Chapter 2
Cox Proportional Hazards Regression

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

Definition 2.1. A **case** or **observation** consists of k random variables measured for one person or thing. The ith case $\boldsymbol{z}_i = (z_{i1}, ..., z_{ik})^T$. The **training data** consists of $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$. A statistical model or method is fit (trained) on the training data. The **test data** consists of $\boldsymbol{z}_{n+1}, ..., \boldsymbol{z}_{n+m}$, and the test data is often used to evaluate the quality of the fitted model.

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

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

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

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

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

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

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

2.1 Proportional Hazards Regression

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

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

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

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

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

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

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

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

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

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

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

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

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

The survival function is

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

If $\boldsymbol{x} = \boldsymbol{0}$ 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

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

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

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

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

Then the hazard ratio

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

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

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

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

2.2 Visualizing the Cox PH Regression Model

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

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

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

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

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

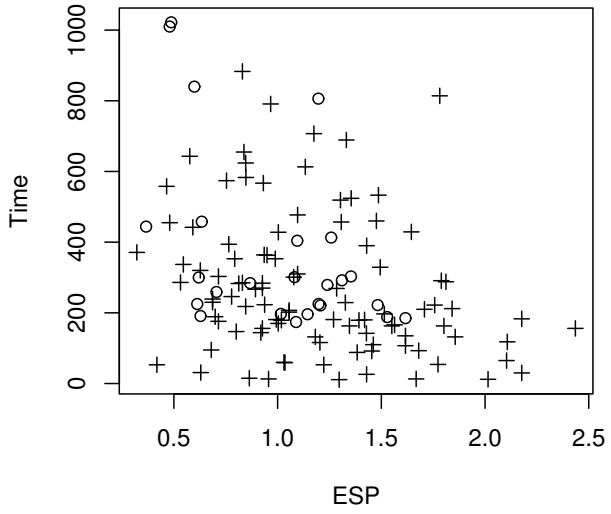

Fig. 2.1 Censored Response Plot for R Lung Cancer Data

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

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

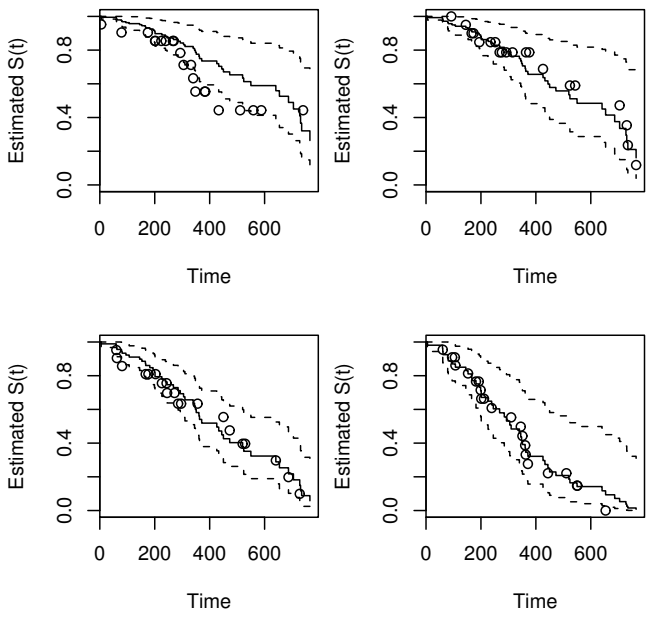

Fig. 2.2 Slice Survival Plots for R Lung Cancer Data

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

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

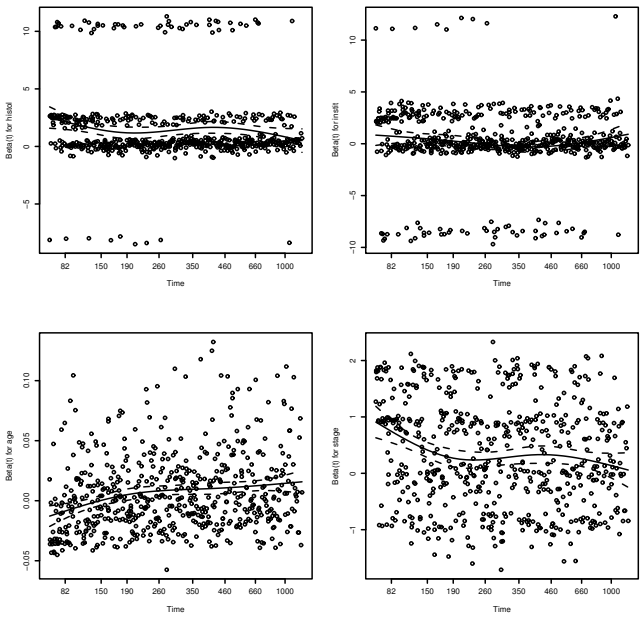

Fig. 2.3 Grambsch and Therneau Plots for NWTCO Data

curves in that the circles do not fall very far outside the pointwise CI bands.

Example 2.2. *R* contains a data set *nwtco* where the response variable Y is the time until relapse with $n = 4028$. The model used predictors *histol* = tumor histology from central lab, *instit* = tumor histology from local institution, *age* in months, and *stage* of disease from 1 to 4 (treated as a continuous variable). In Figure 2.3, the Grambsch and Therneau (1994) plots suggest that the Cox model is not valid since not all of the loess curves are flat, and the global test has p-value $\approx 5.66 \times 10^{-11}$. The slice survival plot in Figure 2.4 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. The greatest contributors to lack of fit seem to be the predictors age and stage corresponding to the bottom two plots of Figure 2.3, and survival for small ESP corresponding to the upper left plot in Figure 2.4.

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

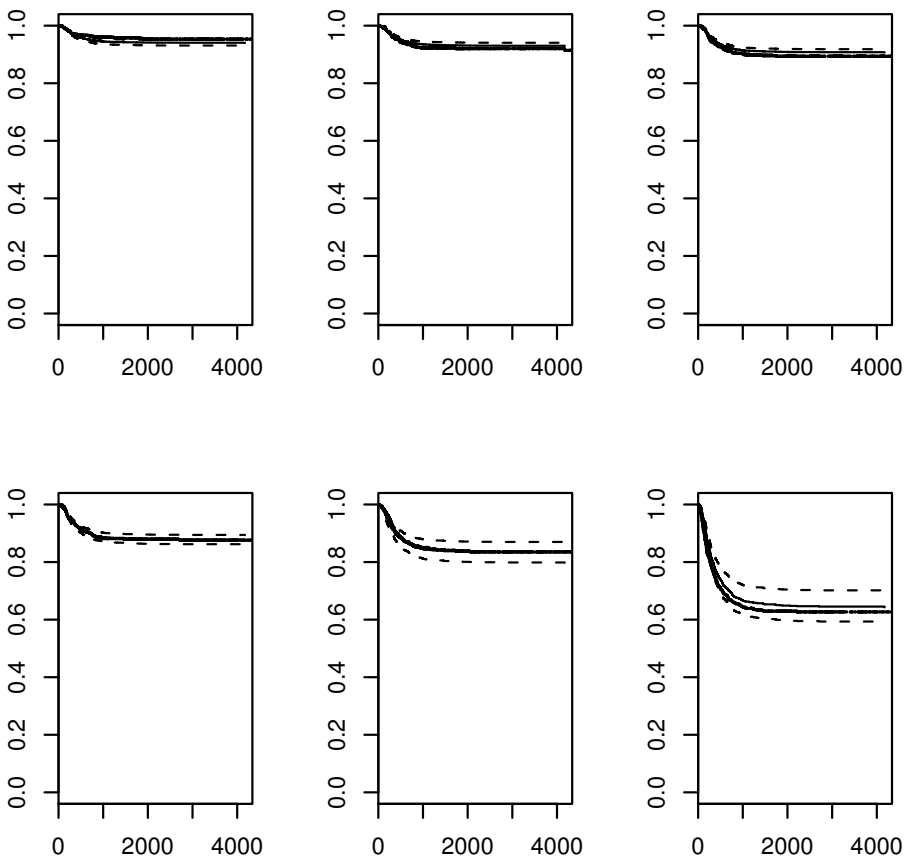

Fig. 2.4 Slice Survival Plot for NWTCO Data: Horizontal Axis is the Estimated Survival Function S(t)

