Chapter 16

Survival Analysis

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.

This chapter will begin with univariate survival analysis: there is a response but no predictors. This model introduces terms also used in the 1D regression models for survival analysis. The survival regression 1D models

differ from the multiple linear regression, experimental design models, generalized linear models and single index 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.

16.1 Univariate Survival Analysis

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 a 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 16.1. Let $Y \geq 0$ be the time until an event occurs. Then Y is called the **survival time**. 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.

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 16.2. i) The **distribution function** (df) 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.

Given one of $F(t), f(t), S(t), h(t)$ or $H(t)$, the following proposition shows how to find the other 4 quantities for $t > 0$. In reliability analysis, the reliability function $R(t) = S(t)$, and in economics, Mill's ratio $= 1/h(t)$.

Proposition 16.1.

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. 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 16.1. 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 $E(T) = 1/\lambda$. 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. 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 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 $\hat{\lambda} \pm 1.96\hat{\lambda}/\sqrt{r}$.

Example 16.2. If $Y \sim$ Weibull(λ, γ) 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. Derive $H(t), S(t), F(t)$ and $f(t)$ if $Y \sim$ Weibull(λ, γ).

Solution:

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

$S(t) = \exp[-H(t)] = \exp[-\lambda t^\gamma]$, 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 $t > 0$.

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 16.3. 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 $\hat{S}_E(t_i) = a_i/n = \sum_{k=1}^{n} I(T_i > t_i)/n = \hat{S}_E(t_{i-1}) - \frac{d_i}{n}$.

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	$\hat{S}_E(t_m) = 0 = \hat{S}_E(t_{m-1}) - \frac{d_m}{n} = \frac{a_{m-1} - d_m}{n} = \frac{a_m}{n}$

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.

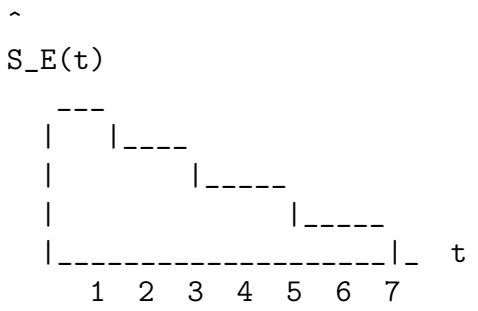
Example 16.3. Smith (2002, p. 68) gives steroid induced remission times for leukemia patients. The $t_{(j)}$, $t - i$ and d_i are given in the following table. The a_i and $\hat{S}_E(t)$ needed to be computed. Note that $a_i = \#$ of cases with $t_{(j)} > t_i$.

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	$\hat{S}_E(23) = (1-1)/21 = 0$

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 $t = 1, 2, ..., 23$.

Example 16.4. 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.



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

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

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

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

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

Example 16.5. 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. $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ 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, 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 th $(j-1)$th interval propagate to the jth interval.
Let $n'_j = n_j - \frac{c_j}{2}$ = average number at risk in I_j.

Definition 16.4. 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

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

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 = P(surviving through I_j| 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

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

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	$n_1 - \frac{c_1}{2}$	$\frac{n'_1 - d_1}{n'_1}$	$\hat{S}_L(t_o) = \hat{S}_L(0) = 1$
$[t_1, t_2)$, 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	$n_m - \frac{c_m}{2}$	$\frac{n'_m - d_m}{n'_m}$	$\hat{S}_L(t_{m-1}) = \hat{S}_L(t_{m-2})\frac{n'_{m-1} - d_{m-1}}{n'_{m-1}}$

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
```

Example 16.6. 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 and censoring times are known. Let $Y_i^* = T_i = \min(Y_i, Z_i)$ where Y_i and Z_i are independent. Let $\delta_i = I(Y_i \leq Z_i)$ so $\delta_i = 1$ if T_i is uncensored and $\delta_i = 0$ if T_i is censored. Let $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ 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 $d_i \geq 2$, then there are **ties**. Let $n_i = \sum_{j=1}^{n} I(t_{(j)} \geq t_i) = \#$ at risk at $t_i = \#$ alive and not yet censored just before t_i.

Definition 16.5. 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 16.7. 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, 487+.

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** $S_Y(t)$.

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
```

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.

16.2 Proportional Hazards Regression

Definition 16.6. 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 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 in survival analysis.

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. For many of the most important survival regression models, 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 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. The survival function is

$$S_{\boldsymbol{x}}(t) \equiv S_{Y|SP}(t) = [S_0(t)]^{\exp(\boldsymbol{\beta}_C^T \boldsymbol{x})} = [S_0(t)]^{\exp(SP)}. \tag{16.1}$$

If $\boldsymbol{x} = \boldsymbol{0}$ is within the range of the predictors, then the baseline survival and hazard functions correspond to the survival and hazard functions of $\boldsymbol{x} = \boldsymbol{0}$. 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

$$\hat{S}_{\boldsymbol{x}}(t) = [\hat{S}_0(t)]^{\exp(\hat{\boldsymbol{\beta}}_C^T \boldsymbol{x})} = [\hat{S}_0(t)]^{\exp(ESP)}. \tag{16.2}$$

16.2.1 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/Splus* function `cox.zph` makes both the plots and tests. See MathSoft (1999, pp. 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 16.7. The **slice survival plot** divides the ESP into J groups of roughly the same size. For each group j, $\hat{S}_j(t)$ is computed using an $\boldsymbol{x}$ corresponding to the middle ESP of the group. (The "middle ESP" is the kth order statistic of the ESP in group j, where $k = 1 + floor[(n_j - 1)/2]$ and n_j is the number of cases in group j.) Let $\hat{S}_{KMj}(t)$ be the Kaplan Meier estimator computed from the survival times (Y_i, δ_i) in the jth group. For each group, $\hat{S}_j(t)$ is plotted and $\hat{S}_{KMj}(t_i)$ as circles at the uncensored event times t_i. The survival regression model is reasonable if the circles "track the curve 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 curves 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, pp. 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 crucial 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.

