MATH 697 Report

Daniel Krasnov

This is a brief summary or abstract of the document. It gives an overview of the main points, stuff and stuff.

1 Overview of Survival Analysis

1.1 Introduction

Survival Analysis (SA) concerns itself with the analysis of data whose research question is concerned with the time until an event occurs. The time leading up to the event is known as the survival time and the event itself is known as a failure. The survival time can be measured on any time scale, years, months, days, etc., and the failure is any event of interest, say when a subject enters remission or dies. For the purposes of this report we assume there is one event of interest at a time, however there could be multiple and these are known as competing risk problems.

SA often must deal with censored data. Censored refers to measures of survival time that are inaccurate. If a the subject experiences failure after the study has completed, this is called right censored; if the subject experiences failure at or before the measured time, it is called left-censored; if the subject experienced failure in a known interval but we do not know the exact time, it is called interval-censored.

We commonly make three assumptions about survival data: independent censoring, random censoring, and non-informative censoring. Independent censoring means that if we take a subset of subjects, censoring is random within that subset. Random censoring means both subjects that have been censored and have not been censored share the same failure rate. Non-informative censoring means the distribution of survival times gives no information about the distribution of censorship times.

In SA there exist two functions of primary interest. Let T, $T \geq 0$, be a random variable denoting a subject's survival time. Also, let $d \in \{0, 1\}$ be an indicator random variable where 0 denotes censorship and 1 denotes failure. Then the survivor function, denoted $S(t)$, and hazard function, denoted $h(t)$, are the greatest subject of study in SA. Mathematically this functions are defined as (Kleinbaum and Klein 1996)

$$S(t) = \exp\left\{-\int_0^t h(u)du\right\}, \tag{1}$$

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = -\left[\frac{dS(t)/dt}{s(t)}\right]. \tag{2}$$

Here we see that $S(t)$ gives us $P(T > t)$ and $h(t)$ gives the instantaneous potential for failure given the subject has survived until time t. It is important to note that $h(t)$ is a rate, not a probability, and as the subject accumulates more hazard, the lower the probability of survival.

Overall there are three goals in SA: the estimation and interpretation of survivor and hazard functions, the comparison of survivor and hazard functions, and the effect of covariates on the survival time.

1.2 Kaplan-Meir Curves and Hypothesis Testing

Consider a SA study where we are interested in assessing the difference in survival for subjects in a treatment group and a placebo group. One way we may assess this is to look at the survivor function $S(t)$. To estimate the survivor function we employ the Kaplan-Meier (KM) Curves. KM curves estimate the survivor function according to the following equations D'Arrigo et al. (2021):

$$\hat{S}(t_{(f)}) = \prod_{i=1}^{f} \hat{P}(T > t_{(i)} | T \geq t_{(i)}), \tag{3}$$

$$\hat{S}(t_{(f)}) = \prod_{i=1}^{f} \frac{n_f - d_f}{n_f} \tag{4}$$

where the KM survival probability at failure time $t_{(f)}$ is the survival probability of the previous failure time multiplied by the conditional probability of surviving past $t_{(f)}$ given the subject has already survived to at least $t_{(f)}$. These probabilities are simply estimated using sample proportions. That is, let n_f be the number of subjects at risk at time t_f and d_f be the number of subjects who fail at time t_f. Then (4) uses the sample proportion to estimate the conditional probabilities.

Confidence intervals for the KM curve are given by

$$\hat{S}_{KM}(t) \pm 1.96\sqrt{\hat{\text{Var}}[\hat{S}_{KM}(t)]}, \tag{5}$$

$$\text{Var}[\hat{S}_{KM}(t)] = (\hat{S}_{KM}(t))^2 \sum_{f:t_{(f)},\leq t} \left[\frac{m_f}{n_f(n_f - m_f)}\right] \tag{6}$$

where $t_{(f)}$ is the fth ordered failure time, m_f is the number of failures at $t_{(f)}$, and n_f is the number of subjects at still at risk at time $t_{(f)}$. We also have access to a confidence interval for the median survival time. Let M be the true unknown median survival time. Then the following holds asymptotically:

$$\frac{(\hat{S}_{KM}(M) - 0.5)^2}{\hat{\text{Var}}[\hat{S}_{KM}(M)]} \sim \chi^2_1. \tag{7}$$

From this a 95% confidence interval for the median survival time is given by

$$(\hat{S}_{KM} - 0.5)^2 < 3.84\hat{\text{Var}}[\hat{S}_{KM}(t)]. \tag{8}$$

As an example of KM curves, consider the toy dataset with 2 groups as shown in Table 1. Calculating the KM cuves leads to the following estimate of $\hat{S}(t)$ for each group as seen in Figure 1.

