2004

Cox Regression Model

Lindsay Sarah Smith
Louisiana State University and Agricultural and Mechanical College, lsmit51@lsu.edu

COX REGRESSION MODEL

A Thesis

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Master of Science

in

The Department of Mathematics

by
Lindsay Smith
B.S., Virginia Polytechnic Institute and State University, 2002
May 2004

Acknowledgments

The author would like to acknowledge the following people for their contributions and support in this work:

- **Dr. Padmanabhan Sundar** for being my thesis advisor. This semester deadlines crept up on us, but we still got it done. I am very grateful for everything you have done for me.
- **Dr. Robert Perlis** for all of your help throughout my graduate school career at LSU. Your teaching style has inspired me to be a great teacher, just as you are. Also thank you very much for your generous help being on my thesis committee.
- **Dr. Lynne LaMotte** for your help on trying to find me data, even though it was not as easy to obtain as I originally thought it would be. Also thank you very much for your generous help being on my committee.
- **Patricio Jara** for all your mathematical expertise in ALL the math classes I have taken at LSU. Also your LaTeXexpertise has been an invaluable help to me. I hope you still remember me when you are a math genius.
- Last but not least my parents, **Roger and Betty Holman**, for everything they have done. Without your never-ending love and support (both emotionally and financially), I would never be where I am today. I appreciate everything you have done for me more that you will ever know. I also want to thank my sister, **Penny Smith**, and boyfriend, **Greg Waddell**, for both of your immeasurable support and love you give me. You both mean the world to me and I hope someday I can repay all the support and encouragement you have given me throughout this whole 'thesis process'.

Table of Contents

List of Tables

List of Figures

Abstract

Cox, in 1972, came up with the Cox Regression Model to deal handle failure time data. This work presents background information leading up to the Cox's regression model for censored survival data. The marginal and partial likelihood approaches to estimate the parameters in this model are presented in detail. The estimation techniques of the hazard and survivor functions are explained. All of these ideas are illustrated using data from the Veteran's Administration lung cancer study.

Chapter 1. Introduction

Statistical analysis of failure-time data is an active and important area of research that has received considerable attention from several applied disciplines. Survival analysis in clinical trials and reliability theory manufacturing systems provide two examples where failure-time data is studied. Historically, failure times are modeled by fitting an exponential, Weibull, or log normal distribution to the data. For instance, an exponential distribution arises when a newly manufactured product is not necessarily excellent (owing to manufacturing defects) while an aged but functioning product is not necessarily bad. If the average longevity of a product is known, then the exponential distribution can be shown to be one with the maximum entropy (i.e. the most unpredictable among all continuous distributions with the given expected value). The model described in the present work owes itself to the exponential distribution in the presence of complex data. The various components of the complexity in the model are described below.

Failure (or response) time data usually arise with measurements of certain auxiliary variables known as covariates. For example, data on the occurrence of a heart-attack of a patient is usually coupled with measurement of blood pressure, weight, age, family history for heart diseases, life-style of patient, cholesterol level, etc. It is important to gauge the weights that must be attributed to these different covariates in order to predict the failure time in general for the patients. The optimal assignment of weights to the covariates is a problem in statistical estimation of parameters.

The second ingredient in the complexity of the model arises from the presence of a time-dependent parameter in the model. If the rate at which failures occur depends on time then the above estimation problem involves the estimation of an infinite-dimensional parameter namely, the estimation of a function. This part of the model is usually alluded to as the non-parametric feature of the model.

In clinical trials, there is a period of observation after which the failure time is censored. Thus the data typically consists of censored failure times where the censoring time may also differ among the subjects. Such censoring techniques arise when subjects quit the study at various times for a variety of reasons. In a failure-time study of automobiles, an automobile bears a censored time if it were to be lost due to an accident. The censoring time itself may happen to be random in certain studies.

The fourth element of complexity is the possible presence of ties in the failure-time, i.e. when two or more observations have the same response time. The model that is described in this work is known as the Cox Regression Model, (CRM), or the Proportional Hazards model, (PHM). The CRM is based on the principle of proportional or marginal likelihood. In chapter two, the required basic probability theory is presented. In chapter three, the set up of the CRM along with the maximum likelihood estimation of parameters is shown. The non-parametric estimation, properties of the estimates, and the connection of CRM to martingale theory are all also mentioned in chapter three. An analysis of the CRM with data is presented in chapter 4. Conclusions of the thesis appear in chapter 5.

Chapter 2. Fundamentals in Probability and Stochastic Processes

Let $(\Omega, \Im, P)$ be a probability space such that Ω is a set, $\Im$ is a sigma-algebra, and P is a probability measure.

Definition 2.1: The *conditional probability* that an event B occurs given that an event A has already happened is denoted $\mathrm{P}(B|A)$, and

$$\mathrm{P}(B|A) = \frac{P(B \cap A)}{P(A)}, \quad \text{if } P(A) > 0.$$

where $\mathrm{P}(A \cap B)$ is the probability of that both event A and event B both occur.

Definition 2.2: Let T be a random variable. Then T is *exponentially distributed* with parameter λ ($\lambda > 0$), mean $1/\lambda$, and its probability density function, $f(t)$, which is given by

$$\begin{aligned} f(t) &= \lambda \exp(-\lambda t) \quad \text{if } t \geq 0 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

The cumulative distribution function $F(t)$, the survival function, $S(t)$, and hazard function, $\lambda(t)$ of T are respectively

$$F(t) = 1 - \exp(-\lambda t), \quad S(t) = \exp(-\lambda t), \text{and} \quad \lambda(t) = \lambda \text{ [7]}.$$

Definition 2.3: The *Weibull distribution* is a two parameter distribution, λ and p, with hazard function, $\lambda(t)$, probability density function, $f(t)$, and survivor function ,$S(t)$, respectively

$$\lambda(t) = \lambda p(\lambda t)^{p-1}, \quad f(t) = \lambda p(\lambda t)^{p-1} \exp\left[-(\lambda t)^p\right], \quad \text{and}$$

$$S(t) = \exp(-(\lambda t)^p).$$

for λ and $p > 0$. The Weibull distribution is a generalization of the exponential distribution.

Definition 2.4: The *log normal distribution* is a two parameter distribution with parameters λ and p. Let

$$\phi(w) = \frac{\exp((-w^2)/2)}{\sqrt{2\pi}}, \text{and} \quad \Phi(w) = \int_{-\infty}^{w} \phi(u)du. \tag{2.0.1}$$

The density function for a failure time t can be written as

$$f(t) = (2\pi)^{-1/2} p t^{-1} \exp\left(\frac{-p^2(\log \lambda t)^2}{2}\right).$$

The survival function and hazard function are respectively

$$S(t) = 1 - \Phi(p \log \lambda t), \text{and} \quad \lambda(t) = \frac{f(t)}{S(t)}.$$

This model is comparable to the normal linear regression survival model except that it assumes that the survival times in the population are logarithmically distributed. For more information on these distributions see [7].

2.1 Poisson Process

A stochastic process, $N(t)$, is said to be a *Poisson Process* with rate λ ($\lambda > 0$) if the following conditions apply to the process.

1. $N(0) = 0$.
2. The number of events that occur in disjoint time intervals are independent.
3. The distribution of the number of the number of events that occur in a given interval depends only on the length of the interval and not on its location.
4. $\mathrm{P}(N(t) = 1) = \lambda t + o(t)$.
5. $\mathrm{P}(N(t) \geq 2) = o(t)$.

where $o(t)$ is the order of t. A function f is said to be of order $o(t)$ if $\lim_{t\to 0} f(t)/t = 0$ [14].

Another important property of a Poisson process is

$$\mathrm{P}(\text{an event occurs in } (t_1, t_1 + h_1), \ldots, \text{an event occurs in } (t_n, t_n + h_n))$$

$$= \prod_{j=1}^{n} (\lambda h_j + o(h_j)),$$

where $(t_j, t_j + h_h)$, $j = 1, 2, \ldots, n$, are disjoint intervals. The following theorem describes a fundamental theorem about Poisson Processes.

Theorem 2.1.1. *Let $N(t)$ be a Poisson process with a mean rate of λ and $t \geq 0$. The length of the interval from a fixed time to the next event occurring has a probability density function $f(x) = \lambda \exp(-\lambda x)$.*

Proof. Let T_0 denote any fixed time and T the next event occurring after T_0, where t is the time between T_0 and T. Then,

$$\begin{aligned} \mathrm{P}(T > t) &= \mathrm{P}(\text{the first event after } T_0 \text{ occurs after a } t \text{ time interval}) \\ &= \mathrm{P}(N(T_0 + t) - N(T_0) = 0) \\ &= \exp(-\lambda t) \quad \text{from the definition of a Poisson process} \end{aligned}$$

□

Another way to state this theorem is

$$\mathrm{P}(\text{No events occur in } [0, t]) = \exp(-\lambda t).$$

The next theorem relates a Poisson Process back to our definition of an exponential distribution.

Theorem 2.1.2. *Let $N(t)$, $t \geq 0$, be a Poisson process with rate λ and let $T_1, T_2, \ldots$ be time intervals. Then T_i, $n = 1, 2, \ldots$ are independent and identically distributed exponential random variables having a failure rate of λ.*

Proof. From the last proof,

$$\mathrm{P}(T_1 > t) = \mathrm{P}(N(t) = 0) = \exp(-\lambda t).$$

By the definitions, T_1 is an exponential random variable with failure rate λ. Then for some random variable s, we have

$$\mathrm{P}(T_2 > t | T_1 = s) = \mathrm{P}(N(t + s) - N(s) = 0) = \exp(-\lambda t).$$

Thus, T_2 is also an exponential random variable with failure rate λ and is independent of T_1. Repeating this process over and over again, it is easy to see that we get the results that we wanted to show that T_i, $n = 1, 2, \ldots$ are independent identically distributed exponential random variables having a failure rate of λ [2]. □

A Poisson probability distribution arises where 'events' occur at certain points of time. It is well known from basic statistics that the Poisson distribution approximates the binomial distributions. The next theorem is relates this idea to theorem (2.1.1).

Theorem 2.1.3. $P(N(t) = n) = \left(\frac{(\lambda t)^n}{n!}\right) \exp(-\lambda t)$.

Proof. Note the nth event of the Poisson process will occur before or at time t if and only if the number of events that occur by time t is at least n, i.e.

$$N(t) \geq n \Leftrightarrow B_n(t) \leq t$$

where B_n is the arrival time of the nth event. Therefore,

$$\begin{aligned} \mathrm{P}(N(t) = n) &= \mathrm{P}(N(t) \geq n) - \mathrm{P}(N(t) \geq n+1) \\ &= \mathrm{P}(B_n \leq t) - \mathrm{P}(B_{n+1} \leq t) \\ &= \int_0^t \lambda \exp(-\lambda x) \frac{(\lambda x)^{n-1}}{(n-1)!} dx - \int_0^t \lambda \exp(-\lambda x) \frac{(\lambda x)^n}{(n)!} dx. \end{aligned}$$

By using integration by parts from basic calculus (the formula is $\int u dv = uv - \int v du$, letting $u = \exp(-\lambda x)$ and $dv = \lambda[(\lambda x)^{n-1}/(n-1)!]dx$), yields

$$\int_0^t \lambda \exp(-\lambda x) \frac{(\lambda x)^{n-1}}{(n-1)!} dx = \exp(-\lambda t) \frac{(\lambda x)^n}{(n)!} + \int_0^t \lambda \exp(-\lambda x) \frac{(\lambda x)^n}{(n)!} dx \text{ [14].}$$

□

Chapter 4. Cox Regression Model

The Cox regression model, CRM, or the proportional hazards model, PHM, is a statistical theory of counting processes that unifies and extends nonparametric censored survival analysis. The approach integrates the benefits on nonparametric and parametric approaches to statistical inferences. The data in a CRM includes (T_i, z_i), $i = 1, 2, \ldots, n$, where n is the number of observations in the study, T_i is the time of failure of the ith observation, and z_i is the p-dimensional vector of covariates. In the presence of censoring of data, T_i is replaced by $T_i \wedge c_i$ where c_i is the censoring time for the ith observation. z_i can also be a time-dependent covariate in which case z_i will be replaced by $z_i(t)$.

4.1 Simple Linear Regression

A simple linear regression model is a model with a single explanatory variable and is represented as

$$\hat{Y_i} = \beta_0 + \beta_1 X_i.$$

In this equation, $\hat{Y_i}$ is the predicted value of Y_i, or the predicted response variable given the value of X_i, the treatment variable. It is worthwhile to consider a simple linear regression model because it captures the essential properties of multiple linear regression models. When making a statistical model it is important to make sure that the underlying assumptions hold. Plotting residuals versus the x values and other residual diagnostics are useful to check the normality of data. Interpretation of censored data must be done carefully because it is not normal and thus complicate the fitting of the distribution. Since failure-time data is almost always censored, one would need to find a model without the underlying assumption of normality.