The slice survival plot tailored to the Cox model is closely related to the May and Hosmer (1998) test, and the plot has been suggested by several

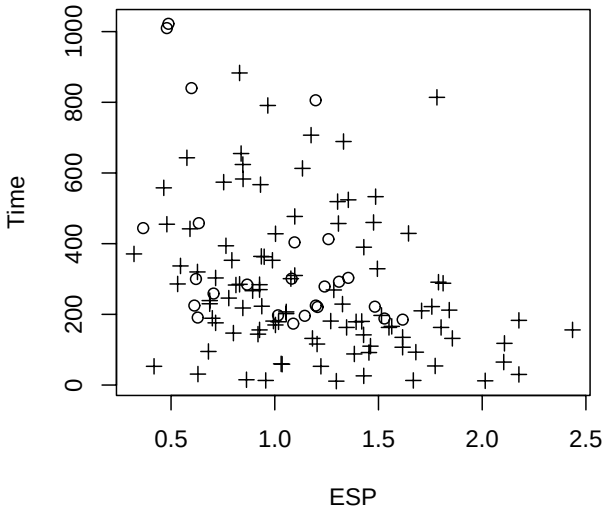


Figure 16.1: Censored Response Plot for R Lung Cancer Data

authors with $\boldsymbol{x}$ divided into J groups instead of the ESP. For example, see Miller (1981, p. 168). Hosmer and Lemeshow (1999, pp. 141–145) suggests making plots based on the quartiles of the ith predictor x_i, and note that a problem with Cox survival curves (16.2) 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.

Definition 16.7. 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.

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.

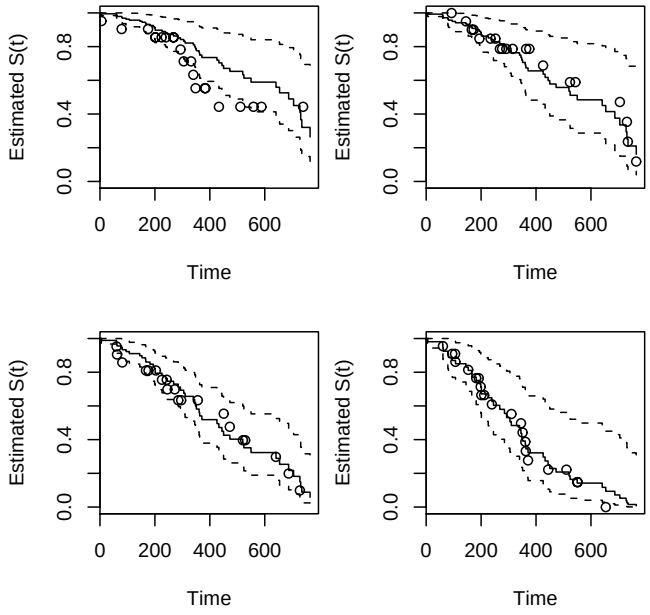


Figure 16.2: Slice Survival Plots for R Lung Cancer Data

Example 16.8 *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 (1999, 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 16.1 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 16.2 shows the slice survival plots. The ESP was divided into 4 groups and correspond to the upper left, upper right, lower right and lower left 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.

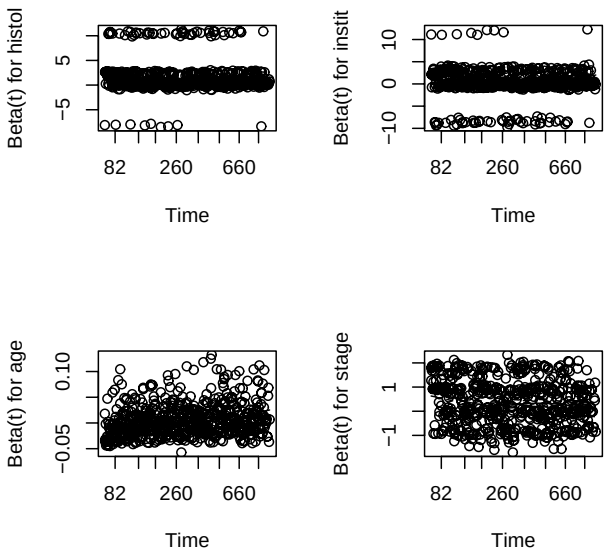


Figure 16.3: Grambsch and Therneau Plots for NWTCO Data

Example 16.9. *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 an continuous variable). Figure 16.3 shows the Grambsch and Therneau (1994) plots which look fairly flat, but with such a large sample, all slopes are significantly different from zero, and the global test has p-value $\approx 5.66 \times 10^{-11}$. The slice survival plot in Figure 16.4 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.