Table 1: Toy dataset for KM curve example.

time	status	group
5	1	A
8	1	A
12	0	B
4	1	B
6	1	A
15	0	A
20	1	B
9	1	A
10	0	B
8	1	B
13	1	A

After obtaining estimates of the survival function, the next logical question is whether these KM curves statistically differ across strata. To do this, we employ the log-rank test, a large-sample chi-square test. The log-rank test statistic for two groups is given by

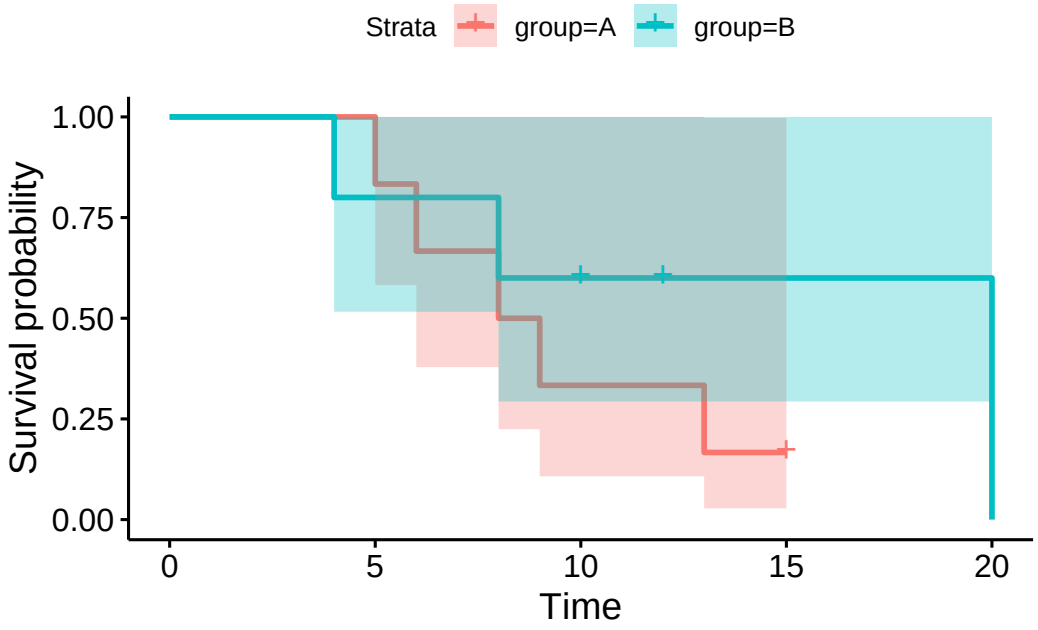


Figure 1: Explain Kaplan-Meir survival function estimate. Shaded regions indicate confidence intervals for each group.

$$\text{Log-rank statistic} = \frac{(O_1 - E_1)^2}{\text{Var}(O_1 - E_1)} \sim \chi^2_1, \tag{9}$$

where

$$e_{1f} = \left(\frac{n_{1f}}{n_{1f} + n_{2f}}\right) \times (m_{1f} + m_{2f}), \tag{10}$$

$$e_{2f} = \left(\frac{n_{2f}}{n_{1f} + n_{2f}}\right) \times (m_{1f} + m_{2f}), \tag{11}$$

$$O_i - E_i = \sum_{f=1}^{n} (m_{if} - e_{if}), \tag{12}$$

$$\text{Var}(O_i - E_i) = \sum_j \frac{n_{1f} n_{2f} (m_{1f} + m_{2f})(n_{1f} + n_{2f} - m_{1f} - m_{2f})}{(n_{1f} + n_{2f})^2 (n_{1f} + n_{2f} - 1)}, \tag{13}$$