Consider a set of observations y_i, $i = 1, 2, \ldots, n$, possibly censored, such that their cumulative distribution functions are $f[(y_i - \mu_i)/\sigma_i]$ where μ_i and σ_i are the respective population mean and population standard deviation. Standardized residuals, for this model, are defined to be

$$\hat{\epsilon_i} = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}_i},$$

where $\hat{\mu}_i$ and $\hat{\sigma}_i$ are the maximum likelihood estimates of μ and σ respectively.

These residuals should look like a possibly censored random sample from a standardized location-scale distribution, i.e. $\mu = 0$, $\sigma = 1$, and the distribution should be normal or logistic. Residual plots can be made in several different ways. If the residuals are plotted against a certain explanatory variable, the plot can show the variable's explanatory power.

When the observation, y_i, is censored then the residual, $\hat{\epsilon}_i$, is also censored. Thus for censored data, all we can say is that the actual residual would have been larger than the censored residual. Plotting these standardized residuals, $\hat{\epsilon}_i$, versus the predicted values of $\hat{y}_i$ will help detect nonlinearity, if the data is not heavily censored [9].

A cumulative distribution function, c.d.f., is often a common way to summarize and display data. If one plots the c.d.f. versus x, the graph produced will provided information on percentiles, dispersion, and general features of the distribution of the data. The c.d.f. can also be the basis for construction of goodness of fit tests for the hypothesized probability models.

4.2 Hazard Function

A hazard rate specifies the stochastic intensity of a multivariate counting process, i.e. by counting the occurrences of an event failure for each individual under observation. There are a number of representing the distribution of a continuous nonnegative random variables, such as $\lambda(t)$, the hazard function and $S(t)$, the survivor function. The hazard function, $\lambda(t)$, is a special function in the context of survival analysis. An observation has a hazard, or risk, of failure equal to

$$\lambda(t) = \lambda_0(t) \exp(\beta z), \tag{4.2.1}$$

where β is the transpose of the column vector β of p unknown regression coefficients, z_i is the column vector of p possibly time-varying covariates, and λ_0 is a fixed unknown baseline hazard rate for an individual with $z \equiv 0$ [5]. Each individual hazard rate has dependence on what was happened before, thus having a conditional probability basis.

The hazard function is very useful because it is practical to consider the immediate risk attaching to an individual known to be alive at age b. If you want to compare two groups of individuals, it may be more insightful to compare them via hazard functions. Hazard-based models are easier to manipulate and work with when dealing with censored or multi-defined failure data. Comparison of the hazard model and the exponential model are easily done [3].

The hazard function is often unknown, thus one needs to estimate it. The cumulative hazard function, $\Lambda(t) = \int_0^t \lambda(s)ds$ is easier to estimate and is sufficient for the our purposes. For the non-covariate case, the Nelson-Aalen estimate of the hazard function is the most commonly used for right-censored failures. An example is an engineering study of time to failure of diesel generator fans. The study detects if fans that are currently running need to be replaced by higher

quality fans to prevent future failures. For each fan, the number of total running hours from first being in service until fan failure was recorded.

Let T_i be the running time until fan failure for the ith generator. Assume that the T_is are independent and identically distributed with the probability density function, $f(t)$ and survivor function $S(t)$. The failure rate $\lambda(t)$ is what will be estimated. We have no fan-specific information and thus must work with the marginal distribution of failure times. Thus $S(t)$ is a mixture of survivor distributions for different qualities of fan. Let $T_i = \min(T_i, c_i)$, where c_i is the fan's censoring time. If we introduce a counting process, Y_i such that $Y_i = \mathrm{I}_{\{T_i \geq t, c \geq t\}}$. This means that $Y_i = 1$ if fan i is under observation just before time t else $Y_i = 0$ thus fan i has failed before time t. Let

$$\begin{aligned} N_i &= 0 \quad \text{until time } T_i \text{ (i.e. until observation } i \text{ fails)} \\ &= 1 \quad \text{time } T_i \text{ and after (i.e. after observation } i \text{ fails)} \end{aligned}$$

The cumulative hazard estimator is based on combined processes $\overline{Y}(t) = \sum_i Y_i(t)$ and $\overline{N}(t) = \sum_i N_i(t)$. The number of fans 'at risk' at time t is $\overline{Y}(t)$ and the number of fans that have failed at or before time t is $\overline{N}(t)$. Consider a time interval:

$$\Lambda(s+h) - \Lambda(s) \approx \lambda(s)h,$$

where $\Lambda(s+h) - \Lambda(s)$ is the probability that an event in $(s, s+h]$ failing when we know that the event is at risk at time s. Sum these amounts over subintervals $(0, t]$, intervals small enough that contain at most one event failure. The Nelson-Aalen estimator is

$$\hat{\Lambda}(t) = \int_0^t \frac{d\overline{N}(s)}{\overline{Y}(s)}.$$

The elements of $d\overline{N}$ is a shorthand that allows mixed continuous and discrete processes to be handled by one single notation, i.e. $d\overline{N}(t) = \Delta\overline{N}(t)$. Let T_1 denote the first failure time, T_2 denote the second failure time, and then so on, another representation of the Nelson-Aalen estimate is:

$$\hat{\Lambda}(t) = \sum_{i:t_i \leq t} \frac{\Delta\overline{N}(t_i)}{\overline{Y}(t_i)}.$$

4.3 Proportional Hazards Model

Let an arbitrary hazard rate be

$$\lambda(t; z) = \lambda_0(t) exp(\beta z),$$

where $\lambda_0(t)$ is an arbitrary unspecified base-line hazard function for a continuous t, β is the regression coefficients, and $\lambda(t, z)$ is the hazard function at time t for an individual with covariates z. The density function, $s(t)$ is

$$s(t; z) = \lambda(t; z)S(t; z), \tag{4.3.1}$$

where $S(t)$ is the survival function defined by

$$S(t; z) = \exp\left(-\int_0^t \lambda_0(u)\exp(z\beta)du\right).$$

The regression coefficients, β, may or may not be estimated with assumptions made about the hazard function. If β is estimated with assumptions made about the hazard function then one would maximize the likelihood functions and would consider contributions made to the hazard rate by censored data.

Covariates act multiplicatively on the hazard function. If $\lambda_0 = \lambda$, our hazard function reduces to the exponential regression model. The Weibull distribution is a special case where $\lambda_0(t) = \lambda p(\lambda t)^{p-1}$. The conditional density function of T given z is

$$f(t; z) = \lambda_0(t)\exp(z\beta)\exp\left[-\exp(z\beta)\int_0^t \lambda_0(u)du\right].$$

The conditional survivor function for T given z is

$$S(t; z) = [S_0(t)]^{\exp(z\beta)}, \quad \text{where}$$

$$S_0(t) = \exp\left[-\int_0^t \lambda_0(u)du\right].$$

Then we can see that the survivor function of t for a covariate value, z, is obtained raising the base-line survivor function, $S_0(t)$, to the power $\exp(z\beta)$ [7].

4.4 Marginal Likelihood

Suppose n individuals are observed to fail at times T_i, $i = 1, \ldots, n$, with corresponding covariates $z_1, \ldots, z_n$. Assume that all failure times are distinct, i.e. no two people (or more) fail or are censored at the same time. The *order statistic* is defined to be $O(t) = [T_{(1)}, T_{(2)}, \ldots, T_{(n)}]$ and refers to the T_is being ordered increasingly (i.e. $T_{(1)} < T_{(2)} < \cdots < T_{(n)}$). The *rank statistic* is defined to be $r(t) = [(1), (2), \ldots, (n)]$ and refers to the label attached to the order. Consider a sample of five failure times, say $T_1 = 6, T_2 = 23, T_3 = 10, T_4 = 67$, and $T_5 = 1$. Then $O(t) = [1, 6, 10, 23, 67]$ and $r(t) = [5, 1, 3, 2, 4]$.

Let $u = g^{-1}$, where g is a strictly increasing and differentiable transformation of $(0, \infty)$ onto $(0, \infty)$. The conditional distribution of u given z is

$$\lambda_0(g(u))g'(u)\exp(z\beta).$$

If one observed data of the form $u_1, u_2, \ldots, u_n$ and with corresponding $z_1, \ldots, z_n$ where $g(u_i) = T_i$ then the inference problem for β would be the same even if $\lambda_0(\cdot)$ were unknown.

By going back to the previous example, and letting $u = t^3$, then $O(u) = [1, 36, 100, 529, 4489]$ and $r(u) = [5, 1, 3, 2, 4] = r(t)$. Any specified order statistic can clearly be obtained by u though the rank statistic is invariant under changes of time. Thus the estimation problem for β are the same under any transform. Only the rank statistic $r(\cdot)$ can carry information about β when λ_0 is completely unknown. It follows that the rank statistic is marginally sufficient to estimate β.

To apply the rank statistic to get inferences about β, one would use the marginal distribution of the ranks and the marginal likelihood. The marginal likelihood is proportional to the probability that the rank vector is observed, i.e.

$$\begin{aligned} \mathrm{P}(r, \beta) &= \mathrm{P}\left\{r = [(1), (2), \ldots, (n)]; \beta\right\} \\ &= \int_0^\infty \int_{T_{(1)}}^\infty \cdots \int_{T_{(n-1)}}^\infty \prod_{i=1}^n s(T_{(i)}; z_{(i)}) dT_{(n)} \ldots dT_{(1)} \\ &= \frac{\exp\left(\sum_{i=1}^n z_i\beta\right)}{\prod_{i=1}^n \left\{\sum_{l \in R(T_{(i)})} \exp(z_l\beta)\right\}}, \end{aligned} \tag{4.4.1}$$

where $R(T_{(i)})$ is

$$R(t) = \{i : T_{(i)} \geq t \quad \text{and} \quad c_{(i)} \geq t\}$$

the risk set at time $T_{(i)}$, that is the group of individuals i that are under observation at time t, i.e., $T_{(i)}) = \{(i), (i+1), \ldots, (n)\}$.

To deal with censored data one must modify this last argument. If censoring takes place, the group then acts transitively on the censoring time and the invariant in the sample space is the first k rank variables, i.e. $(1), (2), \ldots, (k)$. In general, the censored model will not possess the group of invariant properties (this is due to the fact that when censoring occurred only partial information is observed on ranks). For example, if we observe the following failures: $110, 70, 64^*, 90$, with * symbolizing a censored observation. Then the following rank statistics are possible: $[3, 2, 4, 1], [2, 3, 4, 1], [2, 4, 3, 1], [2, 4, 1, 3]$.

The observed part of the statistic r is generating the likelihood. The exact time of censoring is ignored but the invariance of the uncensored model suggests that the lengths of intervals between successive failures is irrelevant for the inferences about β.

Suppose k items are observed, labeled $(1), (2), \ldots, (k)$, and have failure time $T_{(1)}, T_{(2)}, \ldots, T_{(n)}$ with corresponding covariates $z_1, z_2, \ldots, z_k$. Suppose further that

m_i observations with covariates $z_{i1}, z_{i2}, \ldots, z_{im_i}$ are censored in the ith interval $\left[T_{(i)}, T_{(i+1)}\right), i = 1, 2, \ldots, k$, where $T_{(0)} = 0$ and $T_{(k+1)} = \infty$. The marginal likelihood of β is computed as the probability that the rank statistic should be one of the possibilities, which is then the sum of the large number of terms as in equation (4.4.1). The possible rank vectors can be characterized as

$$T_{(1)} < T_{(2)} < \cdots < T_{(k)}$$

$$T_{(i)} < T_{(i1)}, T_{(i2)}, \ldots < T_{(im_i)} < T_{(i+1)}, i = 0, 1, \ldots, k \tag{4.4.2}$$

where $T_{(i1)}, \ldots, T_{(im_i)}$ are (the unobserved failure times) associated with the censored individuals in $[T_{(i)}, T_{(i+1)})$.

The previous marginal likelihood equation allows a simple computation of the marginal likelihood since given $T_{(i)}$ the event $T_{(i)} < T_{(i1)}, T_{(i2)}, \ldots < T_{(im_i)}$, has the conditional probability

$$h(T_{(i)}) = \exp\left[-\sum_{j=1}^{m_i} \exp(z_{ij}\beta) \int_0^{T_{(i)}} \lambda_0(u)du\right], i = 1, 2, \ldots, k.$$

One can now see that the marginal likelihood is proportional to the probability of the event (4.4.2). This probability is

$$\int_0^\infty \int_{T_{(1)}}^\infty \cdots \int_{T_{(k-1)}}^\infty \prod_{i=1}^k s(T_{(i)}; z_{(i)}) h(T_{(i)}) dT_{(k)} \ldots dT_{(1)}$$

$$= \frac{\exp\left(\sum_{i=1}^k z_i\beta\right)}{\prod_{i=1}^k \left\{\sum_{l \in R(T_{(i)})} \exp(z_l\beta)\right\}}, \tag{4.4.3}$$

where $R(T_{(i)}) = \{[(i), (i1), \ldots, (im_j)], j = 1, \ldots, k\}$ is the risk set at time $T_{(i)} - 0$. Note that equation (4.4.3) is not the probability of observing the event (4.4.2) in the censored experiment. The probability would depend on $\lambda_0(t)$ and the censoring mechanism. The expression (4.4.3) is the probability that in the underlying uncensored version (4.4.2) would occur.