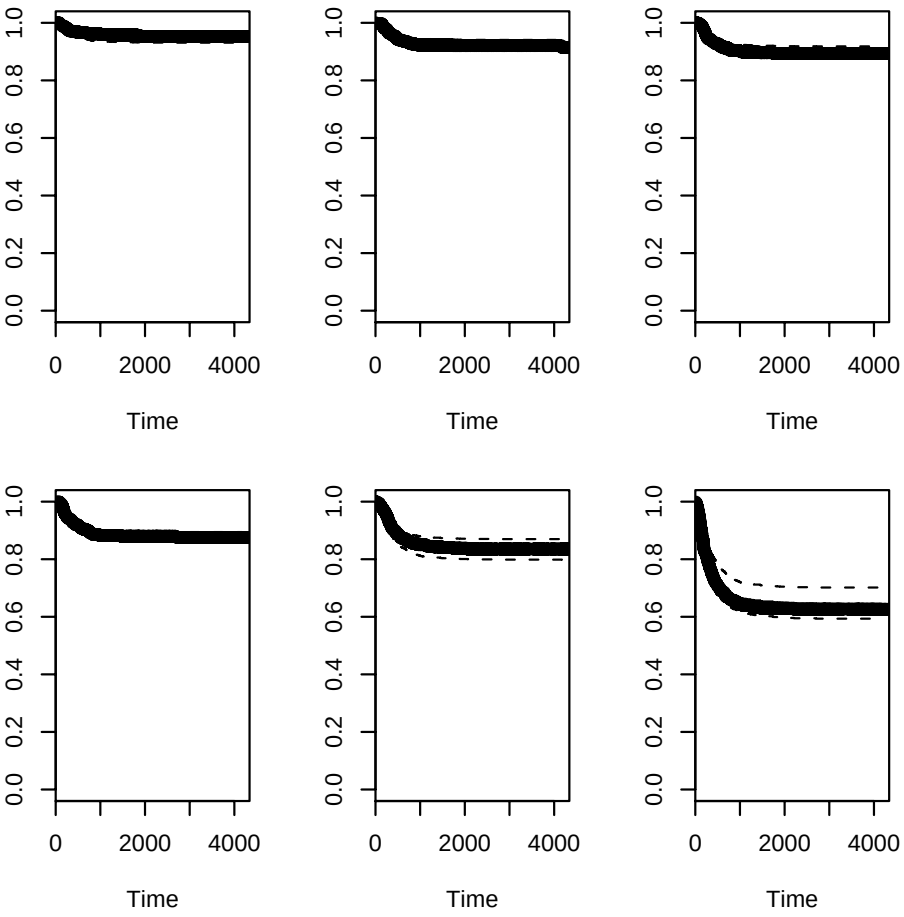


Figure 16.4: Slice Survival Plot for NWTCO Data: Horizontal Axis is the Estimated Survival Function S(t)

16.2.2 Testing and Variable Selection

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. Sometimes "Std. Err." replaces "SE."

The estimated sufficient predictor $\mathbf{ESP} = \hat{\boldsymbol{\beta}}'\boldsymbol{x}_j = \sum_{i=1}^{p} \hat{\beta}_i x_{ij}$. 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. 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.

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 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 that X_j is not needed in the PH survival model given that the other $p-1$ predictors are in the model. Note that X_j could be a very useful PH survival predictor, but 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 $\boldsymbol{\beta} = \boldsymbol{0}$. If this is the case, then the predictors are not needed in the PH model (so survival times $Y \perp\!\!\!\perp \boldsymbol{x}$). If $H_o : \boldsymbol{\beta} = \boldsymbol{0}$ is not rejected, then the Kaplan Meier estimator should be used. If H_o is rejected, use the PH model.

Know that the 4 step **PLRT** is
i) $H_o : \boldsymbol{\beta} = \boldsymbol{0} \quad H_A : \boldsymbol{\beta} \neq \boldsymbol{0}$
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 PH survival relationship between Y and the predictors $\boldsymbol{x}$. If p–value $\geq \delta$, then fail to

reject H_o and conclude that there is not a PH survival relationship between Y and the predictors $\boldsymbol{x}$.