and n_{1f} and n_{2f} are the numbers of subjects in the risk set for each group and m_{1f} and m_{2f} are the number of subjects failing in that group. Under the null hypothesis of no overall difference between the survival curves, the test statistics is ch-squared distributed with one degree of freedom. If we instead wish to test multiple KM curves at once, the test statistic becomes the matrix product

$$\text{Log-rank statistic} = \mathbf{d^T V^{-1} d} \sim \chi^2_{G-1}, \tag{14}$$

$$\mathbf{d} = (O_1 - E_1, \dots, O_{G-1} - E_{G-1})^T, \tag{15}$$

$$\mathbf{V} = ((v_{il})),\ v_{ii} = \text{Var}(O_i - E_i),\ v_{il} = \text{Cov}(O_i - E_i, O_l - E_l), \tag{16}$$

where the number of groups being compares is $G \geq 2$. The log-rank statistic can be thought of as assigning uniform weights to each failture time. Consider the following formulation of the log-rank statistic:

$$\text{Weighted log-rank statistic} = \frac{\left(\sum_f w(t_{(f)})(m_{if} - e_{if})\right)^2}{\text{Var}\left(\sum_f w(t_{(f)})(m_{if} - e_{if})\right)}, \tag{17}$$

where $w(.)$ is some weight function. The regular log-rank statistic takes $w(t_{(f)}) = 1$ however, if we alter this we can emphasize certain failure times. For example the Wilcoxon test sets $w(t_f) = n_f$, the number at risk. This causes earlier failures to receive more weight. This is used to assess the effect of a treatment on survival when changes are best seen early own in the trial. The Tarone-Ware test sets $w(t_f) = \sqrt{n_f}$, the square root of the number at risk. The Peto test sets $w(t_f) = \tilde{s}(t_{(f)})$, a survival estimate that differs slightly from the KM estimator. The Flemington-Harrington test sets $w(t_f) = \hat{S}(t_{(f-1)})^p \times [1 - \hat{S}(t_{(f-1)})]^q$. This statistic allows the user to specify if they want more weight on earlier or later survival times.

1.3 The Cox Proportional Hazards Model

The Cox Proportional Hazards (PH) model is used to model the effect of covariates on survival time. It does so through the estimation of hazard functions. Let $\mathbf{X}$ be a vector of covariates, then the Cox PH model is given by

$$h(t, \mathbf{X}) = h_0(t) e^{\sum_{i=1}^p \beta_i X_i}, \tag{18}$$

where $h_(t)$ is the baseline hazard function. From (18) we see this model has two important assumptions. First, notice that the variable t appears only in the baseline hazard function and not in the exponential term. This is called time-independence because the covariates are independent of time. Variables like sex can be considered time-independent as they do not vary with time. If one wishes to model variables with time dependence this assumption may be relaxed in which case the extended Cox model may be used (Kleinbaum and Klein 1996). The second key assumption is that the hazard for one subject is proportional to the hazard of another subject. That is consider two subjects with covariates $\mathbf{X} = (X_1, \dots, X_p)$ and $\mathbf{X}^* = (X_1^*, \dots, X_p^*)$ then we assume

$$\hat{h}(t, \mathbf{X}^*) = \hat{\theta}\hat{h}(t, \mathbf{X}), \qquad \hat{\theta} = \exp\left\{\sum_{i=1}^{p} \hat{\beta}_i (X_i^* - X_i)\right\}. \tag{19}$$

Given this model, we are interested in estimating the regression coefficient$\hat{} = (\hat{\beta}_1, ..., \hat{\beta}_p)$. We do so through maximum likelihood estimation. Consider the partial likelihood given by (Kumar and Klefsjö 1994)

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp\{s_i \beta\}}{\left[\sum_{m \in R(t_{(i)})} \exp\{X_m \beta\}\right]^{d_i}}, \tag{20}$$

where $R(t_{(i)})$ gives the risk set at $t_{(i)}$, d_i is the number of tied failures at $t_{(i)}$, m is number of items in $F(t_{(i)})$, X_m's are covariates, and $s_i = \sum X_{iq}$ the sum of covariates observed at time t_i. This formulation is of particular interest as although we only consider the likelihoods of each failure time, we still use the information of censored data though the risk set $R(t_{(f)})$. Thus, we have a work around for our incomplete data. The actual algorithm to maximize this function is outside the scope of this report, however various iterative algorithms exist which in general work through first guessing a value and then moving the solution towards an optima (Kleinbaum and Klein 1996).