If a tied failure occurred at T_i, the number of 'ties' would be denote d_i. Note the ranks assigned to d_i individuals who fail at T_i are not affected by the ranks associated with the d_j individuals who fail at T_j. Thus the sum reduces to the product of k terms, one for each failure time. Let Q_i be the set of permutations of the symbols $i1, i2, \ldots, id_i$. Let $p = (p_1, p_2, \ldots, p_{d_i})$ be an element of Q_i. Since we are dealing strictly with the case of no ties, $d_i = 1$ and $s_i = z_{ij}$ where s_i is the sum of covariates of those who fail at time $T_{(i)}$. The marginal likelihood for β is now

$$\prod_{i=1}^k \left\{\exp(z_i\beta) \sum_{P \in Q_i} \left[\sum_{l \in R(T_{(i)}, p_r)} \exp(z_l\beta)\right]^{-1}\right\}. \tag{4.4.4}$$

Equation (4.4.4) is general enough to handle censored data. Also the previous likelihood equations are special cases of (4.4.4). Equation (4.4.4) can be approximated by

$$\ell = \prod_{i=1}^{k} \frac{\exp(z_i\beta)}{\sum_{l\in R(T_{(i)})} \exp(z_l\beta)}. \tag{4.4.5}$$

Note that (4.4.5) is a special case of (4.4.3).

The maximum likelihood estimate of β is $\hat{\beta}$ and can be obtained as a solution to the system of the following equations

$$U_j(\beta) = \frac{\partial \log(\ell)}{\partial \beta_j} = \sum_{i=1}^{k} [z_{ji} - A_{ji}(\beta)] = 0, \tag{4.4.6}$$

where $j = 1, \ldots, s$ and z_{ij} is the jth element of the s-vector z_i and

$$A_{ji}(\beta) = \frac{\sum_{l\in R(T_{(i)})} z_{jl} \exp(z_l\beta)}{\sum_{l\in R(T_{(i)})} \exp(z_l\beta)}.$$

Similarly one can get

$$I_{hj}(\beta) = \frac{-\partial^2 \log(\ell)}{\partial \beta_h \partial \beta_j} = \sum_{i=1}^{k} C_{hji}, \tag{4.4.7}$$

for $h, j = 1, 2, \ldots, s$ and

$$C_{hji}(\beta) = \frac{\sum_{l\in R(T_{(i)})} z_{hl} z_{jl} \exp(z_l\beta)}{\sum_{l\in R(T_{(i)})} \exp(z_l\beta)} - A_{hi}(\beta) A_{ji}(\beta) \quad h, j = 1, 2, \ldots, s. \tag{4.4.8}$$

The value of $\hat{\beta}$ that maximizes (4.4.5) is obtained by solving equation (4.4.6). When dealing with data with no censored ties, mild conditions on the covariates and censoring are required to ensure the asymptotic normality of $\hat{\beta}$ with mean β and estimated covariance matrix $I(\hat{\beta})^{-1} = [I_{ih}(\hat{\beta})]^{-1}$ where $I_{hj}(\beta)$ is from equation (4.4.7) [7].

For some models, optimization software cannot be used to find the maximum of a likelihood function. The problem comes from finding explanatory variables that do not have strong multicollinearity. Software that solves regression models that uses least-squares approach is not able to optimize likelihood functions because it is unable to invert some matrices. The maximum likelihood iteration approach, ML, does the reparameterization internally and does not always allow the user to see this information. In order to have a high probability of finding the maximum of a likelihood function, if one exists, requires the starting values that do not differ greatly from the maximum and the likelihood shape that is not much different from a quadratic with the axes corresponding to the transformed model parameters [9].

4.5 Partial Likelihood

The partial likelihood method proposed by Cox in 1975 gives essentially the same results as the last section for the model (4.2.1). Let us look at the set $R(T_{(i)})$, the set of individuals at risk at time $T_{(i)} - 0$. The conditional probability that item (i) fails at time $T_{(i)}$ given that the items $R(T_{(i)})$ are at risk (and that exactly one failure occurs at $T_{(i)}$ since we are still dealing with no ties in failure times) is

$$\frac{\lambda(T_{(i)}; z_{(i)})}{\sum_{l \in R(T_{(i)})} \lambda(T_{(i)}; z_{(i)})} = \frac{\exp(z_{(i)}\beta)}{\sum_{l \in R(T_{(i)})} \exp(z_l \beta)}, \tag{4.5.1}$$

for $i = 1, 2, \ldots, k$. Note that $\lambda_0(t)$ is unspecified. Thus no additional information about β is obtained from the observation that no failure occurs in the time interval $(T_{(i-1)}, T_{(i)}), i = 1, 2, \ldots, k$. This occurs because we can account for this observation by taking $\lambda_0(t)$ to be very close to zero over this interval.

The partial likelihood for β is now created by taking the product over all failure times T_i of equation (4.5.1) to get

$$L(\beta) = \prod_{i=1}^{k} \left(\frac{\exp(z_i \beta)}{\sum_{j \in R(T_{(i)})} \exp(z_j \beta))} \right),$$

which is identical to (4.4.3). Note that a partial likelihood is not a usual likelihood because the general construction does not result in quantity that is proportional to the conditional or marginal probability of an observed event. But Cox did informally show that this method does give a maximum 'partial' likelihood estimate that is consistent and asymptotically normally distributed. This estimate also has an asymptotic covariance matrix that is estimated by the inverse of the matrix of second partial derivatives of the log likelihood function. The hazard relationship of this function is

$$\frac{\lambda(t; z)dt}{1 - \lambda(t; z)dt} = \frac{\lambda_d(t)dt \exp(z\beta)}{\lambda_d(t)dt} \tag{4.5.2}$$

where $\lambda_d(t)$ is the discrete hazard function giving only positive contributions at the observed failure times $T_{(1)}, T_{(2)}, \ldots, T_{(k)}$. The conditional probability as the ith term in the product is

$$\prod_{i=1}^{k} \left(\frac{\exp(z_i \beta)}{\sum_{l \in R(T_{(i)})} \exp(z_l \beta)} \right). \tag{4.5.3}$$

One may note that equation (4.5.3) is more difficult to compute than equation (4.4.4). Thus it may also be approximated to make calculations friendlier. The partial likelihood, equation (4.5.3), does not derive a consistent estimate of the parameter β for the model (4.2.1) if ties in failure times occur.

4.6 Log-Rank Test

Before looking at the results of the model, there will be a brief discussion about score function tests and a Savage test that arise from marginal and partial likelihoods. The score test statistic for testing the global hypothesis $H_0 : \beta = 0$ (and alternative hypothesis $H_1 : \beta \neq 0$). It arises from substituting 0 for β in equation (4.4.6). For a sample of uncensored, no-ties data, $U(0)$ is

$$U(0) = \sum_{i=1}^{n} z'_{(i)} \left\{ 1 - \left[n^{-1} + (n-1)^{-1} + \cdots + (n-i+1)^{-1} \right] \right\}, \tag{4.6.1}$$

which is of the form of a linear rank statistic, called the Savage test. The score corresponds to the ith ordered failure time being one minus the expected ith order statistic from a unit exponential distribution. Generalizations of the Savage test can be made for tied or censored data and are obtained from the score tests of the appropriate marginal and partial likelihoods. From the approximate likelihood, equation (4.4.5), the score test statistic (for the testing the global null hypothesis $\beta = 0$) is

$$U(0) = \sum_{i=1}^{k} \left(s'_i - n_i^{-1} \sum_{l \in R(T_{(i)})} z'_l \right) \tag{4.6.2}$$

which is equivalent to equation (4.6.1) when dealing with the case of no tied failures and no censoring.

The special case of comparing $s+1$ survival curves labels $0, 1, \ldots, s$ arises when defining $z_i = (z_{1i}, \ldots, z_{si})$, where z_{ij} is equal to either one or zero according to whether or not the ith observation is in the jth sample. Thus equation (4.6.2) is the log-rank or generalized Savage statistic and can be written as

$$U(0) = O - E$$

where O ($O = s'_1 + s'_2 + \cdots + s'_k$) gives the number of failures in each sample $(1, 2, \ldots, s)$ and

$$E = \sum_{i=1}^{k} n_i^{-1} \sum_{l \in R(T_{(i)})} s'_l,$$

where E is sum over failure times of the conditional expected number of failures in each sample.

Note that the score statistic (4.6.2), arises from the partial likelihood (from equation (4.5.3)) as a result of equation (4.5.2). The construction of the partial likelihood relates to the $O - E$ interpretation of the log-rank statistic. Thus the asymptotic results for partial likelihood shows that the log-rank statistic, $O - E$, is also asymptotically normal with estimated covariance matrix given by V, such

that

$$V_{hj} = \sum_{i=1}^{k} \left[(n_{ji} n_i^{-1} - n_{hi} n_{ji} n_i^{-2})(n_i)(n_i - 1)^{-1} \right]$$

and is obtained from the second partial derivative $\partial^2/(\partial\beta_h \partial\beta_j)$ of the logarithm of equation (4.5.3). In the equation for V_{hj}, n_{ji} represents the size of the risk set in sample j just prior to $T_{(i)}$. The appropriate test statistic, for testing the hypothesis $\beta = 0$, is

$$U(0)' V^{-1} U(0)$$

which under the hypothesis will have a $\chi^2_{(s)}$ distribution.

The covariance matrix of $U(0)$ can be estimated from (4.4.5). The estimate gives $I(0)$ with h, j elements

$$I_{hj}(0) = \sum_{i=1}^{k} \left(n_{ji} n_i^{-1} - n_{hi} n_{ji} n_i^{-2} \right).$$

Note since we are dealing with no-tied failures, that $I(0)$ and V agree. If there are ties this estimate will over estimate the score statistic variance. Thus a better statistic would be

$$U(0)' I(0)^{-1} U(0).$$

This statistic is less than or equal to the log-rank statistic. Therefore it provides a conservative test.

The χ^2 statistic is

$$\chi^2 = \sum_{i=1}^{s} \frac{(O_i - E_i)^2}{E_i},$$

where E is the expected value and O is the observed value. If we look at X for our problem where $X = \sum_{i=1}^{s} (O_i - E_i)^2 / E_i$. One can show that $X \leq U(0)' V^{-1} U(0)$. If we compare X with a χ^2 distribution, we will have a good, slightly conservative test [7].

4.7 Survivor Function

When dealing with a failure time distribution, certain properties of the variables must hold in order to apply a failure time regression model. The T_i, the time of failure, must be a nonnegative random variable from a homogenous population. A probability distribution of T_i must be in the form of a probability density function, hazard function and the survival function. The survival, also known as the survivor, function, as denoted in a previous section $S_i(t)$, can be defined for discrete and continuous distributions as the probability that T_i is at least as great as a value of t. Thus

$$S_i(t) = \mathrm{P}(T_i \geq t), \quad 0 < t < \infty.$$

When dealing with survival data, it is often very useful to summarize the data in terms of sample c.d.f., or in terms of the survivor function. If a sample of n distinct uncensored observations come from a homogenous population then the sample survivor function is a step function with a n^{-1} decrease immediately after each observed failure ($S(0) = 1$ and $S(\infty) = 0$). In the event that a tie occurs, simple adjustments can be made. However most failure time data contains censored data and adjustments must be made to take this into account.