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.

```
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).

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	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	$X^2_{o,1} = z^2_{o,1}$	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: $\boldsymbol{\beta}_R = \mathbf{0}$

Know that 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_o and conclude that the reduced model is good.

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).

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 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 $\max(x_i)/\min(x_i) > 10$.

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).

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

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.

Given a list of submodels along with the number of predictors and AIC, be able to find the "best starting submodel" I_o. Let I_{min} be the minimum AIC model. Then I_o is the submodel with the fewest predictors such that $AIC(I_o) \leq AIC(I_{min}) + 2$ (for a given number of predictors r_I, only consider the submodel with the smallest AIC). Also look at models I_j with fewer predictors than I_o such that $AIC(I_j) \leq AIC(I_{min}) + 7$.

Submodels I with more predictors than I_{min} 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 the min AIC model I_{min}.
ii) The slice survival plots for I looks like the slice survival plot for the full model.
iii) corr(ESP,ESP(I))≥ 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.

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 the "best starting submodel."
Tips: i) submodels with more predictors then the min(AIC) submodel have too many predictors.
ii) The best starting submodel I_o has $AIC(I_o) \leq AIC(I_{min}) + 2$.
iii) Submodels I with $AIC(I) > AIC(I_{min}) + 2$ are not the best starting

submodel.
iv) Submodels I with a pvalue < 0.01 for the change in PLR test have too few predictors.
v) The full model may be the best starting submodel if it is the min(AIC) model and M2–M5 satisfy iii). Similarly, then min(AIC) model may be the best starting submodel.

In addition to the best starting submodel I_o, submodels I with fewer predictors than I_o and $AIC(I) \leq AIC(I_{min}) + 7$ are worth considering.

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 $t(x_j)$ (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.

16.3 Weibull and Exponential Regression

Definition 16.8. 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 depends on k unknown parameters but does not depend on the predictors $\boldsymbol{x}$. The survival function is

$$S_{\boldsymbol{x}}(t) \equiv S_{Y|SP}(t) = [S_{0,P}(t)]^{\exp(\boldsymbol{\beta}_P^T \boldsymbol{x})} = [S_{0,P}(t)]^{\exp(SP)}, \tag{16.3}$$

and

$$\hat{S}_{\boldsymbol{x}}(t) = [\hat{S}_{0,P}(t)]^{\exp(\hat{\boldsymbol{\beta}}_P^T \boldsymbol{x})} = [\hat{S}_{0,P}(t)]^{\exp(ESP)}. \tag{16.4}$$

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 $E(Y) = 1/\lambda$. Now V has a smallest extreme value distribution, $V \sim SEV(\theta, \sigma)$, if

$$S_V(v) = P(V > t) = \exp\left(-\exp\left(\frac{v-\theta}{\sigma}\right)\right)$$

where $\sigma > 0$ while v 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 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 16.3.

Definition 16.9. The **Weibull proportional hazards regression (WPH) model** or **Weibull regression model** is a parametric proportional hazards model with $Y \sim W(\gamma = 1/\sigma, \lambda_{\boldsymbol{x}})$ 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 $Y|\boldsymbol{x} \sim W(1, \lambda_{\boldsymbol{x}}) \sim EXP(\lambda_{\boldsymbol{x}})$.

Definition 16.10. 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 (2003, 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 16.11. 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

$$ESPE = -\hat{\boldsymbol{\beta}}_A^T \boldsymbol{x} \text{ versus } \text{ESPC} = \hat{\boldsymbol{\beta}}_{\text{C}}^{\text{T}} \boldsymbol{x}.$$

For the Weibull regression model, $\boldsymbol{\beta}_C = -\boldsymbol{\beta}_A/\sigma$, and the Weibull EE plot is a plot of

$$ESPW = \frac{-1}{\hat{\sigma}} \hat{\boldsymbol{\beta}}_A^T \boldsymbol{x} \text{ versus } \text{ESPC} = \hat{\boldsymbol{\beta}}_{\text{C}}^{\text{T}} \boldsymbol{x}.$$

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 (16.4) as crosses. If the parametric proportional hazards model holds, then (16.1) = (16.3). Thus if (16.2) $\approx$ (16.4) for any

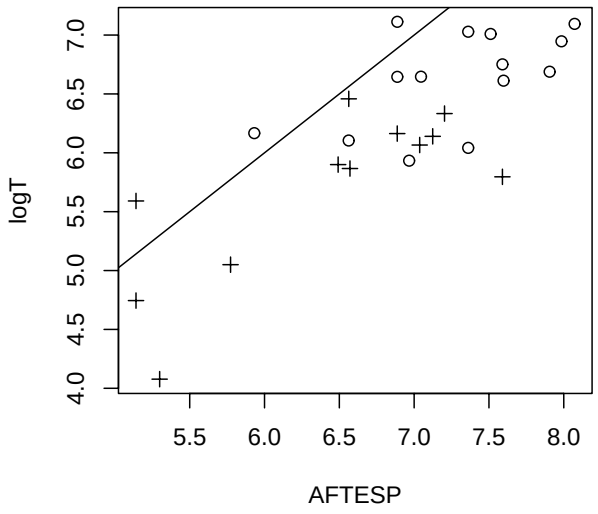


Figure 16.5: LCR Plot for Ovarian Cancer Data

$\boldsymbol{x}_i$, then $S_{0,P}(t) \approx S_0(t)$, (16.2) $\approx$ (16.4) for all $\boldsymbol{x}_i$, and the parametric proportional hazards model is reasonable.

Thus checking parametric proportional hazards models has 3 steps: i) check that the proportional hazards assumption is reasonable 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 (16.2) $\approx$ (16.4) for the $\boldsymbol{x}$ used in each of the J slices.

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 plot for the Cox model is used because of the ease of making pointwise CI bands.

Example 16.10. The ovarian cancer data is from Collett (2003, 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 adri-

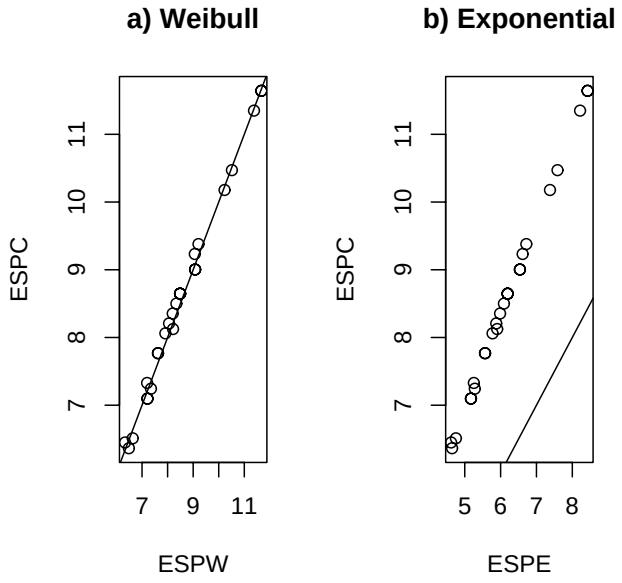


Figure 16.6: EE Plots for Ovarian Cancer Data

amycin). Figure 16.5 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 16.6 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 16.7 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 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.

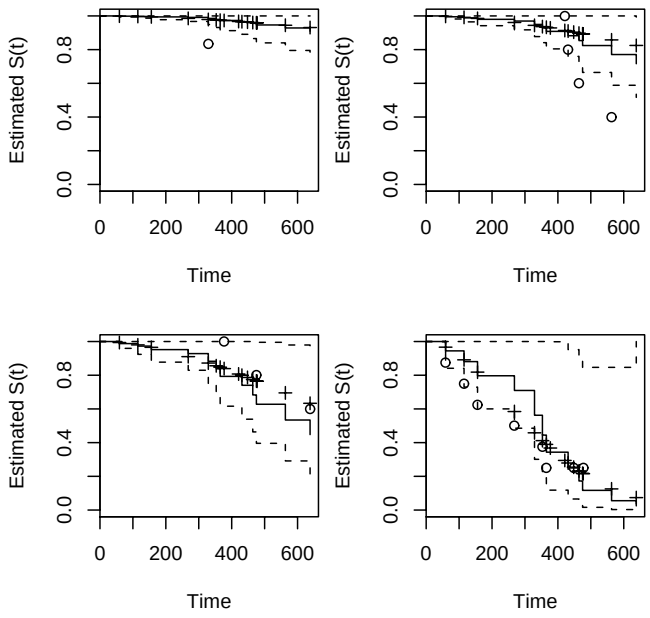


Figure 16.7: Slice Survival Plots for Ovarian Cancer Data

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 Weibull shape	log scale or scale				

For SAS only.
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 $= X^2_{o,j}$ while p and "pr $>$ chisqu" are both p-values.

16.4 Accelerated Failure Time Models

Definition 16.12. For a parametric *accelerated failure time* model,

$$\log(Y_i) = \alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}_i + \sigma e_i \tag{16.5}$$

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, $h_i(t) = e^{-SP} h_o(t/e^{SP})$ and $S_i(t) = S_0(t/\exp(SP))$. 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 LCR plot of Definition 16.10 is still useful for finding influential cases for AFT models. If the Weibull PH regression model holds for Y_i, then $\log(Y_i) = \alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}_i + e_i$ 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. Thus 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 $\hat{\alpha} + \hat{\boldsymbol{\beta}}_A^T \boldsymbol{x} = w$ are approximately a censored sample from an $SEV(w, \hat{\sigma})$ distribution if the fitted model is a good approximation to the data.

