldsymbol{\beta}}_P^T \boldsymbol{x}$ for the observed $\boldsymbol{x}_i$, and slicing on the Cox ESP is nearly the same as slicing on the parametric ESP. Make the slice survival plot for the Cox model and add the estimated parametric survival function (3.3) as crosses. If the parametric proportional hazards model holds, then (2.2) = (3.2). Thus if (2.3) $\approx$ (3.3) for any $\boldsymbol{x}_i$, then $S_{0,P}(t) \approx S_0(t)$, (2.3) $\approx$ (3.3) for all $\boldsymbol{x}_i$, and the parametric proportional hazards model is reasonable.

Remark 3.2. Checking parametric proportional hazards models has 3 steps: i) check that the proportional hazards assumption is reasonable, e.g. with the slice survival plot for the Cox model, ii) check that the parametric and semiparametric ESPs are approximately the same, $\hat{\boldsymbol{\beta}}_P^T \boldsymbol{x} \approx \hat{\boldsymbol{\beta}}_C^T \boldsymbol{x}$ with the EE plot, and iii) using the slice survival plot, check that (2.3) $\approx$ (3.3) for the $\boldsymbol{x}$ used in each of the J slices. Since the Weibull proportional hazards model (Def. 3.3) is valid for $(Y, \boldsymbol{x})$ if and only if the Weibull accelerated failure time model (Def. 3.7) is valid for $(\log(Y), \boldsymbol{x})$, the above procedure can be used to simultaneously check the goodness of fit of both models.

This technique avoids the mistake of comparing quantities from the semiparametric and parametric proportional hazards models without checking that the proportional hazards assumption is reasonable. The slice survival

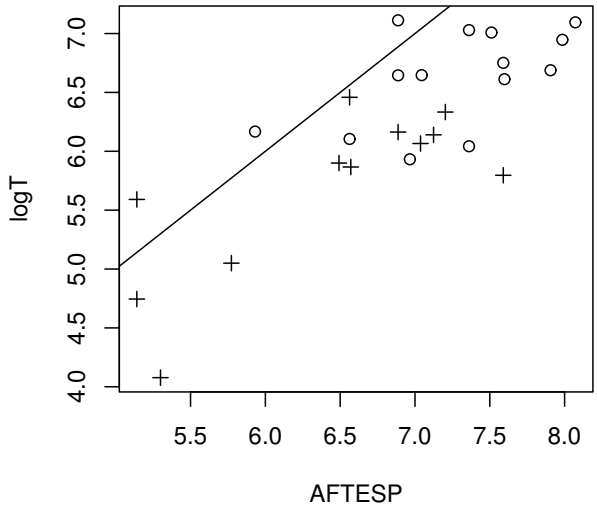

Fig. 3.1 LCR Plot for Ovarian Cancer Data

plot for the Cox model is used because of the ease of making pointwise CI bands.

Example 3.2. The ovarian cancer data is from Collett (2003b, p. 187-190) and Edmunson et al. (1979). The response variable is the survival time of $n = 26$ patients in days with predictors *age* in years and *treat* (1 for cyclophosphamide alone and 2 for cyclophosphamide combined with adriamycin). Figure 3.1 shows that most of the plotted points in the LCR plot for the ovarian cancer data are below the identity line. If a Weibull regression model is a good approximation to the data, then the plotted points in a narrow vertical slice centered at $\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x} = w$ are approximately a censored sample from an $SEV(w, \hat{\sigma})$ distribution. Figure 3.2 shows the Weibull and Exponential regression EE plots. Notice that the estimated risk scores from the Cox regression and Weibull regression are nearly the same with correlation = 0.997. The points from the Exponential regression do not cluster about the identity line. Hence Exponential regression should not be used. Figure 3.3 gives the slice survival plot for the Cox model with the Weibull survival function $\hat{S}_{\boldsymbol{x}}(t) = \exp[-\exp(-\hat{\gamma}\hat{\boldsymbol{\beta}}_A^T \boldsymbol{x}) \ \exp(-\hat{\gamma}\hat{\alpha}) \ t^{\hat{\gamma}} \ ]$ represented by crosses where $\hat{\gamma} = 1/\hat{\sigma}$. Notice that the Weibull and Cox estimated survival functions are close and thus similar. Again the circles corresponding to the

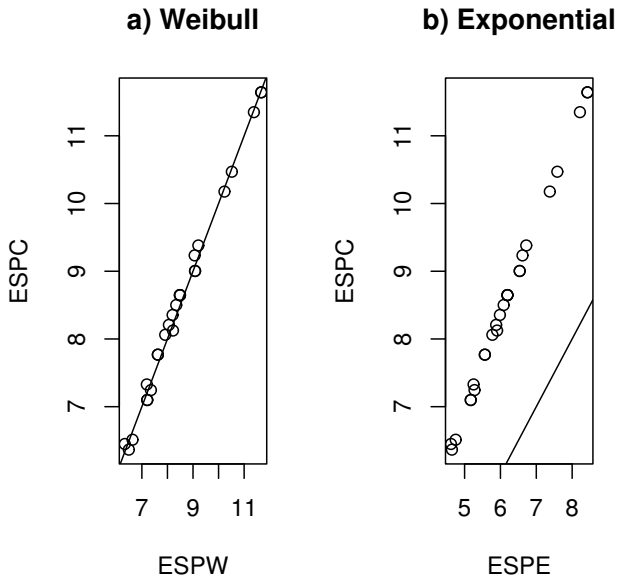

Fig. 3.2 EE Plots for Ovarian Cancer Data

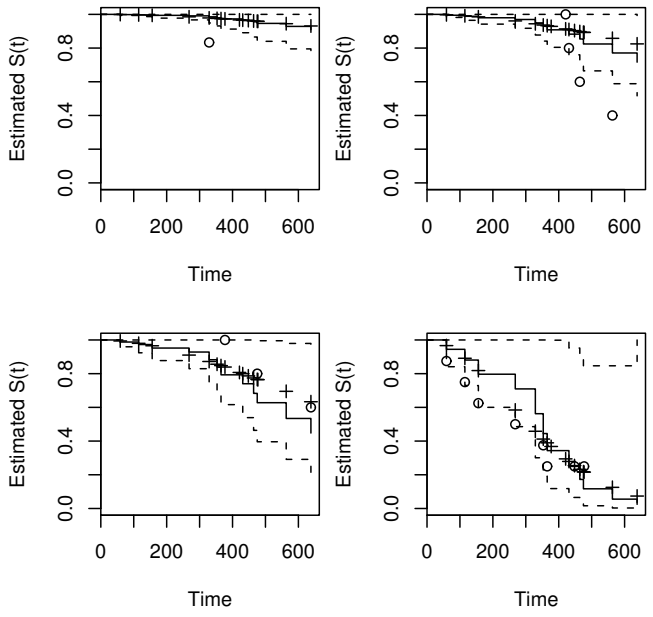

Fig. 3.3 Slice Survival Plots for Ovarian Cancer Data

Kaplan Meier estimator are "close" to the Cox survival curves in that the circles do not fall very far outside the pointwise CI bands.

Output for the Weibull and Exponential regression models is shown below. The output is often from software for accelerated failure time models. The tests are the same for the parametric PH model and the equivalent AFT model, but for Weibull regression the ESP and confidence intervals tend to be for $\hat{\boldsymbol{\beta}}_A = (\beta_i)$, which differs from $\hat{\boldsymbol{\beta}}_P$ by a constant. Output for AFT models will include an intercept $\hat{\alpha}$ and an estimate of scale $\hat{\sigma}$. *SAS* and *R* give output for the AFT.

For *SAS* or *R*.

variable	Est.	SE	Est/SE	or $(Est/SE)^2$	pvalue for
intercept					
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	$X^2_{o,1} = z^2_{o,1}$	Ho: $\beta_1 = 0$
$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$z_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	$X^2_{o,p} = z^2_{o,p}$	Ho: $\beta_p = 0$
scale or	log scale				
Weibull shape	or scale				

Output for the null model for SAS is shown below.
log likelihood log L(none)

variable	Est.	SE	Est/SE	or $(Est/SE)^2$	pvalue
intercept					
scale					
Weibull shape					

For the full model, SAS will have Log Likelihood = log L(full).
For the full model, R will have log L(full), log L(none) and
chisq = [-2 log L(none)] - [-2 log L(full)] on p degrees of freedom with pvalue.

Replace full by reduced for the reduced model.

The SAS and R log likelihood, log L, differ by a constant.

```
SAS Log Likelihood = -29.7672 null model
variable      df Estimate  SE     chi square  pr > chisqu
intercept     1   7.1110   0.2927  590.12       < 0.0001
Weibull Scale 1   1225.4   358.7
Weibull Shape 1   1.1081   0.2810

SAS Log Likelihood = -29.1775  reduced model
variable      df Estimate  SE     chi square  pr > chisqu
```

```
intercept     1   7.3838   0.4370 285.45         < 0.0001
treat         1  -0.5593   0.5292   1.12           0.2906
Scale         1   0.8857   0.2227
Weibull Shape 1   1.1291   0.2840

SAS Log Likelihood = -20.5631    full model
variable      df Estimate  SE     chi square  pr > chisqu
intercept     1  11.5483   1.1970 93.07         < 0.0001
age           1  -0.0790   0.0198 15.97         < 0.0001
treat         1  -0.5615   0.3399 2.73            0.0986
Scale         1   0.5489   0.1291
Weibull Shape 1   1.8218   0.4286


R reduced  model Value Std. Error      z        p
(Intercept)      7.384      0.437 16.895 4.87e-64
treat           -0.559      0.529 -1.057 2.91e-01
Log(scale)      -0.121      0.251 -0.483 6.29e-01
Scale= 0.886
Loglik(model)= -97.4   Loglik(intercept only)= -98
        Chisq= 1.18 on 1 degrees of freedom, p= 0.28

R full model     Value Std. Error     z        p
(Intercept)     11.548     1.1970  9.65 5.04e-22
treat           -0.561     0.3399 -1.65 9.86e-02
age             -0.079     0.0198 -4.00 6.43e-05
Log(scale)      -0.600     0.2353 -2.55 1.08e-02
Scale= 0.549
Loglik(model)= -88.7   Loglik(intercept only)= -98
        Chisq= 18.41 on 2 degrees of freedom, p= 1e-04
```

Shown above is output in symbols from and *SAS* and *R*. The estimated coefficient is $\hat{\beta}_j$. The Wald chi square $= X^2_{0,j}$ while p and "pr > chisqu" are both p-values.

Inference from output is much like that for the Cox PH regression model. Find the ESP, $h_0(t)$, 95% CI for β_i, do a Wald test for $H_0 : \beta_i = 0$, do a likelihood ratio test (LRT) for $H_0 : \boldsymbol{\beta} = \mathbf{0}$versus $H_A : \boldsymbol{\beta} \neq \mathbf{0}$, and do a change in LRT for H_0: the reduced model is good versus H_A: use the full model. The Cox PH regression model used a PLRT and a change in PLRT.

Given $\hat{\boldsymbol{\beta}}$ from output and given $\boldsymbol{x}$, be able to find ESP $= \hat{\boldsymbol{\beta}}^T \boldsymbol{x} = \sum_{i=1}^p \hat{\beta}_i x_i = \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$.

A large sample 95% CI for β_j is $\hat{\beta}_j \pm 1.96\ se(\hat{\beta}_j)$.

4 step Wald test of hypotheses:
i) State the hypotheses Ho: $\beta_j = 0$ Ha: $\beta_j \neq 0$.

ii) Find the test statistic $z_{0,j} = \hat{\beta}_j / se(\hat{\beta}_j)$ or $X^2_{0,j} = z^2_{0,j}$ or obtain it from output.

iii) The p–value $= 2P(Z < -|z_{0j}|) = P(\chi^2_1 > X^2_{0,j})$. Find the p–value from output or use the standard normal table.

iv) If pval $< \delta$, reject Ho and conclude that x_j is needed in the Weibull survival model given that the other $p-1$ predictors are in the model. If pval $\geq \delta$, fail to reject Ho, and conclude that the values of x_j do not (significantly) affect the WPH survival model given that the other $p-1$ predictors are in the model. (Or state that there is not enough evidence to conclude that the values of x_j affect the WPH survival model.)

The 4 step likelihood ratio test **LRT** is

i) $H_0 : \boldsymbol{\beta} = \mathbf{0} \quad H_A : \boldsymbol{\beta} \neq \mathbf{0}$

ii) test statistic $X^2(N|F) = [-2 \log L(none)] - [-2 \log L(full)]$ is often obtained from output.

iii) The p–value $= P(\chi^2_p > X^2(N|F))$ where χ^2_p has a chi–square distribution with p degrees of freedom. The p–value is often obtained from output.

iv) Reject H_0 if the p–value $< \delta$ and conclude that there is a WPH survival relationship between Y and the predictors $\boldsymbol{x}$. If p–value $\geq \delta$, then fail to reject H_0, and conclude that the values of the predictors $\boldsymbol{x}$ do not (significantly) affect the WPH survival model. (Or state that there is not enough evidence to conclude that the values of $\boldsymbol{x}$ affect the WPH survival model.)

For the above test, $X^2(N|F)$ is Chisq from R. Both R and SAS give $\log L$, but for R, $\log L_R = \log L + d_R$ and for SAS, $\log L_{SAS} = \log L + d_{SAS}$. So $\log L$ differs by a constant for R and SAS, but the constant cancels with subtraction.

Also note that there could be a PH survival relationship but not a WPH survival relationship. Check WPH assumptions before doing inference.

The 4 step **change in LR test** is

i) H_0: the reduced model is good $\quad H_A$: use the full model

ii) test statistic $X^2(R|F) = X^2(N|F) - X^2(N|R) = [-2 \log L(red)] - [-2 \log L(full)]$.

iii) The p–value $= P(\chi^2_{p-r} > X^2(R|F))$ where χ^2_{p-r} has a chi–square distribution with $p-r$ degrees of freedom.

iv) Reject H_0 if the p–value $< \delta$ and conclude that the full model should be used. If p–value $\geq \delta$, then fail to reject H_0 and conclude that the reduced model is good (the values of $\boldsymbol{x}_O$ do not (significantly) affect the survival model, or there is not enough evidence to conclude that the values of $\boldsymbol{x}_O$ affect the survival model).

Example 3.3. Between points 1) and 2) in the summary Section 3.5, is output for the ovarian cancer data of Example 3.2. This output is also shown in this section.

a) Find ESP if treat $= 1$ and age $= 60$.

Solution: $ESP = \hat{\boldsymbol{\beta}}^T \boldsymbol{x} = -0.561(1) - 0.079(60) = -5.301$.

b) Find a 95% CI for β_1 corresponding to treat.

Solution: Using output for the R full model, the 95% CI is $\hat{\beta}_1 \pm 1.96\ se(\hat{\beta}_1 = -0.561 \pm 1.96(0.3399) = -0.561 \pm 0.662 = [-1.2272, 0.1052]$. SAS and R output differs slightly.

c) Test $\beta_1 = 0$ corresponding to treat.

Solution: i) $H_0 : \beta_1 = 0\ H_1 : \beta_1 \neq 0$

ii) $Z_{01} = \dfrac{-0.561}{0.3399} = -1.6504$ or use output

or $X_{01}^2 = Z_{01}^2 = 2.7241$ (2.73 from output)

iii) pval $= 2P(Z < 1.65) = 2(0.0495) = 0.099$ (0.0986 from output)

or pval $= P(\chi_1^2 > 2.72)$

```
df| .100  .05
---------------
 1| 2.71   3.84
```

so $0.05 <$ pval < 0.100

iv) Fail to reject H_0. The values of Treatment do not affect the Weibull survival model given age is in the model.

d) Test $\beta_2 = 0$ corresponding to age.

Solution: i) $H_0 : \beta_2 = 0\ H_1 : \beta_2 \neq 0$

ii) $Z_{02} = -4.00$

or $X_{02}^2 = 15.97$

iii) pval $= 0.0000643$ or pval < 0.001

iv) Reject H_0. Age is needed in the Weibull survival model given treat is in the model.

e) Test $\boldsymbol{\beta} = \mathbf{0}$.

Solution: i) $H_0 : \boldsymbol{\beta} = \mathbf{0}\ H_1 : \boldsymbol{\beta} \neq \mathbf{0}$

ii) R: $X^2(N|F) = 18.41$ or

$X^2(N|F) = [-2\log(L(none)] - [-2\log L(Full)] = [-2(-98)] - [-2(-88.7)] = 196 - 177.4 = 18.6$ due to rounding

or SAS: $X^2(N|F) = [-2(-29.7672)] - [-2(-20.5631)] = 59.5344 - 41.1262 = 18.4082$

iii) pval $= P(\chi_2^2 > 18.41)$

```
df| .001
----------
 2| 13.82
```

so pval < 0.001 (0.0001 from output)

iv) Reject H_0: there is a WPH survival relationship between time Y and the predictors age and treat.

f) Test whether the reduced model with treat is good.

Solution: i) H_0 : the reduced model is good H_1 : use the full model

ii) R: $X^2(R|F) = X^2(N|F) - X^2(N|R) = 18.41 - 1.18 = 17.23$ or

$X^2(R|F) = [-2\log(L(Red)] - [-2\log L(Full)] = [-2(-97.4)] - [-2(-88.7)] = 194.8 - 177.4 = 17.4$ due to rounding

SAS: $X^2(R|F) = [-2(-29.1775)] - [-2(-20.5631)] = 58.355 - 41.1262 = 17.2288$

iii) pval $= P(\chi^2_1 > 17.23)$

```
df| .001
----------
 1| 10.83
```

so pval < 0.001

iv) Reject H_0: use the full model.

Warning: Remarks 2.1–2.4 also apply for the models in this chapter.

3.3 Accelerated Failure Time Models

Definition 3.6. For a parametric *accelerated failure time* model,

$$\log(Y_i) = \alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}_i + \sigma e_i \tag{3.4}$$

where the e_i are iid from a location scale family. Let $SP = \boldsymbol{\beta}_A^T \boldsymbol{x}$. Then as a 1D regression model, $\log(Y)|SP = \alpha + SP + e$. The parameters are again estimated by maximum likelihood and the survival function is

$$S_{\boldsymbol{x}}(t) \equiv S_{Y|\boldsymbol{x}}(t) = S_0\left(\frac{t}{\exp(\boldsymbol{\beta}_A^T \boldsymbol{x})}\right),$$

and

$$\hat{S}_{\boldsymbol{x}}(t) = \hat{S}_0\left(\frac{t}{\exp(\hat{\boldsymbol{\beta}}_A^T \boldsymbol{x})}\right)$$

where $\hat{S}_0(t)$ depends on $\hat{\alpha}$ and $\hat{\sigma}$.

For the AFT model, $h_i(t) = h_{\boldsymbol{x}}(t) = e^{-SP} h_0(t/e^{SP})$ and $S_i(t) = S_{\boldsymbol{x}}(t) = S_0(t/\exp(SP))$ where $SP = \boldsymbol{\beta}_A^T \boldsymbol{x}$. If $S_{\boldsymbol{x}}(t_{\boldsymbol{x}}(\rho)) = 1 - \rho$ for $0 < \rho < 1$, then $t_{\boldsymbol{x}}(\rho)$ is the ρth percentile. For the accelerated failure time model,

$$t_{\boldsymbol{x}}(\rho) = t_0(\rho) \exp(\boldsymbol{\beta}_A^T \boldsymbol{x})$$

where $t_0(\rho) = \exp(\sigma e_i(\rho) + \alpha)$ and $S_{e_i}(e_i(\rho)) = P(e_i > e_i(\rho)) = 1 - \rho$. Note that the estimated percentile ratio is free of ρ, $\hat{\sigma}$ and $\hat{\alpha}$

$$\frac{\hat{t}_{\boldsymbol{x}_1}(\rho)}{\hat{t}_{\boldsymbol{x}_2}(\rho)} = \exp(\hat{\boldsymbol{\beta}}_A^T(\boldsymbol{x}_1 - \boldsymbol{x}_2)).$$

The *acceleration factor* $= e^{-SP}$ and $t_{0,\rho} = e^{-SP} t_{\boldsymbol{x},\rho}$. The median survival times are related by $t_{0,0.5} = e^{-SP} t_{\boldsymbol{x},0.5}$. If $e^{-SP} < 1$, then the median survival

time of $\boldsymbol{x} >$ the median survival time of $\mathbf{0}$, a result that is good if the event is death, but bad if the event is time until recovery. Note that $H_{\boldsymbol{x}}(t) = -\log S_{\boldsymbol{x}}(t) = -\log(S_0(t/e^{SP})) = H_0(t/e^{SP})$.

Remark 3.3. Assume $x_i > 0$. Then $\beta_i > 0$ increases $\log(Y_i)$ and Y_i, while $\beta_i < 0$ decreases $\log(Y_i)$ and Y_i. For the Cox PH regression model, $h_{\boldsymbol{x}}(t) = \exp(\boldsymbol{\beta}^T\boldsymbol{x})h_0(t)$. Hence $\beta_i > 0$ increases hazard and decreases Y_i, while $\beta_i < 0$ decreases hazard and increases Y_i. In the WPH model, $\boldsymbol{\beta}_P = -\boldsymbol{\beta}_A/\sigma$.

The LCR plot of Definition 3.4 is still useful for finding influential cases for AFT models. If the Weibull PH regression model holds for Y_i, then $\log(Y_i) = \alpha + \boldsymbol{\beta}_A^T\boldsymbol{x}_i + \sigma e_i$ where $e_i \sim SEV(0,1)$. Thus $\log(Y)|\boldsymbol{x} \sim SEV(\alpha + \boldsymbol{\beta}_A^T\boldsymbol{x}, \sigma)$, and the $\log(Y_i)$ follows a parametric accelerated failure time model. Two other important AFTs are i) the lognormal AFT where $\log(Y)|\boldsymbol{x} \sim N(\alpha+\boldsymbol{\beta}_A^T\boldsymbol{x}, \sigma^2)$ where the Y_i are lognormal and the $e_i \sim N(0,1)$ are normal, and ii) the loglogistic AFT where $\log(Y)|\boldsymbol{x} \sim L(\alpha + \boldsymbol{\beta}_A^T\boldsymbol{x}, \sigma)$ where the Y_i are loglogistic and the $e_i \sim L(0,1)$ are logistic. For the loglogistic AFT, Y follows a proportional odds model. Y does not follow a proportional hazards regression model for the loglogistic and lognormal AFTs. The residuals r_i are the vertical deviations from the identity line in the LCR plot, and should behave like a censored sample from the distribution of σe_i. Hence the r_i are like a censored sample from i) a $SEV(0,\sigma)$ distribution for a Weibull AFT, ii) a $N(0,\sigma^2)$ distribution for a lognormal AFT, and iii) a $L(0,\sigma)$ distribution for a loglogistic distribution. The normal and logistic distributions are symmetric.

Definition 3.7. The *Weibull AFT* satisfies $\log(Y)|(\alpha+\boldsymbol{\beta}_A^T\boldsymbol{x}) \sim SEV(\alpha + \boldsymbol{\beta}_A^T\boldsymbol{x}, \sigma)$. Thus points in a narrow vertical slice about $\hat{\alpha}+\hat{\boldsymbol{\beta}}_A^T\boldsymbol{x} = w$ in the LCR plot are approximately a censored sample from an $SEV(w, \hat{\sigma})$ distribution if the fitted model is a good approximation to the data. The *Exponential AFT* is the special case with $\sigma = 1$.

Theorem 3.1. Weibull regression models, including Exponential regression models, are the only models where Y follows a proportional hazards regression model and $\log(Y)$ follows an accelerated failure time model.

Censoring causes the bulk of the data to be below the identity line in the LCR plot. For example, Hosmer and Lemeshow (1999, p. 226) state that for the Exponential regression model, $\hat{\alpha}$ forces

$$\sum_{i=1}^n \delta_i = \sum_{i=1}^n \frac{T_i}{\exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}_A^T\boldsymbol{x}_i)}.$$

Hence $\hat{T}_i = \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}_A^T\boldsymbol{x}_i) \approx (n/\sum_{i=1}^n \delta_i)T_i$ (roughly). With no censoring, the bulk of the data will still be lower than the identity line if the e_i are left skewed as for the Weibull regression model where the $e_i \sim SEV(0,1)$.

Remark 3.4. Since the Weibull proportional hazards model is valid for $(Y, \boldsymbol{x})$ if and only if the Weibull accelerated failure time model is valid for $(\log(Y), \boldsymbol{x})$, fit the data using Cox regression. Then the graphical procedure described in Remark 3.2 can be used to simultaneously check the goodness of fit of both the Weibull PH and AFT models. Similarly, the Exponential proportional hazards model is valid for $(Y, \boldsymbol{x})$ if and only if the Exponential accelerated failure time model is valid for $(\log(Y), \boldsymbol{x})$.

For Weibull and Exponential regression, instead of fitting a PH model, R and SAS fit an accelerated failure time model $\log(Y_i) = \alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}_i + \sigma e_i$ where the e_i are iid from a smallest extreme value distribution. The Exponential AFT is the special case of the Weibull AFT with $\sigma = 1$. As in Definition 3.10, $\lambda_0 = \exp(-\alpha/\sigma)$ and $\boldsymbol{\beta}_P = -\boldsymbol{\beta}_A/\sigma$ where $\boldsymbol{\beta}_P$ is the vector of coefficients for the WPH model and $\boldsymbol{\beta}_A$ is the vector of coefficients for the Weibull AFT model. Since the AFT is parametric, $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}_A$ are MLEs found from the censored data $(T_i, \delta_i, \boldsymbol{x}_i)$, not from $(Y_i, \boldsymbol{x}_i)$.

If the $Y_i|\boldsymbol{x}_i$ are Weibull, the e_i are from a smallest extreme value distribution. The statement that "*the Weibull regression model is both a proportional hazards model and an accelerated failure time model*" means that the $Y_i|\boldsymbol{x}_i$ follow a Weibull PH model while the $\log(Y_i)|\boldsymbol{x}_i$ follow a Weibull AFT, although the $\log(Y_i)$ are actually from a smallest extreme value distribution. If a Weibull or Exponential AFT is a useful model for the $\log(Y_i)|\boldsymbol{x}_i$, then the Weibull or Exponential PH model is a good approximation for the $Y_i|\boldsymbol{x}_i$. Hence to check the goodness of fit for the Weibull AFT, transform the Weibull AFT into the Weibull PH model. Then use the LCR, EE and slice survival plots as in Example 3.2.

Inference for the AFT model is performed almost in the same way as for the WPH or Weibull AFT. See Section 3.2. But the conclusions change slightly if the AFT is not the Weibull AFT. Change (if necessary) "Weibull survival model" to the appropriate model, e.g. "lognormal survival model". In the LRT, replace "WPH" by "AFT." Given $\hat{\boldsymbol{\beta}} \equiv \hat{\boldsymbol{\beta}}_A$ from output and given $\boldsymbol{x}$, know how to find ESP $= \hat{\boldsymbol{\beta}}^T \boldsymbol{x} = \sum_{i=1}^p \hat{\beta}_i x_i = \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$.

A large sample 95% CI for β_j is $\hat{\beta}_j \pm 1.96\ se(\hat{\beta}_j)$.

Know how to do the **4 step Wald test of hypotheses**:
i) State the hypotheses $H_0 : \beta_j = 0 \quad H_1 : \beta_j \neq 0$.
ii) Find the test statistic $z_{0,j} = \hat{\beta}_j / se(\hat{\beta}_j)$ or $X_{0,j}^2 = z_{0,j}^2$ or obtain it from output.
iii) The p–value $= 2P(Z < -|z_{0j}|) = P(\chi_1^2 > X_{0,j}^2)$. Find the p–value from output or use the standard normal table.
iv) If p-value $< \delta$, reject H_0 and conclude that x_j is needed in the AFT survival model given that the other $p-1$ predictors are in the model. If pval $\geq \delta$, fail to reject Ho, and conclude that the values of x_j do not (significantly)

affect the AFT survival model given that the other $p-1$ predictors are in the model. (Or state that there is not enough evidence to conclude that the values of x_j affect the AFT survival model.)

Know how to do the 4 step likelihood ratio test **LRT**:

i) $H_0 : \boldsymbol{\beta} = \mathbf{0} \quad H_A : \boldsymbol{\beta} \neq \mathbf{0}$

ii) test statistic $X^2(N|F) = [-2 \log L(none)] - [-2 \log L(full)]$ is often obtained from output.

iii) The p–value $= P(\chi^2_p > X^2(N|F))$ where χ^2_p has a chi–square distribution with p degrees of freedom. The p–value is often obtained from output.

iv) Reject H_0 if the p–value $< \delta$ and conclude that there is an AFT survival relationship between Y and the predictors $\boldsymbol{x}$. If p–value $\geq \delta$, then fail to reject H_0, and conclude that the values of the predictors $\boldsymbol{x}$ do not (significantly) affect the AFT survival model. (Or state that there is not enough evidence to conclude that the values of $\boldsymbol{x}$ affect the AFT survival model.)

Know how to do the 4 step **change in LR test**:

i) H_0: the reduced model is good $\quad H_A$: use the full model

ii) test statistic $X^2(R|F) = X^2(N|F) - X^2(N|R) = [-2 \log L(red)] - [-2 \log L(full)]$.

iii) The p–value $= P(\chi^2_{p-r} > X^2(R|F))$ where χ^2_{p-r} has a chi–square distribution with $p-r$ degrees of freedom.

iv) Reject H_0 if the p–value $< \delta$ and conclude that the full model should be used. If p–value $\geq \delta$, then fail to reject H_0, and conclude that the reduced model is good (the values of $\boldsymbol{x}_O$ do not (significantly) affect the survival model, or there is not enough evidence to conclude that the values of $\boldsymbol{x}_O$ affect the survival model).

3.4 Variable Selection

Since the Weibull proportional hazards model is valid for $(Y, \boldsymbol{x})$ if and only if the Weibull accelerated failure time model is valid for $(\log(Y), \boldsymbol{x})$, fit the data using Cox PH regression and perform variable selection such as forward selection, backward elemination, and relaxed lasso. Then fit each candidate submodel with WPH software and check the WPH assumptions. Transform the PH model to a Weibull AFT if the AFT is desired. The following chapter shows how to do inference after variable selection.

3.5 Summary

1) The **Weibull proportional hazards regression (WPH) model** is

$$h_i(t) = h_{Y_i|\boldsymbol{x}_i}(t) = h_{Y_i|\boldsymbol{\beta}_P^T\boldsymbol{x}_i}(t) = \exp(\boldsymbol{\beta}_P^T\boldsymbol{x}_i)h_0(t)$$

where $h_0(t) = h_0(t|\boldsymbol{\theta}) = \lambda_0\gamma t^{\gamma-1}$ is the **baseline function**. So $Y|SP \sim W(\gamma, \lambda_0 \exp(SP))$.

Assume that the WPH model is appropriate.

For SAS only.
log likelihood log L(none)

variable	Est.	SE	Est/SE	or $(Est/SE)^2$	pvalue
intercept					
scale					
Weibull shape					

For SAS or R

variable	Est.	SE	Est/SE	or $(Est/SE)^2$	pvalue for
intercept					
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	$X^2_{o,1} = z^2_{o,1}$	Ho: $\beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$z_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	$X^2_{o,p} = z^2_{o,p}$	Ho: $\beta_p = 0$
scale or	log scale				
Weibull shape	or scale				

For the full model, SAS will have Log Likelihood = log L(full).
For the full model, R will have log L(full), log L (none) and chisq = [-2 log L(none)] - [-2 log L(full)] on p degrees of freedom with pvalue

Replace full by reduced for the reduced model.

The SAS and R log likelihood, log L, differ by a constant.

```
SAS Log Likelihood = -29.7672 null model
variable      df Estimate  SE     chi square  pr > chisqu
intercept     1   7.1110   0.2927  590.12       < 0.0001
Weibull Scale 1   1225.4   358.7
Weibull Shape 1   1.1081   0.2810

SAS Log Likelihood = -29.1775  reduced model
variable      df Estimate  SE     chi square  pr > chisqu
intercept     1   7.3838   0.4370 285.45         < 0.0001
treat         1  -0.5593   0.5292   1.12           0.2906
Scale         1   0.8857   0.2227
Weibull Shape 1   1.1291   0.2840
```

```
SAS Log Likelihood = -20.5631    full model
variable      df Estimate  SE     chi square  pr > chisqu
intercept     1  11.5483   1.1970 93.07         < 0.0001
age           1  -0.0790   0.0198 15.97         < 0.0001
treat         1  -0.5615   0.3399 2.73            0.0986
Scale         1   0.5489   0.1291
Weibull Shape 1   1.8218   0.4286


R reduced  model Value Std. Error      z        p
(Intercept)      7.384      0.437 16.895 4.87e-64
treat           -0.559      0.529 -1.057 2.91e-01
Log(scale)      -0.121      0.251 -0.483 6.29e-01
Scale= 0.886
Loglik(model)= -97.4   Loglik(intercept only)= -98
        Chisq= 1.18 on 1 degrees of freedom, p= 0.28

R full model     Value Std. Error     z        p
(Intercept)     11.548     1.1970  9.65 5.04e-22
treat           -0.561     0.3399 -1.65 9.86e-02
age             -0.079     0.0198 -4.00 6.43e-05
Log(scale)      -0.600     0.2353 -2.55 1.08e-02
Scale= 0.549
Loglik(model)= -88.7   Loglik(intercept only)= -98
        Chisq= 18.41 on 2 degrees of freedom, p= 1e-04
```

Shown above is output in symbols from and *SAS* and *R* . The estimated coefficient is $\hat{\beta}_j$. The Wald chi square $= X^2_{o,j}$ while p and "pr > chisqu" are both p-values.

2) Instead of fitting the WHP model of 1), R and SAS fit an accelerated failure time model $\log(Y_i) = \alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}_i + \sigma\epsilon_i$ where $\mathrm{Var}(\epsilon_i) = 1$ and the ϵ_i are iid from a smallest extreme value distribution. Also $\boldsymbol{\beta}_A \neq \boldsymbol{\beta}_P$ from 1).

$\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ are MLEs found from the censored data $(T_i, \delta_i, \boldsymbol{x}_i)$ not from $(Y_i, \boldsymbol{x}_i)$.

3) Let $\log(T_i) = \hat{\alpha} + \hat{\boldsymbol{\beta}}_A^T \boldsymbol{x}_i + r_i$. A *log censored response (LCR) plot* is a plot of $\hat{\alpha} + \hat{\boldsymbol{\beta}}_A^T \boldsymbol{x}_i$ vs $\log(T_i)$ with plotting symbol 0 for censored cases and + for uncensored cases. The vertical deviations from the identity line $= r_i$. The least squares line based on the +'s can be added to the plot, and should have slope not too far from 1 for the Weibull AFT if $\gamma \geq 1$. The plotted points should be linear with roughly constant variance. The censoring and long left tails of the smallest extreme value distribution make judging linearity and detecting outliers from the left tail difficult. Try to ignore the bottom of the plot where there are few cases when assessing linearity.

4) Given $\hat{\boldsymbol{\beta}}$ from output and given $\boldsymbol{x}$, be able to find ESP $= \hat{\boldsymbol{\beta}}^T \boldsymbol{x} = \sum_{i=1}^p \hat{\beta}_i x_i = \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$.

5) A large sample 95% CI for β_j is $\hat{\beta}_j \pm 1.96\ se(\hat{\beta}_j)$.