Another quantity of interesting in this model is the hazard ratio:

$$\hat{HR} = \frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \exp\left\{\sum_{i=1}^{p} \hat{\beta}_i (X_i^* - X_i)\right\}, \tag{21}$$

where $\mathbf{X}^* = (X_1^*, ..., X_p^*)$ denotes a vector of covariates for one subject and $\mathbf{X} = (X_1, ..., X_p)$ denotes a vector for another. The hazard ratio is usually calculated as experimental group's hazard divded by control group's hazard: if done this way $HR = 1$ can be interpreted as no difference in hazard between groups, $HR < 1$ means experimental treatment reduced hazard, and $HR > 1$ means experimental treatment increased hazard (Barraclough, Simms, and Govindan 2011). A large sample 95% confidence interval for the hazard ration is given by

$$95\%\ CI = \exp\left\{\hat{\ell} \pm 1.96\sqrt{Var(\hat{\ell})}\right\}, \tag{22}$$

$$\hat{\ell} = \hat{\beta}_1 + \hat{\delta}_1 W_1 + \cdots + \delta_k W_k, \tag{23}$$

$$\tag{24}$$

where W_j is an interaction effect $X_i \times X_j$, $i \neq j$, and $\hat{\delta}_j$ is the corresponding estimated coefficient.

Hazard functions estimated using the Cox PH model are adjusted for covaraites. Thus, using the relationship in (2) we may obtain estimated survival curves adjusted for covariates. That is, the survival function for the Cox PH model is given by

$$\hat{S}(t, \mathbf{X}) = \hat{S}_0^{\exp\{\sum_{i=1}^{p} \hat{\beta}_i X_i\}}. \tag{25}$$

First, the SCAD and LASSO methods are introduced for variable selection in the Cox's Proportional Hazards Model. Then results from (Fan and Li 2002) are replicated which compares these variable selection methods with AIC and BIC best subset selection. Finally a dataset is introduced to which SCAD and LASSO will be applied for my MATH 686 project.

2 Variable Selection

Selecting significant predictors is an important issue in any modeling paradigm. In this section we discuss 4 well known criteria for variable and compare them in a simulation study. First we discuss the Akaike's information criterion (AIC) and the Bayesian information criterion (BIC). Consider a dataset D and let n be the amount of information in the data—the number of samples in our case. Consider a set of models where the kth model has likelihood $p(D|\theta_k; M_k)$. In general we are interested in a trade-off between parsimony and predictive power. That is, we would like a model to have the least complexity possible while still accounting for the variance seen in the data. A popular way to achieve this is to make use of a penalized model selection criteria, generally of the form (Kuha 2004)

$$2[l(\hat{\theta_2}) - l(\hat{\theta_1})] - a(p_2 - p_1) \sim \chi^2_{(p_2-p_1)}, \tag{26}$$

where for two candidate models M_1 and M_2 with parameter vectors $\theta_1,\ \dim \theta_1 = p_1$ and $\theta_2,\ \dim \theta_2 = p_2$, and positive constant a, we perform a likelihood ratio test with M_1 as the null model. This formulation can be thought of as having two components, a fit component and a complexity component. The difference in log-likelihoods assess how well the model fits the data and increases with the number of predictors. The difference in the the dimensions of θ_k penalizes increasing the number of predictors in M_2 and thus penalizes increasing the complexity of the alternative model. There are many information criteria of the above form, however the two most popular, AIC and BIC, are given by

$$\text{AIC} = 2[l(\hat{\theta_2}) - l(\hat{\theta_1})] - 2(p_2 - p_1), \text{ and} \tag{27}$$

$$\text{BIC} = 2[l(\hat{\theta_2}) - l(\hat{\theta_1})] - \log\, n(p_2 - p_1). \tag{28}$$

In regression contexts, best AIC/BIC subset selection refers to fitting all possible combinations of models and selecting the one that minimizes either criterion.