Censoring causes the bulk of the data to be below the identity line in the LCR plot. For example, Hosmer and Lemeshow (1998, 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 $e_i \sim SEV(0, 1)$.

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 16.9, $\lambda_0 = \exp(-\alpha/\sigma)$ 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 $\log(Y_i)|bx_i$ 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 16.10.

Similarly, R and SAS Weibull AFT 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. Then transform the PH model to a Weibull AFT.

In addition to the Weibull and Exponential AFTs, there are lognormal and loglogistic AFT models. If the $Y_i|\boldsymbol{x}_i$ are lognormal, the e_i are normal. If the $Y_i|\boldsymbol{x}_i$ are loglogistic, the e_i are logistic. The loglogistic and lognormal AFT models are not PH models. The loglogistic AFT is a proportional odds model.

Inference for the AFT model is performed exactly in the same way as for the WPH = Weibull AFT. See points Section 16.2. But the conclusions change slightly if the AFT is not the Weibull AFT. Change (if necessary) "Weibull survival model" to the appropriate model, eg "lognormal survival model". Change (if necessary) "WPH" to the appropriate model, eg "lognormal 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 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) 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 X_j is not needed in the Weibull survival model given that the other $p-1$ predictors are in the model.

Know how to do the 4 step likelihood ratio test **LRT**:

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

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_o and conclude that there is not a WPH survival relationship between Y and the predictors $\boldsymbol{x}$.

Know how to do the 4 step **change in LR test**:

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_o and conclude that the reduced model is good.

16.5 Stratified Proportional Hazards Regression

Definition 16.12. 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.

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

16.6 Summary

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

Then the **distribution function** (df) $F(t) = P(Y \leq t)$. 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 $Y \sim$ Weibull(λ, γ) 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) $= a/n =$

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

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

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

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

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 $\delta_i = I(Y_i \leq Z_i)$ so $\delta_i = 1$ if T_i is uncensored and $\delta_i = 0$ if T_i is censored. Let $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ 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 $d_i \geq 2$, then there are **ties**.

12) Let $n_i = \sum_{j=1}^{n} I(t_{(j)} \geq t_i) = \#$ 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 $t_{i-1} \leq t < t_i$ 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 $t > t_{(n)}$.

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) 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.

20) 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.

21) 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.

22) The Wald confidence interval (CI) for β_j can also be obtained from

the output: the large sample 95% CI for β_j is

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

23) 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 that X_j is not needed in the PH survival model given that the other $p-1$ predictors are in the model. Note that X_j could be a very useful PH survival predictor, but 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 $\boldsymbol{\beta} = \boldsymbol{0}$. If this is the case, then the predictors are not needed in the PH model (so survival times $Y \perp\!\!\!\perp \boldsymbol{x}$). If $H_o : \boldsymbol{\beta} = \boldsymbol{0}$ is not rejected, then the Kaplan Meier estimator should be used. If H_o is rejected, use the PH model.

24) The 4 step **PLRT** is
i) $H_o : \boldsymbol{\beta} = \boldsymbol{0} \quad H_A : \boldsymbol{\beta} \neq \boldsymbol{0}$
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 PH survival relationship between Y and the predictors $\boldsymbol{x}$. If p–value $\geq \delta$, then fail to reject H_o and conclude that there is not a PH survival relationship between Y and the predictors $\boldsymbol{x}$.

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

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

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

25) 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_o and conclude that the reduced model is good.

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).

26) 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.

27) 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).

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

29) 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.

30) Given a list of submodels along with the number of predictors and AIC, be able to find the "best starting submodel" I_o. Let I_{min} be the minimum AIC model. Then I_o is the submodel with the fewest predictors such that $AIC(I_o) \leq AIC(I_{min}) + 2$ (for a given number of predictors r_I, only consider the submodel with the smallest AIC). Also look at models I_j with fewer predictors than I_o such that $AIC(I_j) \leq AIC(I_{min}) + 7$.

31) Submodels I with more predictors than I_{min} should not be used.

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

33) 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.

34) 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 largest ESP in the 1st $J-1$ groups and the $\boldsymbol{x}$ corresponding to the smallest ESP in the Jth 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.

35) 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 the min AIC model I_{min}.
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 $\geq$ 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.

36) 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 the "best starting submodel."
Tips: i) submodels with more predictors then the min(AIC) submodel have too many predictors.