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

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

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

2.3 Testing

For regression models, we want to test i) whether the predictors $\boldsymbol{x}$ are needed in the model: $H_0 : \boldsymbol{\beta} = \mathbf{0}$ versus $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$, ii) whether a reduced model that that does not use predictors $x_{i_1}, ..., x_{ik}$ is good: $H_0 : (\beta_{i1}, ..., \beta_{ik})^T = \mathbf{0}$ versus $H_1 : (\beta_{i1}, ..., \beta_{ik})^T \neq \mathbf{0}$, and iii) whether predictor x_i is needed in the model given that the other predictors are needed in the model $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0$. Note that tests i) and iii) are special cases of test ii). We also want confidence intervals for β_i. We also want to find $ESP = \hat{\boldsymbol{\beta}}_C^T \boldsymbol{x}_i$ and $\hat{h}_i(t) = e^{ESP} \hat{h}_0(t)$ given $\boldsymbol{x}_i$. Often the hypothesis $H_1 = H_A$.

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

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

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

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

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

The estimated sufficient predictor $\mathbf{ESP} = \hat{\boldsymbol{\beta}}' \boldsymbol{x}_j = \sum_{i=1}^{p} \hat{\beta}_i x_{ij}$. 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. The hypothesis H_0 is rejected if the p–value $< \delta$. A p–value between 0.07 and 1.0 provides little evidence that H_0 should be rejected, a p–value between 0.01 and 0.07 provides moderate evidence and a p–value less than 0.01 provides strong statistical evidence that H_0 should be rejected. Statistical evidence is not necessarily practical evidence, and reporting the p–value along with a statement of the strength of the evidence is more informative than stating that the p–value is

less than some chosen value such as $\delta = 0.05$. Nevertheless, as a **homework convention**, use $\delta = 0.05$ if δ is not given.

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

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

Investigators often test whether a predictor x_j is needed in the model given that the other $p-1$ predictors $x_1, ..., x_{i-1}, x_{i+1}, ..., x_p$ are in the model with a **4 step Wald test of hypotheses**:
i) State the hypotheses $H_0 : \beta_j = 0$ $H_1 : \beta_j \neq 0$.
ii) Find the test statistic $z_{0,j} = \hat{\beta}_j / se(\hat{\beta}_j)$ or $X^2_{0,j} = z^2_{0,j}$ or obtain it from output.
iii) The p–value $= 2P(Z < -|z_{0j}|) = P(\chi^2_1 > X^2_{0,j})$. Find the p–value from output or use the standard normal table.
iv) State whether you reject Ho or fail to reject Ho and give a nontechnical sentence restating your conclusion in terms of the story problem.

If Ho is rejected, then conclude that x_j is needed in the PH survival model given that the other $p-1$ predictors are in the model. If you fail to reject Ho, then conclude that 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.

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

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

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

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

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

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

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

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

ii)

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

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

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

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

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

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

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

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

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

Know that the 4 step **PLRT** is

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

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

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

iv) Reject H_0 if the p–value $< \delta$ and conclude that there is a PH survival relationship between Y and the predictors $\boldsymbol{x}$. If p–value $\geq \delta$, then fail to reject H_0 and conclude that there is not a PH survival relationship between Y and the predictors $\boldsymbol{x}$.

Output in symbols is often given in three ways.

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

or

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

or

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

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

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

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

Example 2.4.

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

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

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

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

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

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

iii) $pval = 0.00959$

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

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

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

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

iii) pval $= 0.25$

iv) Fail to reject H_0: there is not a PH survival relationship between relapse times and x (so no difference between treatments A and B for survival times).