Another philosophy in variable selection is to penalize the likelihood function of the model such that variables are automatically selected during the model fitting process. Consider independent samples $(\mathbf{x}_i, Y_i)$ with conditional density $f_i(y_i; \mathbf{X}_i^T, \beta)$ and let $\ell_i = \log f_i$. Then a general form of penalized likelihood, given by (Fan and Li 2002), is

$$\sum_{i=1}^{n} \ell_i y_i; \mathbf{x}_i^T \beta - n \sum_{j=1}^{d} p_\lambda(|\beta_j|) \tag{29}$$

where d is the dimension of β, $p_\lambda(.)$ is some penalty function and λ is a tuning parameter. Selecting a function $p_\lambda(.)$ amounts to selecting a variable selection method. In this section

we will consider two penalties, the Least Absolute Shrinkage and Selection Operator (LASSO) and Smoothly Clipped Absolute Deviation (SCAD), given by

$$p_\lambda(|\theta|) = \lambda|\theta| \text{ and} \tag{30}$$

$$p_\lambda(\theta) = I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda), \tag{31}$$

where $a > 2$ and $\theta > 0$. In general, a value of $a = 3.7$ is suggested by (Fan and Li 2002). It has been shown that the SCAD penalty is an improvment upon the LASSO penalty in that it display oracle properties when the correct tuning parameter is selected. That is, regression coefficients that are 0 in the true model are estimated as such when using the SCAD penalty.

3 Simulation Study

In this section, we mimic (Fan and Li 2002), and assess variable selection techniques through a simulation study. Consider covariate vector $\mathbf{X}$ and prediction $\hat{\mu}(\mathbf{X})$. Then, the prediction error is defined as

$$\text{PE}(\hat{\mu}) = \mathbb{E}_{(\mathbf{X},Y)}[Y - \hat{\mu}(\mathbf{X})]^2, \tag{32}$$

where $(\mathbf{X}, Y)$ is some new observation. It can be shown, with baseline hazard set to 1, that

$$\mu(\mathbf{X}) = \exp\{-\mathbf{X}^T\beta\} \tag{33}$$

implying model error is given by

$$\mathbb{E}\left[\exp\{-\mathbf{X}^{\mathcal{T}}\} - \exp\{-\mathbf{X}^T\beta_0\}\right]^2 \tag{34}$$

In this section 100 datasets are simulated with $n = 75$ and $n = 100$ observations from the exponential hazard model

$$h(t|\mathbf{x}) = \exp(\mathbf{x}^T\beta) \tag{35}$$

where $\beta = (0.8, 0, 0, 1, 0, 0, 0.6, 0)^T$ and x_i's are marginally standard normal with correlation $\rho = 0.5$, Furthermore, censoring times are exponentially distributed with mean $U\exp(\mathbf{x^T}_{\mathbf{0}})$, $U \sim$ Uniform$(1, 3)$. Our variable selection procedures are compared to the maxium partial likelihood estimated model using the Median of Relative Model Errors (MRME), the ratio of of the model errors defined in (34). Additionally we record the average number of correct and incorrect coefficients estimated to be 0. The results of these simulations can be seen in Table 2. Standard deviations of regression coefficients can be seen in Table 3 for each variable selection technique as well.

Table 2: test

Method	MRME(%)	Aver. no. cor. 0 coeff.	Aver. no. incor. 0 coeff.
n=75			
SCAD	60.9554	3.68	0.01
LASSO	30.3892	2.53	0
AIC	65.8362	4.06	0.06
BIC	53.48	4.62	0.12
n=100			
SCAD	52.1683	4.18	0
LASSO	58.8546	2.7	0
AIC	69.2571	4.3	0.02
BIC	55.0899	4.73	0.03

Table 3: test

Method	Beta1 SD	Beta4 SD	Beta7 SD
n=75			
SCAD	0.1973	0.2137	0.243
LASSO	0.1786	0.2086	0.1947
AIC	0.1649	0.1908	0.14
BIC	0.1926	0.2011	0.1639
n=100			
SCAD	0.1801	0.1895	0.1868
LASSO	0.1843	0.1943	0.1731
AIC	0.1697	0.1474	0.1427
BIC	0.1678	0.1499	0.1456

4 Application

Next we consider the applications of LASSO, SCAD, AIC, and BIC variable selection methods on real data. We use a subset of the Mayo Clinic Primary Biliary Cholangitis (PBC) data. PBC is an autoimmune disease which damages the liver's bile ducts leading to cirrhosis and eventually death (Therneau et al. 2000). The dataset contains 418 cases of PBC, 312 of which are from a randomized trial and 106 cases of patients not present in the trial, but agreed to be tracked. The data used are available in the `survival` R package under the variable `pbc`. A table of covariates present in the data is available below (Therneau and Lumley 2015).