ii) The best starting submodel I_o has $AIC(I_o) \leq AIC(I_{min}) + 2$.
iii) Submodels I with $AIC(I) > AIC(I_{min}) + 2$ are not the best starting submodel.
iv) Submodels I with a pvalue < 0.01 for the change in PLR test have too few predictors.
v) The full model may be the best starting submodel if it is the min(AIC) model and M2–M5 satisfy iii). Similarly, then min(AIC) model may be the best starting submodel.

37) In addition to the best starting submodel I_o, submodels I with fewer predictors than I_o and $AIC(I) \leq AIC(I_{min}) + 7$ are worth considering.

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.

38) 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.

39) 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.

40) 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 $h_0(t) = h_0(t|\boldsymbol{\theta}) = \lambda\gamma t^{\gamma-1}$ 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: $\beta_1 = 0$
$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
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.

41) Instead of fitting the WHP model of 40), R and SAS fit an accelerated failure time model $\log(Y_i) = \alpha + \boldsymbol{\beta}'\boldsymbol{x}_i + \sigma\epsilon_i$ where $\text{Var}(\epsilon_i) = 1$ and the ϵ_i are iid from a smallest extreme value distribution. Also $\boldsymbol{\beta} \neq \boldsymbol{\beta}_W$ from 40).

$\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ are MLEs found from the censored data $(T_i, \delta_i, \boldsymbol{x}_i)$ not from $(Y_i, \boldsymbol{x}_i)$.

42) 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.

43) Given $\hat{\boldsymbol{\beta}}$ from output and given $\boldsymbol{x}$, be able to find ESP $= \hat{\boldsymbol{\beta}}'\boldsymbol{x} = \sum_{i=1}^{p} \hat{\beta}_i x_i = \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$.

44) A large sample 95% CI for β_j is $\hat{\beta}_j \pm 1.96\ se(\hat{\beta}_j)$.

45) **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) 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 X_j is not needed in the Weibull survival model given that the other $p-1$ predictors are in the model.

46) 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_o and conclude that there is not a WPH survival relationship between Y and the predictors $\boldsymbol{x}$.

47) 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.

48) 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.

49) 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.

50) 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.

51) For the AFT model, $h_i(t) = e^{-SP} h_o(t/e^{SP})$ and $S_i(t) = S_0(t/\exp(SP))$.

52) Inference for the AFT model is performed exactly in the same way as for the WPH = 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.

53) 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 $n \to \infty$.

54) 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 $ESPC = \hat{\boldsymbol{\beta}}_C' \boldsymbol{x}$.

55) For the Weibull regression model, $\sigma = 1$, and $\boldsymbol{\beta}_C = -\boldsymbol{\beta}_A/\sigma$. The Weibull EE plot is a plot of

$$-ESPW/\hat{\sigma} = -\frac{1}{\hat{\sigma}} \hat{\boldsymbol{\beta}}_A' \boldsymbol{x} \text{ vs ESPC} = \hat{\boldsymbol{\beta}}_C' \boldsymbol{x}.$$

56) 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.

57) Inference is done exactly as for the PH model. See points 21), 22), 23), 24), and 25). Except the conclusion is changed slightly: in 23) and 24) replace "PH" by "SPH".

16.7 Complements

Excellent texts on survival analysis include Allison (1995), Collett (2003), Klein and Moeschberger (1998), Kleinbaum and Klein (2005b), Hosmer and Lemeshow (1999) and Smith (2002). Graduate level texts include Kalbfleisch and Prentice (2002) and Lawless (2002). A review is given by Freedman (2008). Oakes (2000) notes that the proportional hazards model is not preserved when variables are added or deleted from the model, eg by variable selection.

From the `CRAN` website, eg (www.stathy.com/cran/), 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/STAT User's Guide (1999). The chapters on PHREG, LIFEREG and LIFETEST procedures are useful. These chapters can be found on line 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 and EE plots are due to Olive (2009c). 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 survival 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. See 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 *regpack* available from (www.math.siu.edu/olive/regpack.txt) contains functions for reproducing simulations and some of the plots. The functions `vlung2`, `vovar` and `vnwtco` were used to produce plots in Examples 1, 2 and 3. 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.

16.8 Problems

Problems with an asterisk * are especially important.

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

16.2. 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.

16.3. 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 $t > 0$.

16.4. 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) $H(t)$

16.5. 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 $F(t)$.

16.6. 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) $H(t)$

16.7. Eleven death times from Collett (2003, 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 16.3, 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 $\hat{S}_E(t)$.

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

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

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

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

16.12. 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			$\hat{S}_E(109) =$

16.13. 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

16.14. 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 $S_Y(34)$?

16.15. 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			

16.16. 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 (2003, 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				

16.17. 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			

16.18. 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			$\hat{S}_K(45) =$

16.19. 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

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

16.20. 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 16.21

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

16.21. The Collett (2003, 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 16.22
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
```