6) **4 step Wald test of hypotheses**:
i) State the hypotheses Ho: $\beta_j = 0$ Ha: $\beta_j \neq 0$.
ii) Find the test statistic $z_{0,j} = \hat{\beta}_j / se(\hat{\beta}_j)$ or $X^2_{0,j} = z^2_{0,j}$ or obtain it from output.
iii) The p–value $= 2P(Z < -|z_{0j}|) = P(\chi^2_1 > X^2_{0,j})$. Find the p–value from output or use the standard normal table.
iv) If pval $< \delta$, reject Ho and conclude that x_j is needed in the Weibull survival model given that the other $p-1$ predictors are in the model. If pval $\geq \delta$, fail to reject Ho, and conclude that the values of x_j do not (significantly) affect the WPH survival model given that the other $p-1$ predictors are in the model. (Or state that there is not enough evidence to conclude that the values of x_j affect the WPH survival model.)

7) The 4 step likelihood ratio test **LRT** is
i) $H_o : \boldsymbol{\beta} = \mathbf{0} \quad H_A : \boldsymbol{\beta} \neq \mathbf{0}$
ii) test statistic $X^2(N|F) = [-2 \log L(none)] - [-2 \log L(full)]$ is often obtained from output.
iii) The p–value $= P(\chi^2_p > X^2(N|F))$ where χ^2_p has a chi–square distribution with p degrees of freedom. The p–value is often obtained from output.
iv) Reject H_o if the p–value $< \delta$ and conclude that there is a WPH survival relationship between Y and the predictors $\boldsymbol{x}$. If p–value $\geq \delta$, then fail to reject H_0, and conclude that the values of the predictors $\boldsymbol{x}$ do not (significantly) affect the WPH survival model. (Or state that there is not enough evidence to conclude that the values of $\boldsymbol{x}$ affect the WPH survival model.)

8) The 4 step **change in LR test** is
i) H_o: the reduced model is good $\quad H_A$: use the full model
ii) test statistic $X^2(R|F) = X^2(N|F) - X^2(N|R) = [-2 \log L(red)] - [-2 \log L(full)]$.
iii) The p–value $= P(\chi^2_{p-r} > X^2(R|F))$ where χ^2_{p-r} has a chi–square distribution with $p-r$ degrees of freedom.
iv) Reject H_o if the p–value $< \delta$ and conclude that the full model should be used. If p–value $\geq \delta$, then fail to reject H_o and conclude that the reduced model is good (the values of $\boldsymbol{x}_O$ do not (significantly) affect the survival model, or there is not enough evidence to conclude that the values of $\boldsymbol{x}_O$ affect the survival model).

9) R and SAS programs do not have a variable selection option, but the WPH model is a PH model, so use SAS Cox PH variable selection to suggest good submodels. Then fit each candidate with WPH software and check the WPH assumptions.

10) The **accelerated failure time (AFT) model** has $\log(Y_i) = \alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}_i + \sigma e_i$ where the e_i are iid from a location scale family.

If the Y_i are Weibull, the e_i are from a smallest extreme value distribution. The Weibull regression model is often said to be "both a proportional hazards model and an accelerated failure time model." Actually the Y_i follow a PH models and the $\log(Y_i)$ follow an AFT model.

If the Y_i are lognormal, the e_i are normal.

If the Y_i are loglogistic, the e_i are logistic.

11) Still use the *log censored response (LCR) plot* of 42). The LCR plot is easier to use when the ϵ_i are normal or logistic since these are symmetric distributions.

12) For the AFT model, $h_i(t) = e^{-SP} h_o(t/e^{SP})$ and $S_i(t) = S_0(t/\exp(SP))$.

13) Inference for the AFT model is performed exactly in the same way as for the WPH or Weibull AFT. See points 43) – 47). But the conclusion change slightly if the AFT is not the Weibull AFT. In point 45, change (if necessary) "Weibull survival model" to the appropriate model, eg "lognormal survival model". In point 46, change (if necessary) "WPH" to the appropriate model, eg "lognormal AFT".

In principle, the slice survival plot can be made for parametric AFT models, but the programming may be difficult.

The loglogistic and lognormal AFT models are not PH models. The loglogistic AFT is a proportional odds model.

14) Let $\boldsymbol{\beta}_C$ correspond to the Cox regression and $\boldsymbol{\beta}_A$ correspond to the AFT. An EE plot is a plot of the parametric ESP vs a semiparamtric ESP with the identity line added as a visual aid. The plotted points should follow the identity line with a correlation tending to 1.0 as $n \to \infty$.

15) For the Exponential regression model, $\sigma = 1$, and $\boldsymbol{\beta}_C = -\boldsymbol{\beta}_A$. The Exponential EE plot is a plot of $-ESPE = -\hat{\boldsymbol{\beta}}_A' \boldsymbol{x}$ vs $ESPC = \hat{\boldsymbol{\beta}}_C' \boldsymbol{x}$.

16) For the Weibull regression model, $\sigma = 1$, and $\boldsymbol{\beta}_C = -\boldsymbol{\beta}_A/\sigma$. The Weibull EE plot is a plot of

$$-ESPW/\hat{\sigma} = -\frac{1}{\hat{\sigma}} \hat{\boldsymbol{\beta}}_A' \boldsymbol{x} \ \text{ vs } \ \text{ESPC} = \hat{\boldsymbol{\beta}}_C' \boldsymbol{x}.$$

3.6 Complements

The Weibull PH regression model is the most widely used parametric PH regression model, but the Cox semiparametric PH regression model is used much more often for survival analysis. When the Weibull PH regression model

holds, the parametric inference is slightly better than the Cox PH regression model inference, but the Cox PH regression model gives good results for many data sets where the Weibull PH regression model does not hold.

A Weibull stratified PH regression model can be used where a Weibull PH regression model with the same $\boldsymbol{\beta}$ is used for each of the J strata. Then the α_j and σ_j depend on the strata for $j = 1, ..., J$.

For survival regression plots, see Olive (2011). Inference after variable selection and prediction intervals will be covered in Chapter 4.

A *proportional odds (PO) regression model* has

$$\frac{S_{\boldsymbol{x}}(t)}{1 - S_{\boldsymbol{x}}(t)} = e^{SP} \frac{S_{\mathbf{0}}(t)}{1 - S_{\mathbf{0}}(t)}$$

where $SP = \boldsymbol{\beta}_{PO}^T \boldsymbol{x}$. The logogistic regression model is the only model where $\log(Y)$ follows an AFT and Y follows a proportional odds regression model. For the loglogistic model, $\boldsymbol{\beta}_{PO} = \boldsymbol{\beta}_A/\sigma$.

For a proportional odds regression model, note that

$$\frac{S_{\boldsymbol{x}}(t)}{1 - S_{\boldsymbol{x}}(t)} = \frac{P(Y > t|\boldsymbol{x})}{1 - P(Y > t|\boldsymbol{x})} =$$

odds of survival beyond time t. Then the log odds ratio is

$$\log \left[\frac{\left(\frac{S_{\boldsymbol{x}}(t)}{1 - S_{\boldsymbol{x}}(t)} \right)}{\left(\frac{S_{\mathbf{0}}(t)}{1 - S_{\mathbf{0}}(t)} \right)} \right] = \boldsymbol{\beta}_{PO}^T \boldsymbol{x}.$$

Wei (1992) and Zeng and Lin (2007) give nonparametric methods for AFTs. These methods could be used to check a parametric AFT much like the Cox PH regression model can be used to check a parametric PH regression model like the Weibull PH regression model. Similarly, Bennett (1983) and Yang and Prentice (1999) give nonparametric methods for the proportional odds (PO) regression model, and these method could be used to check the parametric loglogistic PO regression model.

3.7 Problems

Problems with an asterisk * are especially important.

3.1. Leemis (1995, p. 190, 205-6) gives data on $n = 21$ leukemia patients taking the drug 6-MP. Suppose that the remission times given below follow an exponential (λ) distribution.

6, 6, 6, 6+, 7, 9+, 10, 10+, 11+, 13, 16, 17+,

19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+

a) Find $\hat{\lambda}$.

b) Find a 95% CI for λ.

3.2. Suppose that the lifetimes of a certain brand of lightbulb follow an exponential (λ) distribution. 20 light bulbs are tested for 1000 hours. The failure times are below.

71, 88, 254, 339, 372, 403, 498, 499, 593, 774, 935,
1000+, 1000+, 1000+, 1000+, 1000+, 1000+, 1000+, 1000+, 1000+

a) Find $\hat{\lambda}$.

b) Find a 95% CI for λ.

3.3. The following output is from a Weibull Regression for the Allison (1995, p. 270) recidivism data. The response variable *week* is time in weeks until arrest after release from prison (right censored if week = 52). The 7 variables are *Fin* (1 for those who received financial aid, 0 else), *Age* at time of release, *Race* (1 if black, 0 else), *Wexp*(1 if inmate had full time work experience prior to conviction, 0 else), *Mar* (1 if married at time of release, 0 else), *Paro* (1 if released on parole, 0 else), *Prio* (the number of prior convictions).

a) For the reduced model, find a 95% CI for β_1.

b) Test whether the reduced model is good.

```
Output for Problem 3.3  Null Model
Log Likelihood -336.08436 Standard 95% Confidence Chi-
Parameter  DF Estimate Error         Limits     Square Pr>ChiSq
Intercept  1   4.8177  0.1079   4.6062   5.0291 1994.47 <.0001
Scale      1   0.7325  0.0661   0.6138   0.8742
Weib Scale 1 123.6771 13.3417 100.1072 152.7964
Weib Shape 1   1.3651  0.1232   1.1438   1.6293

Full Model Log Likelihood -319.3765238
                        Standard 95% Confidence Chi-
Parameter   DF Estimate Error     Limits      Square Pr>ChiSq
Intercept   1  3.9901  0.4191  3.1687  4.8115 90.65  <.0001
fin         1  0.2722  0.1380  0.0018  0.5426  3.89  0.0485
age         1  0.0407  0.0160  0.0093  0.0721  6.47  0.0110
race        1 -0.2248  0.2202 -0.6563  0.2067  1.04  0.3072
wexp        1  0.1066  0.1515 -0.1905  0.4036  0.49  0.4820
mar         1  0.3113  0.2733 -0.2244  0.8469  1.30  0.2547
paro        1  0.0588  0.1396 -0.2149  0.3325  0.18  0.6735
prio        1 -0.0658  0.0209 -0.1069 -0.0248  9.88  0.0017
Scale       1  0.7124  0.0634  0.5983  0.8482
Weib. Shape 1  1.4037  0.1250  1.1789  1.6713
```

```
Reduced Model Log Likelihood -321.5012378
                        Standard 95% Confidence  Chi-
Parameter  DF Estimate Error    Limits        Square Pr>ChiSq
Intercept   1  3.7738 0.3581   3.0720  4.4755 111.08  <.0001
fin         1  0.2495 0.1372  -0.0194  0.5184   3.31  0.0690
age         1  0.0478 0.0154   0.0176  0.0779   9.66  0.0019
prio        1 -0.0698 0.0201  -0.1092 -0.0304  12.08  0.0005
Scale       1  0.7141 0.0637   0.5995  0.8506
Weib. Shape 1  1.4004 0.1250   1.1756  1.6681
```

```
Output for Problem 3.4
Log Likelihood -321.50124 Standard 95% Confidence Chi-
Parameter    DF Estimate  Error      Limits      Square Pr>ChiSq
Intercept     1  3.7738  0.3581  3.0720  4.4755  111.08  <.0001
fin           1  0.2495  0.1372 -0.0194  0.5184    3.31  0.0690
age           1  0.0478  0.0154  0.0176  0.0779    9.66  0.0019
prio          1 -0.0698  0.0201 -0.1092 -0.0304   12.08  0.0005
Scale         1  0.7141  0.0637  0.5995  0.8506
Weibull Shape 1  1.4004  0.1250  1.1756  1.6681
```

3.4. Above is output from a Weibull Regression for the Allison (1995, p. 270) recidivism data described in Problem 3.3. The full model has 3 predictors, *fin*, *age* and *prio*.

a) Suppose that the log likelihood for the null model is -336.08436. Test whether $\boldsymbol{\beta} = \mathbf{0}$.

b) Test whether $\beta_1 = 0$.

c) Test whether $\beta_2 = 0$.

```
 Output for 3.5
               Value Std. Error     z        p
(Intercept)  5.32632    0.66298  8.03 9.44e-16
age         -0.00891    0.00711 -1.25 0.210
sex          0.37019    0.12796  2.89 0.00382
ph.karno     0.00926    0.00446  2.08 0.0379
Log(scale)  -0.28085    0.06171 -4.55 5.33e-06

Scale= 0.755

Weibull distribution
Loglik(model)= -1138.7   Loglik(intercept only)= -1147.5
Chisq= 17.59 on 3 degrees of freedom, p= 0.00053
n=227 (1 observation deleted due to missingness)
```

3.5. A Weibull regression model was fit to the *R* `lung` data resulting in the above output.

a) Test whether $\boldsymbol{\beta} = \mathbf{0}$.

b) Test whether $\beta_1 = 0$.

c) Test whether $\beta_2 = 0$.

d) Sketch the Weibull EE plot if the Weibull model is good.

```
 Output for 3.6, n = 26
          coef exp(coef) se(coef)     z     p  full model
age      0.121      1.13   0.0484 2.500 0.012
resid.ds 0.792      2.21   0.8078 0.980 0.330
```

```
ecog.ps  0.087      1.09   0.6592 0.132 0.890
Likelihood ratio test= 13.7  on 3 df,   p=0.00333

     coef exp(coef) se(coef)   z      p       reduced model
age 0.137      1.15   0.0474 2.9 0.0038
Likelihood ratio test= 12.7  on 1 df,   p=0.000368
```

3.6. The *R* `ovarian` data gives survival times for patients with ovarian cancer. Predictors are *age* in years, *resid.ds* (residual disease present 1=no,2=yes), and *ecog.ps* (ECOG performance status: 1 is better than 2). A stratified proportional hazards model is fit where the stratification variable *rx* is the treatment group.

a) Test whether $\beta_3 = 0$.

b) Test whether $\boldsymbol{\beta} = \mathbf{0}$ for the full model.

c) Test whether the reduced model is good.

3.7. The *R* `lung` cancer data has the *time* until death or censoring. *ph.ecog* = Ecog performance score 0-4, *pat.karno* = patient's assessment of their karno score and *wt.loss* = weight loss in last 6 months. A stratified proportional hazards model is used and stratification is on *sex*.

a) Find the ESP and $\hat{h}_i(t)$ if $\boldsymbol{x} = (1.0, 80.0, 7.0)$ and $sex = F$.

b) Find a 95% CI for β_2.

c) Do a 4 step test for $Ho : \beta_2 = 0$.

d) Do a 4 step test for $Ho : \beta_3 = 0$.

e) *R* output says Likelihood ratio test=22.8.

Do a 4 step test for $Ho : \boldsymbol{\beta} = \mathbf{0}$.

```
 output for f)
             coef exp(coef) se(coef)      z    p
age       0.01444      1.01 0.010508  1.374 0.17
meal.cal -0.00016      1.00 0.000240 -0.666 0.51

Likelihood ratio test=2.97  on 2 df, p=0.227  n=181
(47 observations deleted due to missingness)
```

f) Now the SPH model uses the predictors *age* and *meal.cal* = calories consumed at meals excluding beverages and snacks.

Do a 4 step test for $Ho : \boldsymbol{\beta} = \mathbf{0}$.

R Problems

3.8. This problem considers the ovarian data from Collett (2003, p. 344-346).

a) Obtain the *R* code for 3.8 from (http://parker.ad.siu.edu/Olive/survhw.txt). Click on the left screen then hit *Enter*. Copy and paste both the

output. (It should be very similar to that on Section 3.5 between points 1) and 2).) Also copy and paste the plot into *Word*.

b) The plot is a log censored response plot. The top line is the identity line and the bottom line the least squares line. Is the slope of the least squares line near 1?

3.9. Use the source commands near the top of (http://parker.ad.siu.edu/Olive/survhw.txt) to get `survpack` into *R*. The programs `phdata`, `weyp` and `wregsim` will be used.

The program `wregsim` generates Weibull proportional hazards regression data with baseline hazard function $h_0(t) = \gamma t^{\gamma-1}$.

a) Type the command *wregsim(gam=1)* 5 times (or use the "up arrow" after typing the command once). This gives 5 simulated Weibull regression data sets with $\gamma = 1$. Hence the Weibull regression is also an exponential regression. Include the last plot in *Word*.

b) Type the command *wregsim(gam=5)* 5 times. To judge linearity, ignore the cases on the bottom of the plot with low density (points with log(time) less than -2). (These tend to be censored cases because time $Y = W^{1/\gamma}$ where $W \sim EXP(\lambda = \exp(SP))$ where $E(W) = 1/\lambda$. $Z \sim EXP(0.1)$ has mean 10 and if $Z_i < Y_i$ then Z_i is usually very small.) Do the plots seem linear ignoring the cases on the bottom of the plot? Do not include the plot.

c) Type the command *wregsim(gam=0.5)* 5 times. (Now censored cases tend to be large because time $Y = W^{1/\gamma} = W^2$ where $W \sim EXP(\lambda)$. $Z \sim EXP(0.1)$ has mean 10 and if $Z_i < Y_i$ then $Y_i > 10$, usually.) Do the plots seem linear (ignoring cases on the bottom of the plot)? (The plot is linear if it is roughly box shaped or ellipsoidal, possibly ignoring some of the points with $\log(time) < -9$. Since the error distribution is left skewed, most of the plotted points will fall below the identity line, even if the plot is linear.) Do not include the plot.

3.10. This problem considers the ovarian data from Collett (2003, pp. 189, 344-346).

a) Obtain the *R* code for 3.10a from (http://parker.ad.siu.edu/Olive/survhw.txt). Copy and paste the plot into *Word*.

b) Now obtain the R code for 3.10b and put the plot into *Word*.

c) Can the Exponential regression model be used or should the more complicated Weibull regression model be used?

3.11. Copy and paste the two source commands from the top of (http://parker.ad.siu.edu/Olive/survhw.txt) to get programs `phdata` and `wregsim2` into *R*.

Make the left window small by moving the cursor to the lower right corner of the window, then hold the right mouse button down and drag the window to the left.

The program `wregsim2` generates Weibull proportional hazards regression data with baseline hazard function $h_0(t) = \gamma t^{\gamma-1}$.

a) Type the command *wregsim2(n=10, gam=1)* 5 times (or use the "up arrow" after typing the command once). This gives 5 simulated Weibull regression data sets with $\gamma = 1$. Increase n by 10 until the plotted points cluster tightly about the identity line in at least 4 out of 5 times. How big is n?

b) Type the command *wregsim2(n =10, gam=5)* 5 times. Increase n by 10 until the plotted points cluster tightly about the identity line in at least 4 out of 5 times. How big is n?

c) Type the command *wregsim2(n=10, gam=0.5)* 5 times. Increase n by 10 until the plotted points cluster tightly about the identity line in at least 4 out of 5 times. How big is n?

3.12. If necessary copy and paste the two source commands as done for Problem 3.11 to get programs `phdata` and `wregsim3` into *R*.

Make the left window small by moving the cursor to the lower right corner of the window, then hold the right mouse button down and drag the window to the left.

The program `wregsim3` generates Weibull proportional hazards regression data with baseline hazard function $h_0(t) = \gamma t^{\gamma-1}$. This is also an AFT model with $\alpha = 0$, $\boldsymbol{\beta}' = -(1/\gamma, ..., 1/\gamma)$ and $\sigma = 1/\gamma$. The program generate 100 Weibull AFT data sets and for each run i computes $\hat{\alpha}_i$, $\hat{\boldsymbol{\beta}}_i$ and $\hat{\sigma}_i$. Then the averages are reported. Want mnint ≈ 0, mncoef $\approx -(1/\gamma, ..., 1/\gamma)$ and mnscale $\approx 1/\gamma$.

a) Make a table (by hand) with headers

```
n    gamma    mnint   mncoef                          mnscale
```

Fill in the table for $n = 20, \gamma = 1; n = 100, \gamma = 1; n = 200, \gamma = 1; n = 20, \gamma = 5; n = 100, \gamma = 5; n = 200, \gamma = 5; n = 20, \gamma = 0.5; n = 100, \gamma = 0.5; n = 200, \gamma = 0.5$ by using the commands *wregsim3(n=20, gam=1)*, ..., *wregsim3(n=200, gam=0.5)*.

b) Are the estimators close to parameters $\alpha, \boldsymbol{\beta}$ and σ for $n = 20$? How about for $n = 100$?

3.13. If necessary copy and paste the two source commands as done for problem 3.11 to get programs `wphsim` and `swhat` into *R*. Type the command *wphsim(n=999)* to make a slice survival plot based on the WPH survival function. Are the KM curve and Weibull estimated survival function close for the plot in the bottom right corner? Include the plot in *Word*. Recent versions of *R* may make 3 curves of cirles. The center curve is the KM curve while the 2 outer curves are pointwise CI bands. (When 3 curves of cirles are made, if the plusses are near or within the circles, then the plots suggest that the WPH model is good.)

3.14. The *R* `lung` cancer data has the *time* until death or censoring and *status* = 0 for censored and 1 for uncensored. Then the covariates are *age*, *sex* = 1 for M and 2 for F, *ph.ecog* = Ecog performance score 0-4, *ph.karno* = a competitor to ph.ecog, *pat.karno* = patient's assessment of their karno score, *meal.cal* = calories consumed at meals excluding beverages and snacks and *wt.loss* = weight loss in last 6 months. The *R* output will use a stratified proportional hazards model that is stratified on *sex* with variables *ph.ecog*, *pat.karno* and *wt.loss*.

a) Copy and paste commands from (http://parker.ad.siu.edu/Olive/survhw.txt) for this problem into *R*. Click on the left window and hit *Enter*. Include the plot in *Word*. Also include the *R* output in *Word*.

b) Test whether $\boldsymbol{\beta} = \mathbf{0}$.

c) Based on the plot, do females or males appear to have better survival rates?

SAS Problem

3.15. This problem considers the ovarian data from Collett (2003, p. 344-346).

a) Obtain the SAS program for 3.15 from (http://parker.ad.siu.edu/Olive/survhw.txt). Print the output. (It should be very similar to that on Section 3.5 between points 1) and 2).)

b) Find the ESP if *age* = 40 and *treat 1* = 1. (Comment: treatment takes on 2 levels so only one indicator is needed. SAS output includes a 2nd indicator *treat 2* but its coefficient is $\hat{\beta}_3 = 0$ and hence can be ignored. In general if the category takes on J levels, SAS will give nonzero output for the first J − 1 levels and a line of 0s for the Jth level. This means level J was omitted and the line of 0s should be ignored.)

c) Give a 95% CI for β_1 corresponding to age from output and the CI using the formula.

d) Give a 95% CI for β_2 corresponding to treat 1 from output and the CI using the formula.

e) If the model statement in the SAS program is changed to

model survtime*status(0)=;

then the null model is fit and the SAS output says Log Likelihood −29.76723997.

Test $\boldsymbol{\beta} = \mathbf{0}$ with the LR test.
(Hint: The full model log likelihood log(L) = −20.56313339. Want −2 log(L) for both the full and null models for the LR test.)

f) Suppose the reduced model does not include *treat*. Then SAS output says Log Likelihood −21.7830. Test whether the reduced model is good.
(Hint: The log likelihood for the full model is log(L) = −20.56313339. Want −2 log(L) for the full and reduced models for the change in LR test.)

Problems from Quizzes and Exams

```
Output for Problem 3.16
Variable   Estimate Std Err   Chi-Square Pr > ChiSq
Intercept  3.9915   0.4349    84.2322       0.0001
fin        0.2724   0.1401    3.7806        0.0518
age        0.04066  0.01655   6.0398        0.0140
race      -0.2255   0.2280    0.9782        0.3226
wexp       0.1073   0.1660    0.4184        0.5177
mar        0.3118   0.2769    1.2679        0.2602
paro       0.05879  0.1398    0.1768        0.6741
prio      -0.06586  0.02130   9.5584        0.0020
scale      0.7151   0.2396
shape      0.9943   0.4849
```

3.16. The recidivism data is from Allison (1995, p. 75). A generalized gamma AFT is fit and has intercept, scale and shape parameters which are not predictors (the other AFTs in this class had two extra parameters). The response variable Y is time until arrest. The testing is like that of the WPH model, except used "generalized gamma AFT" instead of WPH in the appropriate conclusion.

a) Test $\beta_1 = 0$ which corresponds to fin.

b) Test $\beta_2 = 0$ which corresponds to age.

c) An EE plot could be made with the generalized gamma ESP on the vertical axis and the Weibull AFT ESP on the horizontal axis since the Weibull distribution is a special case of the genearlized gamma distribution. Suppose the plotted points cluster about the identity line. Is the Weibull AFT good (or bad)?

```
Output for Problem 3.17
              Value Std. Error      z        p
(Intercept) 15.1449    16.0795  0.942 3.46e-01
age         -0.1291     0.2186 -0.590 5.55e-01
quant       -0.0455     0.0583 -0.782 4.34e-01
Log(scale)   1.7179     0.3103  5.536 3.10e-08

Scale= 5.57   n =20
Loglik(model)= -28.9   Loglik(intercept only)= -29.5
        Chisq= 1.1 on 2 degrees of freedom, p= 0.58
```

3.17. The R data set Tobin Data uses a lognormal AFT. (Handled like a WPH or Weibull AFT except use "lognormal AFT" instead of WPH in the appropriate conclusion.) The predictors are *age*, and *quant*.

a) Test $\boldsymbol{\beta} = \mathbf{0}$.

b) Test $\beta_2 = 0$.

c) Find the ESP $= \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ if $x_1 = age = 50$ and $x_2 = quant = 270$.

Chapter 4
Inference After Variable Selection

This chapter considers inference after variable selection including prediction intervals and bootstrap hypothesis testing. Prediction regions and prediction intervals applied to a bootstrap sample can result in confidence regions and confidence intervals. The bootstrap confidence regions will be used for inference after variable selection. Several of the sections of this chapter are much more technical than the rest of the book.

4.1 Variable Selection

Review Section 2.4 for variable selection. Simpler models are easier to explain and use than more complicated models, and there are several other important reasons to perform variable selection.

When there is a sequence of M submodels, the final submodel I_d needs to be selected with a_d terms. Let the candidate model I contain a terms, including a constant, and let $\boldsymbol{x}_I$ and $\hat{\boldsymbol{\beta}}_I$ be $a \times 1$ vectors. Then there are many criteria used to select the final submodel I_d. Forward selection or backward elimination with the Akaike (1973) AIC criterion or Schwarz (1978) BIC criterion are often used for variable selection. The relaxed lasso or relaxed elastic net estimator fits the regression method, such as the Cox (1972) proportional hazards regression, to the predictors than had nonzero lasso or elastic net coefficients. See Tibshirani (1997) and Simon et al. (2011) for lasso and elastic net.

Forward selection and backward elimination both form a sequence of submodels $I_1, ..., I_p$ where I_j uses j predictors. Heuristically, backward elimination tries to delete the variable that will increase AIC the least, while forward selection tries to add the variable that will decrease AIC the most. Let I_{min} minimize the criterion such as AIC, BIC, or lasso. Often I_{min} from forward selection will differ from I_{min} from backward elimination, especially if the predictors are correlated.

Now suppose $p = 6$ and S in Equation (2.4) corresponds to x_1, x_2, and x_3. Suppose the data set is such that underfitting (omitting a predictor in S) does not occur. Then there are eight possible submodels that contain S: i) x_1, x_2, x_3; ii) x_1, x_2, x_3, x_4; iii) x_1, x_2, x_3, x_5; iv) x_1, x_2, x_3, x_6; v) x_1, x_2, x_3, x_4, x_5; vi) x_1, x_2, x_3, x_4, x_6; vii) x_1, x_2, x_3, x_5, x_6; and the full model viii) $x_1, x_2, x_3, x_4, x_5, x_6$. The possible submodel sizes are $k = 3, 4, 5$, or 6. Suppose $I_{min} = I_d$. Compared to selecting a model I_d before examining the data, the model I_{min} fits the data a bit too well. The fact that the selected model I_{min} from variable selection cannot be used as the full model for classical inference is known as **selection bias**.

If $\hat{\boldsymbol{\beta}}_{I_{min}}$ is $a \times 1$, form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I_{min},0}$ from $\hat{\boldsymbol{\beta}}_{I_{min}}$ by adding 0s corresponding to the omitted variables. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$, then $\hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$.

This chapter offers two remedies: i) use the large sample theory of $\hat{\boldsymbol{\beta}}_{I_{min},0}$ (defined two paragraphs below) and the bootstrap for inference after variable selection, and ii) use data splitting for inference after variable selection.

4.2 Some Tools for Large Sample Theory

This section gives some tools that are useful for inference after variable selection. The multivariate normal distribution is important. The last four subsections are more technical than most of this book. They can be omitted on first reading and refer to relevant theorems as needed.

4.2.1 The Multivariate Normal Distribution

For much of this book, $\boldsymbol{X}$ is an $n \times p$ design matrix, but this subsection will usually use the notation $\boldsymbol{X} = (X_1, ..., X_p)^T$ and $\boldsymbol{Y}$ for the random vectors, and $\boldsymbol{x} = (x_1, ..., x_p)^T$ for the observed value of the random vector. This notation will be useful to avoid confusion when studying conditional distributions such as $\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}$. It can be shown that $\boldsymbol{\Sigma}$ is positive semidefinite and symmetric.

Definition 4.1: Rao (1965, p. 437). A $p \times 1$ random vector $\boldsymbol{X}$ has a $p-$dimensional *multivariate normal distribution* $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ iff $\boldsymbol{t}^T\boldsymbol{X}$ has a univariate normal distribution for any $p \times 1$ vector $\boldsymbol{t}$.

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

$$f(\boldsymbol{z}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(1/2)(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})} \tag{4.1}$$

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

Definition 4.2. The *population mean* of a random $p \times 1$ vector $\boldsymbol{X} = (X_1, ..., X_p)^T$ is

$$E(\boldsymbol{X}) = (E(X_1), ..., E(X_p))^T$$

and the $p \times p$ *population covariance matrix*

$$\text{Cov}(\boldsymbol{X}) = E(\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{X} - E(\boldsymbol{X}))^T = (\sigma_{ij}).$$

That is, the ij entry of $\text{Cov}(\boldsymbol{X})$ is $\text{Cov}(X_i, X_j) = \sigma_{ij}$.

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

$$E(\boldsymbol{a} + \boldsymbol{X}) = \boldsymbol{a} + E(\boldsymbol{X}) \quad \text{and} \quad E(\boldsymbol{X} + \boldsymbol{Y}) = E(\boldsymbol{X}) + E(\boldsymbol{Y}) \tag{4.2}$$

and

$$E(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}E(\boldsymbol{X}) \quad \text{and} \quad E(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}. \tag{4.3}$$

Thus

$$\text{Cov}(\boldsymbol{a} + \boldsymbol{A}\boldsymbol{X}) = \text{Cov}(\boldsymbol{A}\boldsymbol{X}) = \boldsymbol{A}\text{Cov}(\boldsymbol{X})\boldsymbol{A}^T. \tag{4.4}$$

Some important properties of multivariate normal (MVN) distributions are given in the following three theorems. These theorems can be proved using results from Johnson and Wichern (1988, pp. 127-132) or Severini (2005, ch. 8).

Theorem 4.1. a) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and

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

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

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

d) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and if $\boldsymbol{A}$ is a $q \times p$ matrix, then $\boldsymbol{AX} \sim N_q(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^T)$. If $\boldsymbol{a}$ is a $p \times 1$ vector of constants and b is a constant, then $\boldsymbol{a} + b\boldsymbol{X} \sim N_p(\boldsymbol{a} + b\boldsymbol{\mu}, b^2\boldsymbol{\Sigma})$. (Note that $b\boldsymbol{X} = b\boldsymbol{I}_p\boldsymbol{X}$ with $\boldsymbol{A} = b\boldsymbol{I}_p$.)

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

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

Theorem 4.2. a) **All subsets of a MVN are MVN:** $(X_{k_1}, ..., X_{k_q})^T \sim N_q(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ where $\tilde{\boldsymbol{\mu}}_i = E(X_{k_i})$ and $\tilde{\boldsymbol{\Sigma}}_{ij} = \text{Cov}(X_{k_i}, X_{k_j})$. In particular, $\boldsymbol{X}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{X}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$.

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

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

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

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

Theorem 4.3. The conditional distribution of a MVN is MVN. If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. That is,

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

Example 4.1. Let $p = 2$ and let $(Y, X)^T$ have a bivariate normal distribution. That is,

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

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

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

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

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

and the conditional variance

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

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

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

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

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

Remark 4.1. There are several common misconceptions. First, **it is not true that every linear combination $\boldsymbol{t}^T\boldsymbol{X}$ of normal random variables is a normal random variable,** and **it is not true that all uncorrelated normal random variables are independent**. The key condition in Theorem 4.1b and Theorem 4.2c is that the joint distribution of $\boldsymbol{X}$ is MVN. It is possible that $X_1, X_2, ..., X_p$ each has a marginal distribution that is univariate normal, but the joint distribution of $\boldsymbol{X}$ is not MVN. Examine the following example from Rohatgi (1976, p. 229). Suppose that the joint pdf of X and Y is a mixture of two bivariate normal distributions both with $EX = EY = 0$ and $\text{VAR}(X) = \text{VAR}(Y) = 1$, but $\text{Cov}(X,Y) = \pm\rho$. Hence $f(x,y) =$

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

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

where x and y are real and $0 < \rho < 1$. Since both marginal distributions of $f_i(x,y)$ are N(0,1) for $i = 1$ and 2 by Theorem 4.2 a), the marginal distributions of X and Y are N(0,1). Since $\int\int xyf_i(x,y)dxdy = \rho$ for $i = 1$ and $-\rho$ for $i = 2$, X and Y are uncorrelated, but X and Y are not independent since $f(x,y) \neq f_X(x)f_Y(y)$.

Remark 4.2. In Theorem 4.3, suppose that $\boldsymbol{X} = (Y, X_2, ..., X_p)^T$. Let $X_1 = Y$ and $\boldsymbol{X}_2 = (X_2, ..., X_p)^T$. Then $E[Y|\boldsymbol{X}_2] = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ and $\text{VAR}[Y|\boldsymbol{X}_2]$ is a constant that does not depend on $\boldsymbol{X}_2$. Hence $Y|\boldsymbol{X}_2 = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p + e$ follows the multiple linear regression model.

4.2.2 The CLT and the Delta Method

The next three subsections will review large sample theory for the univariate case, then multivariate theory will be given.

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

Theorem 4.4: the Central Limit Theorem (CLT). Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and $\text{VAR}(Y) = \sigma^2$. Let the sample mean $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$. Then

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

Hence

$$\sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) = \sqrt{n}\left(\frac{\sum_{i=1}^n Y_i - n\mu}{n\sigma}\right) \xrightarrow{D} N(0,1).$$

Note that the sample mean is estimating the *population mean* μ with a $\sqrt{n}$ convergence rate, the asymptotic distribution is normal, and the SE $= S/\sqrt{n}$ where S is the *sample standard deviation.* For distributions "close" to the normal distribution, the central limit theorem provides a good approximation if the sample size $n \geq 30$. Hesterberg (2014, pp. 41, 66) suggests $n \geq 5000$ is needed for moderately skewed distributions. A special case of the CLT is proven after Theorem 4.17.