Table 4: Description of the Mayo Clinic Primary Biliary Cholangitis dataset.

Variable	Description
age	in years
albumin	serum albumin (g/dl)
ast	aspartate aminotransferase, once called SGOT (U/ml)
bili	serum bilirubin (mg/dl)
chol	serum cholesterol (mg/dl)
copper	urine copper (ug/day)
edema	0 no edema, 0.5 untreated or successfully treated, 1 edema despite diuretic therapy
protime	standardized blood clotting time
sex	m/f
stage	histologic stage of disease (needs biopsy)
status	status at endpoint, 0/1/2 for censored, transplant, dead
time	number of days between registration and the earlier of death, transplantation, or study analysis in July, 1986

Fitting a full Cox Proportional Hazards model to the data yields the following results:

Look at Table 5 and Table 6

Next we consider using our various variable selection methods and show the results:

Note tha diagnostic plots are avaiable in Section 6.

Table 5: Test

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	p_value
age	0.0306155	1.0310889	0.0110093	2.7808636	0.0054215	**5.42e-03**
sexf	-0.2969957	0.7430472	0.2980594	-0.9964311	0.3190407	3.19e-01
edema	1.0242170	2.7849139	0.3650750	2.8054970	0.0050239	**5.02e-03**
bili	0.0747229	1.0775856	0.0223787	3.3390179	0.0008408	**8.41e-04**
chol	0.0005966	1.0005967	0.0004310	1.3842810	0.1662724	1.66e-01
albumin	-0.7330353	0.4804485	0.2784487	-2.6325685	0.0084742	**8.47e-03**
copper	0.0026904	1.0026940	0.0010660	2.5237943	0.0116096	**1.16e-02**
ast	0.0036099	1.0036164	0.0018321	1.9703884	0.0487939	**4.88e-02**
protime	0.2327870	1.2621126	0.1027797	2.2649131	0.0235180	**2.35e-02**
stage	0.4530538	1.5731088	0.1486060	3.0486909	0.0022984	**2.30e-03**

Table 6: Coefficient Estimates with Various Estimation Methods

	LASSO	SCAD	AIC	BIC
age	0.0196	0.0083	0.0314	0.0287
sex	-0.0063	-	-	-
edema	0.7028	0.3264	0.8218	0.9777
bili	0.0843	0.1241	0.0851	0.1033
chol	-	-	-	-
albumin	-0.6131	-0.4204	-0.7186	-0.7632
copper	0.0029	0.0020	0.0029	0.0034
ast	0.0019	-	0.0044	-
protime	0.1541	0.0474	0.2275	-
stage	0.3057	0.2273	0.4328	0.4420

5 References

Barraclough, Helen, Lorinda Simms, and Ramaswamy Govindan. 2011. "Biostatistics Primer: What a Clinician Ought to Know: Hazard Ratios." *Journal of Thoracic Oncology* 6 (6): 978–82.

D'Arrigo, Graziella, Daniela Leonardis, Samar Abd ElHafeez, Maria Fusaro, Giovanni Tripepi, and Stefanos Roumeliotis. 2021. "Methods to Analyse Time-to-Event Data: The Kaplan-Meier Survival Curve." *Oxidative Medicine and Cellular Longevity* 2021 (1): 2290120.

Fan, Jianqing, and Runze Li. 2002. "Variable Selection for Cox's Proportional Hazards Model and Frailty Model." *The Annals of Statistics* 30 (1): 74–99.

Kleinbaum, David G, and Mitchel Klein. 1996. *Survival Analysis a Self-Learning Text*. Springer.

Kuha, Jouni. 2004. "AIC and BIC: Comparisons of Assumptions and Performance." *Sociological Methods & Research* 33 (2): 188–229.