16.22. 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 16.23
             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
```

16.23. 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 16.24.
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
```

16.24. The *R* `veteran` lung cancer data has Y = survival time. The predictors are *trt* (1=standard, 2=test), the factor *celltype* (1=squamous, 2=smallcell, 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 16.25
 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
```

16.25. 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.

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

16.26. 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 best starting submodel I_0?

c) Are there any other 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

16.27. 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 model should be considered as the first starting submodel I_0? Explain briefly why each of the other 3 submodels should not be used as the starting submodel.

16.28. 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.

16.29. 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.

16.30. 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.

16.31. 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 16.31 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 16.32
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
```

16.32. Above is output from a Weibull Regression for the Allison (1995, p. 270) recidivism data described in problem 16.31. 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 16.33
               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)
```

16.33. 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 16.34, 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
```

16.34. 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.

16.35. 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 $Ho : \boldsymbol{\beta} = \mathbf{0}$.

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) Double click on the *Math Progs* icon and after a window appears, double click on the *SAS* icon. If your computer does not have SAS, go to another computer.

ii) A window should appear with 3 icons. Double click on *The SAS System for ....*

iii) Like Minitab a window with a split screen will open. The top screen says
Log-(Untitled) while the bottom screen says *Editor-Untitled1*. Press the spacebar and an asterisk appears: *Editor-Untitled1**.

iv) Go to the webpage (www.math.siu.edu/olive/reghw.txt) to copy and paste the program for Problem 16.36 into *Notepad*. The *ls* stands for linesize so *l* is a lowercase *L*, not the number one. Save your file as **h16d36.sas** on your diskette (A: drive). (On the top menu of the editor, use the commands "File > Save as". A window will appear. Use the upper right arrow to locate "31/2 Floppy A" and then type the file name in the bottom box. Click on OK.)

v) Get back into SAS, and from the top menu, use the "File> Open" command. A window will open. Use the arrow in the NE corner of the window to navigate to "31/2 Floppy(A:)". (As you click on the arrow, you should see My Documents, C: etc, then 31/2 Floppy(A:).) Double click on **hw16d36.sas**.

(Alternatively cut and paste the program into the SAS editor window.) To execute the program, use the top menu commands "Run>Submit". An output window will appear if successful.

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.

vi) 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".

You may want to save your SAS output as the file **hw16d36.doc**

vii) 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.

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

16.36. 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. See Allison (1995, p. 49-50).

a) Do i) through v) above. But instead of vi), click on the SAS output, then click on the printer icon. This will produce 2 pages of output. Then click on the graph of the survival function and click on the printer icon.

Include these 3 pages of output as part of your homework.

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.

16.37. 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 (www.math.siu.edu/olive/reghw.txt). Obtain the output from the program in the same manner as i) through v) above Problem 16.36.

a) But instead of vi), click on the SAS output, then click on the printer icon. This will produce 3 pages of output (perhaps). Then click on the graph of the survival function and click on the printer icon.

Include these 4 pages of output as part of your homework.

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 $S_Y(13)$.

e) Check this 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) SAS does not compute a hazard estimator for the KM estimator, but from the plot of $\hat{S}_K(t)$, briefly explain survival for days 0–250 and for days 250–2250.

16.38. 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 (www.math.siu.edu/olive/reghw.txt). Obtain the output from the program in the same manner as i) through vi) above Problem 16.36.

But instead of vi), click on the SAS output, then click on the printer icon. This will produce 1 page of output.

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).)

16.39. Data is from SAS/STAT User's Guide (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 (www.math.siu.edu/olive/reghw.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?

16.40. 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) This is a large data file. SAS needs an "end of file" marker to determine when the data ends. SAS uses a period as the end of file marker, and the period has already been added to the file. Obtain the file from (www.math.siu.edu/olive/recid.txt) and save the file as *recid.txt* using the

commands "File>Save as." A window will appear, in the top box make *3 1/2 Floppy (A:)* appear while in the *File name* box type *recid.txt*.

b) Obtain the SAS program for this problem from (www.math.siu.edu/olive/reghw.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 A drive, then you need to change** the *infile* command in the SAS code to the drive that you are using, eg change *infile "a:redic.txt";* to *infile "f:recid.txt";* if you are using 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.

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

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

a) Obtain the SAS program for this problem from (www.math.siu.edu/olive/reghw.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.)

16.42. Copy and paste commands for this problem from
(www.math.siu.edu/olive/reghw.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

R is the free version of *Splus*. The website (www.stat.umn.edu) has a link under the icon *Rweb*. The icon *other links* has the link **Cran** that gives *R* support. Click on the *Rgui* icon to get into *R*. Then typing *q()* gets you out of *R*.

16.43. 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 16.3, 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 (www.math.siu.edu/olive/reghw.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 $\hat{S}_K(23) \pm 1.96 SE(\hat{S}_K(23))$.

16.44. Copy and paste commands for parts a) and b) for this problem from (www.math.siu.edu/olive/reghw.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?

16.45. Go to (www.math.siu.edu/olive/regpack.txt) and cut and paste the program `kmsim2` into *R*. a) 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 $t_{(n)}$.

16.46. 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 (www.math.siu.edu/olive/reghw.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 16.47 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 p is 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 $Ho : \beta_1 = 0$.