Let $S_0(t)$ be an arbitrary survivor function, then the function of t given z is

$$S(t; z) = (S_0(t))^{\exp(z\beta)}. \tag{4.7.1}$$

All of the data available from the model (4.7.1) and the calculations that go into the maximum likelihood estimate of $S_0(t)$ are necessary. Let $T_1 < T_2 < T_3 < \cdots < T_n$ be the failure times from of a sample of n observations from a homogeneous population with survival function $S(t)$. Assume that there are no ties in the failures and no ties in the censoring times. Let C_i denote the set of observations that are censored in the time interval $\left[T_{(j)}, T_{(j+1)}\right)$, where $j = 1, 2, \ldots, k)$, $T_{(0)} = 0$, and $T_{(k+1)} = \infty$. The probability of failure at time T_j is

$$S(T_j) - S(T_j + 0), \tag{4.7.2}$$

where $S(T_j + 0) = \lim_{x \to 0^+} S(T_j + x)$ and $j = 0, 1, 2, \ldots, k$. In(4.7.2), the contribution of a censored observation at time time t is

$$S_0(t + 0)^{\exp(z\beta)}.$$

We are assuming that the observed censored time, $T_{(l)}$, tells us that the actual unobserved failure time is greater than $T_{(l)}$. A likelihood function on the space of survivor function $S(t)$, is

$$L = \prod_{j=0}^{k} \left\{ \left[S_0(T_{(j)})^{\exp(z\beta)} - S_0(T_{(j)} + 0)^{\exp(z\beta)} \right] \prod_{l \in C_i} S_0(T_l + 0)^{\exp(z_l\beta)} \right\}. \tag{4.7.3}$$

A generalized version of the parametric model maximum likelihood estimate is the survivor function $\hat{S}(t)$ that maximizes L. It can be difficult to maximize L. Inconsistent estimates are made if one is not cautious. Clearly $\hat{S}(T)$ is discontinuous at the observed failure times since otherwise $L = 0$. The function $\hat{S}(t)$ is found by taking $S_0(t) = S_0(T_{(i)} + 0)$ for $T_{(i)} < t \leq T_{(i+1)}$ and letting the probability mass function fall only at the observed failure times $T_{(1)}, T_{(2)}, \ldots, T_{(k)}$.

Consider the discrete model, with hazard contribution $1 - \alpha_j$ at $T_{(j)}, j = 1, 2, \ldots, k$. Let $\alpha_0 = 1$, then

$$S_0(T_{(i)}) = S_0(T_{(i-1)} + 0) = \prod_{j=0}^{i-1} \alpha_j, \quad i = 1, 2, \ldots, k. \tag{4.7.4}$$

By substituting 4.7.4 into equation (4.7.1) and simplifying, one gets

$$\prod_{i=1}^{k}\left[\left(1-\alpha_i^{\exp(z\beta)}\right)\prod_{l\in R(T_{(i)})-D_i}\alpha_i^{\exp(z_l\beta)}\right], \tag{4.7.5}$$

as the likelihood function that is to be maximized.

The estimator of the survivor function can be done by making estimations of β and then estimating α's in equation (4.7.5). Make the estimate $\beta=\hat{\beta}$ from the marginal likelihood function. Then maximize equation (4.7.5) with respect to $\alpha_1,\alpha_2,\ldots,\alpha_k$. By differentiating the log of equation (4.7.5) with respect to the α_is, one obtains the maximum likelihood estimate of α_i as a solution to

$$\frac{\exp(z\hat{\beta})}{1-\hat{\alpha}_i^{\exp(z\hat{\beta})}}=\sum_{j\in R(T_{(i)})}\exp(z_j\hat{\beta}). \tag{4.7.6}$$

Since we are dealing with no ties in the failure times, equation (4.7.6) yeilds $\hat{\alpha}_i$ as

$$\hat{\alpha}_i=\left(1-\frac{\exp(z_{(i)}\hat{\beta})}{\sum_{j\in R(T_{(i)})}\exp(z_l\hat{\beta})}\right)^{\exp(-z_{(i)}\hat{\beta})}.$$

The maximum likelihood estimate of base-line survivor function is then

$$\hat{S}_0(t)=\prod_{i|T_{(i)}<t}\hat{\alpha}_i, \tag{4.7.7}$$

which is a step-function with discontinuities at each observed failure, $T_{(i)}$. From equation (4.7.7), the estimated survivor function, with covariates $\tilde{z}$ is

$$\hat{S}(t;\tilde{z})=\prod_{i|T_{(i)}<t}\hat{\alpha}_i^{\exp(\tilde{z}\hat{\beta})}. \tag{4.7.8}$$

4.8 Counting Processes

Another way to look at Cox's Regression model is as a model of random intensity of a multivariate counting process. A multivariate counting process

$$\tilde{N}=\{N_i(t):0\le t<\infty;i=1,2,\ldots,n\}$$

is a stochastic process with n components that can be thought of as counting the occurrences (as time t proceeds) of n different types of events (or the same event for n different individuals. Suppose these events occur singly. If each component $N_i(\cdot)$ is viewed as a function of t, one obtains an integer-valued step function. The function is zero at time zero with jumps of size +1 only. Assume the jumps to be

right continuous, so that $N_i(t)$ is the random number of events of type i in the time interval $[0, t]$. No two components jump at the same time. The process $\tilde{N}$ has an intensity process

$$\tilde{\Lambda} = \Lambda_i(t) : 0 \leq t < \infty; i = 1, 2, \ldots, n, \text{defined by}$$

$$\Lambda_i(t)dt = \mathrm{P}[N_i \text{ jumps in a time interval of length } dt \text{ around time} t | \Im_{t^-}] \quad (4.2.1)$$

where $\Im_{t^-}$ denoted the past up to the beginning of the small time interval dt (i.e. everything that has happened until just before time t).

An example of a simple multivariate counting process, where each component jumps at most once, can be defined as

$$N_i(t) = I\left\{T_i \leq t, T_i \leq c_i\right\}.$$

In this example, N_i jumps once, if at all, at time $T_i \leq c_i$ if individual i is observed death. Given what has happened before the time interval dt, it is known that individual i either died at the observed time T_i (thus T_i is less than t and less than the censoring time c_i) or was censored at time $c_i < t$. Thus N_i either has made its only jump or will never jump, so the probability of a jump in the interval dt is zero. Thus defining

$$\begin{aligned} Y_i(t) &= I\left\{T_i \geq t, c_i \geq t\right\} \\ &= 1 \text{ if individual } i \text{ is under observation just before time } t \\ &= 0 \text{ otherwise.} \end{aligned}$$

Therefore equations (4.2.1) and (4.2.1) yeild

$$\Lambda_i(t)dt = Y_i(t)\lambda_0(t)\exp(\beta z)dt.$$

Note that given the past up to, but not including, time t, $Y_i(t)$ and $\Lambda_i(t)$ are fixed (not random). In such a case, one says that $Y_i(t)$ and $\Lambda_i(t)$ are predictable or previsible.

An extension to Cox's regression model is $\tilde{N}$, a multivariate counting process with intensity process $\tilde{\Lambda}$, satisfying

$$\Lambda_i(t)dt = Y_i(t)\lambda_0(t)\exp(\beta Z)dt,$$

where z (the fixed covariates of individual i at time t) is replaced with Z, random covariates. One no longer needs to require each N_i to make at most one jump. It is also no longer required $Y_i(t)$ to be in the special form given just before, but Y_i has to be nonnegative. It is required that N_i, Y_i, and Z_i are processes that can be observed and that Y_i and Z_i are predictable (i.e. $Y_i(t)$ and $Z_i(t)$ are fixed given what has happened before time t) [5].

4.3 Properties of $\hat{\beta}$

Historically asymptotic normality of a consistent maximum likelihood estimator was derived from a Taylor's expansion of the first derivative of the log likelihood about the true value of $\beta = \beta_0$ and was evaluated at $\beta = \hat{\beta}$. The $D\log(L(\beta))$ is the vector of partial derivatives $(\partial/\partial\beta_i)\log(L(\beta))$ evaluated at β. When writing $D\log(L(\beta))$, it is important to show that $n^{-1/2}D\log(L(\beta))$ is asymptotically multivariate normally distributed. The set up includes independent and identically distributed observations from a density function $f(\cdot, \beta)$. This results in the central limit theorem since $n^{-1/2}D\log(L(\beta))$ is $n^{-1/2}$ times the sum of n random vectors that are independent and identically distributed.

The same results can be derived from a martingale central limit theorem approach. First an introduction to martigales will be discussed.

4.3.1 Martingales

A *martingale* is an adapted stochastic process, $M = \{M(t) : t \geq 0\}$, whose increment over any interval $(u, v]$ has expectation zero. That is

$$E[M(v) - M(u)|\Im_u] = 0, \quad \text{for all } 0 \leq u < v < \infty, \tag{4.3.1}$$

where $(\Im_t : t \geq 0)$. If $\Im_u$ is given then $M(u)$ is fixed. A martingale theorem states that integrating a predictable process with respect to a martingale yields a new martingale. The martingale central limit theorem gives conditions under which the whole process, M, is approximately normally distributed, with independent time increments.

If we let u denote the time just before time t and v, the time right after time t, we can rewrite equation (4.3.1) to be

$$E[dM(t)|\Im_{t^-}] = 0, \tag{4.3.2}$$

where $dM(t) = M(v) - M(u)$. One can relate this to the intensity of a process. In a small time interval, dt, a counting process, N, will either jump once or not at all. Therefore the probability of a jump in that interval is close to the expected number of jumps in the interval. Thus (4.2.1) states $\Lambda(t) = E[dN(t)|\Im_{t^-}]$. Thus defining $dM(t) = dN(t) - \Lambda(t)$, such that

$$M(t) = N(t) - \int_0^t \Lambda(s)ds. \tag{4.3.3}$$

Note, by this definition $M(t)$ is a martingale.

A *predictable process* of a martingale, M, is a non-decreasing process $< M >= \{< M > (t) : t \geq 0\}$ and defined by

$$d < M > (t) = E[(dM(t))^2|\Im_{t^-}] = var[dM(t)|\Im_{t^-}].$$

In general, let $\{M(t)\}$ be a jump Markov process with stationary independent increments (i.e. $M(t)-M(s)$ is independent of $M(s)-M(t)$), $t \geq 0$, and the magnitude of the jumps (i.e. jump size) can be anywhere from $[-J, J]$ with 'intensity' measure λ, then

$$M(t) - \int_o^t \int_{[-J,J]} z\lambda(dz)ds$$

is a martingale, where the predictable process part is

$$\int_0^t \int_{[-J,J]} z\lambda(dz)ds.$$

A predictable process is the sum of conditional variances of increments over M over small time intervals dt in $[0, t]$. Consider M given by equation (4.3.3). Each conditional variance is being taken given the past. Thus N_i is a zero-one variable. Its conditional expectation is $\Lambda(t)dt$ and condtional variance is $\Lambda(t)dt(1-\Lambda(t)dt) = \Lambda(t)dt$. Thus we would expect (and it turns out to be true) that

$$< M > (t) = \int_0^t \Lambda(s)ds.$$

Assuming that there are not ties in failure times, the counting processes N_i and N_j never jump simultaneously. Thus $dN_i(t)dN_j(t)$ is equal to zero. Hence the predictable covariance between dN_j and dN_i, $i \neq j$, is $-\Lambda_i(t)dt \cdot \Lambda_j(t)dt \approx 0$. Let M_i, M_j be processes defined by equation (4.3.3) corresponding to N_i, N_j respectively. Thus the predictable covariance between M_j and M_i, $i \neq j$, denoted $< M_i, M_j >$ is

$$< M_i, M_j > (t) = 0 \quad \text{for all } t \text{ and } i \text{ such that } i \neq j.$$

Theorem 4.3.1. *(Martingale transform theorem)Integrating a predictable process with respect to a martingale yields a new martingale.*

Proof. Suppose M is a martingale, K is a predictable process, and M' is a process such that $M' = \{M'(t) : t \geq 0\}$ and is defined by $dM'(t) = K(t)dM(t)$. Therefore $M'(t) = \int_0^t K(s)dM(s)$. Thus

$$\begin{aligned} E[dM'(t)|\Im_{t^-}] &= E[K(t)dM(t)|\Im_{t^-}] \\ &= K(t)E[dM(t)|\Im_{t^-}] \quad \text{because } K \text{ is predictable} \\ &= 0 \quad \text{because } M \text{ is a martingale.} \end{aligned}$$

Therefore $M'(t)$ is also a martingale. Note

$$\begin{aligned} var[dM'(t)|\Im_{t^-}] &= var[K(t)dM(t)|\Im_{t^-}] \\ &= K(t)^2 E[dM(t)|\Im_{t^-}] \\ &= K(t)^2 d < M > (t). \end{aligned}$$

Thus

$$< M' > (t) = \int_0^t K(s)^2 d < M > (s).$$

□

An application of the *martingale central limit theorem* gives conditions under which the whole process, M, is approximately normally distributed, with independent time increments. Define W such that $W = \{W(t) : t \geq 0\}$ is a process with the following properties:

1. $W(\cdot)$ is a continuous function.
2. $W(0) = 0$.
3. For any $T_1, T_2, \ldots, T_n$, $W(T_1), W(T_2), \ldots, W(T_n)$ is multivariate normally distributed with means zero.
4. $W(\cdot)$ has independent increments.

Since $W(t)$ has independent increments, the conditional variance $dW(t)$ given the path of W on $[0, t)$ does not depend on the past at all. The conditional expectation is zero. Therefore W is a continuous martingale with predictable variation process $< W >$ equal to some deterministic function, say H.