Notation. The notation $X \sim Y$ and $X \stackrel{D}{=} Y$ both mean that the random variables X and Y have the same distribution. Hence $F_X(x) = F_Y(y)$ for all real y. The notation $Y_n \xrightarrow{D} X$ means that for large n we can approximate the cdf of Y_n by the cdf of X. The distribution of X is the limiting distribution or asymptotic distribution of Y_n. For the CLT, notice that

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

is the z–score of $\overline{Y}$. If $Z_n \xrightarrow{D} N(0,1)$, then the notation $Z_n \approx N(0,1)$, also written as $Z_n \sim AN(0,1)$, means approximate the cdf of Z_n by the standard normal cdf. See Definition 4.3. Similarly, the notation

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

also written as $\overline{Y}_n \sim AN(\mu, \sigma^2/n)$, means approximate the cdf of $\overline{Y}_n$ as if $\overline{Y}_n \sim N(\mu, \sigma^2/n)$. The distribution of X does not depend on n, but the approximate distribution $\overline{Y}_n \approx N(\mu, \sigma^2/n)$ does depend on n.

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

Example 4.2. a) Let $Y_1, ..., Y_n$ be iid Ber(ρ). Then $E(Y) = \rho$ and $\mathrm{VAR}(Y) = \rho(1-\rho)$. (The Bernoulli ($\rho$) distribution is the binomial (1,ρ) distribution.) Hence

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

by the CLT.

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

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

since

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

by a).

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

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

or

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

Theorem 4.5: the Delta Method. If g does not depend on n, $g'(\theta) \neq 0$, and

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

then

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

Example 4.3. Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and $\mathrm{VAR}(Y) = \sigma^2$. Then by the CLT,

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

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

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

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

Example 4.4. Let $X \sim \text{Binomial}(n, p)$ where the positive integer n is large and $0 < p < 1$. Find the limiting distribution of $\sqrt{n}\ \left[\ \left(\frac{X}{n}\right)^2 -\ p^2\ \right]$.

Solution. Example 4.2b gives the limiting distribution of $\sqrt{n}(\frac{X}{n} - p)$. Let $g(p) = p^2$. Then $g'(p) = 2p$ and by the delta method,

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

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

Example 4.5. Let $X_n \sim \text{Poisson}(n\lambda)$ where the positive integer n is large and $\lambda > 0$.

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

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

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

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

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

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

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

Example 4.6. Let $Y_1, ..., Y_n$ be independent and identically distributed (iid) from a Gamma(α, β) distribution.

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

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

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

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

4.2.3 Modes of Convergence and Consistency

Definition 4.3. Let $\{Z_n, n = 1, 2, ...\}$ be a sequence of random variables with cdfs F_n, and let X be a random variable with cdf F. Then Z_n **converges in distribution to** X, written

$$Z_n \xrightarrow{D} X,$$

or Z_n *converges in law to* X, written $Z_n \xrightarrow{L} X$, if

$$\lim_{n\to\infty} F_n(t) = F(t)$$

at each continuity point t of F. The distribution of X is called the **limiting distribution** or the **asymptotic distribution** of Z_n.

An important fact is that **the limiting distribution does not depend on the sample size** n. Notice that the CLT and delta method give the limiting distributions of $Z_n = \sqrt{n}(\overline{Y}_n - \mu)$ and $Z_n = \sqrt{n}(g(T_n) - g(\theta))$, respectively.

Convergence in distribution is useful if the distribution of X_n is unknown or complicated and the distribution of X is easy to use. Then for large n we can approximate the probability that X_n is in an interval by the probability that X is in the interval. To see this, notice that if $X_n \xrightarrow{D} X$, then $P(a < X_n \leq b) = F_n(b) - F_n(a) \to F(b) - F(a) = P(a < X \leq b)$ if F is continuous at a and b.

Warning: convergence in distribution says that the cdf $F_n(t)$ of X_n gets close to the cdf of $F(t)$ of X as $n \to \infty$ provided that t is a continuity point of F. Hence for any $\epsilon > 0$ there exists N_t such that if $n > N_t$, then $|F_n(t) - F(t)| < \epsilon$. Notice that N_t depends on the value of t. Convergence in distribution does not imply that the random variables $X_n \equiv X_n(\omega)$ converge to the random variable $X \equiv X(\omega)$ for all ω.

Example 4.7. Suppose that $X_n \sim U(-1/n, 1/n)$. Then the cdf $F_n(x)$ of X_n is

$$F_n(x) = \begin{cases} 0, & x \leq \frac{-1}{n} \\ \frac{nx}{2} + \frac{1}{2}, & \frac{-1}{n} \leq x \leq \frac{1}{n} \\ 1, & x \geq \frac{1}{n}. \end{cases}$$

Sketching $F_n(x)$ shows that it has a line segment rising from 0 at $x = -1/n$ to 1 at $x = 1/n$ and that $F_n(0) = 0.5$ for all $n \geq 1$. Examining the cases $x < 0$, $x = 0$, and $x > 0$ shows that as $n \to \infty$,

$$F_n(x) \to \begin{cases} 0, \ x < 0 \\ \frac{1}{2} \ \ x = 0 \\ 1, \ x > 0. \end{cases}$$

Notice that the right hand side is not a cdf since right continuity does not hold at $x = 0$. Notice that if X is a random variable such that $P(X = 0) = 1$, then X has cdf

$$F_X(x) = \begin{cases} 0, \ x < 0 \\ 1, \ x \geq 0. \end{cases}$$

Since $x = 0$ is the only discontinuity point of $F_X(x)$ and since $F_n(x) \to F_X(x)$ for all continuity points of $F_X(x)$ (i.e. for $x \neq 0$),

$$X_n \xrightarrow{D} X.$$

Example 4.8. Suppose $Y_n \sim U(0, n)$. Then $F_n(t) = t/n$ for $0 < t \leq n$ and $F_n(t) = 0$ for $t \leq 0$. Hence $\lim_{n\to\infty} F_n(t) = 0$ for $t \leq 0$. If $t > 0$ and $n > t$, then $F_n(t) = t/n \to 0$ as $n \to \infty$. Thus $\lim_{n\to\infty} F_n(t) = 0$ for all t, and Y_n does not converge in distribution to any random variable Y since $H(t) \equiv 0$ is not a cdf.

Definition 4.4. A sequence of random variables X_n *converges in distribution to a constant* $\tau(\theta)$, written

$$X_n \xrightarrow{D} \tau(\theta), \quad \text{if} \quad \mathrm{X_n} \xrightarrow{\mathrm{D}} \mathrm{X}$$

where $P(X = \tau(\theta)) = 1$. The distribution of the random variable X is said to be *degenerate at* $\tau(\theta)$ or to be a *point mass at* $\tau(\theta)$.

Definition 4.5. A sequence of random variables X_n *converges in probability to a constant* $\tau(\theta)$, written

$$X_n \xrightarrow{P} \tau(\theta),$$

if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - \tau(\theta)| < \epsilon) = 1 \ \ \text{or, equivalently,} \ \ \lim_{\mathrm{n}\to\infty} \mathrm{P}(|\mathrm{X_n} - \tau(\theta)| \geq \epsilon) = 0.$$

The sequence X_n **converges in probability to** X, written

$$X_n \xrightarrow{P} X,$$

if $X_n - X \xrightarrow{P} 0$.

Notice that $X_n \xrightarrow{P} X$ if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1, \quad \text{or, equivalently,} \quad \lim_{\mathrm{n}\to\infty} \mathrm{P}(|\mathrm{X_n} - \mathrm{X}| \geq \epsilon) = 0.$$

Definition 4.6. Let the *parameter space* Θ be the set of possible values of θ. A sequence of estimators T_n of $\tau(\theta)$ is **consistent** for $\tau(\theta)$ if

$$T_n \xrightarrow{P} \tau(\theta)$$

for every $\theta \in \Theta$. If T_n is consistent for $\tau(\theta)$, then T_n is a **consistent estimator** of $\tau(\theta)$.

Consistency is a weak property that is usually satisfied by good estimators. T_n is a consistent estimator for $\tau(\theta)$ if the probability that T_n falls in any neighborhood of $\tau(\theta)$ goes to one, regardless of the value of $\theta \in \Theta$.

Definition 4.7. For a real number $r > 0$, Y_n *converges in rth mean* to a random variable Y, written

$$Y_n \xrightarrow{r} Y,$$

if

$$E(|Y_n - Y|^r) \to 0$$

as $n \to \infty$. In particular, if $r = 2$, Y_n **converges in quadratic mean** to Y, written

$$Y_n \xrightarrow{2} Y \quad \text{or} \quad \mathrm{Y_n} \xrightarrow{\mathrm{qm}} \mathrm{Y},$$

if

$$E[(Y_n - Y)^2] \to 0$$

as $n \to \infty$.

Theorem 4.6: Generalized Chebyshev's Inequality. Let $u : \mathbb{R} \to [0, \infty)$ be a nonnegative function. If $E[u(Y)]$ exists then for any $c > 0$,

$$P[u(Y) \geq c] \leq \frac{E[u(Y)]}{c}.$$

If $\mu = E(Y)$ exists, then taking $u(y) = |y - \mu|^r$ and $\tilde{c} = c^r$ gives
Markov's Inequality: for $r > 0$ and any $c > 0$,

$$P[|Y - \mu| \geq c] = P[|Y - \mu|^r \geq c^r] \leq \frac{E[|Y - \mu|^r]}{c^r}.$$

If $r = 2$ and $\sigma^2 = \mathrm{VAR}(Y)$ exists, then we obtain
Chebyshev's Inequality:

$$P[|Y - \mu| \geq c] \leq \frac{\mathrm{VAR}(Y)}{c^2}.$$

Proof. The proof is given for pdfs. For pmfs, replace the integrals by sums. Now

$$E[u(Y)] = \int_{\mathbb{R}} u(y)f(y)dy = \int_{\{y:u(y)\geq c\}} u(y)f(y)dy + \int_{\{y:u(y)<c\}} u(y)f(y)dy$$

$$\geq \int_{\{y:u(y)\geq c\}} u(y)f(y)dy$$

since the integrand $u(y)f(y) \geq 0$. Hence

$$E[u(Y)] \geq c \int_{\{y:u(y)\geq c\}} f(y)dy = cP[u(Y) \geq c]. \quad \square$$

The following theorem gives sufficient conditions for T_n to be a consistent estimator of $\tau(\theta)$. Notice that $E_\theta[(T_n - \tau(\theta))^2] = MSE_{\tau(\theta)}(T_n) \to 0$ for all $\theta \in \Theta$ is equivalent to $T_n \stackrel{qm}{\to} \tau(\theta)$ for all $\theta \in \Theta$.

Theorem 4.7. a) If

$$\lim_{n\to\infty} MSE_{\tau(\theta)}(T_n) = 0$$

for all $\theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$.

b) If

$$\lim_{n\to\infty} \mathrm{VAR}_\theta(T_n) = 0 \quad \text{and} \quad \lim_{n\to\infty} E_\theta(T_n) = \tau(\theta)$$

for all $\theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$.

Proof. a) Using Theorem 4.6 with $Y = T_n$, $u(T_n) = (T_n - \tau(\theta))^2$ and $c = \epsilon^2$ shows that for any $\epsilon > 0$,

$$P_\theta(|T_n - \tau(\theta)| \geq \epsilon) = P_\theta[(T_n - \tau(\theta))^2 \geq \epsilon^2] \leq \frac{E_\theta[(T_n - \tau(\theta))^2]}{\epsilon^2}.$$

Hence

$$\lim_{n\to\infty} E_\theta[(T_n - \tau(\theta))^2] = \lim_{n\to\infty} MSE_{\tau(\theta)}(T_n) \to 0$$

is a sufficient condition for T_n to be a consistent estimator of $\tau(\theta)$.

b) Recall that

$$MSE_{\tau(\theta)}(T_n) = \mathrm{VAR}_\theta(T_n) + [\mathrm{Bias}_{\tau(\theta)}(\mathrm{T_n})]^2$$

where $\mathrm{Bias}_{\tau(\theta)}(\mathrm{T_n}) = \mathrm{E}_\theta(\mathrm{T_n}) - \tau(\theta)$. Since $MSE_{\tau(\theta)}(T_n) \to 0$ if both $\mathrm{VAR}_\theta(T_n) \to 0$ and $\mathrm{Bias}_{\tau(\theta)}(\mathrm{T_n}) = \mathrm{E}_\theta(\mathrm{T_n}) - \tau(\theta) \to 0$, the result follows from a). $\square$

The following result shows estimators that converge at a $\sqrt{n}$ rate are consistent. Use this result and the delta method to show that $g(T_n)$ is a consistent

estimator of $g(\theta)$. Note that b) follows from a) with $X_\theta \sim N(0, v(\theta))$. The WLLN shows that $\overline{Y}$ is a consistent estimator of $E(Y) = \mu$ if $E(Y)$ exists.

Theorem 4.8. a) Let X_θ be a random variable with distribution depending on θ, and $0 < \delta \leq 1$. If

$$n^\delta (T_n - \tau(\theta)) \xrightarrow{D} X_\theta$$

then $T_n \xrightarrow{P} \tau(\theta)$.

b) If

$$\sqrt{n}(T_n - \tau(\theta)) \xrightarrow{D} N(0, v(\theta))$$

for all $\theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$.

Definition 4.8. A sequence of random variables X_n *converges almost everywhere* (or *almost surely*, or *with probability 1*) to X if

$$P(\lim_{n\to\infty} X_n = X) = 1.$$

This type of convergence will be denoted by

$$X_n \xrightarrow{ae} X.$$

Notation such as "X_n converges to X ae" will also be used. Sometimes "ae" will be replaced with "as" or "wp1." We say that X_n *converges almost everywhere* to $\tau(\theta)$, written

$$X_n \xrightarrow{ae} \tau(\theta),$$

if $P(\lim_{n\to\infty} X_n = \tau(\theta)) = 1$.

Theorem 4.9. Let Y_n be a sequence of iid random variables with $E(Y_i) = \mu$. Then

a) **Strong Law of Large Numbers (SLLN):** $\overline{Y}_n \xrightarrow{ae} \mu$, and

b) **Weak Law of Large Numbers (WLLN):** $\overline{Y}_n \xrightarrow{P} \mu$.

Proof of WLLN when $V(Y_i) = \sigma^2$: By Chebyshev's inequality, for every $\epsilon > 0$,

$$P(|\overline{Y}_n - \mu| \geq \epsilon) \leq \frac{V(\overline{Y}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$$

as $n \to \infty$. $\square$

In proving consistency results, there is an infinite sequence of estimators that depend on the sample size n. Hence the subscript n will be added to the estimators.

Definition 4.9. Lehmann (1999, pp. 53-54): a) A sequence of random variables W_n is *tight* or *bounded in probability,* written $W_n = O_P(1)$, if for every $\epsilon > 0$ there exist positive constants D_ϵ and N_ϵ such that

$$P(|W_n| \le D_\epsilon) \ge 1 - \epsilon$$

for all $n \ge N_\epsilon$. Also $W_n = O_P(X_n)$ if $|W_n/X_n| = O_P(1)$.

b) The sequence $W_n = o_P(n^{-\delta})$ if $n^\delta W_n = o_P(1)$ which means that

$$n^\delta W_n \xrightarrow{P} 0.$$

c) W_n has the *same order as* X_n *in probability*, written $W_n \asymp_P X_n$, if for every $\epsilon > 0$ there exist positive constants N_ϵ and $0 < d_\epsilon < D_\epsilon$ such that

$$P\left(d_\epsilon \le \left|\frac{W_n}{X_n}\right| \le D_\epsilon\right) = P\left(\frac{1}{D_\epsilon} \le \left|\frac{X_n}{W_n}\right| \le \frac{1}{d_\epsilon}\right) \ge 1 - \epsilon$$

for all $n \ge N_\epsilon$.

d) Similar notation is used for a $k \times r$ matrix $\boldsymbol{A}_n = \boldsymbol{A} = [a_{i,j}(n)]$ if each element $a_{i,j}(n)$ has the desired property. For example, $\boldsymbol{A} = O_P(n^{-1/2})$ if each $a_{i,j}(n) = O_P(n^{-1/2})$.

Definition 4.10. Let $W_n = \|\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}\|$.

a) If $W_n \asymp_P n^{-\delta}$ for some $\delta > 0$, then both W_n and $\hat{\boldsymbol{\mu}}_n$ have (tightness) **rate** n^δ.

b) If there exists a constant κ such that

$$n^\delta(W_n - \kappa) \xrightarrow{D} X$$

for some nondegenerate random variable X, then both W_n and $\hat{\boldsymbol{\mu}}_n$ have *convergence rate* n^δ.

Theorem 4.10. Suppose there exists a constant κ such that

$$n^\delta(W_n - \kappa) \xrightarrow{D} X.$$

a) Then $W_n = O_P(n^{-\delta})$.

b) If X is not degenerate, then $W_n \asymp_P n^{-\delta}$.

The above result implies that if W_n has convergence rate n^δ, then W_n has tightness rate n^δ, and the term "tightness" will often be omitted. Part a) is proved, for example, in Lehmann (1999, p. 67).

The following result shows that if $W_n \asymp_P X_n$, then $X_n \asymp_P W_n$, $W_n = O_P(X_n)$, and $X_n = O_P(W_n)$. Notice that if $W_n = O_P(n^{-\delta})$, then n^δ is a lower bound on the rate of W_n. As an example, if the CLT holds then $\overline{Y}_n = O_P(n^{-1/3})$, but $\overline{Y}_n \asymp_P n^{-1/2}$.

Theorem 4.11. a) If $W_n \asymp_P X_n$, then $X_n \asymp_P W_n$.

b) If $W_n \asymp_P X_n$, then $W_n = O_P(X_n)$.

c) If $W_n \asymp_P X_n$, then $X_n = O_P(W_n)$.

d) $W_n \asymp_P X_n$ iff $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$.

Proof. a) Since $W_n \asymp_P X_n$,

$$P\left(d_\epsilon \le \left|\frac{W_n}{X_n}\right| \le D_\epsilon\right) = P\left(\frac{1}{D_\epsilon} \le \left|\frac{X_n}{W_n}\right| \le \frac{1}{d_\epsilon}\right) \ge 1-\epsilon$$

for all $n \ge N_\epsilon$. Hence $X_n \asymp_P W_n$.

b) Since $W_n \asymp_P X_n$,

$$P(|W_n| \le |X_n D_\epsilon|) \ge P\left(d_\epsilon \le \left|\frac{W_n}{X_n}\right| \le D_\epsilon\right) \ge 1-\epsilon$$

for all $n \ge N_\epsilon$. Hence $W_n = O_P(X_n)$.

c) Follows by a) and b).

d) If $W_n \asymp_P X_n$, then $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$ by b) and c). Now suppose $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$. Then

$$P(|W_n| \le |X_n| D_{\epsilon/2}) \ge 1-\epsilon/2$$

for all $n \ge N_1$, and

$$P(|X_n| \le |W_n| 1/d_{\epsilon/2}) \ge 1-\epsilon/2$$

for all $n \ge N_2$. Hence

$$P(A) \equiv P\left(\left|\frac{W_n}{X_n}\right| \le D_{\epsilon/2}\right) \ge 1-\epsilon/2$$

and

$$P(B) \equiv P\left(d_{\epsilon/2} \le \left|\frac{W_n}{X_n}\right|\right) \ge 1-\epsilon/2$$

for all $n \ge N = \max(N_1, N_2)$. Since $P(A \cap B) = P(A) + P(B) - P(A \cup B) \ge P(A) + P(B) - 1$,

$$P(A \cap B) = P(d_{\epsilon/2} \le \left|\frac{W_n}{X_n}\right| \le D_{\epsilon/2}) \ge 1-\epsilon/2 + 1-\epsilon/2 - 1 = 1-\epsilon$$

for all $n \ge N$. Hence $W_n \asymp_P X_n$. $\square$

The following result is used to prove the following Theorem 4.13 which says that if there are K estimators $T_{j,n}$ of a parameter $\boldsymbol{\beta}$, such that $\|T_{j,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ where $0 < \delta \le 1$, and if T_n^* picks one of these estimators, then $\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta})$.

Theorem 4.12: Pratt (1959). Let $X_{1,n}, ..., X_{K,n}$ each be $O_P(1)$ where K is fixed. Suppose $W_n = X_{i_n,n}$ for some $i_n \in \{1, ..., K\}$. Then

$$W_n = O_P(1). \tag{4.5}$$

Proof.

$$P(\max\{X_{1,n}, ..., X_{K,n}\} \leq x) = P(X_{1,n} \leq x, ..., X_{K,n} \leq x) \leq$$

$$F_{W_n}(x) \leq P(\min\{X_{1,n}, ..., X_{K,n}\} \leq x) = 1 - P(X_{1,n} > x, ..., X_{K,n} > x).$$

Since K is finite, there exists $B > 0$ and N such that $P(X_{i,n} \leq B) > 1-\epsilon/2K$ and $P(X_{i,n} > -B) > 1 - \epsilon/2K$ for all $n > N$ and $i = 1, ..., K$. Bonferroni's inequality states that $P(\cap_{i=1}^K A_i) \geq \sum_{i=1}^K P(A_i) \ - \ (K-1)$. Thus

$$F_{W_n}(B) \geq P(X_{1,n} \leq B, ..., X_{K,n} \leq B) \geq$$

$$K(1 - \epsilon/2K) - (K-1) = K - \epsilon/2 - K + 1 = 1 - \epsilon/2$$

and

$$-F_{W_n}(-B) \geq -1 + P(X_{1,n} > -B, ..., X_{K,n} > -B) \geq$$

$$-1 + K(1 - \epsilon/2K) - (K-1) = -1 + K - \epsilon/2 - K + 1 = -\epsilon/2.$$

Hence

$$F_{W_n}(B) - F_{W_n}(-B) \geq 1 - \epsilon \ \text{ for } \ \mathrm{n > N}. \ \square$$

Theorem 4.13. Suppose $\|T_{j,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ for $j = 1, ..., K$ where $0 < \delta \leq 1$. Let $T_n^* = T_{i_n,n}$ for some $i_n \in \{1, ..., K\}$ where, for example, $T_{i_n,n}$ is the $T_{j,n}$ that minimized some criterion function. Then

$$\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta}). \tag{4.6}$$

Proof. Let $X_{j,n} = n^\delta \|T_{j,n} - \boldsymbol{\beta}\|$. Then $X_{j,n} = O_P(1)$ so by Theorem 4.12, $n^\delta \|T_n^* - \boldsymbol{\beta}\| = O_P(1)$. Hence $\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta})$. $\square$

4.2.4 Slutsky's Theorem and Related Results

Theorem 4.14: Slutsky's Theorem. Suppose $Y_n \xrightarrow{D} Y$ and $W_n \xrightarrow{P} w$ for some constant w. Then

a) $Y_n + W_n \xrightarrow{D} Y + w$,

b) $Y_n W_n \xrightarrow{D} wY$, and

c) $Y_n / W_n \xrightarrow{D} Y/w$ if $w \neq 0$.

Theorem 4.15. a) If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{D} X$.

b) If $X_n \xrightarrow{ae} X$, then $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{D} X$.

c) If $X_n \xrightarrow{r} X$, then $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{D} X$.

d) $X_n \xrightarrow{P} \tau(\theta)$ iff $\mathrm{X_n} \xrightarrow{\mathrm{D}} \tau(\theta)$.

e) If $X_n \xrightarrow{P} \theta$ and τ is continuous at θ, then $\tau(X_n) \xrightarrow{P} \tau(\theta)$.

f) If $X_n \xrightarrow{D} \theta$ and τ is continuous at θ, then $\tau(X_n) \xrightarrow{D} \tau(\theta)$.

Suppose that for all $\theta \in \Theta$, $T_n \xrightarrow{D} \tau(\theta)$, $T_n \xrightarrow{r} \tau(\theta)$, or $T_n \xrightarrow{ae} \tau(\theta)$. Then T_n is a consistent estimator of $\tau(\theta)$ by Theorem 4.15. We are assuming that the function τ does not depend on n.

Example 4.9. Let $Y_1, ..., Y_n$ be iid with mean $E(Y_i) = \mu$ and variance $V(Y_i) = \sigma^2$. Then the sample mean $\overline{Y}_n$ is a consistent estimator of μ since i) the SLLN holds (use Theorems 4.9 and 4.15), ii) the WLLN holds, and iii) the CLT holds (use Theorem 4.8). Since

$$\lim_{n\to\infty} \text{VAR}_\mu(\overline{Y}_n) = \lim_{n\to\infty} \sigma^2/n = 0 \quad \text{and} \quad \lim_{n\to\infty} E_\mu(\overline{Y}_n) = \mu,$$

$\overline{Y}_n$ is also a consistent estimator of μ by Theorem 4.7b. By the delta method and Theorem 4.8b, $T_n = g(\overline{Y}_n)$ is a consistent estimator of $g(\mu)$ if $g'(\mu) \neq 0$ for all $\mu \in \Theta$. By Theorem 4.15e, $g(\overline{Y}_n)$ is a consistent estimator of $g(\mu)$ if g is continuous at μ for all $\mu \in \Theta$.

Theorem 4.16. Assume that the function g does not depend on n.

a) **Generalized Continuous Mapping Theorem:** If $X_n \xrightarrow{D} X$ and the function g is such that $P[X \in C(g)] = 1$ where $C(g)$ is the set of points where g is continuous, then $g(X_n) \xrightarrow{D} g(X)$.

b) **Continuous Mapping Theorem:** If $X_n \xrightarrow{D} X$ and the function g is continuous, then $g(X_n) \xrightarrow{D} g(X)$.

Remark 4.3. For Theorem 4.15, a) follows from Slutsky's Theorem by taking $Y_n \equiv X = Y$ and $W_n = X_n - X$. Then $Y_n \xrightarrow{D} Y = X$ and $W_n \xrightarrow{P} 0$. Hence $X_n = Y_n + W_n \xrightarrow{D} Y + 0 = X$. The convergence in distribution parts of b) and c) follow from a). Part f) follows from d) and e). Part e) implies that if T_n is a consistent estimator of θ and τ is a continuous function, then $\tau(T_n)$ is a consistent estimator of $\tau(\theta)$. Theorem 4.16 says that convergence in distribution is preserved by continuous functions, and even some discontinuities are allowed as long as the set of continuity points is assigned probability 1 by the asymptotic distribution. Equivalently, the set of discontinuity points is assigned probability 0.

Example 4.10. (Ferguson 1996, p. 40): If $X_n \xrightarrow{D} X$, then $1/X_n \xrightarrow{D} 1/X$ if X is a continuous random variable since $P(X = 0) = 0$ and $x = 0$ is the only discontinuity point of $g(x) = 1/x$.

Example 4.11. Show that if $Y_n \sim t_n$, a t distribution with n degrees of freedom, then $Y_n \xrightarrow{D} Z$ where $Z \sim N(0, 1)$.

Solution: $Y_n \stackrel{D}{=} Z/\sqrt{V_n/n}$ where $Z \perp\!\!\!\perp V_n \sim \chi^2_n$. If $W_n = \sqrt{V_n/n} \xrightarrow{P} 1$, then the result follows by Slutsky's Theorem. But $V_n \stackrel{D}{=} \sum_{i=1}^n X_i$ where the

iid $X_i \sim \chi^2_1$. Hence $V_n/n \xrightarrow{P} 1$ by the WLLN and $\sqrt{V_n/n} \xrightarrow{P} 1$ by Theorem 4.15e.

Theorem 4.17: Continuity Theorem. Let Y_n be sequence of random variables with characteristic functions $\phi_n(t)$. Let Y be a random variable with characteristic function (cf) $\phi(t)$.

a)

$$Y_n \xrightarrow{D} Y \text{ iff } \phi_{\mathrm{n}}(\mathrm{t}) \to \phi(\mathrm{t})\ \forall \mathrm{t} \in \mathbb{R}.$$

b) Also assume that Y_n has moment generating function (mgf) m_n and Y has mgf m. Assume that all of the mgfs m_n and m are defined on $|t| \leq d$ for some $d > 0$. Then if $m_n(t) \to m(t)$ as $n \to \infty$ for all $|t| < c$ where $0 < c < d$, then $Y_n \xrightarrow{D} Y$.

Application: Proof of a Special Case of the CLT. Following Rohatgi (1984, pp. 569-9), let $Y_1, ..., Y_n$ be iid with mean μ, variance σ^2, and mgf $m_Y(t)$ for $|t| < t_o$. Then

$$Z_i = \frac{Y_i - \mu}{\sigma}$$

has mean 0, variance 1, and mgf $m_Z(t) = \exp(-t\mu/\sigma)m_Y(t/\sigma)$ for $|t| < \sigma t_o$. We want to show that

$$W_n = \sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) \xrightarrow{D} N(0,1).$$

Notice that $W_n =$

$$n^{-1/2}\sum_{i=1}^n Z_i = n^{-1/2}\sum_{i=1}^n \left(\frac{Y_i - \mu}{\sigma}\right) = n^{-1/2}\frac{\sum_{i=1}^n Y_i - n\mu}{\sigma} = \frac{n^{-1/2}}{\frac{1}{n}}\ \frac{\overline{Y}_n - \mu}{\sigma}.$$

Thus

$$m_{W_n}(t) = E(e^{tW_n}) = E[\exp(tn^{-1/2}\sum_{i=1}^n Z_i)] = E[\exp(\sum_{i=1}^n tZ_i/\sqrt{n})]$$

$$= \prod_{i=1}^n E[e^{tZ_i/\sqrt{n}}] = \prod_{i=1}^n m_Z(t/\sqrt{n}) = [m_Z(t/\sqrt{n})]^n.$$

Set $\psi(x) = \log(m_Z(x))$. Then

$$\log[m_{W_n}(t)] = n\log[m_Z(t/\sqrt{n})] = n\psi(t/\sqrt{n}) = \frac{\psi(t/\sqrt{n})}{\frac{1}{n}}.$$

Now $\psi(0) = \log[m_Z(0)] = \log(1) = 0$. Thus by L'Hôpital's rule (where the derivative is with respect to n), $\lim_{n\to\infty}\log[m_{W_n}(t)] =$

$$\lim_{n\to\infty} \frac{\psi(t/\sqrt{n}\ )}{\frac{1}{n}} = \lim_{n\to\infty} \frac{\psi'(t/\sqrt{n}\ )[\frac{-t/2}{n^{3/2}}]}{(\frac{-1}{n^2})} = \frac{t}{2} \lim_{n\to\infty} \frac{\psi'(t/\sqrt{n}\ )}{\frac{1}{\sqrt{n}}}.$$

Now

$$\psi'(0) = \frac{m'_Z(0)}{m_Z(0)} = E(Z_i)/1 = 0,$$

so L'Hôpital's rule can be applied again, giving $\lim_{n\to\infty} \log[m_{W_n}(t)] =$

$$\frac{t}{2} \lim_{n\to\infty} \frac{\psi''(t/\sqrt{n}\ )[\frac{-t}{2n^{3/2}}]}{(\frac{-1}{2n^{3/2}})} = \frac{t^2}{2} \lim_{n\to\infty} \psi''(t/\sqrt{n}\ ) = \frac{t^2}{2}\psi''(0).$$

Now

$$\psi''(t) = \frac{d}{dt}\frac{m'_Z(t)}{m_Z(t)} = \frac{m''_Z(t)m_Z(t) - (m'_Z(t))^2}{[m_Z(t)]^2}.$$

So

$$\psi''(0) = m''_Z(0) - [m'_Z(0)]^2 = E(Z_i^2) - [E(Z_i)]^2 = 1.$$

Hence $\lim_{n\to\infty} \log[m_{W_n}(t)] = t^2/2$ and

$$\lim_{n\to\infty} m_{W_n}(t) = \exp(t^2/2)$$

which is the N(0,1) mgf. Thus by the continuity theorem,

$$W_n = \sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) \xrightarrow{D} N(0,1). \ \square$$

4.2.5 Multivariate Limit Theorems

Many of the univariate results of the previous 3 subsections can be extended to random vectors. For the limit theorems, the vector $\boldsymbol{X}$ is typically a $k \times 1$ column vector and $\boldsymbol{X}^T$ is a row vector. Let $\|\boldsymbol{x}\| = \sqrt{x_1^2 + \cdots + x_k^2}$ be the Euclidean norm of $\boldsymbol{x}$.

Definition 4.11. Let $\boldsymbol{X}_n$ be a sequence of random vectors with joint cdfs $F_n(\boldsymbol{x})$ and let $\boldsymbol{X}$ be a random vector with joint cdf $F(\boldsymbol{x})$.

a) $\boldsymbol{X}_n$ **converges in distribution** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$, if $F_n(\boldsymbol{x}) \to F(\boldsymbol{x})$ as $n \to \infty$ for all points $\boldsymbol{x}$ at which $F(\boldsymbol{x})$ is continuous. The distribution of $\boldsymbol{X}$ is the **limiting distribution** or **asymptotic distribution** of $\boldsymbol{X}_n$.

b) $\boldsymbol{X}_n$ **converges in probability** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, if for every $\epsilon > 0$, $P(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \epsilon) \to 0$ as $n \to \infty$.

c) Let $r > 0$ be a real number. Then $\boldsymbol{X}_n$ **converges in rth mean** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{r} \boldsymbol{X}$, if $E(\|\boldsymbol{X}_n - \boldsymbol{X}\|^r) \to 0$ as $n \to \infty$.

d) $\boldsymbol{X}_n$ **converges almost everywhere** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \stackrel{ae}{\to} \boldsymbol{X}$, if $P(\lim_{n\to\infty} \boldsymbol{X}_n = \boldsymbol{X}) = 1$.

Theorems 4.18 and 4.19 below are the multivariate extensions of the limit theorems in subsection 4.2.2. When the limiting distribution of $\boldsymbol{Z}_n = \sqrt{n}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta}))$ is multivariate normal $N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$, approximate the joint cdf of $\boldsymbol{Z}_n$ with the joint cdf of the $N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$ distribution. Thus to find probabilities, manipulate $\boldsymbol{Z}_n$ as if $\boldsymbol{Z}_n \approx N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. To see that the CLT is a special case of the MCLT below, let $k = 1$, $E(X) = \mu$, and $V(X) = \boldsymbol{\Sigma_x} = \sigma^2$.

Theorem 4.18: the Multivariate Central Limit Theorem (MCLT). If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $k \times 1$ random vectors with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma_x}$, then

$$\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \stackrel{D}{\to} N_k(\boldsymbol{0}, \boldsymbol{\Sigma_x})$$

where the sample mean

$$\overline{\boldsymbol{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i.$$

To see that the delta method is a special case of the multivariate delta method, note that if T_n and parameter θ are real valued, then $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} = g'(\theta)$.

Theorem 4.19: the Multivariate Delta Method. If $\boldsymbol{g}$ does not depend on n and

$$\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{\theta}) \stackrel{D}{\to} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

then

$$\sqrt{n}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta})) \stackrel{D}{\to} N_d(\boldsymbol{0}, \boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} \boldsymbol{\Sigma} \boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\theta})})$$

where the $d \times k$ Jacobian matrix of partial derivatives

$$\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} = \begin{bmatrix} \frac{\partial}{\partial\theta_1} g_1(\boldsymbol{\theta}) & \dots & \frac{\partial}{\partial\theta_k} g_1(\boldsymbol{\theta}) \\ \vdots & & \vdots \\ \frac{\partial}{\partial\theta_1} g_d(\boldsymbol{\theta}) & \dots & \frac{\partial}{\partial\theta_k} g_d(\boldsymbol{\theta}) \end{bmatrix}.$$

Here the mapping $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^d$ needs to be differentiable in a neighborhood of $\boldsymbol{\theta} \in \mathbb{R}^k$.