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

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

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

Likelihood ratio test=14.3
```

16.47. Use the output above which is for the same data as in 16.46 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).)

16.48. 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 (www.math.siu.edu/olive/reghw.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.

16.49. Go to (www.math.siu.edu/olive/regpack.txt) and cut and paste the program `bphgfit` into *R*.

Alternatively, suppose that you download *regpack.txt* onto a disk. (Use *File* and *Save Page as*.) Enter *R* and wait for the curser to appear. Then go to the *File* menu and drag down *Source R Code*. A window should appear. Navigate the *Look in* box until it says *3 1/2 Floppy(A:)*. In the *Files of*

type box choose *All files(*.*)* and then select *regpack.txt*. The following line should appear in the main R window.

```
> source("A:/regpack.txt")
```

a) Copy and paste commands for this problem from (www.math.siu.edu/olive/reghw.txt) into R. Copy and paste the output into *Word*.

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, eg 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 $t_A > 0$.)

16.50. 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 16.48 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.

An estimated sufficient summary plot (ESSP) is a plot of the ESP $= \hat{\boldsymbol{\beta}}' \boldsymbol{x}$ 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 (www.math.siu.edu/olive/reghw.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 16.50b.

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) Go to (www.math.siu.edu/olive/regpack.txt) and cut and paste the program `vlung2` into *R*. Type the following two commands and include the resulting plot in *Word*.

```
vlung2(1)
title("males")
```

d) Type the following two commands 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?

16.51. The lung cancer data is the same as that described in 16.48, 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 (www.math.siu.edu/olive/reghw.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 (www.math.siu.edu/olive/RMLRhw.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.

16.52. Copy and paste the programs from (www.math.siu.edu/olive/regpack.txt) into *R*.

Alternatively, suppose that you download *regpack.txt* onto a disk. (Use *File* and *Save Page as*.) Enter *R* and wait for the curser to appear. Then go to the *File* menu and drag down *Source R Code*. A window should appear. Navigate the *Look in* box until it says *3 1/2 Floppy(A:)*. In the *Files of type* box choose *All files(*.*)* and then select *regpack.txt*. The following line should appear in the main *R* window.

```
> source("A:/regpack.txt")
```

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 "phsim(k=5)" 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=100,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 not very far outside the pointwise CI bands. What value of n do you get?

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

a) Obtain the R code for this problem from (www.math.siu.edu/olive/reghw.txt). Click on the left screen then hit *Enter*. Copy and paste both the output into *Word*. 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?

16.54. Copy and paste the programs `phdata`, `weyp` and `wregsim` from (www.math.siu.edu/olive/regpack.txt) into R (or download `regpack` on a disk and use the source command as in Problem 16.52).

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 `wregsim` generates Weibull proportional hazards regression data with baseline hazard function $h_0(t) = kt^{k-1}$.

a) Type the command *wregsim(k=1)* 5 times (or use the "up arrow" after typing the command once). This gives 5 simulated Weibull regression data sets with $k = 1$. Hence the Weibull regression is also an exponential regression. Include the last plot in *Word*.

b) Type the command *wregsim(k=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 $Y = W^{1/k}$

where $W \sim EXP(\lambda = \exp(SP))$ where $E(W) = 1/\lambda$. $Z \sim EXP(.1)$ 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?

c) Type the command *wregsim(k=0.5)* 5 times. (Now censored cases tend to be large because time $Y = W^{1/k} = W^2$ where $W \sim EXP(\lambda)$. $Z \sim EXP(.1)$ 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)?

16.55. This problem considers the ovarian data from Collett (2003, p. 189, 344-346).

a) Obtain the *R* code for this problem from (www.math.siu.edu/olive/reghw.txt). Copy and paste the plot into *Word*.

b) Now obtain the R code for this problem and put the plot into *Word*.

c) Can the Exponential regression model be used or should the more complicated Weibull regression model be used?

16.56. Copy and paste the programs `phdata` and `wregsim2` from (www.math.siu.edu/olive/regpack.txt) into *R* (or download `regpack` on a disk and use the source command as in 16.52).

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 $h_0(t) = kt^{k-1}$.

a) Type the command *wregsim2(n=10, k=1)* 5 times (or use the "up arrow" after typing the command once). This gives 5 simulated Weibull regression data sets with $k = 1$. 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, k=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, k=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?

16.57. Copy and paste the programs `phdata` and `wregsim3` from (www.math.siu.edu/olive/regpack.txt) into R (or download `regpack` on a disk and use the source command as in 16.52).

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 $h_0(t) = kt^{k-1}$. This is also an AFT model with $\alpha = 0$, $\boldsymbol{\beta}' = -(1/k, ..., 1/k)$ and $\sigma = 1/k$. 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 $\approx -(1/k, ..., 1/k)$ and mnscale $\approx 1/k$.

a) Make a table (by hand) with headers

```
n   k   mnint  mncoef                    mnscale
```

Fill in the table for $n = 20, k = 1; n = 100, k = 1; n = 200, k = 1; n = 20, k = 5; n = 100, k = 5; n = 200, k = 5; n = 20, k = 0.5; n = 100, k = 0.5; n = 200, k = 0.5$ by using the commands *wregsim3(n=20, k=1)*, ..., *wregsim3(n=200, k=0.5)*.

b) Are the estimators close to parameters $\alpha, \boldsymbol{\beta}$ and σ for $n = 20$? How about for $n = 100$?

16.58. Copy and paste the programs `wphsim` and `swhat` from (www.math.siu.edu/olive/regpack.txt) into R (or download `regpack` on a disk and use the source command as in 16.52). 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*.

16.59. 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 for this problem from (www.math.siu.edu/olive/reghw.txt) 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?