These properties also characterize the distribution of W as Gaussian. It is not surprising that a sequence of martingales $M^{(n)}$, $n = 1, 2, \ldots$ is such that

1. jumps of $M^{(n)}$ get smaller as $n \to \infty$ (because $M^{(n)}$ becomes more nearly continuous).
2. the predictable variation process of the process $M^{(n)}$ becomes deterministic (i.e. $< M^{(n)} > (t) \to H(t)$ in probability as $n \to \infty$. Hence, $< M^{(n)} > (t)$ is asymptotically normally distributed with mean zero and variance $H(t)$ where the increments of $M^{(n)}$ are are asymptotically independent) [5].

More details about martingales and these ideas that are discussed here can be found in Gill [5], Meyer [10], and Rebolledo[13].

Consider $D \log(L(\beta))$ as a sum, or integral, over time intervals t rather than individual i's. The variables Y_i and $N_i(t)$ be the same as they were defined in section 4.8 (Counting Processes). Let $\ell(\beta, u)$ be the likelihood of β based upon the

observation of N_i, Y_i, and $z_i, i = 1, 2, \ldots, n$ on the time interval $[0, u]$, and we will define

$$E_0(t) = \frac{\frac{1}{n}\sum_{j=1}^{n} Y_j(t) z_j(t) \exp(\beta z_j(t))}{\frac{1}{n}\sum_{j=1}^{n} Y_j(t) \exp(\beta z_j(t))}. \tag{4.3.4}$$

Recall our partial likelihood equation (4.5.3),

$$\ell(\beta) = \prod_{i:T_i \le c_i} \left(\frac{\exp(\beta z_i(T_i))}{\sum_{j \in R(t)} \exp(\beta z_j(T_j))} \right).$$

Each term in $\ell(\beta)$ is precisely $\mathrm{P}(i)$. When maximum likelihood theory is applied to this equation, we get asymptotic results, but in practice we use the 'approximate' results. Cox's explanation of the proof to get asymptotic results is very complicated and restrictive with some martingale arguments. It is necessary to rewrite $\ell(\beta)$ to make it more practical. The partial likelihood function, if there are no tied failures, is

$$L(\beta) = \prod_{t \ge 0} \prod_{i=1}^{n} \left(\frac{Y_i(t) \exp(\beta z_i(t))}{\sum_{j=1}^{n} Y_j(t) \exp(\beta z_j(t))} \right)^{dN_i(t)}. \tag{4.3.5}$$

In this equation we are introducing $dN_i(t)$ as the increment of N_i over small time intervals, dt, around time t. That is how the equation became a product over disjoint intervals, $\prod_{t \ge 0}$. The second product is a finite product over individual i for which N_i jumps at time t. Thus $dN_i(t) = 1$ and $dN_i(s) = 0$ for $s \ne t$. Let $dM_i(t) = dN_i(t) - \Lambda_i(t)dt$, where $\Lambda_i(t)dt$ is defined, as it was in section 3.8. By equations (4.3.5) and (4.3.4), one can get

$$\begin{aligned}
n^{-1/2} D \log \ell(\beta, u) &= n^{-1/2} \sum_{i=1}^{n} \sum_{t \le u} \left(z_i(t) - \frac{\sum_{j=1}^{n} Y_j(t) z_j(t) \exp(\beta z_j(T))}{\sum_{a=1}^{n} Y_a(t) \exp(\beta z_a(T))} \right) dN_i(t) \\
&= \sum_{i=1}^{n} \int_{t=0}^{u} n^{-1/2} (z_i(t) - E_0(t)) dN_i(t) \\
&= \sum_{i=1}^{n} \int_{t=0}^{u} n^{-1/2} (z_i(t) - E_0(t)) dM_i(t), \quad \text{since}
\end{aligned}$$

$$\sum_{i=(t)}^{n} (z_i(t) - E_0(t)) \Lambda_i(t) =$$

$$= \sum_{i=1}^{n} z_i(t) Y_i(t) \lambda_0(t) \exp(\beta z_j(t)) - E_0(t) \sum_{i=1}^{n} Y_i(t) \lambda_0(t) \exp(\beta z_i(t)). \quad \text{Thus}$$

$$\sum_{i=(t)}^{n} (z_i(t) - E_0(t)) \Lambda_i(t) = 0 \quad \text{from equation (4.3.4).}$$

Note $n^{-1/2}(z_i(t) - E_0(t))$ is a vector of predictable processes. By the martingale transform theorem, from Section 4.3.1, $n^{-1/2} D \log \ell(\beta, u)$ is a stochastic process and is the sum of n vector martingales. Hence $n^{-1/2} D \log \ell(\beta, u)$ is also a martingale. In order to show that $M^{(n)}(t) = n^{-1/2} D \log \ell(\beta, u)$ is asymptotically normally distributed, one must verify the two assumptions of the martingale central limit theorem (of section 4.3.1). In order to do this one would need a vector version of the theorem. Such a version does not exist unless the vectors β and $z_i(t)$ are scalars. For the simplicity of the paper, let us assume this is the case.

Jumps of $M^{(n)}$ get smaller as $n \to \infty$ is the first assumption. Let us consider a special case, for simplicity, such that $|z_i(t)| \leq C < \infty$ for all i and t and for some constant C. In this case it is easy to see that $z_i(t) - E_0(t)$ (the integrand of $n^{-1/2} D \log \ell(\beta, u)$) is also bounded by C. Each M_i has jump size of $+1$ coinciding with the jumps of N_i. Since there are no multiple jumps in M_i, then the jumps of $M^{(n)}$ are bounded by $n^{-1/2}C$, which tends to zero as n approaches infinity. Thus assumption one is fulfilled.

For the second assumption, we must evaluate the process $< M^{(n)} >$. Using algebra and some of the results from section 4.8, one can show

$$\begin{aligned} < M^{(n)} > (t) &= \int_0^t \frac{1}{n}(z_i(s) - E_0(s))^2 \Lambda_i(s) ds \\ &= \int_0^t \{\frac{1}{n}\sum_{i=1}^n z_i(s)^2 Y_i(s) \exp(\beta_0' z_i(s)) - \\ &\quad \frac{\left[\frac{1}{n}\sum_{i=1}^n z_i(s) Y_i(s) \exp(\beta z_i(s))\right]^2}{\frac{1}{n}\sum_{i=1}^n Y_i(s) \exp(\beta z_i(s))}\} \lambda_0(s) ds \end{aligned}$$

Note $< M^{(n)} > (t)$ is expressed in terms of simple averages of $Y_i(s) \cdot z_i(s)^r \cdot \exp(\beta z_i(s))$, where r can be equal $0, 1$, or 2. If these averages converge in probability, then show that $< M^{(n)} > (t)$ converges in probability to some constant. All of the other parts of the classical proof of asymptotic normality of $\hat{\beta}$ follow the same conditions. This applies to the consistency of $\hat{\beta}$ which needs to be established before the above arguments can be applied.

Large-sample maximum likelihood theory is not valid for $\hat{\beta}$ unless n is large enough that the averages

$$\frac{1}{n} Y_i(t) z_i(t) \exp(\beta z_i(t)), r = 0, 1, \text{and } 2$$

are almost nonrandom for all t and for β_0 close to β.

The martingale property of $M^{(n)}$ is implied by Cox's 1975 definition of partial likelihood. It is shown that in each term $D \log L(\beta)$ has expectation zero given the preceding terms. It appears that the definition of partial likelihood contains enough structure to ensure that the large-sample properties of maximum likelihood estimation hold for it too, under the same regularity conditions. [5].

Chapter 5. Data

The data in Prentice's article [12] will be used for illustration. In the Veteran's Administration lung cancer study, males with advanced inoperable lung cancer were randomized to either standard chemotherapy or test chemotherapy. After the randomization of the study, the patients within each group (either standard or test), were then separated by which type of lung cancer they had (squamous cell, small cell, adeno-carcinoma, or large cell carcinoma).

The failure time for each patient was death. Common heterogeneity in a study like this would be disease extent and pathology, previous treatment of the disease, demographic backgrounds and initial health status. Tables 5.1-5.8 are the tables of different types of therapies (either standard or test chemotherapy) and different types of tumor cell types (squamous, small cell, adeno, or large cell). In the individual tables, t stands for the number of days the patients were alive after the study began (failure time). The next variable in the table, x_1, represents the patient's performance status at the beginning of the study. Performance status is measured at the randomization of the study by the Karnofsky rating, which means: 10-30 completely hospitalized, 40-60 partial confined, and 70-90 able to care for self. The next variable in the table ,x_2, represents the number of months from diagnosis to the randomization of the study. Next, x_3 is the patient's age in years at the beginning of the study. The last variable in the table, x_4, is an indicator variable indicating if the patient has had prior therapy (0 =no prior therapy and 10 = prior therapy). Nine of the one hundred and thirty seven patients were censored, and is indicated in tables 5.1-5.8 by a * next to the patient's failure time.

When first looking at the data, a multiple regression model was fitted with eight regressor variables, let us call them $x_1, \ldots, x_8$. The first four regressor variables are the x_1, x_2, x_3, and x_4 from our previous data tables. Three indicator variables for squamous, small cell, and adenocarcinoma took into account the four cell-type classes. Let x_5 be the indicator variable that describes the differences between squamous cell and large cell. Then x_6 is the indicator variable that describes the differences between small cell and large cell. Thus x_7 is the indicator variable that describes the differences between adeno-carcinoma and large cell. Then x_8 is an indicator variable that distinguished between treatment type. This indicator variable shows the difference between the test and standard chemotherapy. The results are summarized in table 5.9 and the SAS program written to get these

Table 5.1: Lung Cancer Trial [12]. Standard Chemotherapy, Adeno-Carcinoma.

t	x_1	x_2	x_3	x_4
8	20	19	61	10
92	70	10	60	0
35	40	6	62	0
117	80	2	38	0
132	80	5	50	0
12	50	4	63	10
162	80	5	64	0
3	30	3	43	0
95	80	4	34	0

Table 5.2: Lung Cancer Trial [12]. Standard Chemotherapy, Squamous Cell.

t	x_1	x_2	x_3	x_4
72	60	7	69	0
411	70	5	64	10
228	60	3	38	0
126	60	9	63	10
118	70	11	65	10
10	20	5	49	0
82	40	10	69	10
110	80	29	68	0
314	50	18	43	0
100*	70	6	70	0
42	60	4	81	0
8	40	58	63	10
144	30	4	63	0
25*	80	9	52	10
11	70	11	48	10

Table 5.3: Lung Cancer Trial [12]. Standard Chemotherapy, Small Cell.

t	x_1	x_2	x_3	x_4
30	60	3	61	0
384	60	9	42	0
4	40	2	35	0
54	80	4	63	10
13	60	4	56	0
123*	40	3	55	0
97*	60	5	67	0
153	60	14	63	10
59	30	2	65	0
117	80	3	46	0
16	30	4	53	10
151	50	12	69	0
22	60	4	68	0
56	80	12	43	10
21	40	2	55	10
18	20	15	42	0
139	80	2	64	0
20	30	5	65	0
31	75	3	65	0
52	70	2	55	0
287	60	25	66	10
18	30	4	60	0
51	60	1	67	0
122	80	28	53	0
27	60	8	62	0
54	70	1	67	0
7	50	7	72	0
63	50	11	48	0
392	40	4	68	0
10	40	23	67	10

Table 5.4: Lung Cancer Trial [12]. Standard Chemotherapy, Large Cell.

t	x_1	x_2	x_3	x_4
177	50	16	66	10
162	80	5	62	0
216	50	15	52	0
553	70	2	47	0
278	60	12	63	0
12	40	12	68	10
260	80	5	45	0
200	80	12	41	10
156	70	2	66	0
182*	90	2	62	0
143	90	8	60	0
105	80	11	66	0
103	80	5	38	0
250	70	8	53	10
100	60	13	37	10

Table 5.5: Lung Cancer Trial [12]. Test Chemotherapy, Adeno-Carcinoma.

t	x_1	x_2	x_3	x_4
24	44	2	60	0
18	40	5	69	10
83*	99	3	57	0
31	80	3	39	0
51	60	5	62	0
90	60	22	50	10
52	60	3	43	0
73	60	3	79	0
8	50	5	66	0
36	70	8	61	0
48	10	4	81	0
7	40	4	58	0
140	70	3	63	0
186	90	3	60	0
84	80	4	62	10
19	50	10	42	0
45	40	3	69	0
80	40	4	63	0

Table 5.6: Lung Cancer Trial [12]. Test Chemotherapy, Squamous Cell.

t	x_1	x_2	x_3	x_4
999	90	12	54	10
112	80	6	60	0
87*	80	3	48	0
231*	50	8	52	10
242	50	1	70	0
991	70	7	50	10
111	70	3	62	0
1	20	21	65	10
587	60	3	58	0
389	90	2	62	0
33	30	6	64	0
25	20	36	63	0
357	70	13	58	0
467	90	2	64	0
201	80	28	52	10
1	50	7	35	0
30	70	11	63	0
44	60	13	70	10
283	90	2	51	0
15	50	13	40	10

Table 5.7: Lung Cancer Trial [12]. Test Chemotherapy, Small Cell.

t	x_1	x_2	x_3	x_4
25	30	2	69	0
103*	70	22	36	10
21	20	4	71	0
13	30	2	62	0
87	60	2	60	0
2	40	36	44	10
20	30	9	54	10
7	20	11	66	0
24	60	8	49	0
99	70	3	72	0
8	80	2	68	0
99	85	4	62	0
61	70	2	71	0
25	70	2	70	0
95	70	1	61	0
80	50	17	71	0
51	30	87	59	10
29	40	8	67	0

Table 5.8: Lung Cancer Trial [12]. Test Chemotherapy, Large Cell.

t	x_1	x_2	x_3	x_4
52	60	4	45	0
164	70	15	68	10
19	30	4	39	10
53	60	12	66	0
15	30	5	63	0
43	60	11	49	10
340	80	10	64	10
133	75	1	65	0
111	60	5	64	0
231	70	18	67	10
378	80	4	65	0
49	30	3	37	0

Table 5.9: Multiple Regression on Lung Cancer Data Using SAS.