Definition 4.12. If the estimator $\boldsymbol{g}(\boldsymbol{T}_n) \stackrel{P}{\to} \boldsymbol{g}(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$, then $\boldsymbol{g}(\boldsymbol{T}_n)$ is a **consistent estimator** of $\boldsymbol{g}(\boldsymbol{\theta})$.

Theorem 4.20. If $0 < \delta \leq 1$, $\boldsymbol{X}$ is a random vector, and

$$n^\delta(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta})) \stackrel{D}{\to} \boldsymbol{X},$$

then $\boldsymbol{g}(\boldsymbol{T}_n) \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta})$.

Theorem 4.21. If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid, $E(\|\boldsymbol{X}\|) < \infty$, and $E(\boldsymbol{X}) = \boldsymbol{\mu}$, then
a) WLLN: $\overline{\boldsymbol{X}}_n \xrightarrow{P} \boldsymbol{\mu}$ and
b) SLLN: $\overline{\boldsymbol{X}}_n \xrightarrow{ae} \boldsymbol{\mu}$.

Theorem 4.22: Continuity Theorem. Let $\boldsymbol{X}_n$ be a sequence of $k \times 1$ random vectors with characteristic functions $\phi_n(\boldsymbol{t})$, and let $\boldsymbol{X}$ be a $k \times 1$ random vector with cf $\phi(\boldsymbol{t})$. Then

$$\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X} \text{ iff } \phi_{\mathrm{n}}(\boldsymbol{t}) \to \phi(\boldsymbol{t})$$

for all $\boldsymbol{t} \in \mathbb{R}^k$.

Theorem 4.23: Cramér Wold Device. Let $\boldsymbol{X}_n$ be a sequence of $k \times 1$ random vectors, and let $\boldsymbol{X}$ be a $k \times 1$ random vector. Then

$$\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X} \text{ iff } \boldsymbol{t}^{\mathrm{T}} \boldsymbol{X}_{\mathrm{n}} \xrightarrow{\mathrm{D}} \boldsymbol{t}^{\mathrm{T}} \boldsymbol{X}$$

for all $\boldsymbol{t} \in \mathbb{R}^k$.

Application: Proof of the MCLT Theorem 4.18. Note that for fixed $\boldsymbol{t}$, the $\boldsymbol{t}^T \boldsymbol{X}_i$ are iid random variables with mean $\boldsymbol{t}^T \boldsymbol{\mu}$ and variance $\boldsymbol{t}^T \boldsymbol{\Sigma} \boldsymbol{t}$. Hence by the CLT, $\boldsymbol{t}^T \sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N(0, \boldsymbol{t}^T \boldsymbol{\Sigma} \boldsymbol{t})$. The right hand side has distribution $\boldsymbol{t}^T \boldsymbol{X}$ where $\boldsymbol{X} \sim N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. Hence by the Cramér Wold Device, $\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. $\square$

Theorem 4.24. a) If $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, then $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$.
b)

$$\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta}) \text{ iff } \boldsymbol{X}_{\mathrm{n}} \xrightarrow{\mathrm{D}} \boldsymbol{g}(\boldsymbol{\theta}).$$

Let $g(n) \geq 1$ be an increasing function of the sample size n: $g(n) \uparrow \infty$, e.g. $g(n) = \sqrt{n}$. See White (1984, p. 15). If a $k \times 1$ random vector $\boldsymbol{T}_n - \boldsymbol{\mu}$ converges to a nondegenerate multivariate normal distribution with convergence rate $\sqrt{n}$, then $\boldsymbol{T}_n$ has (tightness) rate $\sqrt{n}$.

Definition 4.13. Let $\boldsymbol{A}_n = [a_{i,j}(n)]$ be an $r \times c$ random matrix.
a) $\boldsymbol{A}_n = O_P(X_n)$ if $a_{i,j}(n) = O_P(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
b) $\boldsymbol{A}_n = o_p(X_n)$ if $a_{i,j}(n) = o_p(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
c) $\boldsymbol{A}_n \asymp_P (1/(g(n))$ if $a_{i,j}(n) \asymp_P (1/(g(n))$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
d) Let $\boldsymbol{A}_{1,n} = \boldsymbol{T}_n - \boldsymbol{\mu}$ and $\boldsymbol{A}_{2,n} = \boldsymbol{C}_n - c\boldsymbol{\Sigma}$ for some constant $c > 0$. If $\boldsymbol{A}_{1,n} \asymp_P (1/(g(n))$ and $\boldsymbol{A}_{2,n} \asymp_P (1/(g(n))$, then $(\boldsymbol{T}_n, \boldsymbol{C}_n)$ has (tightness) rate $g(n)$.

Theorem 4.25: Continuous Mapping Theorem. Let $\boldsymbol{X}_n \in \mathbb{R}^k$. If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and if the function $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^j$ is continuous, then $\boldsymbol{g}(\boldsymbol{X}_n) \xrightarrow{D} \boldsymbol{g}(\boldsymbol{X})$.

The following two theorems are taken from Severini (2005, pp. 345-349, 354).

Theorem 4.26. Let $\boldsymbol{X}_n = (X_{1n}, ..., X_{kn})^T$ be a sequence of $k \times 1$ random vectors, let $\boldsymbol{Y}_n$ be a sequence of $k \times 1$ random vectors, and let $\boldsymbol{X} = (X_1, ..., X_k)^T$ be a $k \times 1$ random vector. Let $\boldsymbol{W}_n$ be a sequence of $k \times k$ nonsingular random matrices, and let $\boldsymbol{C}$ be a $k \times k$ constant nonsingular matrix.

a) $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$ iff $X_{in} \xrightarrow{P} X_i$ for $i = 1, ..., k$.

b) **Slutsky's Theorem**: If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and $\boldsymbol{Y}_n \xrightarrow{P} \boldsymbol{c}$ for some constant $k \times 1$ vector $\boldsymbol{c}$, then i) $\boldsymbol{X}_n + \boldsymbol{Y}_n \xrightarrow{D} \boldsymbol{X} + \boldsymbol{c}$ and

ii) $\boldsymbol{Y}_n^T \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{c}^T \boldsymbol{X}$.

c) If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ and $\boldsymbol{W}_n \xrightarrow{P} \boldsymbol{C}$, then $\boldsymbol{W}_n \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{C}\boldsymbol{X}$, $\boldsymbol{X}_n^T \boldsymbol{W}_n \xrightarrow{D} \boldsymbol{X}^T \boldsymbol{C}$, $\boldsymbol{W}_n^{-1} \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{C}^{-1} \boldsymbol{X}$, and $\boldsymbol{X}_n^T \boldsymbol{W}_n^{-1} \xrightarrow{D} \boldsymbol{X}^T \boldsymbol{C}^{-1}$.

Theorem 4.27. Let W_n, X_n, Y_n, and Z_n be sequences of random variables such that $Y_n > 0$ and $Z_n > 0$. (Often Y_n and Z_n are deterministic, e.g. $Y_n = n^{-1/2}$.)

a) If $W_n = O_P(1)$ and $X_n = O_P(1)$, then $W_n + X_n = O_P(1)$ and $W_n X_n = O_P(1)$, thus $O_P(1) + O_P(1) = O_P(1)$ and $O_P(1) O_P(1) = O_P(1)$.

b) If $W_n = O_P(1)$ and $X_n = o_P(1)$, then $W_n + X_n = O_P(1)$ and $W_n X_n = o_P(1)$, thus $O_P(1) + o_P(1) = O_P(1)$ and $O_P(1) o_P(1) = o_P(1)$.

c) If $W_n = O_P(Y_n)$ and $X_n = O_P(Z_n)$, then $W_n + X_n = O_P(\max(Y_n, Z_n))$ and $W_n X_n = O_P(Y_n Z_n)$, thus $O_P(Y_n) + O_P(Z_n) = O_P(\max(Y_n, Z_n))$ and $O_P(Y_n) O_P(Z_n) = O_P(Y_n Z_n)$.

Theorem 4.28. i) Suppose $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}$ be a $q \times p$ constant matrix. Then $\boldsymbol{A}\sqrt{n}(T_n - \boldsymbol{\mu}) = \sqrt{n}(\boldsymbol{A}T_n - \boldsymbol{A}\boldsymbol{\mu}) \xrightarrow{D} N_q(\boldsymbol{A}\boldsymbol{\theta}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

ii) Let $\boldsymbol{\Sigma} > 0$. If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ where $s > 0$ is some constant, then $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + o_P(1)$, so $D^2_{\boldsymbol{x}}(T, \boldsymbol{C})$ is a consistent estimator of $s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

iii) Let $\boldsymbol{\Sigma} > 0$. If $\sqrt{n}(T - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ and if $\boldsymbol{C}$ is a consistent estimator of $\boldsymbol{\Sigma}$, then $n(T - \boldsymbol{\mu})^T \boldsymbol{C}^{-1} (T - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$. In particular, $n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1} (\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$.

Proof: ii) $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) =$
$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T [\boldsymbol{C}^{-1} - s^{-1} \boldsymbol{\Sigma}^{-1} + s^{-1} \boldsymbol{\Sigma}^{-1}] (\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)$
$= (\boldsymbol{x} - \boldsymbol{\mu})^T [s^{-1} \boldsymbol{\Sigma}^{-1}] (\boldsymbol{x} - \boldsymbol{\mu}) + (\boldsymbol{x} - T)^T [\boldsymbol{C}^{-1} - s^{-1} \boldsymbol{\Sigma}^{-1}] (\boldsymbol{x} - T)$
$+ (\boldsymbol{x} - \boldsymbol{\mu})^T [s^{-1} \boldsymbol{\Sigma}^{-1}] (\boldsymbol{\mu} - T) + (\boldsymbol{\mu} - T)^T [s^{-1} \boldsymbol{\Sigma}^{-1}] (\boldsymbol{x} - \boldsymbol{\mu})$
$+ (\boldsymbol{\mu} - T)^T [s^{-1} \boldsymbol{\Sigma}^{-1}] (\boldsymbol{\mu} - T) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(1)$.

(Note that $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(n^{-\delta})$ if $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ with rate n^{δ} where $0 < \delta \leq 0.5$ if $[\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1}] = O_P(n^{-\delta})$.)

Alternatively, $D^2_{\boldsymbol{x}}(T, \boldsymbol{C})$ is a continuous function of $(T, \boldsymbol{C})$ if $\boldsymbol{C} > 0$ for $n > 10p$. Hence $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) \xrightarrow{P} D^2_{\boldsymbol{x}}(\mu, s\boldsymbol{\Sigma})$.

iii) Note that $\boldsymbol{Z}_n = \sqrt{n}\ \boldsymbol{\Sigma}^{-1/2}(T - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{I}_p)$. Thus $\boldsymbol{Z}_n^T \boldsymbol{Z}_n = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(T - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$. Now $n(T - \boldsymbol{\mu})^T \boldsymbol{C}^{-1}(T - \boldsymbol{\mu}) =$
$n(T - \boldsymbol{\mu})^T[\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}](T - \boldsymbol{\mu}) = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(T - \boldsymbol{\mu}) +$
$n(T - \boldsymbol{\mu})^T[\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1}](T - \boldsymbol{\mu}) = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(T - \boldsymbol{\mu}) + o_P(1) \xrightarrow{D} \chi^2_p$ since $\sqrt{n}(T - \boldsymbol{\mu})^T[\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1}]\sqrt{n}(T - \boldsymbol{\mu}) = O_P(1)o_P(1)O_P(1) = o_P(1)$. $\square$

Example 4.12. Suppose that $\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{y}_n$ for $n = 1, 2, ...$ Suppose $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x}$, and $\boldsymbol{y}_n \xrightarrow{D} \boldsymbol{y}$ where $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$. Then

$$\begin{bmatrix} \boldsymbol{x}_n \\ \boldsymbol{y}_n \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}$$

by Theorem 4.22. To see this, let $\boldsymbol{t} = (\boldsymbol{t}_1^T, \boldsymbol{t}_2^T)^T$, $\boldsymbol{z}_n = (\boldsymbol{x}_n^T, \boldsymbol{y}_n^T)^T$, and $\boldsymbol{z} = (\boldsymbol{x}^T, \boldsymbol{y}^T)^T$. Since $\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{y}_n$ and $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$, the characteristic function

$$\phi_{\boldsymbol{z}_n}(\boldsymbol{t}) = \phi_{\boldsymbol{x}_n}(\boldsymbol{t}_1)\phi_{\boldsymbol{y}_n}(\boldsymbol{t}_2) \to \phi_{\boldsymbol{x}}(\boldsymbol{t}_1)\phi_{\boldsymbol{y}}(\boldsymbol{t}_2) = \phi_{\boldsymbol{z}}(\boldsymbol{t}).$$

Hence $\boldsymbol{g}(\boldsymbol{z}_n) \xrightarrow{D} \boldsymbol{g}(\boldsymbol{z})$ by Theorem 4.25.

4.3 Mixture Distributions

Mixture distributions are useful for model and variable selection since $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a mixture distribution of $\hat{\boldsymbol{\beta}}_{I_j,0}$, and the lasso estimator $\hat{\boldsymbol{\beta}}_L$ is a mixture distribution of $\hat{\boldsymbol{\beta}}_{L,\lambda_i}$ for $i = 1, ..., M$. See the second to last paragraph of Section 4.1 for $\hat{\boldsymbol{\beta}}_{I_{min},0}$. A random vector $\boldsymbol{u}$ has a mixture distribution if $\boldsymbol{u}$ equals a random vector $\boldsymbol{u}_j$ with probability π_j for $j = 1, ..., J$. See Definition 4.2 for the population mean and population covariance matrix of a random vector.

Definition 4.14. The distribution of a $g \times 1$ random vector $\boldsymbol{u}$ is a mixture distribution if the cumulative distribution function (cdf) of $\boldsymbol{u}$ is

$$F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t}) \tag{4.7}$$

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

Theorem 4.29. Suppose $E(h(\boldsymbol{u}))$ and the $E(h(\boldsymbol{u}_j))$ exist. Then

$$E(h(\boldsymbol{u})) = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{u}_j)]. \tag{4.8}$$

Hence

$$E(\boldsymbol{u}) = \sum_{j=1}^{J} \pi_j E[\boldsymbol{u}_j], \tag{4.9}$$

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

$$\sum_{j=1}^{J} \pi_j Cov(\boldsymbol{u}_j) + \sum_{j=1}^{J} \pi_j E(\boldsymbol{u}_j)[E(\boldsymbol{u}_j)]^T - E(\boldsymbol{u})[E(\boldsymbol{u})]^T. \tag{4.10}$$

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

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

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

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

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

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

$$E(\boldsymbol{u})[E(\boldsymbol{u})]^T = \sum_{j=1}^{J} \sum_{k=1}^{J} \pi_j \pi_k E(\boldsymbol{u}_j)[E(\boldsymbol{u}_k)]^T. \tag{4.11}$$

4.4 Large Sample Theory for Some Variable Selection Estimators

Large sample theory is often tractable if the optimization problem is convex. The optimization problem for variable selection is not convex, so new tools are needed. Tibshirani et al. (2018) and Leeb and Pötscher (2006, 2008) note that we can not find the limiting distribution of $\boldsymbol{Z}_n = \sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{I_{min}} - \boldsymbol{\beta}_I)$ after variable selection. One reason is that with positive probability, $\hat{\boldsymbol{\beta}}_{I_{min}}$ does not have the same dimension as $\boldsymbol{\beta}_I$ if AIC is used. Hence $\boldsymbol{Z}_n$ is not defined with positive probability. Also, the dimension of a random vector is $k \times 1$, say, while the dimension of $\hat{\boldsymbol{\beta}}_{I_{min}}$ is $K \times 1$ where K is a random variable. Hence the random quantity $\hat{\boldsymbol{\beta}}_{I_{min}}$ is not a random vector and not a statistic.

We will show that large sample theory becomes simple by using zero padding. If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$, then $\hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. Assume p is fixed, and $n \to \infty$.

The Rathnayake and Olive (2019) theory in this section applies to many regression models including many generalized linear models, some time series models, some survival regression models such as the Cox (1972) proportional hazards survival regression model and AFTs, and the multiple linear regression model where the error distribution is unknown.

Suppose the regression model satisfies $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T\boldsymbol{\beta}$, that model (2.4) holds, and that if $S \subseteq I_j$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{V}_j)$. Also assume that a variable selection criterion, such as AIC or relaxed lasso, is used such that $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Hence

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}_{j,0}) \tag{4.12}$$

where $\boldsymbol{V}_{j,0}$ adds columns and rows of zeros corresponding to the x_i not in I_j. Hence $\boldsymbol{V}_{j,0}$ is singular unless I_j corresponds to the full model. Since fewer than 2^p regression models I contain the true model S, and each such model gives a $\sqrt{n}$ consistent estimator $\hat{\boldsymbol{\beta}}_{I,0}$ of $\boldsymbol{\beta}$, the probability that I_{min} picks one of these models goes to one as $n \to \infty$. Then $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ under model (2.4) if the variable selection criterion is used with forward selection, backward elimination, or all subsets. This result holds since picking from a fixed number of $\sqrt{n}$ consistent estimators results in a $\sqrt{n}$ consistent estimator by Pratt (1959). See Theorem 4.12 and Theorem 4.13. This section will use mixture distributions to find the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_{min},0} - \boldsymbol{\beta})$.

Under regularity conditions, $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$ if BIC or AIC is used with forward selection, backward elimination, or all subsets. See Charkhi and Claeskens (2018), Claeskens and Hjort (2008, pp. 70, 85-86, 101, 102, 114, 232), and Hjort and Claeskens (2006).

Mixture distributions are useful for variable selection since $\hat{\boldsymbol{\beta}}_{I_{min},0}$ has a mixture distribution of the $\hat{\boldsymbol{\beta}}_{I_j,0}$. Review mixture distributions from Section 4.3. The following theorem is due to Pelawa Watagoda and Olive (2019a). Note that the cdf of T_n is $F_{T_n}(\boldsymbol{z}) = \sum_j \pi_{jn} F_{T_{jn}}(\boldsymbol{z})$ where $F_{T_{jn}}(\boldsymbol{z})$ is the cdf of T_{jn}.

Theorem 4.30, Mixture Distribution CLT. Suppose the $g \times 1$ statistic T_n is equal to the estimator T_{jn} with probability π_{jn} for $j = 1, ..., J$ where $\sum_j \pi_{jn} = 1$, $\pi_{jn} \to \pi_j$ as $n \to \infty$, and $\boldsymbol{u}_{jn} = \sqrt{n}(T_{jn} - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}_j$ with $E(\boldsymbol{u}_j) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{u}_j) = \boldsymbol{\Sigma}_j$. Then

$$\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u} \tag{4.13}$$

where the cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{z}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{z})$ and $F_{\boldsymbol{u}_j}(\boldsymbol{z})$ is the cdf of $\boldsymbol{u}_j$. Thus, $\boldsymbol{u}$ is a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j, $E(\boldsymbol{u}) = \boldsymbol{0}$, and $\text{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \sum_j \pi_j \boldsymbol{\Sigma}_j$.

Proof: Note that T_n has a mixture distribution of the T_{jn} with probabilities π_{jn}. Hence $\sqrt{n}(T_n - \boldsymbol{\theta})$ has a mixture distribution of the $\boldsymbol{u}_{jn} = \sqrt{n}(T_{jn} - \boldsymbol{\theta})$, and the cdf of $\sqrt{n}(T_n - \boldsymbol{\theta})$ is $\sum_j \pi_{jn} F_{\boldsymbol{u}_{jn}}(\boldsymbol{z}) \to \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{z})$ at continuity points $\boldsymbol{z}$ of the $F_{\boldsymbol{u}_j}$. $\square$

Applying the above results makes large sample theory for $\hat{\boldsymbol{\beta}}_{I_{min},0}$ simple. The following theorem is due to Rathnayake and Olive (2019), generalizing the Pelawa Watagoda and Olive (2019a) result for multiple linear regression.

Theorem 4.31, Variable Selection CLT. Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{I_{min},0} = \hat{\boldsymbol{\beta}}_{I_k,0}$ 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

$$\boldsymbol{u}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_{min},0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \tag{4.14}$$

where the cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{z}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{z})$. Thus $\boldsymbol{u}$ is a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j, $E(\boldsymbol{u}) = \boldsymbol{0}$, and $\text{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \sum_j \pi_j \boldsymbol{V}_{j,0}$.

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

$$\boldsymbol{v}_n = \boldsymbol{A}\boldsymbol{u}_n = \sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} = \boldsymbol{v} \tag{4.15}$$

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.

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 $F_{\boldsymbol{u}_n}(\boldsymbol{z}) = \sum_k \pi_{kn} F_{\boldsymbol{u}_{kn}}(\boldsymbol{z}) \to F_{\boldsymbol{u}}(\boldsymbol{z}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{z})$ at

continuity points of the $F_{\boldsymbol{u}_j}(\boldsymbol{z})$ as $n \to \infty$.
b) Since $\boldsymbol{u}_n \xrightarrow{D} \boldsymbol{u}$, then $\boldsymbol{A}\boldsymbol{u}_n \xrightarrow{D} \boldsymbol{A}\boldsymbol{u}$. $\square$

Remark 4.4. If $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$ (e.g. for AIC, BIC, or relaxed lasso), the values of π_j depend on the regression variable selection method such as backward elimination, forward selection, all subsets, and lasso. Typically the mixture distribution is not asymptotically normal. There are two exceptions. First, suppose $\pi_d = 1$ with $\boldsymbol{u} \sim \boldsymbol{u}_d \sim N_p(\boldsymbol{0}, \boldsymbol{V}_{d,0})$. This exception occurs if $a_S = p$ so S is the full model, and for methods like BIC that choose I_S with probability going to one under strong regularity conditions.

The second exception occurs for each $\pi_j > 0$, $\boldsymbol{A}\boldsymbol{u}_j \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}_j\boldsymbol{A}^T) = N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. Then $\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. This exception occurs for $\hat{\boldsymbol{\beta}}_S$ 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 $\hat{\boldsymbol{\beta}}_S$ has the same multivariate normal limiting distribution for I_{min} and for the full model.

Remark 4.5. This theory has several applications. First, the theory gives the asymptotic distribution for many variable selection estimators, which are some of the most used estimators in Statistics. Second, the theory is useful for explaining why $\hat{\boldsymbol{\beta}}_{I_{min}}$ is not a good estimator, but $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a good estimator. Suppose $I_{min} = I_j$ is observed. Due to selection bias, the model using predictors I_j underestimates the variability of the responses $Y_1, ..., Y_n$, and $Cov(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j})$ is not the correct covariance matrix for $\boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min}}$. Typically $\hat{\boldsymbol{\beta}}_{I_{min}}$ is not a consistent estimator for any parameter vector $\boldsymbol{\beta}_{I_j}$, since in general $P(I_{min} = I_j)$ does not go to one as $n \to \infty$, and the dimension of I_{min} is a random variable. Selection bias occurs from acting as if $\hat{\boldsymbol{\beta}}_{I_{min}}$ is the "full model" (using large sample theory as if the "full model" was selected before gathering the data), when $\hat{\boldsymbol{\beta}}_{I_{min},0}$ has large sample theory given by Theorem 4.31.

A third application will be bootstrap inference for hypothesis testing. See Section 4.8. Fourth, the theory can be used to justify prediction intervals after variable selection. See Section 4.5, and Olive et al. (2020). Fifth, recall p is fixed. Suppose a shrinkage method, such as lasso or elastic net, does variable selection. Let $\hat{\boldsymbol{\beta}}_{I_{min}}$ be the regression estimator, such as a Cox regression, applied to a constant and the variables with nonzero shrinkage estimator coefficients. If the shrinkage estimator is consistent, then $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and thus the relaxed shrinkage estimator $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is $\sqrt{n}$ consistent. In particular, relaxed lasso and relaxed elastic net are $\sqrt{n}$ consistent if lasso and elastic net are consistent.

Remark 4.6. If $\pi_d = 1$ corresponds to $\boldsymbol{\beta}_d$, then $\hat{\boldsymbol{\beta}}_{I_{min}}$ can give useful information about $\boldsymbol{\beta}_d$, but information is lost about the parameters estimated to be zero if S is not the full model. There is a large literature on *variable selection consistency* and the *oracle property* where $P(I_{min} = S) \to 1$ as $n \to$

∞. See Claeskens and Hjort (2008, pp. 99-114) for references. A necessary condition for $P(I_{min} = S) \to 1$ is that S is one of the models considered with probability going to one. This condition holds for all subsets regression, but only under very strong regularity conditions for fast methods such as forward selection, backward elimination, and lasso.

4.5 Prediction Intervals

Prediction intervals for regression and prediction regions for multivariate data are important topics. Inference after variable selection will consider bootstrap hypothesis testing. Applying certain prediction intervals or prediction regions to the bootstrap sample will result in confidence intervals or confidence regions. The prediction intervals and regions are based on samples of size n, while the bootstrap sample size is $B = B_n$. Hence this section and the following section are important.

Definition 4.15. Consider predicting a future test value Y_f given a $p \times 1$ vector of predictors $\boldsymbol{x}_f$ and training data $(Y_1, \boldsymbol{x}_1), ..., (Y_n, \boldsymbol{x}_n)$. A large sample $100(1-\delta)\%$ *prediction interval* (PI) for Y_f has the form $[\hat{L}_n, \hat{U}_n]$ where $P(\hat{L}_n \leq Y_f \leq \hat{U}_n)$ is eventually bounded below by $1-\delta$ as the sample size $n \to \infty$. A large sample $100(1-\delta)\%$ PI is *asymptotically optimal* if $[\hat{L}_n, \hat{U}_n] \to [L_s, U_s]$ as $n \to \infty$ where $[L_s, U_s]$ is the *population shorth*: the shortest interval covering at least $100(1-\delta)\%$ of the mass.

If $Y_f|\boldsymbol{x}_f$ has a pdf, we often want $P(\hat{L}_n \leq Y_f \leq \hat{U}_n) \to 1-\delta$ as $n \to \infty$. The interpretation of a 100 $(1-\delta)\%$ PI for a random variable Y_f is similar to that of a confidence interval (CI). Collect data, then form the PI, and repeat for a total of k times where the k trials are independent from the same population. If Y_{fi} is the ith random variable and PI_i is the ith PI, then the probability that $Y_{fi} \in PI_i$ for j of the PIs approximately follows a binomial$(k, \rho = 1-\delta)$ distribution. Hence if 100 95% PIs are made, $\rho = 0.95$ and $Y_{fi} \in PI_i$ happens about 95 times.

There are two big differences between CIs and PIs. First, the length of the CI goes to 0 as the sample size n goes to ∞ while the length of the PI converges to some nonzero number J, say. Secondly, many confidence intervals work well for large classes of distributions while many prediction intervals assume that the distribution of the data is known up to some unknown parameters. Usually the $N(\mu, \sigma^2)$ distribution is assumed, and the parametric PI may not perform well if the normality assumption is violated. This section will describe PIs for parametric 1D regression models, which include many parametric survival regression models.

First we will consider the location model, $Y_i = \mu + e_i$, where $Y_1, ..., Y_n, Y_f$ are iid and there are no vectors of predictors $\boldsymbol{x}_i$ and $\boldsymbol{x}_f$. Let $Z_{(1)} \leq Z_{(2)} \leq$

$\cdots \leq Z_{(n)}$ be the order statistics of n iid random variables $Z_1, ..., Z_n$. Let a future random variable Z_f be such that $Z_1, ..., Z_n, Z_f$ are iid. Let $k_1 = \lceil n\delta/2 \rceil$ and $k_2 = \lceil n(1-\delta/2) \rceil$ where $\lceil x \rceil$ is the smallest integer $\geq x$. For example, $\lceil 7.7 \rceil = 8$. Then a common nonparametric large sample $100(1-\delta)\%$ prediction interval for Z_f is

$$[Z_{(k_1)}, Z_{(k_2)}] \tag{4.16}$$

where $0 < \delta < 1$. See Frey (2013) for references.

The shorth(c) estimator of the population shorth is useful for making asymptotically optimal prediction intervals. With the Z_i and $Z_{(i)}$ as in the above paragraph, let the shortest closed interval containing at least c of the Z_i be

$$\mathrm{shorth(c)} = [\mathrm{Z}_{(\mathrm{s})}, \mathrm{Z}_{(\mathrm{s+c-1})}]. \tag{4.17}$$

Let

$$k_n = \lceil n(1-\delta) \rceil. \tag{4.18}$$

Frey (2013) showed that for large $n\delta$ and iid data, the shorth(k_n) prediction interval has maximum undercoverage $\approx 1.12\sqrt{\delta/n}$, and used the shorth($c$) estimator as the large sample $100(1-\delta)\%$ PI where

$$c = \min(n, \lceil n[1 - \delta + 1.12\sqrt{\delta/n}\ ]\ \rceil). \tag{4.19}$$

An interesting fact is that the maximum undercoverage occurs for the family of uniform $U(\theta_1, \theta_2)$ distributions where such a distribution has pdf $f(y) = 1/(\theta_2 - \theta_1)$ for $\theta_1 \leq y \leq \theta_2$ where $f(y) = 0$, otherwise, and $\theta_1 < \theta_2$.

A problem with the prediction intervals that cover $\approx 100(1-\delta)\%$ of the training data cases Y_i (such as (4.8) using $c = k_n$ given by (4.9)), is that they have coverage lower than the nominal coverage of $1-\delta$ for moderate n. This result is not surprising since empirically statistical methods perform worse on test data. For iid data, Frey (2013) used (4.10) to correct for undercoverage.

Example 4.13. Given below were votes for preseason 1A basketball poll from Nov. 22, 2011 WSIL News where the 778 was a typo: the actual value was 78. As shown below, finding shorth(3) from the ordered data is simple. If the outlier was corrected, shorth(3) = [76,78].

```
111   89   778   78   76

order data: 76 78 89 111 778

                    13 = 89 - 76

                        33 = 111 - 78

                           689 = 778 - 89
shorth(3) = [76,89]
```

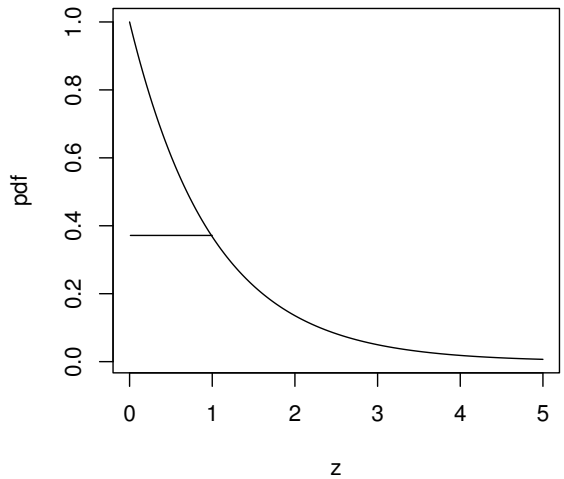

Fig. 4.1 The 36.8% Highest Density Region is [0,1]

For a random variable Y, the $100(1-\delta)\%$ highest density region is a union of $k \geq 1$ disjoint intervals such that the mass within the intervals $\geq 1-\delta$ and the sum of the k interval lengths is as small as possible. Suppose that $f(z)$ is a unimodal pdf that has interval support, and that the pdf $f(z)$ of Y decreases rapidly as z moves away from the mode. Let $[a, b]$ be the shortest interval such that $F_Y(b) - F_Y(a) = 1 - \delta$ where the cdf $F_Y(z) = P(Y \leq z)$. Then the interval $[a, b]$ is the $100(1-\delta)$ highest density region. To find the $100(1-\delta)\%$ highest density region of a pdf, move a horizontal line down from the top of the pdf. The line will intersect the pdf or the boundaries of the support of the pdf at $[a_1, b_1], ..., [a_k, b_k]$ for some $k \geq 1$. Stop moving the line when the areas under the pdf corresponding to the intervals is equal to $1-\delta$. As an example, let $f(z) = e^{-z}$ for $z > 0$. See Figure 4.1 where the area under the pdf from 0 to 1 is 0.368. Hence [0,1] is the 36.8% highest density region. The shorth PI estimates the highest density interval which is the highest density region for a distribution with a unimodal pdf. Often the highest density region is an interval $[a, b]$ where $f(a) = f(b)$, especially if the support where $f(z) > 0$ is $(-\infty, \infty)$.

A parametric 1D regression model is $Y|\boldsymbol{x} \sim D(h(\boldsymbol{x}), \boldsymbol{\gamma})$ for some real valued function, such as $h(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$, where D is a parametric distribution that depends on the $p \times 1$ vector of predictors $\boldsymbol{x}$ only through $h(\boldsymbol{x})$, and $\boldsymbol{\gamma}$ is a $q \times 1$ vector of parameters.

The first new large sample $100(1-\delta)\%$ prediction interval for Y_f applies the shorth(c) prediction interval to the parametric bootstrap sample $Y_1^*, ..., Y_B^*$ where the Y_i^* are iid from the distribution $D(\hat{h}(\boldsymbol{x}_f), \hat{\boldsymbol{\gamma}})$ with

$$c = \min(B, \lceil B[1 - \delta + 1.12\sqrt{\delta/B}\ ]\ \rceil). \tag{4.20}$$

If $Y|\boldsymbol{x}_f \sim D(h(\boldsymbol{x}_f), \boldsymbol{\gamma})$ and the regression method produces a consistent estimator $(\hat{h}(\boldsymbol{x}_f), \hat{\boldsymbol{\gamma}})$ of $(h(\boldsymbol{x}_f), \boldsymbol{\gamma})$, then this new prediction interval is a large sample $100(1-\delta)\%$ PI.

For models with a linear predictor, we will want prediction intervals after variable selection or model selection. The prediction interval (4.20) can have undercoverage if n is small compared to the number of estimated parameters. The modified shorth PI (4.21) inflates PI (4.20) to compensate for parameter estimation and model selection. Let d be the number of variables $x_1^*, ..., x_d^*$ used by the full model, forward selection, lasso, or relaxed lasso. We want $n \geq 10d$, and the prediction interval length will be increased (penalized) if n/d is not large. For the second new prediction interval, let $q_n = \min(1-\delta + 0.05, 1 - \delta + d/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta d/n), \quad \text{otherwise.}$$

If $1 - \delta < 0.999$ and $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Then compute the shorth PI with

$$c_{mod} = \min(B, \lceil B[q_n + 1.12\sqrt{\delta/B}\ ]\ \rceil). \tag{4.21}$$

Olive (2007, 2018) and Pelawa Watagoda and Olive (2019b) used similar correction factors for regression models with an additive error since the maximum simulated undercoverage was about 0.05 when $n = 20d$. If a $q \times 1$ vector of parameters $\boldsymbol{\gamma}$ is also estimated, we may need to replace d by $d_q = d + q$.

Hong et al. (2018) explain why classical PIs after AIC variable selection may not work. Fix p and let I_{min} correspond to the predictors used after variable selection. To show that (4.20) and (4.21) are large sample prediction intervals, we need to show that $(\hat{\boldsymbol{\beta}}_{I_{min},0}, \hat{\gamma}_{I_{min}})$ is a consistent estimator of $(\boldsymbol{\beta}, \gamma)$. Theorem 4.31 shows that $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a consistent estimator of $\boldsymbol{\beta}$. Suppose $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Suppose model (2.4) holds with $S \subseteq I_j$. Then under regularity conditions that are often mild, $(\hat{\boldsymbol{\beta}}_{I_j}, \hat{\gamma}_{I_j})$ is a consistent estimator of $(\boldsymbol{\beta}_{I_j}, \gamma)$. Then $\hat{\gamma}_{I_{min}}$ is a consistent estimator of γ. Hence if $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$ (AIC, BIC, or relaxed lasso), then (4.20) and (4.21) are large sample PIs.