Let the **full model** be

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

let the **reduced model**

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

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

Assume that the full model is useful. Then we want to test H_0: the reduced model is good (can be used instead of the full model, so $\boldsymbol{x}_O$ 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	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{0,1} = \hat{\beta}_1 / se(\hat{\beta}_1)$	$X^2_{0,1} = z^2_{0,1}$	$H_0 : \beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$z_{0,p} = \hat{\beta}_p / se(\hat{\beta}_p)$	$X^2_{0,p} = z^2_{0,p}$	Ho $H_0 : \beta_p = 0$

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

`SAS: Testing Global Null Hypotheses: BETA = 0`

Test	Chi-Square	DF	Pr > Chisq
Likelihood ratio	$X^2(N\|F)$	p	pval for $H_0 : \boldsymbol{\beta} = \mathbf{0}$

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

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

SAS: Testing Global Null Hypotheses: BETA = 0			
Test	Chi-Square	DF	Pr > Chisq
Likelihood ratio	$X^2(N\|R)$	r	pval for Ho: $\boldsymbol{\beta}_R = \mathbf{0}$

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

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

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

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

iv) Reject H_0 if the p–value $< \delta$ and conclude that the full model should be used. If p–value $\geq \delta$, then fail to reject H_0 and conclude that the reduced model is good.

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

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

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

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

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

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

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

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

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

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

iv) Reject H_0, use the full model.

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

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

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

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

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

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

In R, there is a useful shortcut for doing the change in PLR test. In the code below let "fit" be for the full model and "fitR" be for the reduced model.

The anova command gives the following output in symbols. Values left blank are not needed for the test.

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

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

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

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

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

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

Remark 2.1. For testing, $\boldsymbol{\beta} = \mathbf{0}$ means changing values of $\boldsymbol{x}$, within the observed range of $\boldsymbol{x}$ or of $\hat{\boldsymbol{\beta}}^T\boldsymbol{x}$, does not affect survival. For example, suppose $p = 1$ and $x = 1$ for treatment 1 and $x = 0$ for treatment 0. If treatments 1 and 0 are both very and equally effective, then $h_1(t) = h_0(t) = e^{\beta x}h_0(t)$ with $\beta = 0$. For this example, x is important for survival times Y, in that survival could be poor if neither treatment were given, but the value of x 0 or 1 did not affect the value of Y. Hence $\boldsymbol{\beta} = \mathbf{0}$ could imply that the survival relationship between $\boldsymbol{x}$ and Y is the same for all observed values of $\hat{\boldsymbol{\beta}}^T\boldsymbol{x}$. Hence concluding $\boldsymbol{\beta} = \mathbf{0}$ does not necessarily mean that the predictors $\boldsymbol{x}$ are not important for survival times. Similarly, $\beta_i = 0$ means changing values of x_i, within the observed range of x_i, does not affect the survival times. If $\boldsymbol{\beta} = (\boldsymbol{\beta}_R^T, \boldsymbol{\beta}_O^T)^T$, then $\boldsymbol{\beta}_O = \mathbf{0}$ means changing the values of $\boldsymbol{x}_O$, within the observed values of $\boldsymbol{x}_O$ or $\hat{\boldsymbol{\beta}}_O^T\boldsymbol{x}_O$, does not affect the survival times. Then the reduced model is good in that you get the "same survival model" regardless

of the $\boldsymbol{x}_O$ values. So "no survival relationship" between Y and $\boldsymbol{x}$ or $\boldsymbol{x}_O$ or x_i means within the observed range of $\boldsymbol{x}$, or $\boldsymbol{x}_O$, or x_i. This remark for testing applies to the other models in Chapters 2 and 3.

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

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

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

An **interaction** is a product of two or more main effects, but for a factor include products for the $K-1$ indicator variables of the factor. Hence an interaction between a variate x_1 and a factor A with indicator variables $x_{2A}, ..., x_{KA}$ is incorporated into the model with $x_1x_{2A}, ..., x_1x_{KA}$. An interaction between factor A and factor B with indicators $x_{2B}, ..., x_{bB}$ is incorporated into the model with the $(K-1)(b-1)$ pairs

$x_{2A}x_{2B}, ..., x_{2A}x_{bB}$

$\vdots$

$x_{KA}x_{KB}, ..., x_{KA}x_{bB}$.