Regression Variable	Reg. Coefficient ($\hat{\beta}$)
Performance Status (Karnofsky scale)	2.9168
Disease Duration (months)	-0.4579
Age (years)	0.1363
Prior Therapy(0 no, 10 yes)	37.7350
Cell Type	
Small vs. Large	20.1080
Squamous Cell vs. Large	-55.6310
Adeno vs. Large	-83.3434
Treatment (0 standard, 1 test)	31.0765

results is the first appendix. The resulting regression equation would then be:

$$t = -40.1555 + 2.9168(\text{Performance Status}) - 0.4579(\text{Disease Duration})$$

$$+0.1363(\text{Age}) + 37.7350(\text{Prior Therapy}) + 20.1080(\text{Small vz. Large})$$

$$-55.6310(\text{Squamous vs. Large}) - 83.3434(\text{Adeno vs. Large}) + 31.0765(\text{Treatment}). \tag{5.3.6}$$

The problem with this model can be noticed when one looks at the model's coefficient of multiple determination, the R^2 value. The R^2 is the portion of the sum of squared deviations from the mean accounted for in the regression line. Thus $R^2 = SSR/SST$ and if we have a good model this value will be close to one. The R^2 value obtained for this model is 0.2317, which is not even close to one.

Interpretation of a regression model when all of the explanatory variables are measured in different units is difficult. The only way to check if one variable has more significance than another variable is to obtain the standardized regression coefficients. For a full description of how to calculate standardized regression coefficients consult Neter, Kuter, Nachtsheim, and Wasserman [11]. The obtained standardized regression coefficients for (5.3.6), can be found in table 5.10. The first column in the table is the variable name, the second is the regular regression coefficient from above, and the third is the standardized regression coefficient for the multiple linear regression model. The larger the standardized regression coefficient, the more influence that variable has on the model. The results of table 5.10 appear to support the results from table 5.9. In both tables performance status has a large positive effect on time. Prior therapy has a positive effect on time. Disease duration, age, and the indicator varables for squamous vs. large cell and adneo vs. large cell all have a small negative effect on time. Treatment and the indicator variable for squamous vs. large cell both have a small positive effect on time.

Table 5.10: Standardized Regression Coefficients, Found in SAS.

Regression Variable	Reg. Coeff. ($\hat{\beta}$)	Std. Reg. Coeff.
Performance Status (Karnofsky scale)	2.9168	0.3697
Disease Duration (months)	-0.4579	-0.03075
Age (years)	0.1363	-.0092
Prior Therapy(0 no, 10 yes)	37.7350	0.1090
Cell Type		
Small vs. Large	20.1080	0.0609
Squamous Cell vs. Large	-55.6310	-0.1527
Adeno vs. Large	-83.3434	-0.2108
Treatment (0 standard, 1 test)	31.0765	0.1005

Next, a proportional hazards model was run in SAS to obtain the values table 5.11. The SAS program that was created to find these solutions is the first appendix. The model fits much better with the covariates included in the model. This is shown by the model fit statistics found in SAS. The model fit statistics criterion -2 LOG L and AIC were both 1011.768 without the covariates and 949.484 with the covariates. When evaluating these statistics small is better. It can easily be see that with the covariates the statistic is 46.284 less.

Table 5.11: PHM on Lung Cancer Data Using SAS.

Regression Variable	Reg. Coeff.($\hat{\beta}$)	p-value	Hazard Ratio
Performance Status (K. scale)	-0.03193	$<$0.0001	0.969
Disease Duration (months)	0.0005	0.9549	1.001
Age (years)	-0.01068	0.2429	0.989
Prior Therapy (0 no, 10 yes)	0.09726	0.6730	1.102
Cell Type			
Small vs. Large	0.48611	0.0677	1.626
Squamous Cell vs. Large	-0.39849	0.1590	0.671
Adeno vs. Large	0.82694	0.0062	2.286
Treatment (0 std, 1 test)	0.30660	0.1392	1.359

The conditional hazard function for all patients in the study would be

$$\lambda(t;z) = \lambda_0(t)\exp\{-0.032(\text{Performance Status}) + 0.001(\text{Disease Duration})$$

$$-0.011(\text{Age}) + 0.097(\text{Prior Therapy}) - 0.486(\text{Small vs. Large Cell})$$

$$-0.398(\text{Squamous vs.Large Cell})+0.827(\text{Adeno vs.Large Cell})+0.307(\text{Treatment})\}. \tag{5.3.7}$$

Thus the conditional hazard function, for a patient with large cell lung cancer and receiving the standard chemotherapy, would be

$$\lambda(t;z) = \lambda_0(t)\exp\{-0.032(\text{Performance Status}) + 0.001(\text{Disease Duration})$$

$$-0.011(\text{Age}) + 0.097(\text{Prior Therapy})\}.$$

Since the patient has large cell cancer, variables for 'Squamous vs. Large Cell', 'Small vs. Large Cell' and 'Adeno vs. Large Cell' all equal zero and the variable for 'Treatment' equals zero because the patient is receiving the standard chemotherapy.

Therefore if a patient has squamous cell lung cancer and is receiving the test chemotherapy, his/ hers conditional hazard rate would be:

$$\lambda(t;z) = \lambda_0(t)\exp\{-0.032(\text{Performance Status}) + 0.001(\text{Disease Duration})$$

$$-0.011(\text{Age}) + 0.097(\text{Prior Therapy}) + 0.091\}.$$

The 0.091 comes from the 0.307 for the test chemotherapy and the -0.398 for having squamous cell cancer. Following the same procedures you can form personal conditional hazard rate, depending on what type of cancer the patient has and what type of treatment he/she is receiving.

The proportional hazard model function that we want to estimate here is the survivor function. The estimated survivor function for the all of the patients in the study is:

$$S(t;z) = \exp[-\int_0^t \lambda_0(u)\exp\{-0.032(\text{Performance Status})+0.001(\text{Disease Duration})$$

$$-0.011(\text{Age})+0.097(\text{Prior Therapy})-0.486(\text{Small vs. Large Cell})-0.398(\text{Squamous}$$

$$\text{vs.Large Cell}) + 0.827(\text{Adeno vs.Large Cell}) + 0.307(\text{Treatment})\}du]. \tag{5.3.8}$$

The main objective in a study such as this one is to test whether or not the test chemotherapy is better than the standard chemotherapy. With a Log-Rank χ^2 statistic of 0.0120, we would have to fail to reject H_0 (where $H_0 : \mu_{std} = \mu_{test}$ and $H_a : \mu_{std} \neq \mu_{test}$). Thus there is no statistical evidence to show a difference in the test chemotherapy and the standard chemotherapy.

Performance status significantly impacts our model which can be easily seen by making a scatter plot of time versus performance status (x_1). Figure 5.1 illustrates the strong trend; where there is an increase in performance status, there is also an increase in failure time. It is reasonable to say that a person will live long (i.e. larger failure time) if they are healthier.

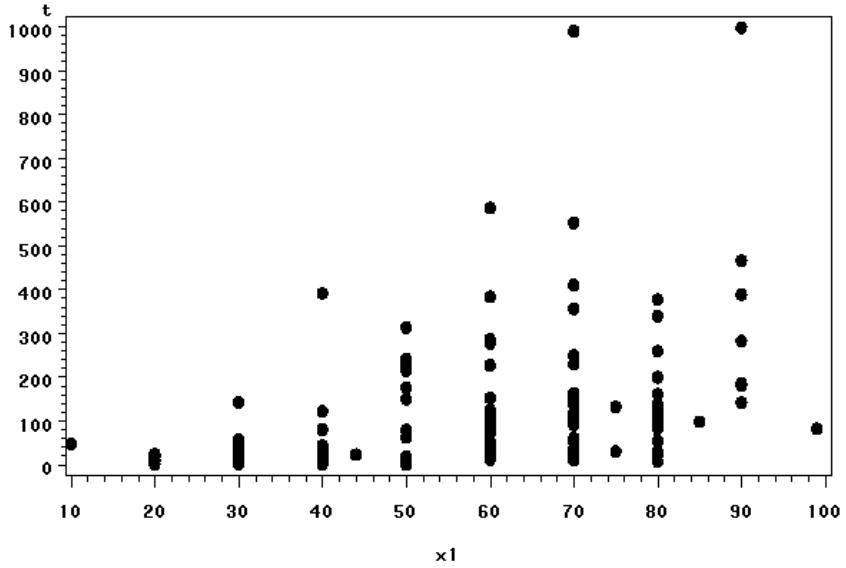

Figure 5.1: Time vs. Performance Status

The next two factors are not as significant. It appears that age and disease duration do not have a large impact on failure time. One can easily see this by looking at figures 5.2 and 5.3. These scatter plots seem random and not to have an specified relationship.

By looking at test significance levels and the hazard ratios prior therapy seems to make a difference when being put into the model. If one plots time vs. prior therapy, the plot is hard to interpret due to the fact that prior therapy only obtains two values, one if the patient has had no prior therapy and zero if the patient has had prior therapy. This plot is figure 5.4. Thus we will now run separate analysis on the different groups for further information. We will first make a proportional hazard model for all patients that have not had prior therapy. Figure 5.6 was generated by the program in the second appendix. It is a graph of what would happen if we divided the data into two separate survivor functions. These curves do not look as different as originally thought to be. Thus looking at a χ^2 test of difference, there is a log-rank statistic of 0.6013. Thus the two models are not significantly different. Thus we should use the single model for both patients that have received prior therapy and those that have had no prior therapy.

Graphs are helpful to view relationship, but a statistician/ mathematician needs more concrete values. A test of significance is done by dividing the regression

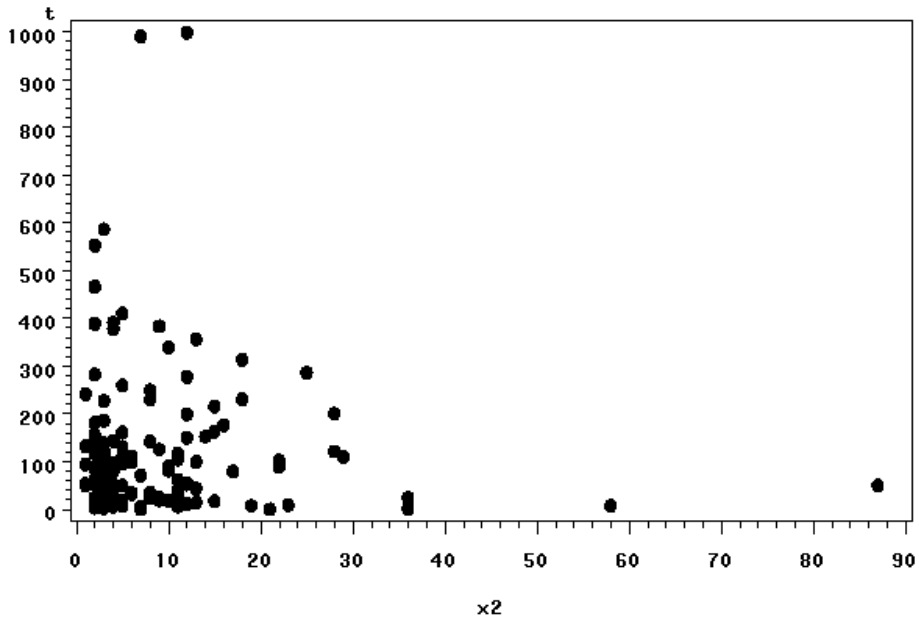

Figure 5.2: Time vs. Disease Duration

coefficient by its standard error; then comparing these results with the standard normal distribution. If this ratio is greater than 1.96, the coefficient is significant at the 5 percent level. The optimal model is found by the stepwise model selection. When a stepwise model selection was run on this data, only variables x_1, x_5, and x_7 are significant enough to remain in the model at a significance level of 0.05. Our new proportional hazards model with just these covariates can be found in table 5.12.