As an example, consider the *Weibull proportional hazards regression model*

$$Y|SP \sim W(\gamma = 1/\sigma, \lambda_0 \exp(SP)),$$

where $\lambda_0 = \exp(-\alpha/\sigma)$, and Y has a Weibull $W(\gamma, \lambda)$ distribution if the pdf of Y is

$$f(y) = \lambda\gamma y^{\gamma-1} \exp[-\lambda y^{\gamma}]$$

for $y > 0$. Note that the PI is for survival times Y, not censored survival times.

The *survpack* function `wpisim` simulates PI (4.21) for the WPH full model with $d = p$. WPH data $(\boldsymbol{x}_1, T_1), ..., (\boldsymbol{x}_n, T_n)$ is generated as described for variable selection in Section 4.8. The T_i are right censored survival times corresponding to Y_i. Hence for the output below, $\boldsymbol{\beta} = (1, 1, 1, 1, 0, ..., 0)^T$ with $p = 10$, and psi $= 0.9$ means the ten predictor variables are highly correlated. The Weibull AFT is fit and used to get $\hat{\gamma}$ and $\hat{\boldsymbol{\beta}}_W$ for the WPH. Then $B = 1000$ values $Y_1^*, ..., Y_B^*$ are generated for $Y|\boldsymbol{x}_f \sim W(\hat{\gamma}, \hat{\lambda}_0 \exp(\hat{\boldsymbol{\beta}}_W^T \boldsymbol{x}_f))$. The large sample 95% PI (4.21) is used for Y_f with $d = p = 10$. 5000 WPH data sets are generated with 5000 values of $(\boldsymbol{x}_f, Y_f)$. The values of Y_f and Y_i^* are not censored. Then 94.76% of the 5000 PIs contained Y_f, with an average length of 1.0554.

```
wpisim(n=1000,p=10,k=4,nruns=5000,psi=0.9,gam=4,B=1000)
$int
(Intercept)
  0.0169485
$beta
 [1] 1 1 1 1 0 0 0 0 0 0
$fullpicov
[1] 0.9476
$fullpimenlen
[1] 1.0554
```

4.6 Prediction Regions

Consider predicting a $p \times 1$ future test value $\boldsymbol{x}_f$, given past training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ where $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid. Much as confidence regions and intervals give a measure of precision for the point estimator $\hat{\boldsymbol{\theta}}$ of the parameter $\boldsymbol{\theta}$, prediction regions and intervals give a measure of precision of the point estimator $T = \hat{\boldsymbol{x}}_f$ of the future random vector $\boldsymbol{x}_f$.

Definition 4.16. A *large sample* $100(1-\delta)\%$ *prediction region* is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n)$ is eventually bounded below by $1 - \delta$ as $n \to \infty$. A prediction region is *asymptotically optimal* if its volume converges in probability to the volume of the minimum volume covering region or the highest density region of the distribution of $\boldsymbol{x}_f$.

If $\boldsymbol{x}_f$ has a pdf, we often want $P(\boldsymbol{x}_f \in \mathcal{A}_n) \to 1 - \delta$ as $n \to \infty$. A PI is a prediction region where $p = 1$. Highest density regions are usually hard to estimate for p much larger than four, but many elliptically contoured

distributions with a nonsingular population covariance matrix, including the multivariate normal distribution, have highest density regions that can be estimated by the nonparametric prediction region (4.29). For more about highest density regions, see Olive (2017b, pp. 148-155) and Hyndman (1996).

For multivariate data, sample Mahalanobis distances play a role similar to that of residuals in multiple linear regression. Let the observed training data be collected in an $n \times p$ matrix $\boldsymbol{W}$. Let the $p \times 1$ column vector $T = T(\boldsymbol{W})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C} = \boldsymbol{C}(\boldsymbol{W})$ be a dispersion estimator.

Definition 4.17. Let $x_{1j}, ..., x_{nj}$ be measurements on the jth random variable X_j corresponding to the jth column of the data matrix $\boldsymbol{W}$. The jth *sample mean* is $\overline{x}_j = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} x_{kj}$. The *sample covariance* S_{ij} estimates $\mathrm{Cov}(X_i, X_j) = \sigma_{ij} = E[(X_i - E(X_i))(X_j - E(X_j))]$, and

$$S_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j).$$

$S_{ii} = S_i^2$ is the *sample variance* that estimates the population variance $\sigma_{ii} = \sigma_i^2$. The *sample correlation* r_{ij} estimates the population correlation $\mathrm{Cor}(X_i, X_j) = \rho_{ij} = \dfrac{\sigma_{ij}}{\sigma_i \sigma_j}$, and

$$r_{ij} = \frac{S_{ij}}{S_i S_j} = \frac{S_{ij}}{\sqrt{S_{ii} S_{jj}}} = \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{kj} - \overline{x}_j)^2}}.$$

Definition 4.18. Let $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ be the data where $\boldsymbol{x}_i$ is a $p \times 1$ vector. The **sample mean** or *sample mean vector*

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i = (\overline{x}_1, ..., \overline{x}_p)^T = \frac{1}{n}\boldsymbol{W}^T\mathbf{1}$$

where $\mathbf{1}$ is the $n \times 1$ vector of ones. The **sample covariance matrix**

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T = (S_{ij}).$$

That is, the ij entry of $\boldsymbol{S}$ is the sample covariance S_{ij}. The *classical estimator of multivariate location and dispersion* is $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. The **sample correlation matrix**

$$\boldsymbol{R} = (r_{ij}).$$

That is, the ij entry of $\boldsymbol{R}$ is the sample correlation r_{ij}.

It can be shown that $(n-1)\boldsymbol{S} = \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^T - \overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^T =$

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

Hence if the *centering matrix* $\boldsymbol{G} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T$, then $(n-1)\boldsymbol{S} = \boldsymbol{W}^T\boldsymbol{G}\boldsymbol{W}$.

See Definition 4.2 for the population mean and population covariance matrix. The Mahalanobis distance in Definition 4.8 is a random variable that estimates the population Mahalanobis distance defined after Definition 4.8.

Definition 4.19. The ith *Mahalanobis distance* $D_i = \sqrt{D_i^2}$ where the ith *squared Mahalanobis distance* is

$$D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W})) \quad (4.22)$$

for each point $\boldsymbol{x}_i$. Notice that D_i^2 is a random variable (scalar valued). Let $(T, \boldsymbol{C}) = (T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}))$. Then

$$D_{\boldsymbol{x}}^2(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T).$$

Hence D_i^2 uses $\boldsymbol{x} = \boldsymbol{x}_i$.

Let the $p \times 1$ location vector be $\boldsymbol{\mu}$, often the population mean, and let the $p \times p$ dispersion matrix be $\boldsymbol{\Sigma}$, often the population covariance matrix. Notice that if $\boldsymbol{x}$ is a random vector, then the population squared Mahalanobis distance is

$$D_{\boldsymbol{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \quad (4.23)$$

and that the term $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})$ is the $p-$dimensional analog to the z-score used to transform a univariate $N(\mu, \sigma^2)$ random variable into a $N(0,1)$ random variable. Hence the sample Mahalanobis distance $D_i = \sqrt{D_i^2}$ is an analog of the absolute value $|Z_i|$ of the sample Z-score $Z_i = (X_i - \overline{X})/\hat{\sigma}$. Also notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center $T(\boldsymbol{W})$ is $D_i(T(\boldsymbol{W}), \boldsymbol{I}_p)$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix.

Consider the hyperellipsoid

$$\mathcal{A}_n = \{\boldsymbol{x} : D_{\boldsymbol{x}}^2(\overline{\boldsymbol{x}}, \boldsymbol{S}) \le D_{(c)}^2\} = \{\boldsymbol{x} : D_{\boldsymbol{x}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \le D_{(c)}\}. \quad (4.24)$$

If n is large, we can use $c = k_n = \lceil n(1-\delta) \rceil$. If n is not large, using $c = U_n$ where U_n decreases to k_n, can improve small sample performance. U_n will be defined in the paragraph below Equation (4.28). Olive (2013b) showed that (4.24) is a large sample $100(1-\delta)\%$ prediction region under mild conditions, although regions with smaller volumes may exist. Note that the result follows since if $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{S}$ are nonsingular, then the Mahalanobis distance is a continuous function of $(\overline{\boldsymbol{x}}, \boldsymbol{S})$. Let $\boldsymbol{\mu} = E(\boldsymbol{x})$ and $D = D(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. Then $D_i \xrightarrow{D} D$

and $D_i^2 \xrightarrow{D} D^2$. Hence the sample percentiles of the D_i are consistent estimators of the population percentiles of D at continuity points of the cumulative distribution function of D.

A problem with the prediction regions that cover $\approx 100(1-\delta)\%$ of the training data cases $\boldsymbol{x}_i$ (such as (4.24) for $c = k_n$), is that they have coverage lower than the nominal coverage of $1-\delta$ for moderate n. This result is not surprising since empirically statistical methods perform worse on test data. Increasing c will improve the coverage for moderate samples. Empirically for many distributions, for $n \approx 20p$, the prediction region (4.24) applied to iid data using $c = k_n = \lceil n(1-\delta) \rceil$ tended to have undercoverage as high as 5%. The undercoverage decreases rapidly as n increases. Let $q_n = \min(1-\delta+0.05, 1-\delta+p/n)$ for $\delta > 0.1$ and

$$q_n = \min(1-\delta/2, 1-\delta+10\delta p/n), \quad \text{otherwise.} \tag{4.25}$$

If $1-\delta < 0.999$ and $q_n < 1-\delta+0.001$, set $q_n = 1-\delta$. Using

$$c = \lceil n q_n \rceil \tag{4.26}$$

in (4.24) decreased the undercoverage.

If $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, d\ \boldsymbol{\Sigma})$ for some constant $d > 0$ where $\boldsymbol{\Sigma}$ is nonsingular, then $D^2(T, \boldsymbol{C}) = (\boldsymbol{x}-T)^T \boldsymbol{C}^{-1}(\boldsymbol{x}-T) =$

$$(\boldsymbol{x}-\boldsymbol{\mu}+\boldsymbol{\mu}-T)^T[\boldsymbol{C}^{-1} - d^{-1}\boldsymbol{\Sigma}^{-1} + d^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x}-\boldsymbol{\mu}+\boldsymbol{\mu}-T)$$

$$= d^{-1}D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + o_p(1).$$

Thus the sample percentiles of $D_i^2(T, \boldsymbol{C})$ are consistent estimators of the percentiles of $d^{-1}D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (at continuity points $D_{1-\delta}$ of the cdf of $D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$). If $\boldsymbol{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $D_{\boldsymbol{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \chi_m^2$.

Suppose $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}_M, b\ \boldsymbol{S}_M)$ is the sample mean and scaled sample covariance matrix applied to some subset of the data. The classical estimator $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ satisfies this assumption. For $h > 0$, the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z}-T)^T \boldsymbol{C}^{-1}(\boldsymbol{z}-T) \le h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}^2 \le h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \le h\} \tag{4.27}$$

has volume equal to

$$\frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p \sqrt{det(\boldsymbol{C})} = \frac{2\pi^{p/2}}{p\Gamma(p/2)} h^p b^{p/2} \sqrt{det(\boldsymbol{S}_M)}. \tag{4.28}$$

A future observation (random vector) $\boldsymbol{x}_f$ is in the region (4.27) if $D_{\boldsymbol{x}_f} \le h$.

If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ for some constant $d > 0$ where $\boldsymbol{\Sigma}$ is nonsingular, then (4.27) is a large sample $100(1-\delta)\%$ prediction region if $h = D_{(U_n)}$ where $D_{(U_n)}$ is the $100q_n$th sample quantile of the D_i where q_n is defined above (4.26). If $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ and $\boldsymbol{x}_f$ are iid, then prediction region (4.29) is asymptotically optimal for a large class of elliptically contoured

distributions since the volume of (4.29) converges in probability to the volume of the highest density region. (These distributions have a highest density region which is a hyperellipsoid determined by a population Mahalanobis distance.)

The Olive (2013a) nonparametric prediction region uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. For the classical prediction region, see Johnson and Wichern (1988, pp. 134, 151). Refer to the above paragraph for $D_{(U_n)}$.

Definition 4.20. The large sample $100(1-\delta)\%$ *nonparametric prediction region* for a future value $\boldsymbol{x}_f$ given iid data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D^2_{(U_n)}\}, \tag{4.29}$$

while the large sample $100(1-\delta)\%$ *classical prediction region* is

$$\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq \chi^2_{p,1-\delta}\}. \tag{4.30}$$

If p is small, Mahalanobis distances tend to be right skewed with a population shorth that discards the right tail. For $p = 1$ and $n \geq 20$, the finite sample correction factors c/n for c given by (4.19) and (4.26) do not differ by much more than 3% for $0.01 \leq \delta \leq 0.5$. See Figure 4.2 where ol = (Eq. 4.26)$/n$ is plotted versus fr = (Eq. 4.19)$/n$ for $n = 20, 21, ..., 500$. The top plot is for $\delta = 0.01$, while the bottom plot is for $\delta = 0.3$. The identity line is added to each plot as a visual aid. The value of n increases from 20 to 500 from the right of the plot to the left of the plot. Examining the axes of each plot shows that the correction factors do not differ greatly. *R* code to create Figure 4.2 is shown below.

```
cmar <- par("mar"); par(mfrow = c(2, 1))
par(mar=c(4.0,4.0,2.0,0.5))
frey(0.01); frey(0.3)
par(mfrow = c(1, 1)); par(mar=cmar)
```

Remark 4.7. The nonparametric prediction region (4.29) is useful if $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid from a distribution with a nonsingular covariance matrix, and the sample size n is large enough. The distribution could be continuous, discrete, or a mixture. The asymptotic coverage is $1-\delta$ if D has a pdf, although prediction regions with smaller volume may exist. If the $100(1-\delta)$th percentile $D_{1-\delta}$ of D is not a continuity point of the distribution of D, then the asymptotic coverage tends to be $\geq 1-\delta$ since a sample percentile with cutoff q_n that decreases to $1-\delta$ is used and a closed region is used. Often D has a continuous distribution and hence has no discontinuity points for $0 < \delta < 1$. (If there is a jump in the distribution from 0.9 to 0.96 at discontinuity point a, and the nominal coverage is 0.95, we want 0.96 coverage instead of 0.9. So we want the sample percentile to decrease to a.) The nonparametric prediction region (4.29) contains U_n of the training data cases $\boldsymbol{x}_i$ provided

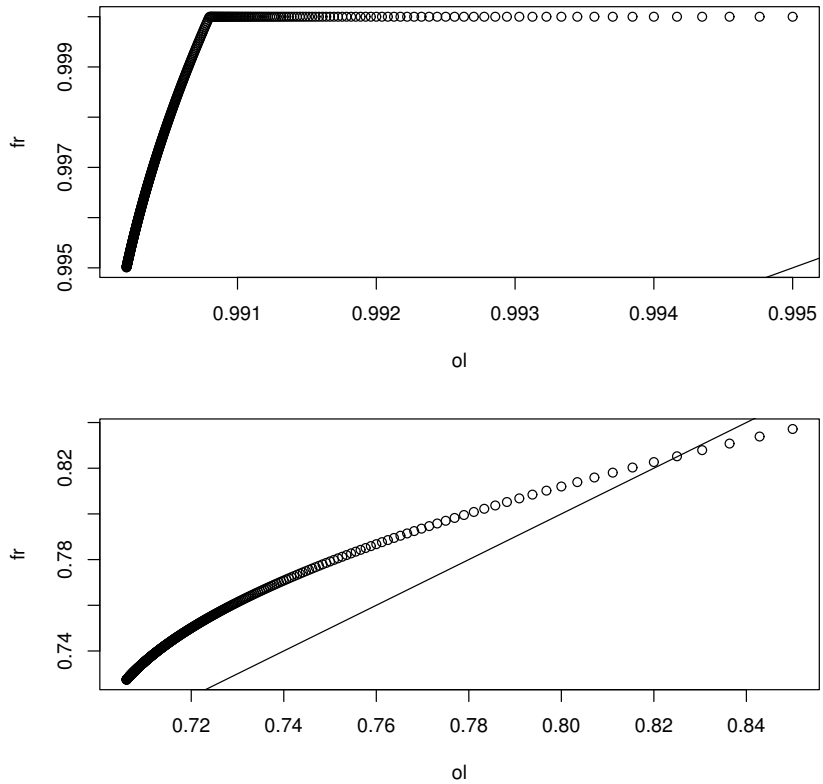

Fig. 4.2 Correction Factor Comparison when $\delta = 0.01$ (Top Plot) and $\delta = 0.3$ (Bottom Plot)

that $\boldsymbol{S}$ is nonsingular, even if the model is wrong. For many distributions, the coverage started to be close to $1 - \delta$ for $n \geq 10p$ where the coverage is the simulated percentage of times that the prediction region contained $\boldsymbol{x}_f$.

Remark 4.8. The most used prediction regions assume that the error vectors are iid from a multivariate normal distribution. Using (4.28), the ratio of the volumes of regions (4.30) and (4.29) is

$$\left(\frac{\chi^2_{p,1-\delta}}{D^2_{(U_n)}}\right)^{p/2},$$

which can become close to zero rapidly as p gets large if the $\boldsymbol{x}_i$ are not from the light tailed multivariate normal distribution. For example, suppose $\chi^2_{4,0.5} \approx 3.33$ and $D^2_{(U_n)} \approx D^2_{\boldsymbol{x},0.5} = 6$. Then the ratio is $(3.33/6)^2 \approx 0.308$. Hence if the data is not multivariate normal, severe undercoverage can occur

if the classical prediction region is used, and the undercoverage tends to get worse as the dimension p increases. The coverage need not to go to 0, since by the multivariate Chebyshev's inequality, $P(D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma_x}) \leq \gamma) \geq 1 - p/\gamma > 0$ for $\gamma > p$ where the population covariance matrix $\boldsymbol{\Sigma_x} = \text{Cov}(\boldsymbol{x})$. See Budny (2014), Chen (2011), and Navarro (2014, 2016). Using $\gamma = h^2 = p/\delta$ in (4.27) usually results in prediction regions with volume and coverage that is too large.

Remark 4.9. The nonparametric prediction region (4.29) starts to have good coverage for $n \geq 10p$ for a large class of distributions. Olive (2013b) suggests $n \geq 50p$ may be needed for the prediction region to have a good volume. Of course for any n there are error distributions that will have severe undercoverage.

For the multivariate lognormal distribution with $n = 20p$, the large sample nonparametric 95% prediction region (4.29) had coverages 0.970, 0.959, and 0.964 for $p = 100, 200$, and 500. Some R code is below.

```
nruns=1000 #lognormal, p = 100, n = 20p = 2000
count<-0
for(i in 1:nruns){
x <- exp(matrix(rnorm(200000),ncol=100,nrow=2000))
xff <- exp(as.vector(rnorm(100)))
count <- count + predrgn(x,xf=xff)$inr}
count #970/1000, may take a few minutes
```

Notice that for the training data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, if $\boldsymbol{C}^{-1}$ exists, then $c \approx 100q_n\%$ of the n cases are in the prediction regions for $\boldsymbol{x}_f = \boldsymbol{x}_i$, and $q_n \to 1-\delta$ even if $(T, \boldsymbol{C})$ is not a good estimator. 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 $(T, \boldsymbol{C})$ is used or if the $\boldsymbol{x}_i$ do not come from an elliptically contoured distribution. Also notice that $q_n = 1 - \delta/2$ or $q_n = 1 - \delta + 0.05$ for $n \leq 20p$ and $q_n \to 1 - \delta$ as $n \to \infty$. If $q_n \equiv 1 - \delta$ and $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$ where $d > 0$ and $\boldsymbol{\Sigma}$ is nonsingular, then (4.27) with $h = D_{(U_n)}$ is a large sample prediction region, but taking q_n given by (4.25) improves the finite sample performance of the prediction region. Taking $q_n \equiv 1 - \delta$ does not take into account variability of $(T, \boldsymbol{C})$, and for $n = 20p$ the resulting prediction region tended to have undercoverage as high as $\min(0.05, \delta/2)$. Using (4.25) helped reduce undercoverage for small $n \geq 20p$ due to the unknown variability of $(T, \boldsymbol{C})$.

4.7 Bootstrapping Hypothesis Tests and Confidence Regions

This section shows that, under regularity conditions, applying the nonparametric prediction region of Section 4.6 to a bootstrap sample results in a confidence region. The volume of a confidence region $\to 0$ as $n \to 0$, while the volume of a prediction region goes to that of a population region that would contain a new $\boldsymbol{x}_f$ with probability $1-\delta$. The nominal coverage is $100(1-\delta)$. If the actual coverage $100(1-\delta_n) > 100(1-\delta)$, then the region is *conservative*. If $100(1-\delta_n) < 100(1-\delta)$, then the region is *liberal*. A region that is 5% conservative is considered "much better" than a region that is 5% liberal.

When teaching confidence intervals, it is often noted that by the central limit theorem, the probability that $\overline{Y}_n$ is within two standard deviations $(2SD(\overline{Y}_n) = 2\sigma/\sqrt{n})$ of $\theta = \mu$ is about 95%. Hence the probability that θ is within two standard deviations of $\overline{Y}_n$ is about 95%. Thus the interval $[\theta - 1.96S/\sqrt{n}, \theta + 1.96S/\sqrt{n}]$ is a large sample 95% prediction interval for a future value of the sample mean $\overline{Y}_{n,f}$ if θ is known, while $[\overline{Y}_n - 1.96S/\sqrt{n}, \overline{Y}_n + 1.96S/\sqrt{n}]$ is a large sample 95% confidence interval for the population mean θ. Note that the lengths of the two intervals are the same. Where the interval is centered, at the parameter θ or the statistic $\overline{Y}_n$, determines whether the interval is a prediction or a confidence interval. See Theorem 4.32 for a similar relationship between confidence regions and prediction regions.

Definition 4.21. A *large sample* $100(1-\delta)\%$ *confidence region* for a vector of parameters $\boldsymbol{\theta}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$.

If $\mathcal{A}_n$ is based on a squared Mahalanobis distance D^2 with a limiting distribution that has a pdf, we often want $P(\boldsymbol{\theta} \in \mathcal{A}_n) \to 1-\delta$ as $n \to \infty$.

There are several methods for obtaining a bootstrap sample $T_1^*, ..., T_B^*$ where the sample size n is suppressed: $T_i^* = T_{in}^*$. The parametric bootstrap, nonparametric bootstrap, and residual bootstrap will be used. Applying prediction region (4.29) to the bootstrap sample will result in a confidence region for $\boldsymbol{\theta}$. When $g = 1$, applying the shorth PI (4.19) or PI (4.16) to the bootstrap sample results in a confidence interval for θ. Section 4.7.2 will help clarify ideas.

When $g = 1$, a confidence interval is a special case of a confidence region. One sided confidence intervals give a lower or upper confidence bound for θ. A large sample $100(1-\delta)\%$ lower confidence interval $(-\infty, U_n]$ uses an upper confidence bound U_n and is in the lower tail of the distribution of $\hat{\theta}$. A large sample $100(1-\delta)\%$ upper confidence interval $[L_n, \infty)$ uses a lower confidence bound L_n and is in the upper tail of the distribution of $\hat{\theta}$. These CIs can be

useful if $\theta \in [a, b]$ and $\theta = a$ or $\theta = b$ is of interest for a hypothesis test. For example, $[a, b] = [0, 1]$ if $\theta = \rho^2$, the squared population correlation. Then use $[0, U_n]$ and $[L_n, 1]$ as CIs, e.g. if we expect $\theta = 0$ we might test $H_0 : \theta \leq 0.05$ versus $H_0 : \theta > 0.05$, and fail to reject H_0 if $U_n < 0.05$. Again we often want the probability to converge to $1 - \delta$ if the confidence interval is based on a statistic with an asymptotic distribution that has a pdf.

Definition 4.22. The interval $[L_n, U_n]$ is a large sample $100(1 - \delta)\%$ *confidence interval* for θ if $P(L_n \leq \theta \leq U_n)$ is eventually bounded below by $1 - \delta$ as $n \to \infty$. The interval $(-\infty, U_n]$ is a large sample $100(1 - \delta)\%$ *lower confidence interval* for θ if $P(\theta \leq U_n)$ is eventually bounded below by $1 - \delta$ as $n \to \infty$. The interval $[L_n, \infty)$ is large sample $100(1-\delta)\%$ *upper confidence interval* for θ if $P(\theta \geq L_n)$ is eventually bounded below by $1 - \delta$ as $n \to \infty$.

Next we discuss bootstrap confidence intervals that are obtained by applying prediction intervals (4.16) and (4.19) to the bootstrap sample. Some additional bootstrap CIs are obtained from bootstrap confidence regions from Section 4.7.2 when $g = 1$. See Efron (1982) and Chen (2016) for the percentile method CI. Let T_n be an estimator of a parameter θ such as $T_n = \overline{Z} = \sum_{i=1}^n Z_i/n$ with $\theta = E(Z_1)$. Let $T_1^*, ..., T_B^*$ be a bootstrap sample for T_n. Let $T_{(1)}^*, ..., T_{(B)}^*$ be the order statistics of the the bootstrap sample. The CI (4.31) is obtained by applying PI (4.16) to the bootstrap sample with B used instead of n. Hence (4.31) is also a large sample prediction interval for a future value of T_f^* if the T_i^* are iid from the empirical distribution discussed in Section 4.5.1.

Definition 4.23. The bootstrap percentile method large sample $100(1 - \delta)\%$ confidence interval for θ is an interval $[T_{(k_L)}^*, T_{(K_U)}^*]$ containing $\approx \lceil B(1 - \delta)\rceil$ of the T_i^*. Let $k_1 = \lceil B\delta/2\rceil$ and $k_2 = \lceil B(1 - \delta/2)\rceil$. A common choice is

$$[T_{(k_1)}^*, T_{(k_2)}^*]. \tag{4.31}$$

The large sample $100(1 - \delta)\%$ *lower percentile method* CI for θ is $(-\infty, T_{(\lceil B(1-\delta)\rceil)}^*]$. The large sample $100(1 - \delta)\%$ *upper percentile method* CI for θ is $[T_{(\lceil B\delta\rceil)}^*, \infty)$.

Definition 4.24. The large sample $100(1 - \delta)\%$ *lower shorth* CI for θ is $(-\infty, T_{(c)}^*]$, while the large sample $100(1 - \delta)\%$ *upper shorth* CI for θ is $[T_{(B-c+1)}^*, \infty)$. The large sample $100(1 - \delta)\%$ *shorth*(c) *CI* uses the interval $[T_{(1)}^*, T_{(c)}^*], [T_{(2)}^*, T_{(c+1)}^*], ..., [T_{(B-c+1)}^*, T_{(B)}^*]$ of shortest length. Here

$$c = \min(B, \lceil B[1 - \delta + 1.12\sqrt{\delta/B}\ ]\ \rceil). \tag{4.32}$$

Applied to a bootstrap sample, the Frey shorth interval can be regarded as the shortest percentile method confidence interval, asymptotically. Hence the shorth confidence interval is a practical implementation of the Hall (1988)

shortest bootstrap interval based on all possible bootstrap samples. See Remark 4.13 for some theory for bootstrap CIs such as (4.31) and (4.32).

4.7.1 The Bootstrap

This subsection illustrates the nonparametric bootstrap with some examples. Suppose a statistic T_n is computed from a data set of n cases. The nonparametric bootstrap draws n cases with replacement from that data set. Then T_1^* is the statistic T_n computed from the sample. This process is repeated B times to produce the bootstrap sample $T_1^*, ..., T_B^*$. Sampling cases with replacement uses the empirical distribution.

Definition 4.25. Suppose that data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ has been collected and observed. Often the data is a random sample (iid) from a distribution with cdf F. The *empirical distribution* is a discrete distribution where the $\boldsymbol{x}_i$ are the possible values, and each value is equally likely. If $\boldsymbol{w}$ is a random variable having the empirical distribution, then $p_i = P(\boldsymbol{w} = \boldsymbol{x}_i) = 1/n$ for $i = 1, ..., n$. The *cdf of the empirical distribution* is denoted by F_n.

Example 4.14. Let $\boldsymbol{w}$ be a random variable having the empirical distribution given by Definition 4.25. Show that $E(\boldsymbol{w}) = \overline{\boldsymbol{x}} \equiv \overline{\boldsymbol{x}}_n$ and $\text{Cov}(\boldsymbol{w}) = \dfrac{n-1}{n}\boldsymbol{S} \equiv \dfrac{n-1}{n}\boldsymbol{S}_n$.

Solution: Recall that for a discrete random vector, the population expected value $E(\boldsymbol{w}) = \sum \boldsymbol{x}_i p_i$ where $\boldsymbol{x}_i$ are the values that $\boldsymbol{w}$ takes with positive probability p_i. Similarly, the population covariance matrix

$$\text{Cov}(\boldsymbol{w}) = E[(\boldsymbol{w} - E(\boldsymbol{w}))(\boldsymbol{w} - E(\boldsymbol{w}))^T] = \sum (\boldsymbol{x}_i - E(\boldsymbol{w}))(\boldsymbol{x}_i - E(\boldsymbol{w}))^T p_i.$$

Hence

$$E(\boldsymbol{w}) = \sum_{i=1}^{n} \boldsymbol{x}_i \frac{1}{n} = \overline{\boldsymbol{x}},$$

and

$$\text{Cov}(\boldsymbol{w}) = \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \frac{1}{n} = \frac{n-1}{n}\boldsymbol{S}. \quad \square$$

Example 4.15. If $W_1, ..., W_n$ are iid from a distribution with cdf F_W, then the empirical cdf F_n corresponding to F_W is given by

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} I(W_i \le y)$$

where the indicator $I(W_i \le y) = 1$ if $W_i \le y$ and $I(W_i \le y) = 0$ if $W_i > y$. Fix n and y. Then $nF_n(y) \sim$ binomial $(n, F_W(y))$. Thus $E[F_n(y)] = F_W(y)$ and $V[F_n(y)] = F_W(y)[1 - F_W(y)]/n$. By the central limit theorem,

$$\sqrt{n}(F_n(y) - F_W(y)) \xrightarrow{D} N(0, F_W(y)[1 - F_W(y)]).$$

Thus $F_n(y) - F_W(y) = O_P(n^{-1/2})$, and F_n is a reasonable estimator of F_W if the sample size n is large.

Suppose there is data $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ collected into an $n \times p$ matrix $\boldsymbol{W}$. Let the statistic $T_n = t(\boldsymbol{W}) = T(F_n)$ be computed from the data. Suppose the statistic estimates $\boldsymbol{\mu} = T(F)$, and let $t(\boldsymbol{W}^*) = t(F_n^*) = T_n^*$ indicate that t was computed from an iid sample from the empirical distribution F_n: a sample $\boldsymbol{w}_1^*, ..., \boldsymbol{w}_n^*$ of size n was drawn with replacement from the observed sample $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. This notation is used for von Mises differentiable statistical functions in large sample theory. See Serfling (1980, ch. 6). The empirical distribution is also important for the influence function (widely used in robust statistics). The *nonparametric bootstrap* draws B samples of size n from the rows of $\boldsymbol{W}$, e.g. from the empirical distribution of $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$. Then T_{jn}^* is computed from the jth bootstrap sample for $j = 1, ..., B$.

Example 4.16. Suppose the data is 1, 2, 3, 4, 5, 6, 7. Then $n = 7$ and the sample median T_n is 4. Using R, we drew $B = 2$ bootstrap samples (samples of size n drawn with replacement from the original data) and computed the sample median $T_{1,n}^* = 3$ and $T_{2,n}^* = 4$.

```
b1 <- sample(1:7,replace=T)
b1
[1] 3 2 3 2 5 2 6
median(b1)
[1] 3
b2 <- sample(1:7,replace=T)
b2
[1] 3 5 3 4 3 5 7
median(b2)
[1] 4
```

The bootstrap has been widely used to estimate the population covariance matrix of the statistic $\text{Cov}(T_n)$, for testing hypotheses, and for obtaining confidence regions (often confidence intervals). An iid sample $T_{1n}, ..., T_{Bn}$ of size B of the statistic would be very useful for inference, but typically we only have one sample of data and one value $T_n = T_{1n}$ of the statistic. Often $T_n = t(\boldsymbol{w}_1, ..., \boldsymbol{w}_n)$, and the bootstrap sample $T_{1n}^*, ..., T_{Bn}^*$ is formed where $T_{jn}^* = t(\boldsymbol{w}_{j1}^*, ..., \boldsymbol{w}_{jn}^*)$. Section 4.7.3 will show that $T_{1n}^* - T_n, ..., T_{Bn}^* - T_n$ is pseudodata for $T_{1n} - \boldsymbol{\theta}, ..., T_{Bn} - \boldsymbol{\theta}$ when n is large in that $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ and $\sqrt{n}(T^* - T_n) \xrightarrow{D} \boldsymbol{u}$.

Suppose there is a statistic T_n that is a $g \times 1$ vector. Let

$$\overline{T}^* = \frac{1}{B}\sum_{i=1}^{B} T_i^* \quad \text{and} \quad \boldsymbol{S}_T^* = \frac{1}{B-1}\sum_{i=1}^{B}(T_i^* - \overline{T}^*)(T_i^* - \overline{T}^*)^T \tag{4.33}$$

be the sample mean and sample covariance matrix of the bootstrap sample $T_1^*, ..., T_B^*$ where $T_i^* = T_{i,n}^*$. Fix n, and let $E(T_{i,n}^*) = \boldsymbol{\theta}_n$ and $\text{Cov}(T_{i,n}^*) = \boldsymbol{\Sigma}_n$.

We will often assume that $\text{Cov}(T_n) = \boldsymbol{\Sigma}_T$, and $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma}_A)$ where $\boldsymbol{\Sigma}_A > 0$ is positive definite and nonsingular. Often $n\hat{\boldsymbol{\Sigma}}_T \xrightarrow{P} \boldsymbol{\Sigma}_A$. Suppose the $T_i^* = T_{i,n}^*$ are iid from some distribution with cdf $\tilde{F}_n$. For example, if $T_{i,n}^* = t(F_n^*)$ where iid samples from F_n are used, then $\tilde{F}_n$ is the cdf of $t(F_n^*)$. With respect to $\tilde{F}_n$, both $\boldsymbol{\theta}_n$ and $\boldsymbol{\Sigma}_n$ are parameters, but with respect to F, $\boldsymbol{\theta}_n$ is a random vector and $\boldsymbol{\Sigma}_n$ is a random matrix. For fixed n, by the multivariate central limit theorem,

$$\sqrt{B}(\overline{T}^* - \boldsymbol{\theta}_n) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma}_n) \quad \text{and} \quad \mathrm{B}(\overline{\mathrm{T}}^* - \boldsymbol{\theta}_\mathrm{n})^\mathrm{T}[\boldsymbol{S}_\mathrm{T}^*]^{-1}(\overline{\mathrm{T}}^* - \boldsymbol{\theta}_\mathrm{n}) \xrightarrow{\mathrm{D}} \chi^2_\mathrm{r}$$

as $B \to \infty$.