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

Suppose x_1 is quantitative and x_2 is qualitative with 2 levels and $x_2 = 1$ for level c_2 and $x_2 = 0$ for level c_1. Then a first order model with interaction is $SP = \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2$. This model yields two unrelated lines in the sufficient predictor depending on the value of x_2: $SP = \beta_2 + (\beta_1 + \beta_3)x_1$ if $x_2 = 1$ and $SP = \beta_1x_1$ if $x_2 = 0$. If $\beta_3 = 0$, then there are two parallel lines: $SP = \beta_2 + \beta_1x_1$ if $x_2 = 1$ and $SP = \beta_1x_1$ if $x_2 = 0$. If $\beta_2 = \beta_3 = 0$, then the two lines are coincident: $SP = \beta_1x_1$ for both values of x_2. If $\beta_2 = 0$, then the two lines both have the intercept at the origin: $SP = (\beta_1 + \beta_3)x_1$ if $x_2 = 1$ and $SP = \beta_1x_1$ if $x_2 = 0$. In general, as factors have more levels and interactions have more terms, e.g. $x_1x_2x_3x_4$, the interpretation of the model rapidly becomes very complex.

A **scatterplot** is a plot of x_i versus x_j. A **scatterplot matrix** is an array of scatterplots. It is used to examine the marginal relationships of the

predictors. Variables with outliers, missing values or strong nonlinearities may be so bad that they should not be included in the full model.

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

2.4 Variable Selection

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Example 2.10. Given summaries on several models, be able to pick out the "best starting model" I_I. In the table below, M1 is the full model and

M3 is the minimum AIC model I_{min}. M2 and M2 have more predictors than the minimum AIC model and the AIC for M4 is to large to be the starting model. So use M3 as the starting model.

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

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

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

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

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

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

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

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

```
fit1 <- coxph(Surv(time,status) ~ ., data=alung)
fit2 <- coxph(Surv(time,status) ~ 1, data=alung)
#fit1 <- coxph(Surv(alung[,1],alung[,2]) .,data=alung)
#fails because it uses time and status as predictors
outb<-stepAIC(fit1,direction="backward")
#Imin has perf and ttype
outf<-stepAIC(fit2,direction="forward",scope=
list(upper=fit1,lower=fit2))

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

2.5 Stratified Proportional Hazards Regression

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

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

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

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

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

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

Since J Cox PH regression models are fit for SPH , one for each group, check each Cox PH model with graphs. Another useful method is to divide the ESP $\hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ into k groups where $4 \leq k \leq 9$. Choose an $\boldsymbol{x}_i$ from near the center of each group. Then plot t versus $\hat{S}_{\boldsymbol{x}_i,j}(t)$ for $j = 1, ..., J$ on the same graph for $\boldsymbol{x}_i$. Make such graphs for $\boldsymbol{x}_1, ..., \boldsymbol{x}_k$.

2.6 Generalized Cox Regression

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

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

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

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

Note that

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

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

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

There are two types of time dependent variables. An *internal time dependent variable* is suject specific and requires the subject to be under periodic

observation. An *external time dependent variable* does not require the subject to be under direct observation, and often only needs one initial measurement. For example, if the patient's birthdate is known, then the patient's age can be computed at any time after the patient enters the study.

Example 2.11.

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

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

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

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

2.7 Summary

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5) Investigators also sometimes test whether a predictor X_j is needed in the model given that the other $k-1$ nontrivial predictors are in the model with a **4 step Wald test of hypotheses**:
i) State the hypotheses Ho: $\beta_j = 0$ Ha: $\beta_j \neq 0$.

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

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

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

If Ho is rejected, then conclude that X_j is needed in the PH survival model given that the other $p-1$ predictors are in the model. If you fail to reject Ho, then conclude 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 $\boldsymbol{\beta} = \mathbf{0}$. If this is the case, then the predictors are not needed in the PH model (so survival times $Y \perp\!\!\!\perp \boldsymbol{x}$). If $H_o : \boldsymbol{\beta} = \mathbf{0}$ is not rejected, then the Kaplan Meier estimator should be used. If H_o is rejected, use the PH model.

6) The 4 step **PLRT** is

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

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

iii) The p–value $= P(\chi^2_p > X^2(N|F))$ 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 $X^2(R|F) \equiv X^2(N|F) - X^2(N|R) =$

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

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

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

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

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

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

iv) Reject H_o if the p–value $< \delta$ and conclude that the full model should be used. If p–value $\geq \delta$, then fail to reject H_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).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

ix) Want hardly any predictors with pvalues > 0.05.
x) Want few predictors with pvalues between 0.01 and 0.05.

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

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

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

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

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

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

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

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

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

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

2.8 Complements

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

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

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

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

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

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

p. The plots are also useful for teaching survival regression to students and for explaining the models to consulting clients.

The slice survival, censored response, LCR, and EE plots are due to Olive (2011). Emphasis was on proportional hazards models since pointwise CI bands are available for the Cox proportional hazards model. Thus the slice survival plot can be made for the Cox model, and then the estimated 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. For some parametric survival models, see Chapter 3, Bennett (1983), Yang and Prentice (1999), Wei (1992), and Zeng and Lin (2007).

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

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

Plots were made in R and the function `coxph` produces the survival curves for Cox regression. The collection of R functions *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 Figures 2.1, 2.2, and 2.4. The function `bphsim3` shows that the Kaplan Meier estimator was close to the Cox survival curves for 2 groups (a single binary predictor) when censoring was light and $n = 10$.

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

2.9 Problems

Problems with an asterisk * are especially important.

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

a) Show

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

b) Show

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

c) Show

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

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

a) Find $H_0(t)$.

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

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

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

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

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

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

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

reduced model  coef    exp(coef) se(coef)    z      p
```

```
age           0.00203     1.002  0.00925  0.220 0.8300
sex          -0.82931     0.436  0.29895 -2.774 0.0055

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

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

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

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

c) Test whether the reduced model is good.

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

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

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

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

b) Find a 95% CI for β_1.

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

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

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

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

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

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

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

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

b) Test whether the reduced model is good.

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

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

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

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

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

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

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

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

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

b) What is the submodel I_I?

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

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

2.9. Data from Klein and Moeschberger (1997, p. 7) is on severely burned patients. The above table gives summary statistics for 4 PH regression models considered as final submodels after performing variable selection. Assume that the PH assumptions hold for all 4 models. The full model was M1, and M2 was the minimum AIC model found. Which submodel is the initial model

to examine I_I? Explain briefly why each of the other 3 submodels should not be used as the starting submodel.

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

SAS Problems

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

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

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

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

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

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

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

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

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

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

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

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

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

i) Are there any other good candidate models?

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

R Problems

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

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

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

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

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

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

e) Find a 95% CI for β_1.

f) Find a 95% CI for β_2.

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

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

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

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

Likelihood ratio test=14.3  output for 2.16
```