Table 5.12: PHM on Lung Cancer Data Using SAS, Significant Coefficients

Reg. Variable	Reg. Coeff.($\hat{\beta}$)
Performance Status	-0.02928
Disease Duration	
Age	
Prior Therapy	
Cell Type	
Small vs. Large	0.58873
Squamous vs. Large	
Adeno vs. Large	1.02136
Treatment	

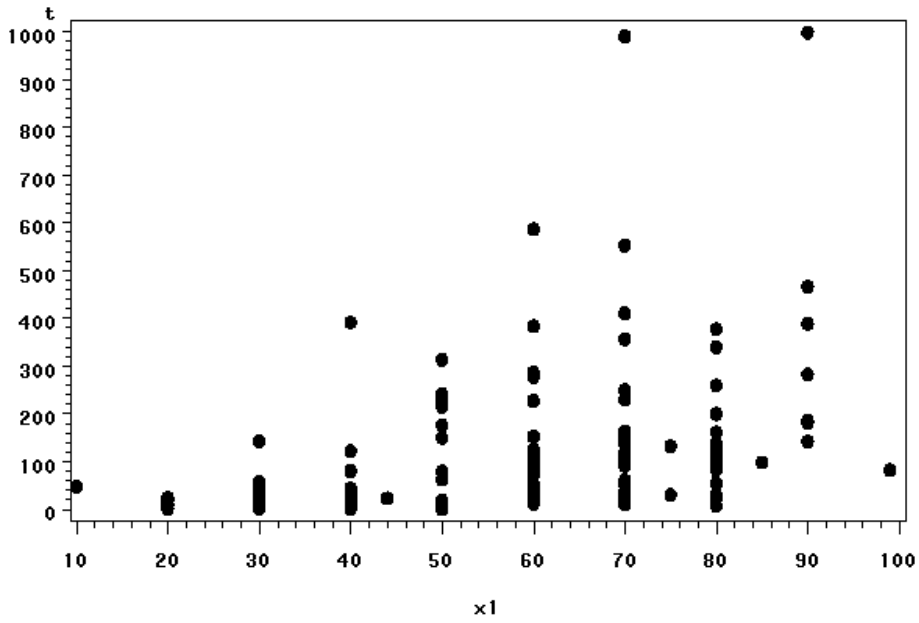

Figure 5.3: Time vs. Age

The conditional hazard function for all patients in the study would now be

$$\lambda(t;z) = \lambda_0(t)\exp\{-0.02928(\text{Performance Status})$$

$$+0.58873(\text{Small vs. Large Cell}) + 1.02136(\text{Adeno vs.Large Cell})\}.$$

The proportional hazard model function that we want to estimate here is the survivor function. The estimated survivor function for the all of the patients in the study is:

$$S(t;z) = \exp[-\int_0^t \lambda_0(u)\exp\{-0.02928(\text{Performance Status})$$

$$+0.58873(\text{Small vs. Large Cell}) + 1.02136(\text{Adeno vs.Large Cell})\}.$$

It is not necessary to calculate the α_i for the estimation of the $S(t)$, but it is now possible. Table 5.13-5.15 is pointwise estimate of the survival function for all of the given observations (the observations are ordered by increasing failure times). Figure 5.6 is a graph of the estimated survivor function versus time.

Now that all of variables in the model are significant, it is time start interpretation. First note the sign of the regression coefficient. A positive sign means the hazard is higher and the prognosis is worse. For explanatory variables, which are continuous (such as age, disease duration, and performance status), the regression coefficient refers to an increase in log hazard for an increase of 1 in the value of the covariate. For example, the estimated hazard of death decreases by $\exp(-0.02928) = 2.89$ percent for every 10 points in the Karnofsky scale.

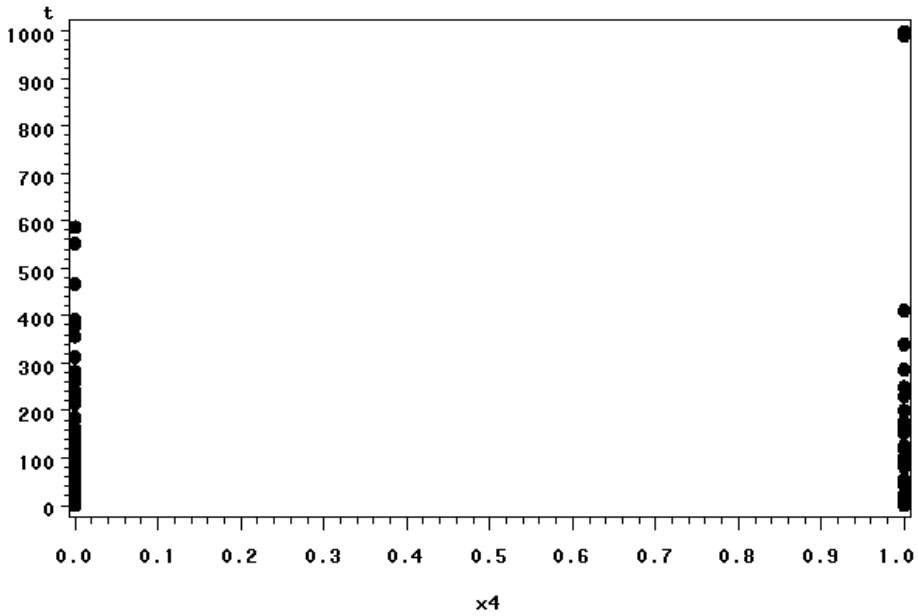

Figure 5.4: Time vs. Prior Therapy

From the analysis of this data, one can conclude that the best survivor and hazard functions are formed from the proportional hazards model for all patients in the study where the explanatory variables are performance status, indicator for small cell versus large cell, and an indicator variable for adeno versus large cell.

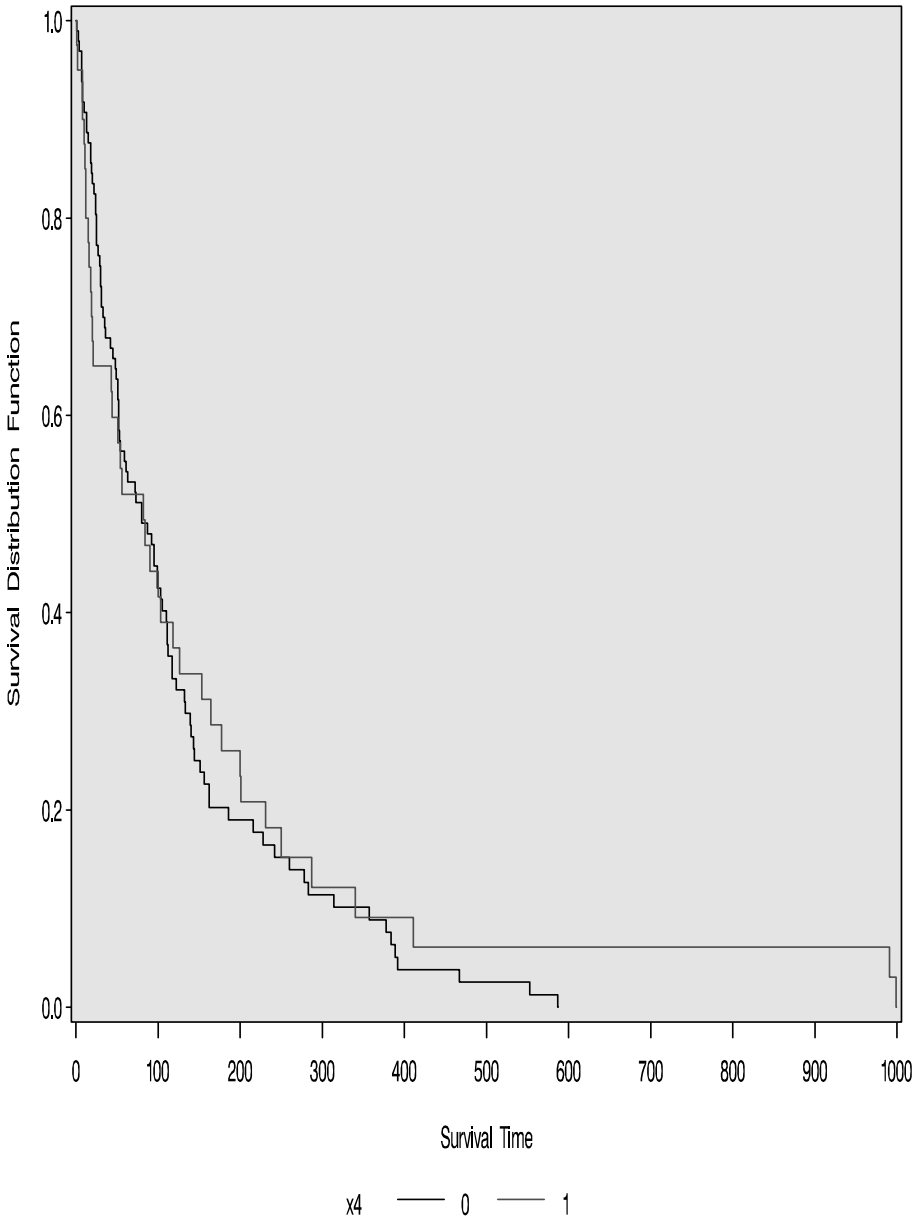

Figure 5.5: Survivor Functions of Prior Therapy and No Prior Therapy

Table 5.13: Estimated Survivor Function (Obs. 1-60).

t	Survivor Function Estimate	No. Failed	No. Left
0.000	1.0000	0	137
1.000	0.9854	2	135
2.000	0.9781	3	134
3.000	0.9708	4	133
4.000	0.9635	5	132
7.000	0.9416	8	129
8.000	0.9124	12	125
10.000	0.8978	14	123
11.000	0.8905	15	122
12.000	0.8759	17	120
13.000	0.8613	19	118
15.000	0.8467	21	116
16.000	0.8394	22	115
18.000	0.8175	25	112
19.000	0.8029	27	110
20.000	0.7883	29	108
21.000	0.7810	30	107
21.000*	.	30	106
22.000	0.7737	31	105
24.000	0.7589	33	103
25.000	0.7368	36	100
25.000*	.	36	99
27.000	0.7294	37	98
29.000	0.7219	38	97
30.000	0.7070	40	95
31.000	0.6922	42	93
33.000	0.6847	43	92
35.000	0.6773	44	91
36.000	0.6698	45	90
42.000	0.6624	46	89
43.000	0.6549	47	88
44.000	0.6475	48	87
45.000	0.6401	49	86
48.000	0.6326	50	85
49.000	0.6252	51	84
51.000	0.6028	54	81
52.000	0.5805	57	78
53.000	0.5731	58	77
54.000	0.5582	60	75

Table 5.14: Estimated Survivor Function (Obs. 61-101).

t	Survivor Function Estimate	No. Failed	No. Left
56.000	0.5507	61	74
59.000	0.5433	62	73
61.000	0.5359	63	72
63.000	0.5284	64	71
72.000	0.5210	65	70
73.000	0.5135	66	69
80.000	0.4987	68	67
82.000	0.4912	69	66
83.000*	.	69	65
84.000	0.4837	70	64
87.000	0.4761	71	63
87.000*	.	71	62
90.000	0.4684	72	61
92.000	0.4607	73	60
95.000	0.4454	75	58
97.000*	.	75	57
99.000	0.4298	77	55
100.000	0.4219	78	54
100.000*	.	78	53
103.000	0.4060	80	51
105.000	0.3981	81	50
110.000	0.3901	82	49
111.000	0.3742	84	47
112.000	0.3662	85	46
117.000	0.3503	87	44
118.000	0.3423	88	43
122.000	0.3344	89	42
123.000*	.	89	41
126.000	0.3262	90	40
132.000	0.3181	91	39
133.000	0.3099	92	38
139.000	0.3017	93	37
140.000	0.2936	94	36
143.000	0.2854	95	35
144.000	0.2773	96	34
151.000	0.2691	97	33
153.000	0.2610	98	32
156.000	0.2528	99	31
162.000	0.2365	101	29

Table 5.15: Estimated Survivor Function (Obs 101-137).