Remark 4.10. For Example 4.14, the bootstrap works but is expensive compared to large sample theory. Fix n, then $\overline{T}^* \xrightarrow{P} \boldsymbol{\theta}_n = \overline{\boldsymbol{x}}$ and $\boldsymbol{S}_T^* \xrightarrow{P} (n-1)\boldsymbol{S}/n$ as $B \to \infty$, but using $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ makes more sense. For Example 4.14, it is known how the bootstrap sample behaves as $B \to \infty$. The bootstrap can be very useful when $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\mathbf{0}, \boldsymbol{\Sigma}_A)$, but it not known how to estimate $\boldsymbol{\Sigma}_A$ without using a resampling method like the bootstrap. The bootstrap may be useful when $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, but the limiting distribution (the distribution of $\boldsymbol{u}$) is unknown.

4.7.2 Bootstrap Confidence Regions for Hypothesis Testing

When the bootstrap is used, a large sample $100(1-\delta)\%$ confidence region for a $g \times 1$ parameter vector $\boldsymbol{\theta}$ is a set $\mathcal{A}_n = \mathcal{A}_{n,B}$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_{n,B})$ is eventually bounded below by $1-\delta$ as $n, B \to \infty$. The B is often suppressed. Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a known $g \times 1$ vector. Then reject H_0 if $\boldsymbol{\theta}_0$ is not in the confidence region $\mathcal{A}_n$. Let the $g \times 1$ vector T_n be an estimator of $\boldsymbol{\theta}$. Let $T_1^*, ..., T_B^*$ be the bootstrap sample for T_n. Let $\boldsymbol{A}$ be a full rank $g \times p$ constant matrix. For variable selection, consider testing $H_0 : \boldsymbol{A\beta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{A\beta} \neq \boldsymbol{\theta}_0$ with $\boldsymbol{\theta} = \boldsymbol{A\beta}$ where often $\boldsymbol{\theta}_0 = \mathbf{0}$. Then let $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ and let $T_i^* = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{I_{min},0,i}$ for $i = 1, ..., B$. The statistic $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is the variable selection estimator padded

with zeroes. See Section 4.4. Let $\overline{T}^*$ and $\boldsymbol{S}_T^*$ be the sample mean and sample covariance matrix of the bootstrap sample $T_1^*, ..., T_B^*$. See Equation (4.33). A useful result is $d_n F_{g,d_n,1-\delta} \to \chi^2_{g,1-\delta}$ as $d_n \to \infty$. Here $P(X \le \chi^2_{g,1-\delta}) = 1-\delta$ if $X \sim \chi^2_g$, and $P(X \le F_{g,d_n,1-\delta}) = 1-\delta$ if $X \sim F_{g,d_n}$. Let $k_B = \lceil B(1-\delta) \rceil$.

Definition 4.26. a) The standard bootstrap large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\boldsymbol{S}_T^*]^{-1} (\boldsymbol{w} - T_n) \le D^2_{1-\delta}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \le D^2_{1-\delta}\} \tag{4.34}$$

where $D^2_{1-\delta} = \chi^2_{g,1-\delta}$ or $D^2_{1-\delta} = d_n F_{g,d_n,1-\delta}$ where $d_n \to \infty$ as $n \to \infty$. b) The Bickel and Ren (2001) large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T [\hat{\boldsymbol{\Sigma}}_A/n]^{-1} (\boldsymbol{w} - T_n) \le D^2_{(k_B,T)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \hat{\boldsymbol{\Sigma}}_A/n) \le D^2_{(k_B,T)}\} \tag{4.35}$$

where the cutoff $D^2_{(k_B,T)}$ is the $100k_B$th sample quantile of the $D_i^2 = (T_i^* - T_n)^T [\hat{\boldsymbol{\Sigma}}_A/n]^{-1} (T_i^* - T_n) = n(T_i^* - T_n)^T [\hat{\boldsymbol{\Sigma}}_A]^{-1} (T_i^* - T_n)$.

Confidence region (4.34) needs $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$ and $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{\Sigma}_A > 0$ as $n, B \to \infty$. See Machado and Parente (2005) for regularity conditions for this assumption. Bickel and Ren (2001) have interesting sufficient conditions for (4.35) to be a confidence region when $\hat{\boldsymbol{\Sigma}}_A$ is a consistent estimator of positive definite $\boldsymbol{\Sigma}_A$. Let the vector of parameters $\boldsymbol{\theta} = T(F)$, the statistic $T_n = T(F_n)$, and the bootstrapped statistic $T^* = T(F_n^*)$ where F is the cdf of iid $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, F_n is the empirical cdf, and F_n^* is the empirical cdf of $\boldsymbol{x}_1^*, ..., \boldsymbol{x}_n^*$, a sample from F_n using the nonparametric bootstrap. If $\sqrt{n}(F_n - F) \xrightarrow{D} \boldsymbol{z}_F$, a Gaussian random process, and if T is sufficiently smooth (has a Hadamard derivative $\dot{T}(F)$), then $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ and $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$ with $\boldsymbol{u} = \dot{T}(F)\boldsymbol{z}_F$. Note that F_n is a perfectly good cdf "F" and F_n^* is a perfectly good empirical cdf from F_n = "F." Thus if n is fixed, and a sample of size m is drawn with replacement from the empirical distribution, then $\sqrt{m}(T(F_m^*) - T_n) \xrightarrow{D} \dot{T}(F_n)\boldsymbol{z}_{F_n}$. Now let $n \to \infty$ with $m = n$. Then bootstrap theory gives $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \lim_{n\to\infty} \dot{T}(F_n)\boldsymbol{z}_{F_n} = \dot{T}(F)\boldsymbol{z}_F \sim \boldsymbol{u}$.

The following three confidence regions will be used for inference after variable selection. The Olive (2017ab, 2018) prediction region method applies prediction region (4.29) to the bootstrap sample. Olive (2017ab, 2018) also gave the modified Bickel and Ren confidence region that uses $\hat{\boldsymbol{\Sigma}}_A = n\boldsymbol{S}_T^*$. The hybrid confidence region is due to Pelawa Watagoda and Olive (2019). Let $q_B = \min(1 - \delta + 0.05, 1 - \delta + g/B)$ for $\delta > 0.1$ and

$$q_B = \min(1 - \delta/2, 1 - \delta + 10\delta g/B), \quad \text{otherwise.} \tag{4.36}$$

If $1-\delta < 0.999$ and $q_B < 1-\delta+0.001$, set $q_B = 1-\delta$. Let $D_{(U_B)}$ be the $100q_B$th sample quantile of the D_i. Use (4.36) as a correction factor for finite $B \geq 50p$.

Definition 4.27. a) The prediction region method large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w}-\overline{T}^*)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w}-\overline{T}^*) \leq D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\} \tag{4.37}$$

where $D^2_{(U_B)}$ is computed from $D_i^2 = (T_i^* - \overline{T}^*)^T[\boldsymbol{S}_T^*]^{-1}(T_i^* - \overline{T}^*)$ for $i = 1, ..., B$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(\overline{T}^* - \boldsymbol{\theta}_0)^T[\boldsymbol{S}_T^*]^{-1}(\overline{T}^* - \boldsymbol{\theta}_0) > D^2_{(U_B)}$. (This procedure is basically the one sample Hotelling's T^2 test applied to the T_i^* using $\boldsymbol{S}_T^*$ as the estimated covariance matrix and replacing the $\chi^2_{g,1-\delta}$ cutoff by $D^2_{(U_B)}$.) b) The modified Bickel and Ren (2001) large sample $100(1-\delta)\%$ confidence region is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \leq D^2_{(U_B,T)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B,T)}\} \tag{4.38}$$

where the cutoff $D^2_{(U_B,T)}$ is the $100q_B$th sample quantile of the $D_i^2 = (T_i^* - T_n)^T[\boldsymbol{S}_T^*]^{-1}(T_i^* - T_n)$. Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(T_n - \boldsymbol{\theta}_0)^T[\boldsymbol{S}_T^*]^{-1}(T_n - \boldsymbol{\theta}_0) > D^2_{(U_B,T)}$. c) Shift region (4.37) to have center T_n, or equivalently, change the cutoff of region (4.38) to $D^2_{(U_B)}$ to get the hybrid large sample $100(1-\delta)\%$ confidence region: $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \leq D^2_{(U_B)}\} =$

$$\{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \leq D^2_{(U_B)}\}. \tag{4.39}$$

Note that the corresponding test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(T_n - \boldsymbol{\theta}_0)^T[\boldsymbol{S}_T^*]^{-1}(T_n - \boldsymbol{\theta}_0) > D^2_{(U_B)}$.

Hyperellipsoids (4.37) and (4.39) have the same volume since they are the same region shifted to have a different center. The ratio of the volumes of regions (4.37) and (4.38) is

$$\frac{|\boldsymbol{S}_T^*|^{1/2}}{|\boldsymbol{S}_T^*|^{1/2}}\left(\frac{D_{(U_B)}}{D_{(U_B,T)}}\right)^g = \left(\frac{D_{(U_B)}}{D_{(U_B,T)}}\right)^g. \tag{4.40}$$

The volume of confidence region (4.38) tends to be greater than that of (4.37) since the T_i^* are closer to $\overline{T}^*$ than T_n on average.

If $g = 1$, then a hyperellipsoid is an interval, and confidence intervals are special cases of confidence regions. Suppose the parameter of interest is θ, and there is a bootstrap sample $T_1^*, ..., T_B^*$ where the statistic T_n is an estimator of θ based on a sample of size n. The percentile method uses an interval that

contains $U_B \approx k_B = \lceil B(1-\delta) \rceil$ of the T_i^*. Let $a_i = |T_i^* - \overline{T}^*|$. Let $\overline{T}^*$ and S_T^{2*} be the sample mean and variance of the T_i^*. Then the squared Mahalanobis distance $D_\theta^2 = (\theta - \overline{T}^*)^2 / S_T^{*2} \leq D_{(U_B)}^2$ is equivalent to $\theta \in [\overline{T}^* - S_T^* D_{(U_B)}, \overline{T}^* + S_T^* D_{(U_B)}] = [\overline{T}^* - a_{(U_B)}, \overline{T}^* + a_{(U_B)}]$, which is an interval centered at $\overline{T}^*$ just long enough to cover U_B of the T_i^*. Hence the prediction region method is a special case of the percentile method if $g = 1$. See Definition 4.23. Efron (2014) used a similar large sample $100(1-\delta)\%$ confidence interval assuming that $\overline{T}^*$ is asymptotically normal. The CI corresponding to (4.38) is defined similarly, and $[T_n - a_{(U_B)}, T_n + a_{(U_B)}]$ is the CI for (4.34). Note that the three CIs corresponding to (4.37)–(4.39) can be computed without finding S_T^* or $D_{(U_B)}$ even if $S_T^* = 0$. The Frey (2013) shorth(c) CI (4.27) computed from the T_i^* can be much shorter than the Efron (2014) or prediction region method confidence intervals. See Remark 4.13 for some theory for bootstrap CIs.

Remark 4.11. We may need $n >> p$ before the $\boldsymbol{S}_T^*$ is a good estimator of $\text{Cov}(T) = \boldsymbol{\Sigma}_T$. The distribution of $\sqrt{n}(T_n - \boldsymbol{\theta})$ is approximated by the distribution of $\sqrt{n}(T^* - T_n)$ or by the distribution of $\sqrt{n}(T^* - \overline{T}^*)$, but n may need to be large before the approximation is good.

Suppose the bootstrap sample mean $\overline{T}^*$ estimates $\boldsymbol{\theta}$, and the bootstrap sample covariance matrix $\boldsymbol{S}_T^*$ estimates $c_n \widehat{\text{Cov}}(T_n) \approx c_n \boldsymbol{\Sigma}_T$ where c_n increases to 1 as $n \to \infty$. For multiple linear regression, this result happens for the residual bootstrap for least squares (OLS) with $c_n = (n-p)/n$. Then $\boldsymbol{S}_T^*$ is not a good estimator of $\widehat{\text{Cov}}(T_n)$ until $c_n \approx 1$ ($n \geq 100p$ for OLS $\hat{\boldsymbol{\beta}}$), but the squared Mahalanobis distance $D_{\boldsymbol{w}}^{2*}(\overline{T}^*, \boldsymbol{S}_T^*) \approx D_{\boldsymbol{w}}^2(\boldsymbol{\theta}, \boldsymbol{\Sigma}_T)/c_n$ and $D_{(U_B)}^{2*} \approx D_{1-\delta}^2/c_n$. Hence the prediction region method has a cutoff $D_{(U_B)}^{2*}$ that estimates the cutoff $D_{1-\delta}^2/c_n$. Thus the prediction region method may give good results for much smaller n than a bootstrap method that uses a $\chi_{g,1-\delta}^2$ cutoff when a cutoff $\chi_{g,1-\delta}^2/c_n$ should be used for moderate n.

Remark 4.12. For bootstrapping the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I_{min},0}$, we will often want $n \geq 20p$ and $B \geq \max(100, n, 50p)$. If T_n is $g \times 1$, we might replace p by g or replace p by d if d is the model degrees of freedom. Sometimes much larger n is needed to avoid undercoverage. We want $B \geq 50g$ so that $\boldsymbol{S}_T^*$ is a good estimator of $Cov(T_n^*)$. Prediction region theory uses correction factors like (4.26) and (4.19) to compensate for finite n. The bootstrap confidence regions (4.37)–(4.39) and the shorth CI use the correction factors (4.36) and (4.32) to compensate for finite $B \geq 50g$. Note that the correction factors make the volume of the confidence region larger as B decreases. Hence a test with larger B will have more power.

4.7.3 Theory for Bootstrap Confidence Regions

Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}$ is $g \times 1$. This section gives some theory for bootstrap confidence regions and for the bagging estimator $\overline{T}^*$, also called the smoothed bootstrap estimator. Empirically, bootstrapping with the bagging estimator often outperforms bootstrapping with T_n. See Breiman (1996), Yang (2003), and Efron (2014). See Büchlmann and Yu (2002) and Friedman and Hall (2007) for theory and references for the bagging estimator. Since (4.38) is a large sample confidence region by Bickel and Ren (2001), (4.37) and (4.39) are too, provided $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \mathbf{0}$.

If i) $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, then under regularity conditions, ii) $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$, iii) $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, iv) $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$, and v) $n\boldsymbol{S}_T^* \xrightarrow{P} \mathrm{Cov}(\boldsymbol{u})$.

Suppose i) and ii) hold with $E(\boldsymbol{u}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma_u}$. With respect to the bootstrap sample, T_n is a constant and the $\sqrt{n}(T_i^* - T_n)$ are iid for $i = 1, ..., B$. Let $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{v}_i \sim \boldsymbol{u}$ where the $\boldsymbol{v}_i$ are iid with the same distribution as $\boldsymbol{u}$. Fix B. Then the average of the $\sqrt{n}(T_i^* - T_n)$ is

$$\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{D} \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim AN_g\left(\mathbf{0}, \frac{\boldsymbol{\Sigma_u}}{B}\right)$$

where $\boldsymbol{z} \sim AN_g(\mathbf{0}, \boldsymbol{\Sigma})$ is an asymptotic multivariate normal approximation. Hence as $B \to \infty$, $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \mathbf{0}$, and iii) and iv) hold. If B is fixed and $\boldsymbol{u} \sim N_g(\mathbf{0}, \boldsymbol{\Sigma_u})$, then

$$\frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim N_g\left(\mathbf{0}, \frac{\boldsymbol{\Sigma_u}}{B}\right) \text{ and } \sqrt{\mathrm{B}}\sqrt{\mathrm{n}}(\overline{\mathrm{T}}^* - \mathrm{T_n}) \xrightarrow{\mathrm{D}} \mathrm{N_g}(\mathbf{0}, \boldsymbol{\Sigma_u}).$$

Hence the prediction region method gives a large sample confidence region for $\boldsymbol{\theta}$ provided that the sample percentile $\hat{D}^2_{1-\delta}$ of the $D^2_{T_i^*}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - \overline{T}^*)^T(n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(T_i^* - \overline{T}^*)$ is a consistent estimator of the percentile $D^2_{n,1-\delta}$ of the random variable $D^2_{\boldsymbol{\theta}}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(\boldsymbol{\theta} - \overline{T}^*)^T(n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(\boldsymbol{\theta} - \overline{T}^*)$ in that $\hat{D}^2_{1-\delta} - D^2_{n,1-\delta} \xrightarrow{P} 0$. Since iii) and iv) hold, the sample percentile will be consistent under much weaker conditions than v) if $\boldsymbol{\Sigma_u}$ is nonsingular. Olive (2017b: ∮ 5.3.3, 2018) proved that the prediction region method gives a large sample confidence region under the much stronger conditions of v) and $\boldsymbol{u} \sim N_g(\mathbf{0}, \boldsymbol{\Sigma_u})$, but the above Pelawa Watagoda and Olive (2019a) proof is simpler.

Remark 4.13. Note that if $\sqrt{n}(T_n - \theta) \xrightarrow{D} U$ and $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} U$ where U has a unimodal probability density function symmetric about zero, then the confidence intervals from the three confidence regions (4.37)–(4.39), the shorth confidence interval (4.32), and the "usual" percentile method confi-

dence interval (4.31) are asymptotically equivalent (use the central proportion of the bootstrap sample, asymptotically).

Assume $n\boldsymbol{S}_T^* \xrightarrow{P} \boldsymbol{\Sigma}_A$ as $n, B \to \infty$ where $\boldsymbol{\Sigma}_A$ and $\boldsymbol{S}_T^*$ are nonsingular $g \times g$ matrices, and T_n is an estimator of $\boldsymbol{\theta}$ such that

$$\sqrt{n}\ \ (T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u} \tag{4.41}$$

as $n \to \infty$. Then

$$\sqrt{n}\ \boldsymbol{\Sigma}_A^{-1/2}\ (T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{\Sigma}_A^{-1/2}\boldsymbol{u} = \boldsymbol{z},$$

$$n\ (T_n - \boldsymbol{\theta})^T\ \hat{\boldsymbol{\Sigma}}_A^{-1}\ (T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{z}^T\boldsymbol{z} = D^2$$

as $n \to \infty$ where $\hat{\boldsymbol{\Sigma}}_A$ is a consistent estimator of $\boldsymbol{\Sigma}_A$, and

$$(T_n - \boldsymbol{\theta})^T\ [\boldsymbol{S}_T^*]^{-1}\ (T_n - \boldsymbol{\theta}) \xrightarrow{D} D^2 \tag{4.42}$$

as $n, B \to \infty$. Assume the cumulative distribution function of D^2 is continuous and increasing in a neighborhood of $D^2_{1-\delta}$ where $P(D^2 \le D^2_{1-\delta}) = 1-\delta$. If the distribution of D^2 is known, then we could use the large sample confidence region (4.34) $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \le D^2_{1-\delta}\}$. Often by a central limit theorem or the multivariate delta method, $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{\Sigma}_A)$, and $D^2 \sim \chi^2_g$. Note that $[\boldsymbol{S}_T^*]^{-1}$ could be replaced by $n\hat{\boldsymbol{\Sigma}}_A^{-1}$.

Remark 4.14. Under reasonable conditions, i) $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, ii) $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} \boldsymbol{u}$, iii) $\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$, and iv) $\sqrt{n}(T_i^* - \overline{T}^*) \xrightarrow{D} \boldsymbol{u}$. Suppose $(n\boldsymbol{S}_T^*)^{-1}$ is "not too ill conditioned." Then

$$D_1^2 = D^2_{T_i^*}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - \overline{T}^*)^T(n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(T_i^* - \overline{T}^*),$$

$$D_2^2 = D^2_{\boldsymbol{\theta}}(T_n, \boldsymbol{S}_T^*) = \sqrt{n}(T_n - \boldsymbol{\theta})^T(n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(T_n - \boldsymbol{\theta}),$$

$$D_3^2 = D^2_{\boldsymbol{\theta}}(\overline{T}^*, \boldsymbol{S}_T^*) = \sqrt{n}(\overline{T}^* - \boldsymbol{\theta})^T(n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(\overline{T}^* - \boldsymbol{\theta}), \quad \text{and}$$

$$D_4^2 = D^2_{T_i^*}(T_n, \boldsymbol{S}_T^*) = \sqrt{n}(T_i^* - T_n)^T(n\boldsymbol{S}_T^*)^{-1}\sqrt{n}(T_i^* - T_n),$$

are well behaved. If $(n\boldsymbol{S}_T^*)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_T^{-1}$, then $D_j^2 \xrightarrow{D} D^2 = \boldsymbol{u}^T\boldsymbol{\Sigma}_T^{-1}\boldsymbol{u}$. If $(n\boldsymbol{S}_T^*)^{-1}$ is "not too ill conditioned" then $D_j^2 \approx \boldsymbol{u}^T(n\boldsymbol{S}_T^*)^{-1}\boldsymbol{u}$ for large n, and the confidence regions (4.37), (4.39), and (4.39) will have coverage near $1 - \delta$. The regularity conditions for (4.37)–(4.39) are weaker when $g = 1$, since $\boldsymbol{S}_T^*$ does not need to be computed.

The following Pelawa Watagoda and Olive (2019a) theorem is very useful. Let $D^2_{(U_B)}$ be the cutoff for the nonparametric prediction region (4.34) com-

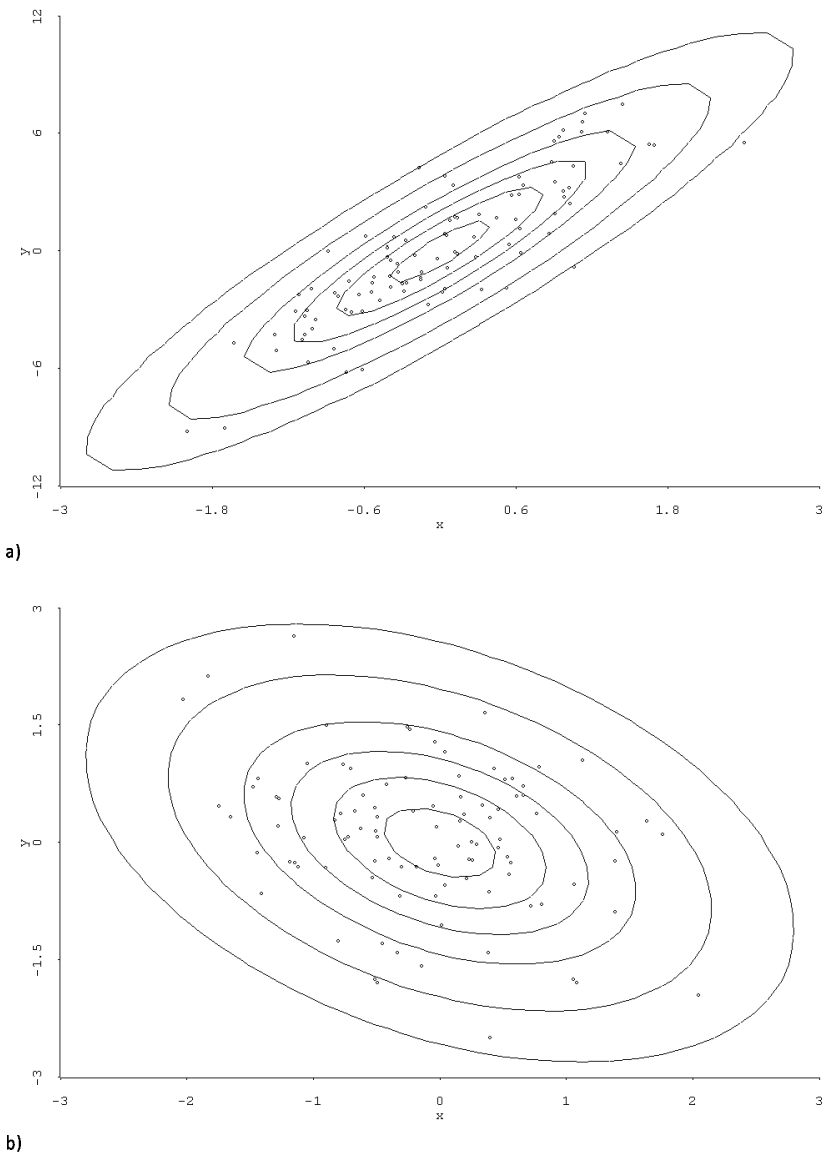

Fig. 4.3 Confidence Regions for 2 Statistics with MVN Distributions

puted from the $D_i^2(\overline{T}, \boldsymbol{S}_T)$ for $i = 1, ..., B$. Hence n is replaced by B. Since T_n depends on the sample size n, we need $(n\boldsymbol{S}_T)^{-1}$ to be fairly well behaved ("not too ill conditioned") for each $n \geq 20g$, say. This condition is weaker than $(n\boldsymbol{S}_T)^{-1} \xrightarrow{P} \boldsymbol{\Sigma}_A^{-1}$. Note that $T_i = T_{in}$.

Theorem 4.32: Geometric Argument. Suppose $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ with $E(\boldsymbol{u}) = \boldsymbol{0}$ and $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma_u}$. Assume $T_1, ..., T_B$ are iid with nonsingular covariance matrix $\boldsymbol{\Sigma}_{T_n}$. Then the large sample $100(1-\delta)\%$ prediction region $R_p = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}, \boldsymbol{S}_T) \leq D^2_{(U_B)}\}$ centered at $\overline{T}$ contains a future value of the statistic T_f with probability $1 - \delta_B \to 1 - \delta$ as $B \to \infty$. Hence the region $R_c = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T) \leq D^2_{(U_B)}\}$ is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ where T_n is a randomly selected T_i.

Proof. The region R_c centered at a randomly selected T_n contains $\overline{T}$ with probability $1 - \delta_B$ which is eventually bounded below by $1 - \delta$ as $B \to \infty$. Since the $\sqrt{n}(T_i - \boldsymbol{\theta})$ are iid,

$$\begin{bmatrix} \sqrt{n}(T_1 - \boldsymbol{\theta}) \\ \vdots \\ \sqrt{n}(T_B - \boldsymbol{\theta}) \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_B \end{bmatrix}$$

where the $\boldsymbol{v}_i$ are iid with the same distribution as $\boldsymbol{u}$. (Use Theorems 4.22 and 4.23, and see Example 4.12.) For fixed B, the average of these random vectors is

$$\sqrt{n}(\overline{T} - \boldsymbol{\theta}) \xrightarrow{D} \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{v}_i \sim AN_g\left(\boldsymbol{0}, \frac{\boldsymbol{\Sigma u}}{B}\right)$$

by Theorem 4.25. Hence $(\overline{T} - \boldsymbol{\theta}) = O_P((nB)^{-1/2})$, and $\overline{T}$ gets arbitrarily close to $\boldsymbol{\theta}$ compared to T_n as $B \to \infty$. Thus R_c is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ as $n, B \to \infty$. $\square$

Examining the iid data cloud $T_1, ..., T_B$ and the bootstrap sample data cloud $T_1^*, ..., T_B^*$ is often useful for understanding the bootstrap. If $\sqrt{n}(T_n - \boldsymbol{\theta})$ and $\sqrt{n}(T_i^* - T_n)$ both converge in distribution to $\boldsymbol{u}$, then the bootstrap sample data cloud of $T_1^*, ..., T_B^*$ is like the data cloud of iid $T_1, ..., T_B$ shifted to be centered at T_n. The nonparametric confidence region (4.37) applies the prediction regon to the bootstrap. Then the hybrid region (4.39) centers that region at T_n. Hence (4.39) is a confidence region by the geometric argument, and (4.37) is a confidence region if $\sqrt{n}(\overline{T}^* - T_n) \xrightarrow{P} \boldsymbol{0}$. Since the T_i^* are closer to $\overline{T}^*$ than T_n on average, $D^2_{(U_B,T)}$ tends to be greater than $D^2_{(U_B)}$. Hence the coverage and volume of (4.38) tend to be at least as large as the coverage and volume of (4.39).

The hyperellipsoid corresponding to the squared Mahalanobis distance $D^2(T_n, \boldsymbol{C})$ is centered at T_n, while the hyperellipsoid corresponding to the squared Mahalanobis distance $D^2(\overline{T}, \boldsymbol{C})$ is centered at $\overline{T}$. Note that

$D^2_{\overline{T}}(T_n, \boldsymbol{C}) = (\overline{T} - T_n)^T \boldsymbol{C}^{-1}(\overline{T} - T_n) = (T_n - \overline{T})^T \boldsymbol{C}^{-1}(T_n - \overline{T}) = D^2_{T_n}(\overline{T}, \boldsymbol{C})$. Thus $D^2_{\overline{T}}(T_n, \boldsymbol{C}) \le D^2_{(U_B)}$ iff $D^2_{T_n}(\overline{T}, \boldsymbol{C}) \le D^2_{(U_B)}$.

The prediction region method will often simulate well even if B is rather small. If the ellipses are centered at T_n or $\overline{T}^*$, Figure 4.3 shows confidence regions if the plotted points are $T_1^*, ..., T_B^*$ where the T_i^* are approximately multivariate normal. If the ellipses are centered at $\overline{T}$, Figure 4.3 shows 10%, 30%, 50%, 70%, 90%, and 98% prediction regions for a future value of T_f for two multivariate normal statistics. Then the plotted points are iid $T_1, ..., T_B$. If $nCov(T) \xrightarrow{P} \boldsymbol{\Sigma}_A$, and the T_i^* are iid from the bootstrap distribution, then $Cov(\overline{T}^*) \approx Cov(T)/B \approx \boldsymbol{\Sigma}_A/(nB)$. By Theorem 4.32, if $\overline{T}^*$ is in the 90% prediction region with probability near 90%, then the confidence region should give simulated coverage near 90% and the volume of the confidence region should be near that of the 90% prediction region. If $B = 100$, then $\overline{T}^*$ falls in a covering region of the same shape as the prediction region, but centered near T_n and the lengths of the axes are divided by $\sqrt{B}$. Hence if $B = 100$, then the axes lengths of this covering region are about one tenth of those in Figure 4.3. Hence when T_n falls within the 70% prediction region, the probability that $\overline{T}^*$ falls in the 90% prediction region is near one. If T_n is just within or just without the boundary of the 90% prediction region, $\overline{T}^*$ tends to be just within or just without of the 90% prediction region. Hence the coverage and volume of prediction region confidence region is near that of the nominal coverage 90% and near the volume of the 90% prediction region.

Hence B does not need to be large provided that n and B are large enough so that $S_T^* \approx \mathrm{Cov}(T^*) \approx \boldsymbol{\Sigma}_A/n$. If n is large, the sample covariance matrix starts to be a good estimator of the population covariance matrix when $B \ge Jg$ where $J = 20$ or 50. For small g, using $B = 1000$ often led to good simulations, but $B = \max(50g, 100)$ may work well.

Remark 4.15. Remark 4.11 suggests that even if the statistic T_n is asymptotically normal so the Mahalanobis distances are asymptotically χ^2_g, the prediction region method can give better results for moderate n by using the cutoff $D^2_{(U_B)}$ instead of the cutoff $\chi^2_{g,1-\delta}$. Theorem 4.32 says that the hyperellipsoidal prediction and confidence regions have exactly the same volume. We compensate for the prediction region undercoverage when n is moderate by using $D^2_{(U_n)}$. If n is large, by using $D^2_{(U_B)}$, the prediction region method confidence region compensates for undercoverage when B is moderate, say $B \ge Jg$ where $J = 20$ or 50. See Remark 4.12. This result can be useful if a simulation with $B = 1000$ or $B = 10000$ is much slower than a simulation with $B = Jg$. The price to pay is that the prediction region method confidence region is inflated to have better coverage, so the power of the hypothesis test is decreased if moderate B is used instead of larger B.

4.8 Bootstrapping Variable Selection

This section considers bootstrapping some survival regression models after variable selection, with emphasis on Cox PH regression. This section will explain why the bootstrap confidence regions (4.37), (4.38), and (4.39) give useful results. Much of the theory in Section 4.7.3 does not apply to the variable selection estimator $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ with $\boldsymbol{\theta} = \boldsymbol{A}\boldsymbol{\beta}$, because T_n is not smooth since T_n is equal to the estimator T_{jn} with probability π_{jn} for $j = 1, ..., J$. Here $\boldsymbol{A}$ is a known full rank $g \times p$ matrix with $1 \le g \le p$. We have $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{v}$ by (4.15) in Theorem 4.31 where $E(\boldsymbol{v}) = \boldsymbol{0}$, and $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \sum_j \sigma^2 \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T$. Hence the geometric argument Theorem 4.32 holds: applying the prediction region (4.29) to an iid sample $T_1, ..., T_B$ and then centering the region at T_n gives a large sample confidence region for $\boldsymbol{\theta}$. Hence if $n\boldsymbol{S}_T^*$ is "not too ill conditioned," there exists a cutoff $\hat{D}^2_{1-\delta}$ such that $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \le \hat{D}^2_{1-\delta}\}$ has coverage close to or higher than $1 - \delta$. See Remark 4.14.

We will denote the ith case by $(Z_i, \delta_i, \boldsymbol{x}_i)$ where $Z_i = Y_i$ if $\delta_i = 1$ so that the survival time is uncensored, and $Z_i = Y_i^*$ if $\delta_i = 0$ so that the survival time is right censored. In R, "time" is often used for the vector of Z_i and "status" for the vector of δ_i. Sometimes $T_i = Z_i$ is used for a possibly censored survival time, but in this chapter $T = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a test statistic.

Suppose the regression model satisfies $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T\boldsymbol{\beta}$, that Equation (2.4) holds, and that if $S \subseteq I_j$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{V}_j)$. Also assume that a variable selection criterion, such as AIC or relaxed lasso, is used such that $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Hence

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V}_{j,0}) \tag{4.43}$$

where $\boldsymbol{V}_{j,0}$ adds columns and rows of zeros corresponding to the x_i not in I_j. Hence $\boldsymbol{V}_{j,0}$ is singular unless I_j corresponds to the full model.

For variable selection with $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, let $T_n = T_{kn} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the π_k with $S \subseteq I_k$ by π_j. The other $\pi_k = 0$. Then Theorem 4.31 holds: $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_{min},0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}$.

Note that $\boldsymbol{V}_{j,0}$ is singular unless I_j corresponds to the full model. For example, if $p = 3$ and model I_j uses a constant $x_1 \equiv 1$ and x_3 with

$$\boldsymbol{V}_j = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}, \quad \text{then} \quad \boldsymbol{V}_{\mathrm{j},0} = \begin{bmatrix} V_{11} & 0 & V_{12} \\ 0 & 0 & 0 \\ V_{21} & 0 & V_{22} \end{bmatrix}.$$

For variable selection, this section will show that the bootstrap sample data cloud $T_1^*, ..., T_B^*$ tends to be slightly more variable than the data cloud of iid $T_1, ..., T_B$ for large n. This result will hold for the parametric bootstrap and

nonparametric bootstrap, which are discussed in the next two subsections. Hence by the geometric argument, we expect $D^2_{(U_B)}$ or $D^2_{(U_B,T)}$ can be used as $\hat{D}^2_{1-\delta}$.

4.8.1 The Parametric Bootstrap

Suppose $Y_i|\boldsymbol{x}_i \sim D(\boldsymbol{x}_i^T\boldsymbol{\beta},\boldsymbol{\gamma})$, $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0},\boldsymbol{V}(\boldsymbol{\beta}))$, 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 maximum likelihood estimator (MLE) $\hat{\boldsymbol{\beta}}$ is used. Then $\boldsymbol{V}(\boldsymbol{\beta}) = \boldsymbol{I}^{-1}(\boldsymbol{\beta})$, the inverse Fisher information matrix. If $\boldsymbol{I}_n(\boldsymbol{\beta})$ is the Fisher information matrix based on a sample of size n, then $\boldsymbol{I}_n(\boldsymbol{\beta})/n \xrightarrow{P} \boldsymbol{I}(\boldsymbol{\beta})$. 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 $\boldsymbol{x}_i^T$.