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

a) Find a 95% CI for β.

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

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

2.17. The *R* `lung` cancer data has the *time* until death or censoring and *status* = 0 for censored and 1 for uncensored. Then the covariates are *age*, *sex* = 1 for M and 2 for F, *ph.ecog* = Ecog performance score 0-4, *ph.karno* = a competitor to ph.ecog, *pat.karno* = patient's assessment of their karno score, *meal.cal* = calories consumed at meals excluding beverages and snacks and *wt.loss* = weight loss in last 6 months. A stratified proportional hazards model with stratification on *sex* will be used.

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

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

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

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

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

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

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

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

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

2.19. An extension of the PH model is the stratified PH model where $h_{\boldsymbol{x},j} = \exp(\boldsymbol{\beta}^T \boldsymbol{x}) h_{0,j}(t)$ for $j = 1, ..., K$ where $K \geq 2$ is the number of strata (groups). Testing is done in exactly the same manner as for the PH model, and the same $\boldsymbol{\beta}$ is used for each strata, only the baseline function

changes. The regression in Problem 2.17 used gender, male and female, as strata. If the model was good, then a PH model should hold for males and a PH model should hold for females. For the lung cancer data, females had a higher survival curve than males for $\boldsymbol{x}$ set to the average values.

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

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

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

How does the variability in the plot for a narrow vertical strip at ESP $= 0.5$ compare to the variability for a narrow vertical strip at ESP $= -1.5$?

c) Copy and paste the commands for this part into *R*, and include the resulting plot in *Word*.

d) Copy and paste the commands for this part into *R*, and include the resulting plot in *Word*.

```
vlung2(2)
title("females")
```

e) The plots in c) and d) divide the ESP into 4 slices. The estimated PH survival function is evaluated at the last point in the first 3 slices and at the first point in the 4th slice. Pointwise confidence intervals are also included (dashed upper and lower lines). The plotted circles correspond to the Kaplan Meier estimator for the points in each slice. The 1st slice is in the NW corner, the 2nd slice in the NE, the 3rd slice in the SW and the 4th slice in the SE. Confidence bands that would include an entire reasonable survival function would be much wider. Hence if the plotted circles are not very far outside the pointwise CI bands, then the PH model is reasonable.