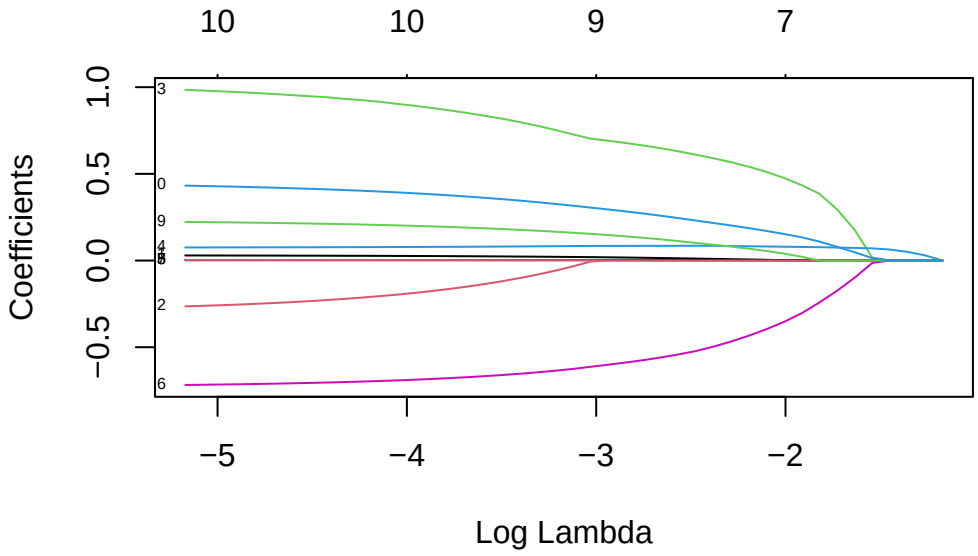
Kumar, Dhananjay, and Bengt Klefsjö. 1994. "Proportional Hazards Model: A Review." *Reliability Engineering & System Safety* 44 (2): 177–88.

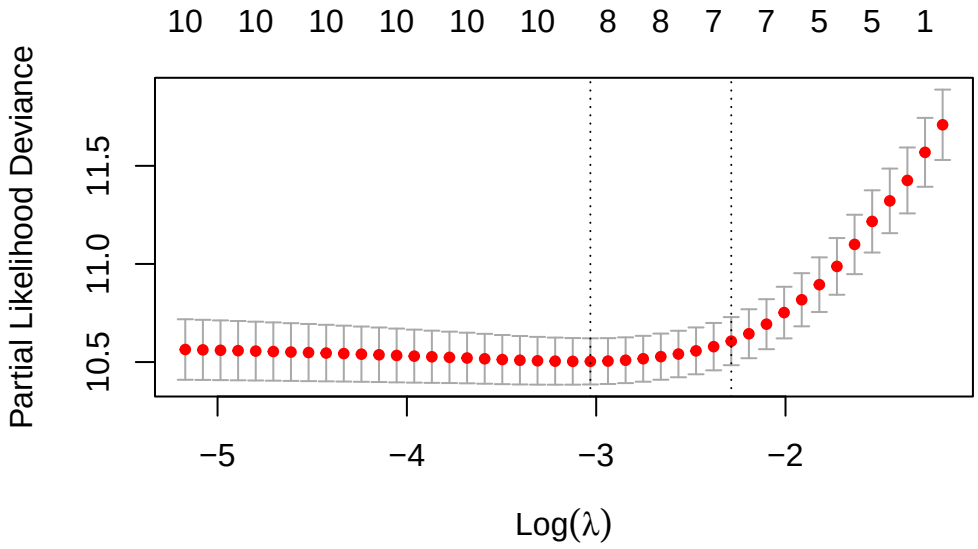
Therneau, Terry M, Patricia M Grambsch, Terry M Therneau, and Patricia M Grambsch. 2000. *The Cox Model*. Springer.

Therneau, Terry M, and Thomas Lumley. 2015. "Package 'Survival'." *R Top Doc* 128 (10): 28–33.

6 Appendix

EXPLAIN LASSO DIAGNOSTIC PLOTS HERE





EXPLAIN SCAD DIAGNOSTIC PLOTS HERE