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

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^*-\hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\mathbf{0},\boldsymbol{V}(\boldsymbol{\beta})) \tag{4.44}$$

as $n\to\infty$.

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 $\hat{\boldsymbol{\beta}}(O)\neq\mathbf{0}$. However, under regularity conditions, $E(\hat{\boldsymbol{\beta}}_I^*)\approx\hat{\boldsymbol{\beta}}_I$ and $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_I^*)-\mathrm{Cov}(\hat{\boldsymbol{\beta}}_I)\to\mathbf{0}$ as $n,B\to\infty$.

The parametric bootstrap should be useful for bootstrapping parametric survival regression models such as the Weibull PH regression model or the Weibull AFT.

4.8.2 The Nonparametric Bootstrap

Suppose a statistic T_n is computed from a data set of n cases. The *nonparametric bootstrap* draws n cases with replacement from that data set. Then T_1^* is the statistic T_n computed from the sample. This process is repeated B times to produce the bootstrap sample $T_1^*, ..., T_B^*$. This procedure is also called the *empirical bootstrap* or *naive bootstrap*.

Under regularity conditions, $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\mathbf{0}, \sigma^2 \boldsymbol{W}) \sim N_p(\mathbf{0}, \boldsymbol{V})$. Hence if $S \subseteq I_j$,

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

as $n, B \to \infty$. (Treat I_j as if I_j is the full model.)

One set of regularity conditions is that the survival regression model holds for $Y_i|\boldsymbol{x}_i$, the $\boldsymbol{x}_i$ are iid from some population with a nonsingular covariance matrix, the cases are independent, and the survival times are right censored. The cases $(Z_i, \delta_i, \boldsymbol{x}_i)$ are sampled with replacement. This method can be useful with proportional hazards regression models. See Burr (1994), Efron and Tibshirani (1986), and Shao, and Tu (1995).

4.8.3 Bootstrapping Variable Selection

Let the $g \times 1$ vector T_n be an estimator of the $g \times 1$ parameter vector $\boldsymbol{\theta}$. 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, consider testing $H_0 : \boldsymbol{A\beta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{A\beta} \neq \boldsymbol{\theta}_0$ with $\boldsymbol{\theta} = \boldsymbol{A\beta}$ where often $\boldsymbol{\theta}_0 = \mathbf{0}$. Then let $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ and let $T_i^* = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{I_{min},0,i}$ for $i = 1, ..., B$.

The explanation for why the bootstrap confidence regions (4.37), (4.38), and (4.39) give useful results after variable selection is due to Rathnayake and Olive (2019). Let the variable selection estimator $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ with $\boldsymbol{\theta} = \boldsymbol{A\beta}$. Then T_n is not smooth since T_n is equal to the estimator T_{jn} with probability π_{jn} for $j = 1, ..., J$. Here $\boldsymbol{A}$ is a known full rank $g \times p$ matrix with $1 \leq g \leq p$. We have $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{v}$ by (4.15) where $E(\boldsymbol{v}) = \mathbf{0}$, and $\boldsymbol{\Sigma_v} = \sum_j \pi_j \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T$. Hence the geometric argument Theorem 4.32 holds: if we had iid data $T_1, ..., T_B$, then R_c would be a large sample confidence region for $\boldsymbol{\theta}$. For variable selection, this section will show that the bootstrap sample data cloud $T_1^*, ..., T_B^*$ tends to be slightly more variable than the data cloud of iid $T_1, ..., T_B$ for large n. Empirically, for a mixture distribution, the bagging estimator $\overline{T}^*$ tends to estimate $\boldsymbol{\theta}$ at least as well as T_n. See Breiman (1996) and Yang (2003).

The full model should be checked before doing variable selection inference. Assume p is fixed and $n \geq 20p$. See Chapter 3 and 4. For the bootstrap, suppose that T_i^* is equal to T_{ij}^* with probability ρ_{jn} for $j = 1, ..., J$ where

$\sum_j \rho_{jn} = 1$, and $\rho_{jn} \to \pi_j$ as $n \to \infty$. 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

$$T_{1,1}^*, ..., T_{B_{1n},1}^*, ..., T_{1,J}^*, ..., T_{B_{Jn},J}^*$$

where the B_{jn} follow a multinomial distribution and $B_{jn}/B \xrightarrow{P} \rho_{jn}$ as $B \to \infty$. 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 $\boldsymbol{S}_{T,j}^*$. Then

$$\overline{T}^* = \frac{1}{B}\sum_{i=1}^{B} T_i^* = \sum_j \frac{B_{jn}}{B}\frac{1}{B_{jn}}\sum_{i=1}^{B_{jn}} T_{ij}^* = \sum_j \hat{\rho}_{jn}\overline{T}_j^*.$$

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 $\boldsymbol{S}_{T,j}$.

Suppose the jth component of an iid sample $T_1, ..., T_B$ and the jth component of the bootstrap sample $T_1^*, ..., T_B^*$ have the same variability asymptotically. Since $E(T_{jn}) \approx \boldsymbol{\theta}$, each component of the iid sample is approximately centered at $\boldsymbol{\theta}$. The bootstrap components are centered at $E(T_{jn}^*)$, and often $E(T_{jn}^*) = T_{jn}$. Geometrically, separating the component clouds so that they are no longer centered at one value makes the overall data cloud larger. Thus the variability of T_n^* is larger than that of T_n for a mixture distribution, asymptotically. Hence the prediction region applied to the bootstrap sample is slightly larger than the prediction region applied to the iid sample, asymptotically (we want $n \geq 20p$). Hence cutoff $\hat{D}_{1,1-\delta}^2 = D_{(U_B)}^2$ gives coverage close to or higher than the nominal coverage for confidence regions (4.37) and (4.38), using the geometric argument. The deviation $T_i^* - T_n$ tends to be larger in magnitude than the deviation and $T_i^* - \overline{T}^*$. Hence the cutoff $\hat{D}_{2,1-\delta}^2 = D_{(U_B,T)}^2$ tends to be larger than $D_{(U_B)}^2$, and region (4.38) tends to have higher coverage than region (4.39) for a mixture distribution.

Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and that $S \subseteq I_j$. The components of the iid sample are centered at $\boldsymbol{A\beta}$ while the components of the bootstrap sample are centered at $\boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}$. Consider regression models with $Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{x}^T\boldsymbol{\beta}$. Assume $\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{\Sigma}_j)$ where $\boldsymbol{\Sigma}_j = \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T$. For the nonparametric bootstrap, assume $\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}^* - \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}) \xrightarrow{D} N_{a_j}(\boldsymbol{0}, \boldsymbol{\Sigma}_j)$. Then the components of the iid sample and bootstrap sample have the same variability asymptotically. The components of iid sample are centered at $\boldsymbol{A\beta}$ while the components of the bootstrap sample are centered at $\boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}$. For the nonparametric bootstrap, the above results tend to hold if $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V})$ and if $\sqrt{n}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V})$. Assumptions for the nonparametric bootstrap tend to be rather strong: often one assumption is that the n cases are iid from some population.

For the parametric bootstrap, Section 4.8.1 noted that under regularity conditions, $\text{Cov}(\hat{\boldsymbol{\beta}}_I^*) - \text{Cov}(\hat{\boldsymbol{\beta}}_I) \to \mathbf{0}$ as $n, B \to \infty$ if $S \subseteq I$. Hence $\text{Cov}(T_{jn}) - \text{Cov}(T_{jn}^*) \to \mathbf{0}$ as $n, B \to \infty$ if $S \subseteq I$. Here $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$, $T_{jn} = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_j,0}$, $T_n^* = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{I_{min},0}$, and $T_{jn}^* = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{I_j,0}$. Then $E(T_{jn}) \approx \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\theta}$ while the $E(T_{jn}^*)$ are more variable than the $E(T_{jn})$ with $E(T_{jn}^*) \approx \boldsymbol{A}\hat{\boldsymbol{\beta}}(I_j, 0)$, roughly, where $\hat{\boldsymbol{\beta}}(I_j, 0)$ is formed from $\hat{\boldsymbol{\beta}}(I_j)$ by adding zeros corresponding to variables not in I_j. Hence the jth component of an iid sample $T_1, ..., T_B$ and the jth component of the bootstrap sample $T_1^*, ..., T_B^*$ have the same variability asymptotically.

In simulations for $n \geq 20p$ for $H_0 : \boldsymbol{A}\boldsymbol{\beta}_S = \boldsymbol{\theta}_0$, the coverage tended to get close to $1 - \delta$ for $B \geq \max(200, 50p)$ so that $\boldsymbol{S}_T^*$ is a good estimator of $\text{Cov}(T^*)$. In the simulations where S is not the full model, inference with the submodel I_{min} was often more precise than inference with the full model if $n \geq 20p$ and $B \geq 50p$. It is possible that $\boldsymbol{S}_T^*$ is singular if a column of the bootstrap sample is equal to $\mathbf{0}$. If the regression model has a $q \times 1$ vector of parameters $\boldsymbol{\gamma}$, we may need to replace p by $p + q$.

Undercoverage can occur if bootstrap sample data cloud is less variable than the iid data cloud, e.g., if $(n-p)/n$ is not close to one. Coverage can be higher than the nominal coverage for two reasons: i) the bootstrap data cloud is more variable than the iid data cloud of $T_1, ..., T_B$, and ii) zero padding.

To see the effect of zero padding, consider $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\beta}_O = \mathbf{0}$ where $\boldsymbol{\beta}_O = (\beta_{i_1}, ..., \beta_{i_g})^T$ and $O \subseteq E$ in Equation (2.4) so that H_0 is true. Suppose a nominal 95% confidence region is used and U_B is the 96th percentile. Hence the confidence region (4.37) or (4.38) covers at least 96% of the bootstrap sample. If $\hat{\boldsymbol{\beta}}_{O,j}^* = \mathbf{0}$ for more than 4% of the $\hat{\boldsymbol{\beta}}_{O,1}^*, ..., \hat{\boldsymbol{\beta}}_{O,B}^*$, then $\mathbf{0}$ is in the confidence region and the bootstrap test fails to reject H_0. If this occurs for each run in the simulation, then the observed coverage will be 100%.

Now suppose $\hat{\boldsymbol{\beta}}_{O,j}^* = \mathbf{0}$ for $j = 1, ..., B$. Then $\boldsymbol{S}_T^*$ is singular, but the singleton set $\{\mathbf{0}\}$ is the large sample $100(1-\delta)$% confidence region (4.37), (4.38), or (4.39) for $\boldsymbol{\beta}_O$ and $\delta \in (0, 1)$, and the pvalue for $H_0 : \boldsymbol{\beta}_O = \mathbf{0}$ is one. (This result holds since $\{\mathbf{0}\}$ contains 100% of the $\hat{\boldsymbol{\beta}}_{O,j}^*$ in the bootstrap sample.) For large sample theory tests, the pvalue estimates the population pvalue. Let I denote the other predictors in the model so $\boldsymbol{\beta} = (\boldsymbol{\beta}_I^T, \boldsymbol{\beta}_O^T)^T$. For the I_{min} model from variable selection, there may be strong evidence that $\boldsymbol{x}_O$ is not needed in the model given $\boldsymbol{x}_I$ is in the model if the "100%" confidence region is $\{\mathbf{0}\}$, $n \geq 20p$, and $B \geq 50p$. (Since the pvalue is one, this technique may be useful for data snooping: applying MLE theory to submodel I may have negligible selection bias.)

Remark 4.16. Note that there are several important variable selection models, including the model given by Equation (2.4) where $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$. Another model is $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_{S_i}^T\boldsymbol{\beta}_{S_i}$ for $i = 1, ..., K$. Then there are $K \geq 2$ competing "true" nonnested submodels where $\boldsymbol{\beta}_{S_i}$ is $a_{S_i} \times 1$. For example, suppose the $K = 2$ models have predictors x_1, x_2, x_3 for S_1 and x_1, x_2, x_4 for

S_2. Then x_3 and x_4 are likely to be selected and omitted often by forward selection for the B bootstrap samples. Hence omitting all predictors x_i that have a $\beta_{ij}^* = 0$ for at least one of the bootstrap samples $j = 1, ..., B$ could result in underfitting, e.g. using just x_1 and x_2 in the above $K = 2$ example. If n and B are large enough, the singleton set $\{\mathbf{0}\}$ could still be the "100%" confidence region for a vector $\boldsymbol{\beta}_O$.

Suppose the predictors x_i have been standardized. Then another important regression model has the β_i taper off rapidly, but no coefficients are equal to zero. For example, $\beta_i = e^{-i}$ for $i = 1, ..., p$.

Another way to look at the bootstrap confidence region for variable selection estimators is to consider the estimator $T_{2,n}$ that chooses I_j with probability equal to the observed bootstrap proportion $\hat{\rho}_{jn}$. The bootstrap sample $T_1^*, ..., T_B^*$ tends to be slightly more variable than an iid sample $T_{2,1}, ..., T_{2,B}$, and the geometric argument suggests that the large sample coverage of the nominal $100(1-\delta)\%$ confidence region will be at least as large as the nominal coverage $100(1-\delta)\%$.

4.8.4 Simulations

For variable selection with the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I_{min},0}$, consider testing $H_0 : \boldsymbol{A\beta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{A\beta} \neq \boldsymbol{\theta}_0$ with $\boldsymbol{\theta} = \boldsymbol{A\beta}$ where often $\boldsymbol{\theta}_0 = \mathbf{0}$. Then let $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ and let $T_i^* = \boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{I_{min},0,i}$ for $i = 1, ..., B$. The shorth estimator can be applied to a bootstrap sample $\hat{\beta}_{i1}^*, ..., \hat{\beta}_{iB}^*$ to get a confidence interval for β_i. Here $T_n = \hat{\beta}_i$ and $\theta = \beta_i$.

Next, we describe a small simulation study that was done using $B = \max(1000, n/25, 50p)$ and 5000 runs. The simulation used $p = 4, 6, 7, 8$, and 10; $n = 25p$ and $50p$; $\psi = 0, 1/\sqrt{p}$, and 0.9; and $k = 1$ and $p - 2$ where k and ψ are defined in the following paragraph. In the simulations, we use $\theta = \boldsymbol{A\beta} = \beta_i$, $\boldsymbol{\theta} = \boldsymbol{A\beta} = \boldsymbol{\beta}_S = \mathbf{1}$ and $\boldsymbol{\theta} = \boldsymbol{A\beta} = \boldsymbol{\beta}_E = \mathbf{0}$.

In the simulations, for $i = 1, ..., n$, we generated $\boldsymbol{w}_i \sim N_p(\mathbf{0}, \boldsymbol{I})$ where the p elements of the vector $\boldsymbol{w}_i$ are iid N(0,1). Let the $p \times p$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ii} = 1$ and $a_{ij} = \psi$ where $0 \leq \psi < 1$ for $i \neq j$. Then the vector $\boldsymbol{z}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $Cov(\boldsymbol{z}_i) = \boldsymbol{\Sigma_z} = \boldsymbol{AA}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = [1 + (p-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = [2\psi + (p-2)\psi^2]$. Then $\sum_{j=1}^k z_j \sim N(0, k\sigma_{ii} + k(k-1)\sigma_{ij}) = N(0, v^2)$. Let $\boldsymbol{x} = a\boldsymbol{z}/v$. Hence the correlations are $Cor(x_i, x_j) = \rho = (2\psi + (p-2)\psi^2)/(1 + (p-1)\psi^2)$ for $i \neq j$. If $\psi = 1/\sqrt{cp}$, then $\rho \to 1/(c+1)$ as $p \to \infty$ where $c > 0$. As ψ gets close to 1, the predictor vectors cluster about the line in the direction of $(1, ..., 1)^T$. Let $SP = \boldsymbol{x}_i^T\boldsymbol{\beta} = 1x_{i,1} + \cdots + 1x_{i,k} \sim N(0, a^2)$ for $i = 1, ..., n$. The simulations use $a = 1$ where $\boldsymbol{\beta} = (1, ..., 1, 0, ..., 0)^T$ with k ones and $p - k$ zeros.

Then the Y_i were generated such that $Y|\boldsymbol{x}$ follows a Weibull proportional hazards regression model with the above $\boldsymbol{\beta}$. We used R code similar to that of Zhou (2001) to obtain $(Z_i, \delta_i, \boldsymbol{x}_i)$ where some of the Z_i are right censored. Some functions from *survpack* were useful. The function `phdata2` generates a data set as described above. We used the nonparametric bootstrap and the Cox PH model. The function `PHboot` bootstraps the full Cox PH model. The function `PHbootsim` is used to simulate the bootstrap for the full Cox PH model. The functions `LPHboot` and `RLPHboot` bootstraps a Cox PH model with lasso and relaxed lasso. The function `RLPHbootsim` is used to simulate the bootstrap for relaxed lasso. The `shorth3` function computes the shorth(c) intervals with the Frey (2013) correction used when $g = 1$. Some R code is shown below.

```
library(survival)
library(MASS)
library(glmnet)

out<- phdata2(n=100,p=4,k=1,psi=0,a=1,gam=1,clam = 0.1)
out$beta
$betaP
[1] 1 0 0 0
#out$x gives the matrix of predictors
out$time
$time
  [1] 10.5015  2.5748  2.1266  0.4238  0.4454
  [6]  0.1165  0.0233  0.3108  0.0856  0.3908
 .
 .
 .
 [96]  5.4669  0.1603  0.1510  0.1206  0.6356
out$status
$status  #0 means right censored
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 0
      0 1 0 0 1 1 1 0 0 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1
      1 1 1 1 1 0 1 0 1 1 1 0 0 1 1 0 1 1 1 1 1 1 1 1
      1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
      0 0 1 1

RLPHbootsim(nruns=100,B=200,k=2) #slow 3 runs per minute
$mndd
[1] 3.01  #relaxed lasso used 3 predictors on average
$cicov
 [1] 0.94 0.96 0.97 0.99 0.95 0.97 0.97 0.93 0.95 0.95
$avelen
 [1] 0.8642748 0.8473142 0.7334978 0.7219106 2.5561583
     2.5561583 2.6622667 2.5124382 2.5124382 2.6253967
```

```
$beta
[1] 1 1 0 0
$k
[1] 2
PHbootsim(nruns=100,B=200,k=2) #fairly fast
$cicov
 [1] 0.96 0.95 0.92 0.92 0.91 0.94 0.94 0.95 0.99 0.99
$avelen
 [1] 0.8571470 0.8582906 0.7541797 0.7416362 2.5247451
     2.5247451 2.5558537 2.5021201 2.5021201 2.6243971
$beta
[1] 1 1 0 0
$k
[1] 2
```

The simulation computed the Frey shorth(c) interval for each β_i and used bootstrap confidence regions to test $H_0 : \boldsymbol{\beta}_S = \mathbf{1}$ (whether first k $\beta_i = 1$) and $H_0 : \boldsymbol{\beta}_E = \mathbf{0}$ (whether the last $p - k$ $\beta_i = 0$). The nominal coverage was 0.95 with $\delta = 0.05$. Observed coverage between 0.94 and 0.96 suggests coverage is close to the nominal value. The number of runs = 100 is tiny since the relaxed lasso simulation is slow. Using 5000 runs would be much better.

The regression models used the nonparamtric bootstrap on the relaxed lasso estimator $\hat{\boldsymbol{\beta}}_{I_{min},0}$. Table 4.1 gives results with $n = 100$, $p = 4$, and $k = 1$. Table 4.1 shows two rows for each model giving the observed confidence interval coverages and average lengths of the confidence intervals. The term "reg" is for the full model regression, and the term "vs" is for variable selection with relaxed lasso. The last six columns give results for the tests. The terms pr, hyb, and br are for the prediction region method (4.37), hybrid region (4.38), and Bickel and Ren region (4.39). The 0 indicates the test was $H_0 : \boldsymbol{\beta}_E = \mathbf{0}$, while the 1 indicates that the test was $H_0 : \boldsymbol{\beta}_S = \mathbf{1}$. The length and coverage = P(fail to reject H_0) for the interval $[0, D_{(U_B)}]$ or $[0, D_{(U_B,T)}]$ where $D_{(U_B)}$ or $D_{(U_B,T)}$ is the cutoff for the confidence region. The cutoff will often be near $\sqrt{\chi^2_{g,0.95}}$ if the statistic T is asymptotically normal. Note that $\sqrt{\chi^2_{2,0.95}} = 2.448$ is close to 2.45 for the full model regression bootstrap tests.

Volume ratios of the three confidence regions can be compared using (4.40), but there is not enough information in Table 4.1 to compare the volume of the confidence region for the full model regression versus that for the relaxed lasso since the two methods have different determinants $|\boldsymbol{S}^*_T|$. Table 4.1 corresponds to the above R output with $k = 2$.

The inference for forward selection was often as precise or more precise than the inference for the full model. The coverages were near 0.95 for the regression bootstrap on the full model, although there was slight undercoverage for the tests since $(n - p)/n = 0.96$ when $n = 25p$. Suppose $\psi = 0$.

Table 4.1 Bootstrapping Cox PH Regression With Relaxed Lasso

ψ	β_1	β_2	β_{p-1}	β_p	pr0	hyb0	br0	pr1	hyb1	br1
reg,0	0.96	0.95	0.92	0.92	0.91	0.94	0.94	0.95	0.99	0.99
len	0.857	0.858	0.754	0.742	2.525	2.525	2.556	2.502	2.502	2.624
vs,0	0.94	0.96	0.97	0.99	0.95	0.97	0.97	0.93	0.95	0.95
len	0.864	0.847	0.733	0.722	2.556	2.556	2.662	2.512	2.512	2.625

Then it may be true that $\hat{\boldsymbol{\beta}}_S$ has the same limiting distribution for I_{min} and the full model. Note that the average lengths and coverages were similar for the full model and forward selection I_{min} for β_1, β_2, and $\boldsymbol{\beta}_S = (\beta_1, \beta_2)^T$. Forward selection inference was more precise for $\boldsymbol{\beta}_E = (\beta_3, \beta_4)^T$. The Bickel and Ren (4.38) cutoffs and coverages were at least as high as those of the hybrid region (4.39).

4.9 Data Splitting

Data splitting is used for inference after model selection. Use a training set to select a full model, and a validation set for inference with the selected full model. Here $p >> n$ is possible. See Hurvich and Tsai (1990, p. 216) and Rinaldo et al. (2019). Typically when training and validation sets are used, the training set is bigger than the validation set or half sets are used, often causing large efficiency loss.

Let J be a positive integer and let $\lfloor x \rfloor$ be the integer part of x, e.g., $\lfloor 7.7 \rfloor = 7$. Initially divide the data into two sets H_1 with $n_1 = \lfloor n/(2J) \rfloor$ cases and V_1 with $n - n_1$ cases. If the fitted model from H_1 is not good enough, randomly select n_1 cases from V_1 to add to H_1 to form H_2. Let V_2 have the remaining cases from V_1. Continue in this manner, possibly forming sets $(H_1, V_1), (H_2, V_2), ..., (H_J, V_J)$ where H_i has $n_i = in_1$ cases. Stop when H_d gives a reasonable model I_d with a_d predictors if $d < J$. Use $d = J$, otherwise. Use the model I_d as the full model for inference with the data in V_d.

This procedure is simple for a fixed data set, but it would be good to automate the procedure. For example, if $n = 500000$ and $p = 90$, using $n_1 = 900$ would result in a much smaller loss of efficiency than $n_1 = 250000$.

4.10 Summary

1) A *model for variable selection* can be described by $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$ where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ is a $p \times 1$ vector of predictors, $\boldsymbol{x}_S$ is an $a_S \times 1$

vector, and $\boldsymbol{x}_E$ is a $(p-a_S)\times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \mathbf{0}$. Assume p is fixed while $n \to \infty$.

2) If $\hat{\boldsymbol{\beta}}_I$ is $a\times 1$, form the $p\times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$, then $\hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. If $S \subseteq I$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I) \xrightarrow{D} N_{a_I}(\mathbf{0}, \boldsymbol{V}_I)$.

3) **Theorem 4.31, Variable Selection CLT.** Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $T_n = \hat{\boldsymbol{\beta}}_{I_{min},0}$ and $T_{jn} = \hat{\boldsymbol{\beta}}_{I_j,0}$. Let $T_n = T_{kn} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the π_k with $S \subseteq I_k$ by π_j. The other $\pi_k = 0$ since $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Assume $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\mathbf{0}, \boldsymbol{V}_j)$ and $\boldsymbol{u}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_p(\mathbf{0}, \boldsymbol{V}_{j,0})$. a) Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_{min},0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}$$

where the cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{z}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{z})$. Thus $\boldsymbol{u}$ is a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j, $E(\boldsymbol{u}) = \mathbf{0}$, and $\text{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \sum_j \pi_j \boldsymbol{V}_{j,0}$.

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

$$\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} = \boldsymbol{v}$$

where $\boldsymbol{A}\boldsymbol{u}$ has a mixture distribution of the $\boldsymbol{A}\boldsymbol{u}_j \sim N_g(\mathbf{0}, \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T)$ with probabilities π_j.

4) For $h > 0$, the hyperellipsoid $\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{z} - T) \le h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \le h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \le h\}$. A future observation (random vector) $\boldsymbol{x}_f$ is in this region if $D_{\boldsymbol{x}_f} \le h$. A large sample $100(1-\delta)\%$ prediction region is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$ where $0 < \delta < 1$. A *large sample* $100(1-\delta)\%$ *confidence region* for a vector of parameters $\boldsymbol{\theta}$ is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\theta} \in \mathcal{A}_n)$ is eventually bounded below by $1 - \delta$ as $n \to \infty$.

5) Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + p/n)$ for $\delta > 0.1$ and $q_n = \min(1-\delta/2, 1-\delta+10\delta p/n)$, otherwise. If $q_n < 1-\delta+0.001$, set $q_n = 1-\delta$. If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$, then $\{\boldsymbol{z} : D_{\boldsymbol{z}}(T, \boldsymbol{C}) \le h\}$ is a large sample $100(1-\delta)\%$ prediction regions if $h = D_{(U_n)}$ where $D_{(U_n)}$ is the $100q_n$th sample quantile of the D_i. The large sample $100(1 - \delta)\%$ nonparametric prediction region $\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \le D^2_{(U_n)}\}$ uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. We want $n \ge 10p$ for good coverage and $n \ge 50p$ for good volume.

6) Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \ne \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a known $g \times 1$ vector. Make a confidence region and reject H_0 if $\boldsymbol{\theta}_0$ is not in the confidence region. Let q_B and U_B be as in 5) with n replaced by B and p replaced by g. Let $\overline{T}^*$ and $\boldsymbol{S}^*_T$ be the sample mean and sample covariance matrix of the bootstrap sample $T^*_1, ..., T^*_B$. a) The prediction region method large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$ is $\{\boldsymbol{w} : (\boldsymbol{w} - \overline{T}^*)^T [\boldsymbol{S}^*_T]^{-1} (\boldsymbol{w} - \overline{T}^*) \le D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}^*, \boldsymbol{S}^*_T) \le D^2_{(U_B)}\}$ where $D^2_{(U_B)}$ is computed from $D^2_i = (T^*_i - \overline{T}^*)^T [\boldsymbol{S}^*_T]^{-1} (T^*_i - \overline{T}^*)$ for $i = 1, ..., B$. Note that the corresponding

test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 if $(\overline{T}^* - \boldsymbol{\theta}_0)^T[\boldsymbol{S}_T^*]^{-1}(\overline{T}^* - \boldsymbol{\theta}_0) > D^2_{(U_B)}$. This procedure applies the nonparametric prediction region to the bootstrap sample. b) The modified Bickel and Ren (2001) large sample $100(1-\delta)\%$ confidence region is $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \le D^2_{(U_B,T)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \le D^2_{(U_B,T)}\}$ where the cutoff $D^2_{(U_B,T)}$ is the $100q_B$th sample quantile of the $D_i^2 = (T_i^* - T_n)^T[\boldsymbol{S}_T^*]^{-1}(T_i^* - T_n)$. c) The hybrid large sample $100(1-\delta)\%$ confidence region: $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T[\boldsymbol{S}_T^*]^{-1}(\boldsymbol{w} - T_n) \le D^2_{(U_B)}\} = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T^*) \le D^2_{(U_B)}\}$.

If $g = 1$, confidence intervals can be computed without $\boldsymbol{S}_T^*$ or D^2 for a), b), and c).

For some data sets, $\boldsymbol{S}_T^*$ may be singular due to one or more columns of zeroes 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 if n and B are large enough. Let $\boldsymbol{\beta}_O = (\beta_{i_1}, ..., \beta_{i_g})^T$, and consider testing $H_0 : \boldsymbol{A}\boldsymbol{\beta}_O = \boldsymbol{0}$. If $\boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{O,i} = \boldsymbol{0}$ for greater than $B\delta$ of the bootstrap samples $i = 1, ..., B$, then fail to reject H_0. (If $\boldsymbol{S}_T^*$ is nonsingular, the $100(1-\delta)\%$ prediction region method confidence region contains $\boldsymbol{0}$.)

7) **Theorem 4.32: Geometric Argument.** Suppose $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{u}$ with $E(\boldsymbol{u}) = \boldsymbol{0}$ and $Cov(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$. Assume $T_1, ..., T_B$ are iid with nonsingular covariance matrix $\boldsymbol{\Sigma}_{T_n}$. Then the large sample $100(1-\delta)\%$ prediction region $R_p = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(\overline{T}, \boldsymbol{S}_T) \le D^2_{(U_B)}\}$ centered at $\overline{T}$ contains a future value of the statistic T_f with probability $1 - \delta_B \to 1 - \delta$ as $B \to \infty$. Hence the region $R_c = \{\boldsymbol{w} : D^2_{\boldsymbol{w}}(T_n, \boldsymbol{S}_T) \le D^2_{(U_B)}\}$ is a large sample $100(1-\delta)\%$ confidence region for $\boldsymbol{\theta}$.

8) Applying the nonparametric prediction region (4.29) to the iid data $T_1, ..., T_B$ results in the $100(1-\delta)\%$ confidence region $\{\boldsymbol{w} : (\boldsymbol{w} - T_n)^T \boldsymbol{S}_T^{-1}(\boldsymbol{w} - T_n) \le D^2_{(U_B)}(T_n, \boldsymbol{S}_T)\}$ where $D^2_{(U_B)}(T_n, \boldsymbol{S}_T)$ is computed from the $(T_i - T_n)^T \boldsymbol{S}_T^{-1}(T_i - T_n)$ provided the $\boldsymbol{S}_T = \boldsymbol{S}_{T_n}$ are "not too ill conditioned." For OLS variable selection, assume there are two or more component clouds. The bootstrap component data clouds have the same asymptotic covariance matrix as the iid component data clouds, which are centered at $\boldsymbol{\theta}$. The jth bootstrap component data cloud is centered at $E(T_{ij}^*)$ and often $E(T_{jn}^*) = T_{jn}$. Confidence region (4.37) is the prediction region (4.29) applied to the bootstrap sample, and (4.37) is slightly larger in volume than (4.29) applied to the iid sample, asymptotically. The hybrid region (4.39) shifts (4.37) to be centered at T_n. Shifting the component clouds slightly and computing (4.29) does not change the axes of the prediction region (4.29) much compared to not shifting the component clouds. Hence by the geometric argument, we expect (4.39) to have coverage at least as high as the nominal, asymptotically, provided the $\boldsymbol{S}_T^*$ are "not too ill conditioned." The Bickel and Ren confidence region (4.38) tends to have higher coverage and volume than (4.39). Since $\overline{T}^*$ tends to be closer to $\boldsymbol{\theta}$ than T_n, (4.37) tends to have good coverage.

9) Suppose m independent large sample $100(1-\delta)\%$ prediction regions are made where $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid from the same distribution for each of the m runs. Let Y count the number of times $\boldsymbol{x}_f$ is in the prediction region. Then $Y \sim$ binomial $(m, 1-\delta_n)$ where $1-\delta_n$ is the true coverage. Simulation can be used to see if the true or actual coverage $1-\delta_n$ is close to the nominal coverage $1-\delta$. A prediction region with $1-\delta_n < 1-\delta$ is liberal and a region with $1-\delta_n > 1-\delta$ is conservative. It is better to be conservative by 3% than liberal by 3%. Parametric prediction regions tend to have large undercoverage and so are too liberal. Similar definitions are used for confidence regions.

10) For the bootstrap, perform variable selection on $\boldsymbol{Y}_i^*$ and $\boldsymbol{X}$ (or $\boldsymbol{X}^*$ for the nonparametric bootstrap), fit the model that minimizes the criterion, and add 0s corresponding to the omitted variables, resulting in estimators $\hat{\boldsymbol{\beta}}_1^*, ..., \hat{\boldsymbol{\beta}}_B^*$ where $\hat{\boldsymbol{\beta}}_i^* = \hat{\boldsymbol{\beta}}_{I_{min},0,i}^*$.

11) Let $Z_1, ..., Z_n$ be random variables, let $Z_{(1)}, ..., Z_{(n)}$ be the order statistics, and let c be a positive integer. Compute $Z_{(c)} - Z_{(1)}, Z_{(c+1)} - Z_{(2)}, ..., Z_{(n)} - Z_{(n-c+1)}$. Let shorth(c) = $[\mathrm{Z}_{(\mathrm{d})}, \mathrm{Z}_{(\mathrm{d+c-1})}]$ correspond to the interval with the shortest length.

The large sample $100(1-\delta)\%$ *shorth*(c) *CI* uses the interval $[T_{(1)}^*, T_{(c)}^*], [T_{(2)}^*, T_{(c+1)}^*], ..., [T_{(B-c+1)}^*, T_{(B)}^*]$ of shortest length. Here $c = \min(B, \lceil B[1-\delta + 1.12\sqrt{\delta/B}\ ]\ \rceil)$. The shorth CI is computed by applying the shorth PI to the bootstrap sample.

4.11 Complements

Some variable selection methods for the Cox PH regression model include Fan and Li (2002), Huang et al. (2013) who give KKT conditions, Simon et al. (2013), and Tibshirani (1997). Also see Claeskens and Hjort (2008). For bootstrapping the Cox PH regression model, see Burr (1994), Efron and Tibshirani (1986), Rathnayake (2019), Rathnayake and Olive (2019), and Shao and Tu (1995). For bootstrapping some other survival analysis models, see Efron (1981), Gross and Lai (1996), and Li and Datta (2001).

This chapter followed Olive (2017b, ch. 5), Pelawa Watagoda and Olive (2019ab) and Rathnayake and Olive (2020) closely. Also see Olive (2013a, 2018), and Rathnayake (2019). Olive (2014: p. 283, 2017ab, 2018) recommended using the shorth(c) estimator for the percentile method. Olive (2017a: p. 128, 2017b: p. 181, 2018) showed that the prediction region method can simulate well for the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I_{min},0}$. Hastie et al. (2009, p. 57) noted that variable selection is a shrinkage estimator: the coefficients are shrunk to 0 for the omitted variables.

Good references for the bootstrap include Efron (1982), Efron and Hastie (2016, ch. 10–11), and Efron and Tibshirani (1993). Also see Chen (2016) and Hesterberg (2014).

There is a massive literature on variable selection and a fairly large literature for inference after variable selection. See, for example, Leeb and Pötscher (2006, 2008). Inference techniques for the variable selection model, other than data splitting, have not had much success. The methods are often inferior to data splitting, or are asymptotically equivalent to using the full model, or find a quantity to test that is not $\boldsymbol{A\beta}$. See Ewald and Schneider (2018).