Is the PH model reasonable for males? Is the PH model reasonable for females?

2.20. The lung cancer data is the same as that described in 2.17, but the PH model is stratified on *sex* with variables *ph.ecog*, *ph.karno*, *pat.karno* and *wt.loss*.

a) Copy and paste commands for this problem from (http://parker.ad.siu.edu/Olive/survhw.txt) into *R*. Click on the left window and hit *Enter*. Then 4 plots should appear. Include the plot in *Word*.

b) The plots are of x_j versus the martingale residuals when x_j is omitted. The loess curve should be roughly linear (or at least not taking on some simple shape such as a quadratic) if x_j is the correct functional form. If the loess curve looks like $t(x_j)$ for some simple t (eg $t(x_j) = x_j^2$), then $t(x_j)$ should be used instead of x_j. Are the loess curves in the 4 plots roughly linear?

c) Copy and paste commands for this problem from (http://parker.ad.siu.edu/Olive/survhw.txt) into *R*. Click on the left window and hit *Enter*. Then 4 plots should appear. Include the plot in *Word*. Also include the output from *cox.zph(lungfit2)* in *Word*.

d) The plots are of survival times vs scaled Schoenfeld residuals for each of the 4 variables. The loess curves should be approximately horizontal (0 slope) lines if the PH assumption is reasonable. Alternatively, the pvalue for Ho slope = 0 from *cox.zph* should be greater than 0.05 for each of the 4 variables. Is the PH assumption is reasonable? Explain briefly.

2.21. Copy and paste the *R* commands for this problem into *R*. This problem shows how to do backward elimination for the PH model in *R* using the Leemis (1995, p. 249-250) and Lawless (1982, p. 286) lung survival data. List the AIC for the model chosen in each step. Some entries are below.

```
                                        model  AIC
perf, age, ttoent, size, type, ttype, trt  189.22  full model
perf, age, ttoent, size,       ttype, trt  187.22
.
.
.
perf,                          ttype       181.52
perf                                       183.12
```

2.22. 16.52: Copy and paste the *R* command

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

from near the top of (http://parker.ad.siu.edu/Olive/survhw.txt) into *R*. **(Do not give any plots for this problem.)**

a) In *R*, type "library(survival)" if necessary. Then type "phsim(k=1)". Hit the up arrow to repeat this command several times. Repeat for "phsim(k=0.5)" and "" to make ET plots. The simulated data follows a PH Weibull regression model with $h_0(t) = kt^{k-1}$. For $k = 1$ the data follows a PH exponential regression model. Did the survival times decrease rapidly as ESP increases?

b) The function `phsim2` slices the ESP into 9 groups and computes the Kaplan Meier estimator for each group. If the PH model is reasonable and n is large enough, the 9 plots should have approximately the same shape. Type "phsim2(n=100,k=1)", then "phsim2(n=200,k=1)" and keep increasing n by 100 until the nine plots look similar (assuming survival decreases from 1 to 0, and ignoring the labels on the horizontal axis and the + signs that correspond to censored times). We will say that the plots look similar if $n = 800$. What value of n did you get?

c) The function `bphsim3` makes the slice survival plots when the single covariate is an indicator for 2 groups. The PH assumption is reasonable if the plotted circles corresponding to the Kaplan Meier estimator track the

solid line corresponding to the PH estimated survival function. Type "bphsim3(n=10,k=1)" and repeat several times (use the up arrow). Do the plotted circle track the solid line fairly well?

d) The function `phsim5` is similar but the ESP takes on many values and is divided into 9 groups. Type "phsim5(n=50,k=1)", then "phsim5(n=60,k=1)" and keep increasing n by 10 until the circles track the solid lines well. We will say that the circles track the solid lines well if they are within or not very far outside the pointwise CI bands. What value of n do you get?