4.12 Problems

4.1. Consider the Cushny and Peebles data set (see Staudte and Sheather 1990, p. 97) listed below. Find shorth(7). Show work.

```
0.0  0.8  1.0  1.2  1.3  1.3  1.4  1.8  2.4  4.6
```

4.2. Find shorth(5) for the following data set. Show work.

```
6   76   90   90   94   94   95   97   97   1008
```

4.3. Find shorth(5) for the following data set. Show work.

```
66  76   90   90   94   94   95   95   97   98
```

4.4. Suppose you are estimating the mean θ of losses with the maximum likelihood estimator (MLE) $\overline{X}$ assuming an exponential (θ) distribution. Compute the sample mean of the fourth bootstrap sample.
actual losses 1, 2, 5, 10, 50: $\overline{X} = 13.6$
bootstrap samples:
2, 10, 1, 2, 2: $\overline{X} = 3.4$
50, 10, 50, 2, 2: $\overline{X} = 22.8$
10, 50, 2, 1, 1: $\overline{X} = 12.8$
5, 2, 5, 1, 50: $\overline{X} =$?

4.5. The data below are a sorted residuals from a least squares regression where $n = 100$ and $p = 4$. Find shorth(97) of the residuals.

```
number     1      2      3      4    ...  97   98   99   100
residual -2.39 -2.34 -2.03 -1.77 ... 1.76 1.81 1.83  2.16
```

4.6. To find the sample median of a list of n numbers where n is odd, order the numbers from smallest to largest and the median is the middle ordered number. The sample median estimates the population median. Suppose the sample is $\{14, 3, 5, 12, 20, 10, 9\}$. Find the sample median for each of the three bootstrap samples listed below.
Sample 1: 9, 10, 9, 12, 5, 14, 3
Sample 2: 3, 9, 20, 10, 9, 5, 14
Sample 3: 14, 12, 10, 20, 3, 3, 5

4.7. Suppose you are estimating the mean μ of losses with $T = \overline{X}$.
actual losses 1, 2, 5, 10, 50: $\overline{X} = 13.6$,

a) Compute $T_1^*, ..., T_4^*$, where T_i^* is the sample mean of the ith bootstrap sample. bootstrap samples:

2, 10, 1, 2, 2:

50, 10, 50, 2, 2:

10, 50, 2, 1, 1:

5, 2, 5, 1, 50:

b) Now compute the bagging estimator which is the sample mean of the T_i^*: the bagging estimator $\overline{T}^* = \dfrac{1}{B}\displaystyle\sum_{i=1}^{B} T_i^*$ where $B = 4$ is the number of bootstrap samples.

R Problems

Use the command *source("G:/linmodpack.txt")* **to download the functions** and the command *source("G:/linmoddata.txt")* **to download the data. See Preface or Section 11.1.** Typing the name of the `linmodpack` function, e.g. *regbootsim2*, will display the code for the function. Use the `args` command, e.g. *args(regbootsim2)*, to display the needed arguments for the function. For the following problem, the *R* command can be copied and pasted from (http://parker.ad.siu.edu/Olive/linmodrhw.txt) into *R*.

4.8. a) Type the *R* command `predsim()` and paste the output into *Word*.

This program computes $\boldsymbol{x}_i \sim N_4(\mathbf{0}, diag(1, 2, 3, 4))$ for $i = 1, ..., 100$ and $\boldsymbol{x}_f = \boldsymbol{x}_{101}$. One hundred such data sets are made, and ncvr, scvr, and mcvr count the number of times $\boldsymbol{x}_f$ was in the nonparametric, semiparametric, and parametric MVN 90% prediction regions. The volumes of the prediction regions are computed and voln, vols, and volm are the average ratio of the volume of the ith prediction region over that of the semiparametric region. Hence vols is always equal to 1. For multivariate normal data, these ratios should converge to 1 as $n \to \infty$.

b) Were the three coverages near 90%?

Chapter 5
Stuff for Students

5.1 R

R is available from the **CRAN** website (https://cran.
r-project.org/). As of January 2020, the author's personal computer has Version 3.3.1 (June 21, 2016) of *R*. *R* is similar to *Splus*, but is free. *R* is very versatile since many people have contributed useful code, often as packages.

Downloading the book's files into R

Many of the homework problems use *R* functions contained in the book's website (http://parker.ad.siu.edu/Olive/survbk.htm) under the file name *survpack.txt*. The following two *R* commands can be copied and pasted into *R* from near the top of the file (http://parker.ad.siu.edu/Olive/
survhw.txt).

Downloading the book's R functions *survpack.txt* and data files *linmoddata.txt* into *R*: the commands

```
source("http://parker.ad.siu.edu/Olive/survpack.txt")
source("http://parker.ad.siu.edu/Olive/survdata.txt")
```

can be used to download the *R* functions and data sets into *R*. Type *ls()*. Nearly 10 *R* functions from *linmodpack.txt* should appear. In *R*, enter the command *q()*. A window asking "*Save workspace image?*" will appear. Click on *No* to remove the functions from the computer (clicking on *Yes* saves the functions in *R*, but the functions and data are easily obtained with the source commands).

Citing packages

We will use *R* packages often in this book. The following *R* command is useful for citing the Friedman et al. (2015) `glmnet` package. Another packages cited in this book is `MASS` from Venables and Ripley (2010).

```
citation("glmnet")
```

This section gives tips on using *R*, but is no replacement for books such as Becker et al. (1988), Crawley (2005, 2013), Fox and Weisberg (2010), or Venables and Ripley (2010). Also see Mathsoft (1999ab) and use the website (www.google.com) to search for useful websites. For example enter the search words *R documentation*.

The command *q()* gets you out of *R*.

Least squares regression can be done with the function *lsfit* or *lm*.

The commands *help(fn)* and *args(fn)* give information about the function fn, e.g. if fn = lsfit.

Type the following commands.

```
x <- matrix(rnorm(300),nrow=100,ncol=3)
y <- x%*%1:3 + rnorm(100)
out<- lsfit(x,y)
out$coef
ls.print(out)
```

The first line makes a 100 by 3 matrix x with N(0,1) entries. The second line makes $y[i] = 0 + 1 * x[i,1] + 2 * x[i,2] + 3 * x[i,2] + e$ where e is N(0,1). The term 1:3 creates the vector $(1,2,3)^T$ and the matrix multiplication operator is $\%*\%$. The function `lsfit` will automatically add the constant to the model. Typing "out" will give you a lot of irrelevant information, but *out$coef* and *out$resid* give the OLS coefficients and residuals respectively.

To make a residual plot, type the following commands.

```
fit <- y - out$resid
plot(fit,out$resid)
title("residual plot")
```

The first term in the plot command is always the horizontal axis while the second is on the vertical axis.

To put a graph in *Word,* hold down the *Ctrl* and *c* buttons simultaneously. Then select "Paste" from the *Word* menu, or hit *Ctrl* and *v* at the same time.

To enter data, open a data set in *Notepad* or *Word*. You need to know the number of rows and the number of columns. Assume that each case is entered in a row. For example, assuming that the file *cyp.lsp* has been saved on your flash drive from the webpage for this book, open *cyp.lsp* in *Word.* It has 76 rows and 8 columns. In *R* , write the following command.

```
cyp <- matrix(scan(),nrow=76,ncol=8,byrow=T)
```

A data frame is a two-dimensional array in which the values of different variables are stored in different named columns.

Then copy the data lines from *Word* and paste them in *R*. If a cursor does not appear, hit *enter*. The command *dim(cyp)* will show if you have entered the data correctly.

Enter the following commands

```
cypy <- cyp[,2]
cypx<- cyp[,-c(1,2)]
lsfit(cypx,cypy)$coef
```

to produce the output below.

```
   Intercept           X1           X2           X3
205.40825985   0.94653718   0.17514405   0.23415181
          X4           X5           X6
  0.75927197  -0.05318671  -0.30944144
```

Making functions in R is easy.

For example, type the following commands.

```
mysquare <- function(x){
# this function squares x
r <- x^2
r }
```

The second line in the function shows how to put comments into functions.

Modifying your function is easy.

Use the fix command.
fix(mysquare)
This will open an editor such as *Notepad* and allow you to make changes. (In *Splus*, the command *Edit(mysquare)* may also be used to modify the function *mysquare*.)

To save data or a function in *R*, when you exit, click on *Yes* when the "*Save worksheet image?*" window appears. When you reenter *R*, type *ls()*. This will show you what is saved. You should rarely need to save anything for this book. To remove unwanted items from the worksheet, e.g. x, type *rm(x)*,
pairs(x) makes a scatterplot matrix of the columns of x,
hist(y) makes a histogram of y,
boxplot(y) makes a boxplot of y,
stem(y) makes a stem and leaf plot of y,
scan(), source(), and *sink()* are useful on a *Unix* workstation.
To type a simple list, use *y <− c(1,2,3.5)*.
The commands *mean(y), median(y), var(y)* are self explanatory.

The following commands are useful for a scatterplot created by the command *plot(x,y)*.
lines(x,y), lines(lowess(x,y,f=.2))

identify(x,y)
abline(out$coef), abline(0,1)

The usual arithmetic operators are $2 + 4$, $3 - 7$, $8 * 4$, $8/4$, and

```
2^{10}.
```

The ith element of vector y is $y[i]$ while the ij element of matrix x is $x[i, j]$. The second row of x is $x[2,]$ while the 4th column of x is $x[, 4]$. The transpose of x is *t(x)*.

The command *apply(x,1,fn)* will compute the row means if fn = mean. The command *apply(x,2,fn)* will compute the column variances if fn = var. The commands *cbind* and *rbind* combine column vectors or row vectors with an existing matrix or vector of the appropriate dimension.

Getting information about a library in R

In *R*, a *library* is an add–on package of *R* code. The command *library()* lists all available libraries, and information about a specific library, such as `leaps` for variable selection, can be found, e.g., with the command *library(help=leaps)*.

Downloading a library into R

Many researchers have contributed a *library* or *package* of *R* code that can be downloaded for use. To see what is available, go to the website (http://cran.us.r-project.org/) and click on the Packages icon.

Following Crawley (2013, p. 8), you may need to "Run as administrator" before you can install packages (right click on the *R* icon to find this). Then use the following command to install the *glmnet* package.

```
install.packages("glmnet")
```

Open *R* and type the following command.

library(glmnet)

Next type *help(glmnet)* to make sure that the library is available for use.

Warning: *R* is free but not fool proof. If you have an old version of *R* and want to download a library, you may need to update your version of *R*. The libraries for robust statistics may be useful for outlier detection, but the methods have not been shown to be consistent or high breakdown. All software has some bugs. For example, Version 1.1.1 (August 15, 2000) of *R* had a random generator for the Poisson distribution that produced variates with too small of a mean θ for $\theta \geq 10$. Hence simulated 95% confidence intervals might contain θ 0% of the time. This bug seems to have been fixed in Versions 2.4.1 and later. Also, some functions in *survpack* may no longer work in new versions of *R*.

5.2 SAS

Allison (1995, 2010) is very useful for using SAS for Survival Analysis. Also see SAS Institute (1999). SAS (www.sas.com) has a free SAS University Edition and free tutorials for SAS programming. You can request materials from the SAS institute as well. They make these available for free for professors to use in teaching. They have some nice examples and data sets. See SAS Global Academic Program (http://support.sas.com/learn/ap/prof/index.html) for information.

There are some nice examples in SAS Statistics 1, this is also now available free as an e-course for anyone.

`(https://support.sas.com/edu/elearning.html?ctry=us&productType=library)`

SAS Training in the United States – e-Learning

This includes a SAS programming course.

Google SAS>Ad (www.sas.com) >How to buy>academic

`http://www.sas.com/en_us/software/trials-demos.html`

5.3 Hints for Selected Problems

Chapter 1

5.4 Tables

Tabled values are F(k,d, 0.95) where $P(F < F(k, d, 0.95)) = 0.95$.
00 stands for ∞. Entries were produced with the `qf(.95,k,d)` command in *R*. The numerator degrees of freedom are k while the denominator degrees of freedom are d.

k	1	2	3	4	5	6	7	8	9	00
d										
1	161	200	216	225	230	234	237	239	241	254
2	18.5	19.0	19.2	19.3	19.3	19.3	19.4	19.4	19.4	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.41
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	1.84
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	1.62
00	3.84	3.00	2.61	2.37	2.21	2.10	2.01	1.94	1.88	1.00

Tabled values are $t_{\alpha,d}$ where $P(t < t_{\alpha,d}) = \alpha$ where t has a t distribution with d degrees of freedom. If $d > 29$ use the $N(0,1)$ cutoffs $d = Z = \infty$.

					alpha					pvalue
d	0.005	0.01	0.025	0.05	0.5	0.95	0.975	0.99	0.995	left tail
1	-63.66	-31.82	-12.71	-6.314	0	6.314	12.71	31.82	63.66	
2	-9.925	-6.965	-4.303	-2.920	0	2.920	4.303	6.965	9.925	
3	-5.841	-4.541	-3.182	-2.353	0	2.353	3.182	4.541	5.841	
4	-4.604	-3.747	-2.776	-2.132	0	2.132	2.776	3.747	4.604	
5	-4.032	-3.365	-2.571	-2.015	0	2.015	2.571	3.365	4.032	
6	-3.707	-3.143	-2.447	-1.943	0	1.943	2.447	3.143	3.707	
7	-3.499	-2.998	-2.365	-1.895	0	1.895	2.365	2.998	3.499	
8	-3.355	-2.896	-2.306	-1.860	0	1.860	2.306	2.896	3.355	
9	-3.250	-2.821	-2.262	-1.833	0	1.833	2.262	2.821	3.250	
10	-3.169	-2.764	-2.228	-1.812	0	1.812	2.228	2.764	3.169	
11	-3.106	-2.718	-2.201	-1.796	0	1.796	2.201	2.718	3.106	
12	-3.055	-2.681	-2.179	-1.782	0	1.782	2.179	2.681	3.055	
13	-3.012	-2.650	-2.160	-1.771	0	1.771	2.160	2.650	3.012	
14	-2.977	-2.624	-2.145	-1.761	0	1.761	2.145	2.624	2.977	
15	-2.947	-2.602	-2.131	-1.753	0	1.753	2.131	2.602	2.947	
16	-2.921	-2.583	-2.120	-1.746	0	1.746	2.120	2.583	2.921	
17	-2.898	-2.567	-2.110	-1.740	0	1.740	2.110	2.567	2.898	
18	-2.878	-2.552	-2.101	-1.734	0	1.734	2.101	2.552	2.878	
19	-2.861	-2.539	-2.093	-1.729	0	1.729	2.093	2.539	2.861	
20	-2.845	-2.528	-2.086	-1.725	0	1.725	2.086	2.528	2.845	
21	-2.831	-2.518	-2.080	-1.721	0	1.721	2.080	2.518	2.831	
22	-2.819	-2.508	-2.074	-1.717	0	1.717	2.074	2.508	2.819	
23	-2.807	-2.500	-2.069	-1.714	0	1.714	2.069	2.500	2.807	
24	-2.797	-2.492	-2.064	-1.711	0	1.711	2.064	2.492	2.797	
25	-2.787	-2.485	-2.060	-1.708	0	1.708	2.060	2.485	2.787	
26	-2.779	-2.479	-2.056	-1.706	0	1.706	2.056	2.479	2.779	
27	-2.771	-2.473	-2.052	-1.703	0	1.703	2.052	2.473	2.771	
28	-2.763	-2.467	-2.048	-1.701	0	1.701	2.048	2.467	2.763	
29	-2.756	-2.462	-2.045	-1.699	0	1.699	2.045	2.462	2.756	
Z	-2.576	-2.326	-1.960	-1.645	0	1.645	1.960	2.326	2.576	
CI						90%	95%		99%	
	0.995	0.99	0.975	0.95	0.5	0.05	0.025	0.01	0.005	right tail
	0.01	0.02	0.05	0.10	1	0.10	0.05	0.02	0.01	two tail

Aelen, O.O., (1978), "Non Parametric Inference for a Family of Counting Processes," *The Annals of Statistics*, 6, 701-726.

Agresti, A., and Coull, B.A. (1998), "Approximate is Better than Exact for Interval Estimation of Binomial Parameters," *The American Statistician*, 52, 119-126.

Akaike, H. (1973), "Information Theory as an Extension of the Maximum Likelihood Principle," in *Proceedings, 2nd International Symposium on Information Theory*, eds. Petrov, B.N., and Csakim, F., Akademiai Kiado, Budapest, 267-281.

Allison, P.D. (1995), *Survival Analysis Using SAS: a Practical Guide*, 1st ed., SAS Institute, Cary, NC.

Allison, P.D. (2010), *Survival Analysis Using SAS: a Practical Guide*, 2nd ed., SAS Institute, Cary, NC.

Andersen, P.K., and Skovgaard, L.T. (2010), *Regression with Linear Predictors*, Springer, New York, NY.

Becker, R.A., Chambers, J.M., and Wilks, A.R. (1988), *The New S Language: a Programming Environment for Data Analysis and Graphics,* Wadsworth and Brooks/Cole, Pacific Grove, CA.

Bennett, S. (1983), "Analysis of Survival Data by the Proportional Odds Model," *Statistics in Medicine*, 2, 273-277.

Bickel, P.J., and Ren, J.–J. (2001), "The Bootstrap in Hypothesis Testing," in *State of the Art in Probability and Statistics: Festschrift for William R. van Zwet*, eds. de Gunst, M., Klaassen, C., and van der Vaart, A., The Institute of Mathematical Statistics, Hayward, CA, 91-112.

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24, 123-140.

Breslow, N.E. (1974), "Covariance Analysis of Censored Survival Data," *Biometrics*, 30, 89-100.

Büchlmann, P., and Yu, B. (2002), "Analyzing Bagging," *The Annals of Statistics*, 30, 927-961.

Budny, K. (2014), "A Generalization of Chebyshev's Inequality for Hilbert-Space-Valued Random Variables," *Statistics & Probability Letters,* 88, 62-65.

Burnham, K.P., and Anderson, D.R. (2002), *Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach*, 2nd ed., Springer, New York, NY.

Burnham, K.P., and Anderson, D.R. (2004), "Multimodel Inference Understanding AIC and BIC in Model Selection," *Sociological Methods & Research*, 33, 261-304.

Burr, D. (1994), "A Comparison of Certain Bootstrap Confidence Intervals in the Cox Model," *Journal of the American Statistical Association*, 89, 1290-1302.

Camilli, S.J., Duncan, I., and London, R.L. (2014), *Models for Quantifying Risk*, 6th ed., ACTEX Publications, Winsted, CT.

Charkhi, A., and Claeskens, G. (2018), "Asymptotic Post-Selection Inference for the Akaike Information Criterion," *Biometrika*, 105, 645-664.

Chen, S.X. (2016), "Peter Hall's Contributions to the Bootstrap," *The Annals of Statistics*, 44, 1821-1836.

Chen, X. (2011), "A New Generalization of Chebyshev Inequality for Random Vectors," see arXiv:0707.0805v2.

Chihara, L., and Hesterberg, T. (2011), *Mathematical Statistics with Resampling and R*, Wiley, Hoboken, NJ.

Claeskens, G., and Hjort, N.L. (2008), *Model Selection and Model Averaging*, Cambridge University Press, New York, NY.

Collett, D. (2003), *Modelling Survival Data in Medical Research*, 2nd ed., Chapman & Hall/CRC, Boca Raton, FL.

Collett, D. (2014), *Modelling Survival Data in Medical Research*, 3rd ed., Chapman & Hall/CRC, Boca Raton, FL.

Cook, R.D., and Weisberg, S. (1999), *Applied Regression Including Computing and Graphics,* Wiley, New York, NY.

Cox, D.R. (1962), *Renewal Theory*, Wiley, New York, NY.

Cox, D.R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society, B,* 34, 187-220.

Crawley, M.J. (2005), *Statistics an Introduction Using R,* Wiley, Hoboken, NJ.

Crawley, M.J. (2007), *The R Book,* Wiley, Hoboken, NJ.

Edmunson, J.H., Fleming, T.R., Decker, D.G., Malkasian, G.D., Jorgenson, E.O., Jeffries, J.A., Webb, M.J., and Kvols, L.K. (1979), "Different Chemotherapeutic Sensitivities and Host Factors Affecting Prognosis in Advanced Ovarian Carcinoma Versus Minimal Residual Disease," *Cancer Treatment Reports*, 63, 241-247.

Efron, B. (1981), "Censored-Data and the Bootstrap," *Journal of the American Statistical Association*, 76, 312-319.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans,* SIAM, Philadelphia, PA.

Efron, B. (2014), "Estimation and Accuracy After Model Selection," (with discussion), *Journal of the American Statistical Association,* 109, 991-1007.

Efron, B., and Hastie, T. (2016), *Computer Age Statistical Inference*, Cambridge University Press, New York, NY.

Efron, B., and Tibshirani, R. (1986), "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Methods of Statistical Accuracy," (with discussion), *Statistical Science*, 1, 54-77.

Efron, B., and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap*, Chapman & Hall/CRC, New York, NY.

Ewald, K., and Schneider, U. (2018), "Uniformly Valid Confidence Sets Based on the Lasso," *Electronic Journal of Statistics*, 12, 1358-1387.

Fan, J., and Li, R. (2002), "Variable Selection for Cox's Proportional Hazard Model and Frailty Model," *The Annals of Statistics*, 30, 74-99.

Feng, C., Wang, H., Zhang, Y., Han, Y., Liang, Y., and Tu, X.M. (2017), "On Testing Proportionality in the Cox Regression Model by Andersen's Plot," *Communications in Statistics: Theory and Methods*, 46, 3489-3500.

Ferguson, T.S. (1996), *A Course in Large Sample Theory*, Chapman & Hall, New York, NY.

Fox, J., and Weisberg, S. (2010), *An R Companion to Applied Regression*, Sage Publications, Thousand Oaks, CA.

Freedman, D.A. (2008), "Survival Analysis: a Primer," *The American Statistician*, 62, 110-119.

Frey, J. (2013), "Data-Driven Nonparametric Prediction Intervals," *Journal of Statistical Planning and Inference*, 143, 1039-1048.

Friedman, J.H., and Hall, P. (2007), "On Bagging and Nonlinear Estimation," *Journal of Statistical Planning and Inference,* 137, 669-683.

Grambsch, P.M., and Therneau, T.M. (1994), "Proportional Hazards Tests and Diagnostics Based on Weighted Residuals," *Biometrika*, 81, 515-526.

Gross, S.T., and Lai, T.L. (1996), "Bootstrap Methods for Truncated and Censored Data," *Statistica Sinica*, 6, 509-530.

Hall, P. (1988), "Theoretical Comparisons of Bootstrap Confidence Intervals," (with discussion), *The Annals of Statistics*, 16, 927-985.

Harrell, F.E. (2015), *Regression Modeling Strategies*, 2nd ed., Springer, New York, NY.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd ed., Springer, New York, NY.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015), *Statistical Learning with Sparsity: the Lasso and Generalizations*, CRC Press Taylor & Francis, Boca Raton, FL.

Hess, K.R. (1995), "Graphical Methods for Assessing Violations of the Proportional Hazards Assumption in Cox Regression," *Statistics in Medicine*, 14, 1707-1723.

Hesterberg, T., (2014), "What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum," available from (http://arxiv.org/pdf/1411.5279v1.pdf). (An abbreviated version was published (2015), *The American Statistician*, 69, 371-386.)

Hjort, N.L., and Claeskens, G. (2006), "Focussed Information Criteria and Model Averaging for Cox's Hazard Regression Model," *Journal of the American Statistical Association*, 101, 1449-1464.

Hogg, R.V., Tanis, E.A., and Zimmerman, D.L. (2015), *Probability and Statistical Inference,* 9th ed., Pearson, Boston, MA.

Hong, L., Kuffner, T.A., and Martin, R. (2018), "On Overfitting and Post-Selection Uncertainty Assessments," *Biometrika*, 105, 221-224.

Hosmer, D.W., and Lemeshow, S. (1999), *Applied Survival Analysis: Regression Modeling of Time to Event Data*, 1st ed., Wiley, New York, NY.

Hosmer, D.W., Lemeshow, S., and May, S. (2008), *Applied Survival Analysis: Regression Modeling of Time to Event Data*, 2nd ed., Wiley, New York, NY.

Huang, J., Sun, T., Ying, Z., Yu, Y., and Zhang, C.-H. (2013), "Oracle Inequalities for the Lasso in the Cox Model," *The Annals of Statistics*, 41, 1142-1165.

Hurvich, C., and Tsai, C.L. (1990), "The Impact of Model Selection on Inference in Linear Regression," *The American Statistician*, 44, 214-217.

Hyndman, R.J. (1996), "Computing and Graphing Highest Density Regions," *The American Statistician,* 50, 120-126.

Johnson, R.A., and Wichern, D.W. (1988, 2007), *Applied Multivariate Statistical Analysis,* 2nd and 6th ed., Prentice Hall, Englewood Cliffs, NJ.

Kalbfleisch, J.D., and Prentice, R.L. (2002), *The Statistical Analysis of Failure Time Data*, 2nd ed., Wiley, New York, NY.

Kaplan, E.L., and Meier, P. (1958), "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, 53, 457-481.

Kellison, S.G. and London, R.L. (2011), *Risk Models and Their Estimation*, ACTEX Publications, Winsted, CT.

Klein, J.P., and Moeschberger, M.L. (1997), *Survival Analysis*, 1st ed., Springer, New York, NY.

Klein, J.P., and Moeschberger, M.L. (2003), *Survival Analysis*, 2nd ed., Springer, New York, NY.

Kleinbaum, D.G., and Klein, M. (2012), *Survival Analysis: a Self-Learning Text*, 3rd ed. Springer, New York, NY.

Klugman, S.A., Panjer, H.H., and Willmot, G.E. (2008), *Loss Models: from Data to Decisions*, 3rd ed., New York, NY, Wiley.

Larsen, R.J., and Marx, M.L. (2017), *Introduction to Mathematical Statistics and Its Applications*, 6th ed., Pearson, Boston, MA.

Lawless, J.F. (2002), *Statistical Models and Methods for Lifetime Data Analysis*, 2nd ed., Wiley, New York, NY.

Lee, E.T. and Wang, J.W. (2003), *Statistical Methods for Survival Data Analysis*, 3rd ed., Wiley, NY.

Leeb, H., and Pötscher, B.M. (2006), "Can One Estimate the Conditional Distribution of Post–Model-Selection Estimators?" *The Annals of Statistics*, 34, 2554-2591.

Leeb, H. and Pötscher, B.M. (2008), "Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators?" *Econometric Theory*, 24, 338-376.

Leemis, L.M. (1995), *Reliability, Probabilistic Models and Statistical Methods*, Prentice Hall, Upper Saddle River, NJ.

Lehmann, E.L. (1999), *Elements of Large–Sample Theory*, Springer, New York, NY.

Li, G., and Datta, S. (2001), "A Bootstrap Approach to Nonparametric Regression for Right Censored Data," *Annals of the Institute of Statistical Mathematics*, 53, 708-729.

Machado, J.A.F., and Parente, P. (2005), "Bootstrap Estimation of Covariance Matrices Via the Percentile Method," *Econometrics Journal,* 8, 70–78.

MathSoft (1999a), *S-Plus 2000 User's Guide,* Data Analysis Products Division, MathSoft, Seattle, WA.

MathSoft (1999b), *S-Plus 2000 Guide to Statistics,* Volume 2, Data Analysis Products Division, MathSoft, Seattle, WA.

May, S., and Hosmer, D.W. (1998), "A Simple Method for Calculating a Goodness-of-Fit Test for the Proportional Hazards Model," *Lifetime Data Analysis*, 4, 109-120.

Meeker, W.Q., and Escobar, L.A. (1998), *Statistical Methods for Reliability Data,* John Wiley and Sons, NY.

Miller, R. (1981), *Survival Analysis*, Wiley, New York, NY.

Navarro, J. (2014), "Can the Bounds in the Multivariate Chebyshev Inequality be Attained?" *Statistics & Probability Letters*, 91, 1-5.

Navarro, J. (2016), "A Very Simple Proof of the Multivariate Chebyshev's Inequality," *Communications in Statistics: Theory and Methods,* 45, 3458-3463.

Nelson, W. (1969), "Hazard Plotting for Incomplete Failure Data," *Journal of Quality Technology*, 1, 27-52.

Nelson, W. (1972), "Theory and Application of Hazard plotting for Censored Failure Data," *Technometrics*, 14, 945-965.

Oakes, D. (2000), "Survival Analysis," *Journal of the American Statistical Association*, 95, 282-285.

Olive, D.J. (2007), "Prediction Intervals for Regression Models," *Computational Statistics & Data Analysis,* 51, 3115-3122.

Olive, D.J. (2010), *Multiple Linear and 1D Regression Models*, online course notes, see (http://parker.ad.siu.edu/Olive/regbk.htm).

Olive, D.J. (2013a), "Plots for Survival Regression," unpublished manuscript, see (http://parker.ad.siu.edu/Olive/ppvsurv.pdf).

Olive, D.J. (2013b), "Asymptotically Optimal Regression Prediction Intervals and Prediction Regions for Multivariate Data," *International Journal of Statistics and Probability*, 2, 90-100.

Olive, D.J. (2014), *Statistical Theory and Inference*, Springer, New York, NY.

Olive, D.J. (2017a), *Linear Regression*, Springer, New York, NY.

Olive, D.J. (2017b), *Robust Multivariate Analysis*, Springer, New York, NY.

Olive, D.J. (2018), "Applications of Hyperellipsoidal Prediction Regions," *Statistical Papers,* 59, 913-931.

Olive, D.J. (2023a), *Prediction and Statistical Learning*, online course notes, see (http://parker.ad.siu.edu/Olive/slearnbk.htm).

Olive (2023b) *Large Sample Theory*: online course notes, (http://parker.ad.siu.edu/Olive/lsampbk.pdf).

Olive, D.J., and Hawkins, D.M. (2005), "Variable Selection for 1D Regression Models," *Technometrics*, 47, 43-50.

Olive, D.J., Rathnayake, R.C., and Haile, M. (2022), "Prediction Intervals for GLMs, GAMs, and Some Survival Regression Models," *Communications in Statistics: Theory and Methods*, , 51, 8012-8026.

Pelawa Watagoda, L. C. R., and Olive, D.J. (2021a), "Bootstrapping Multiple Linear Regression After Variable Selection," *Statistical Papers,* 62, 681-700.

Pelawa Watagoda, L.C.R., and Olive, D.J. (2021b), "Comparing Six Shrinkage Estimators With Large Sample Theory and Asymptotically Optimal Prediction Intervals," *Statistical Papers,* 62, 2407-2431.

Pratt, J.W. (1959), "On a General Concept of 'in Probability'," *The Annals of Mathematical Statistics,* 30, 549-558.

Rao, C.R. (1965, 1973), *Linear Statistical Inference and Its Applications,* 1st and 2nd ed., Wiley, New York, NY.

Rathnayake, R.C. (2019), *Inference for Some GLMs and Survival Regression Models after Variable Selection*, Ph.D. thesis, Southern Illinois University, at (http://parker.ad.siu.edu/Olive/srasanjiphd.pdf).

Rathnayake, R.C., and Olive, D.J. (2023), "Bootstrapping Some GLM and Survival Regression Variable Selection Estimators," *Communications in Statistics: Theory and Methods*, 52, 2625-2645.

R Core Team (2023), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

Rinaldo, A., Wasserman, L., and G'Sell, M. (2019), "Bootstrapping and Sample Splitting for High-Dimensional, Assumption-Lean Inference," *The Annals of Statistics*, 47, 3438-3469.

Rohatgi, V.K. (1976), *An Introduction to Probability Theory and Mathematical Statistics,* Wiley, New York, NY.

Rohatgi, V.K. (1984), *Statistical Inference,* Wiley, New York, NY.

SAS Institute (1985), *SAS User's Guide: Statistics*, Version 5, SAS Institute, Cary, NC.

SAS Institute, (1999), *SAS/STAT User's Guide*, Version 8, SAS Institute, Cary, NC.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics,* 6, 461-464.

Sen, P.K., and Singer, J.M. (1993), *Large Sample Methods in Statistics: an Introduction with Applications,* Chapman & Hall, New York, NY.

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics,* Wiley, New York, NY.

Severini, T.A. (2005), *Elements of Distribution Theory*, Cambridge University Press, New York, NY.

Shao, J., and Tu, D.S. (1995), *The Jackknife and the Bootstrap*, Springer, New York, NY.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011), "Regularization Paths for Cox's Proportional Hazards Model Via Coordinate Descent," *Journal of Statistical Software*, 39, 1-13.

Smith, P.J. (2002), *Analysis of Failure and Survival Data*, Chapman and Hall/CRC, Boca Raton, FL.

Staudte, R.G., and Sheather, S.J. (1990), *Robust Estimation and Testing,* Wiley, New York, NY.

Tableman, M. and Kim, J.S. (2003), *Survival Analysis Using S: Analysis of Time-to-Event Data*, Chapman and Hall/CRC, Boca Rotan, FL.

Tibshirani, R, (1997), "The Lasso Method for Variable Selection in the Cox Model," *Statistics in Medicine*, 16, 385-395.

Tibshirani, R.J., Rinaldo, A., Tibshirani, R., and Wasserman, L. (2018), "Uniform Asymptotic Inference and the Bootstrap After Model Selection," *The Annals of Statistics*, 46, 1255-1287.

van Houwelingen, H.C., Bruinsma, T., Hart, A.A.M., van't Veer, L.J., and Wessels, L.F.A. (2006), "Cross-validated Cox Regression on Microarray Gene Expression Data," *Statistics in Medicine*, 25, 32013216.

Venables, W.N., and Ripley, B.D. (2010), *Modern Applied Statistics with S*, 4th ed., Springer, New York, NY.

Vittinghoff, E., Glidden, D.V., Shiboski, S.C., and McCulloch, C.E. (2012), *Regression Methods in Biostatistics: Linear, Logistic, Survival and Repeated Measures Models*, 2nd ed., Springer, New York, NY.

Wackerly, D.D., Mendenhall, W., and Scheaffer, R.L. (2008), *Mathematical Statistics with Applications,* 7th ed., Thomson Brooks/Cole, Belmont, CA.

Walpole, R.E., Myers, R.H., Myers, S.L., and Ye, K. (2016), *Probability* & *Statistics for Engineers* & *Scientists,* 9th ed., Pearson, Boston, MA.

Wei, L.J. (1992), "The Accelerated Failure Time Model: a Useful Alternative to the Cox Regression Model in Survival Analysis," *Statistics in Medicine*, 11, 1871-1879.

White, H. (1984), *Asymptotic Theory for Econometricians,* Academic Press, San Diego, CA.

Yang, L.H. (2016), *Confidence Intervals for the Survival Function*, MS paper, Southern Illinois University, see (https://pdfs.semanticscholar.org/2405/1ce15958ed4ccc82666085f66bf1140b7c25.pdf).

Yang, S., and Prentice, R.L. (1999), "Semiparametric Inference in the Proportional Odds Regression Model," *Journal of the American Statistical Association,* 94, 124-136.

Yang, Y. (2003), "Regression with Multiple Candidate Models: Selecting or Mixing?" *Statistica Sinica*, 13, 783-809.

Zeng, D., and Lin, D.Y. (2007), "Efficient Estimation for the Accelerated Failure Time Model," *Journal of the American Statistical Association*, 102, 1387-1396.

Zhou, M. (2001), "Understanding the Cox Regression Models with Time–Change Covariates," *The American Statistician*, 55, 153-155.

Index