t	Survivor Function Estimate		
164.000	0.2283	102	28
177.000	0.2202	103	27
182.000*	.	103	26
186.000	0.2117	104	25
200.000	0.2033	105	24
201.000	0.1948	106	23
216.000	0.1863	107	22
228.000	0.1778	108	21
231.000	0.1694	109	20
231.000*	.	109	19
242.000	0.1605	110	18
250.000	0.1515	111	17
260.000	0.1426	112	16
278.000	0.1337	113	15
283.000	0.1248	114	14
287.000	0.1159	115	13
314.000	0.1070	116	12
340.000	0.0981	117	11
357.000	0.0891	118	10
378.000	0.0802	119	9
384.000	0.0713	120	8
389.000	0.0624	121	7
392.000	0.0535	122	6
411.000	0.0446	123	5
467.000	0.0357	124	4
553.000	0.0267	125	3
587.000	0.0178	126	2
991.000	0.00891	127	1
999.000	0	128	0

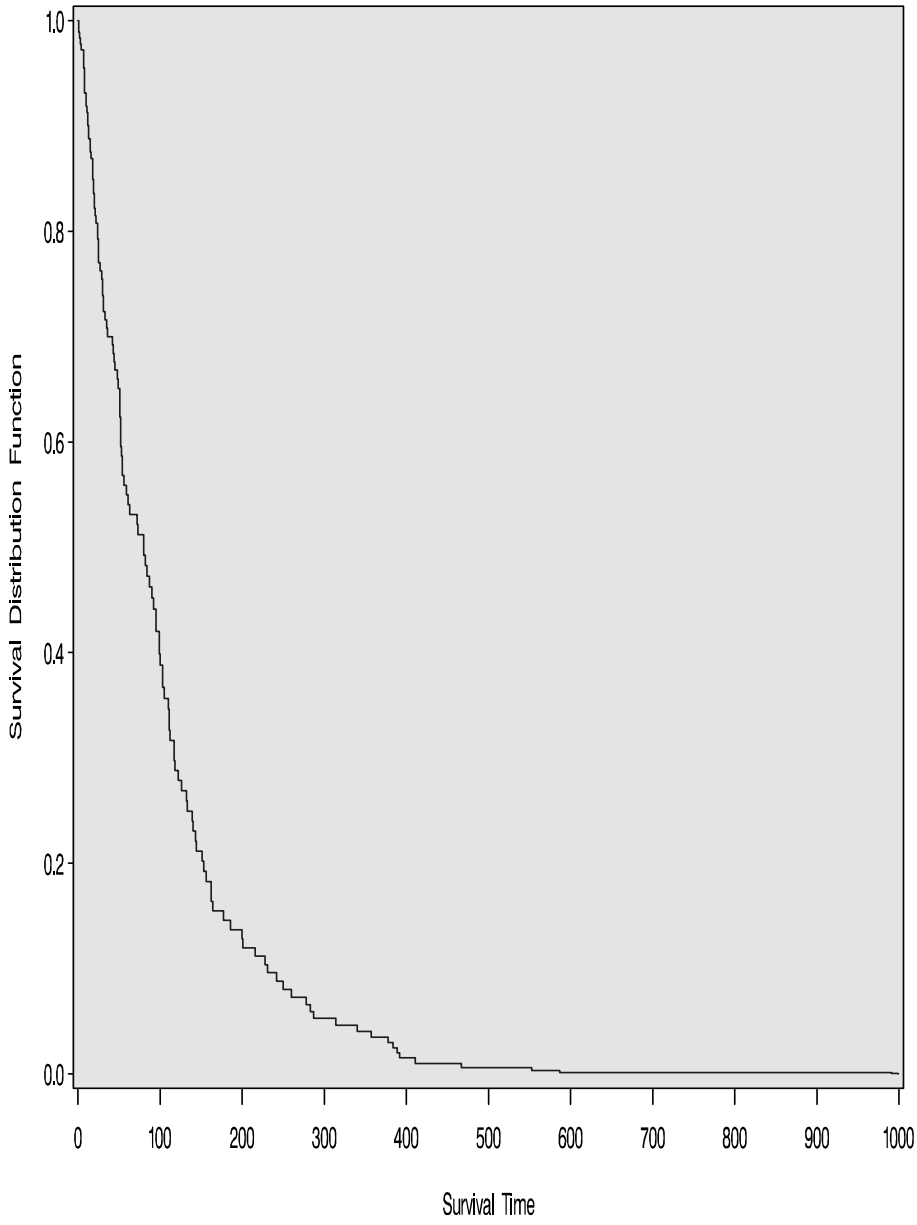

Figure 5.6: Final Survivor Function

Chapter 6. Conclusions

A failure-time regression model is a regression model that contains explanatory variables that predict why some units fail quickly and other units survive longer. These explanatory variables are required to be in the model in order to be able to make unbiased inferences, unbiased confidence intervals, and unbiased failure probabilities.

Survival data is made up of a response variable that measures the time until a specified failure occurs and a set of independent variables thought to be associated with the failure-time variable. Examples of survival data studies include component lifetimes in industrial reliability, durations of jobs, and survival times in clinical trials. Survival analysis' main goal is to model the underlying distributions of failure times and to assess the dependence of the failure time on other explanatory variables. In some situations, the failure time is not observed due to early withdrawal from the study or termination of the study; this phenomenon is known as censoring. Survival analysis models correctly model both the censored and uncensored observations.

The Cox Regression model is a very significant model due to its ability to be a non-parametric model and a parametric model at the same time. It is a parametric model due to the parameter β in the model. The failure time distribution is assumed known except for a few scalar parameters. It is non-parametric in the sense of λ_0, which is an unspecified function in the form of the baseline hazard function which is undefined. Which these together make the model more flexible, but one must be very careful when dealing with approximations and testing.

Another flexiblily of the Cox regression model is the error term. Measurements are always measured with error involved. The PHM can also be carried out when the response time is not known exactly, but is known to lie in an interval. Such a situation arises in clinical trials when errors occur measuring the failure time.

The data in the present context appears as (I_j, Z_j), $j = 1, 2, \ldots, n$ where $I_j = (L_j, R_j]$ is the interval in which the failure occured, and Z_j is the r-dimensional vector of covariates. It is possible that several of these intervals overlap with each other. If a response X_j is right-censored then $R_j = \infty$. Likewise, left-censoring leads to $L_j = 0$.

It is assumed that

1. the censoring mechanism is independent of both the failure time distribution and the covariates.

2. each subject will eventually fail in the absence of censoring.

Then the likelihood is proportional to

$$L = \prod_{i=1}^{n} (G(L_i|Z_i) - G(R_i|Z_i)),$$

where $G(x|z) = P(X > x|Z = z)$.

Let $0 = s_0 < s_2 < \cdots < s_m = \infty$ be an ordering that contains $\{L_j, R_j : j = 1, 2, \ldots, n\}$. Clearly, the likelihood L is maximized by by choosing G as a discrete distribution with masses at a subject of $\{s_0, s_1, \ldots, s_m\}$. The contribution of the ith observation to L can be written

$$\sum_{j=1}^{m} \alpha_{ij} \left[G(s_{j-1}|z_i) - G(s_j|z_i) \right],$$

where $\alpha_{ij} = 1$ if $(s_{j-1}, s_j) \subset I_i (= (L_i, R_i])$ and is zero otherwise. By the PHM, $G(s_j|z_i) = [G(s_j)]^{\exp(z_i\beta)}$ where $G(s_j) = P(S > s_j|z_j = 0)$. Let $\delta_j = \log[-\log G(s_i)]$. Then

$$\begin{aligned} \log L &= \log \prod_{i=1}^{n} \sum_{j=1}^{n} \alpha_{ij} \left((G(s_{j-1}))^{\exp(z_i\beta)} - G((s_j))^{\exp(z_i\beta)} \right) \\ &= \sum_{i=1}^{n} \log \sum_{j=1}^{m} \alpha_{ij} \left(\exp(-\exp(z_j\beta + \delta_{j-1})) - \exp(-\exp(z_j\beta + \delta_j)) \right) \end{aligned}$$

where $G(0) = 1$ and $G(s_m) = 0$. Thus $\log L$ is a function of s_j, $j = 1, 2, \ldots, m-1$, and β. By setting $\frac{\partial \log L}{\partial \delta_j} = 0$ and $\frac{\partial \log L}{\partial \beta_j} = 0$. We can get the MLEs for the survival curve and β as before. This area of the PHM could be a good area to perform further research in.

Bibliography

[1] Andersen P.K. and Gill, R.D., 'Cox's Regression Model for Counting Processes: A Large Sample Study' in *The Annals of Statistics*,Vol. 10, No. 4, p. 1100-1120.

[2] Balakrishnan, N. and Basu, Asit P. editors, *The Exponential Distribution: Theory Methods and Applications*, Gordon and Breach Publishers, Amsterdam, 1995.

[3] Cox, D.R. and Oakes, D., *Analysis of Survival Data*,Chapman and Hall, London, 1984.

[4] Finkelstein, Dianne M., 'A Proportional Hazards Model for Interval-Censored Failure Time Data' in *Biometrics*, Vol.42, No. 4, Dec. 1986, p. 845-854.

[5] Gill, Richard D., 'Understanding Cox's Regression Model: A Martingale Approach' in *Journal of American Statistical Association*, Theory and Methods Section, June 1984, Volume 79, Number 386, p 441-447.

[6] Johansen, Soren, 'An Extension of Cox's Regression Model' in *A Review*, Longman Group Limited, International Statistical Institute, Great Britain, Vol. 51, 1983, p. 165-174.

[7] Kalbfleisch, J.D. and Prentice, R.L., *The Statistical Analysis of Failure Time Data*, John Wiley and Sons, New York, 1980.

[8] Kumar, Dhananjay, *Reliability Analysis Considering Operation Conditions in a Mine*, Licentiate Thesis, Lulea University of Technology, ISSN 0280-8242, 1993:17L.

[9] Meeker, William Q. and Escobar, Luis A., *Statistical Methods for Reliability Data*, New York John Wiley and Sons, Inc., New York , 1998.

[10] Meyer, P.A. , 'Un Cours sur les Integrales Stochastiques,' Seminaire de Probabilies X, *Lecture Notes in Mathematics*, p.245-400, 1976.

[11] Neter, John, Micheal Kutner, Christopher Nachtsheim, and William Wasserman, *Applied Linear Statistical Models*, Boston, WCB McGraw-Hill, 1996.

[12] Prentice, R.L., 'Exponential Survivals with Censoring and Explanatory Variables' in *Biometrika*, Biometrika Trust, Great Britain, Vol.60, No.2, 1973, p.279.

[13] Rebolledo, R. , 'Central Limit Theorems for Local Martingales,' *Zeitschrift fur Wahrscheinlichkeitstheorie and Verwandte Gebieten*, p.51 and 269-286, 1980.

[14] Ross, Sheldon, *A First Course in Probability, 3rd Edistion*, Macmillian Publishing Company, New York, 1988.

[15] Ross, Sheldon, *Introduction to Probability Models, 7th Edition*, Harcourt Academic Press, San Diego, 2000.

[16] Sasieni, Peter, 'Some New Estimators for Cox Regression' in *The Annals of Statistics*,Vol.21, No.4, Dec. 1993, p. 1721-1759.

[17] Therneau, Terry M. and Grambsch, *Modeling Survival Data, Extending the Cox Model*, Springer, New York, 2000.

[18] Tsiatis, Anastasios A.,'A Large Sample Study of Cox's Regression Model' in *The Annals of Statistics*, Vol.9, No. 1, p. 93-108.

Appendix: General SAS Program

This is the SAS program written to analyze the data in the Data chapter. This program was generated in SAS analyst.

```
    title;
footnote;
Proportional Hazards Models *** ;
options pageno=1;
proc phreg data=proj.one2 OUTEST=WORK.ESTIM COVOUT;
model T * X9 (0) = X1 X2 X3 X4 X5 X6 X7 X8;
** Create output data set for saving data **;
output OUT=WORK.SCORE STDXBETA=STDERR SURVIVAL=SURV LOG-
SURV=LSURV
NUMLEFT=LEFT LMAX=LMAX;
baseline out=work.surv survival=surviv upper=sdfucl lower=sdflcl;
run; quit;
goptions reset=all device=WIN;
** Survival plot **;
title;
footnote;
goptions ftext=SWISS ctext=BLACK htext=1 cells;
proc gplot data=work.surv ;
label t = 'Survival Time';
axis2 minor=none major=(number=6)
label=(angle=90 'Survival Distribution Function');
symbol1 i=stepj c=BLUE l=1 width=1;
plot surviv * t=1 /
description="SDF of t"
frame cframe=CXF7E1C2 caxis=BLACK
vaxis=axis2 hminor=0 name='SDF';
run;
quit;
goptions ftext= ctext= htext= reset=symbol;
```

Appendix: SAS Program, No Prior vs. Prior Therapy

This is the SAS program written to analyze the data in the Data chapter of this paper, for 97 patients (6 of which were censored) that have no prior treatment for their cancer versus the 40 patients (3 of them were censored) that have had prior treatment for their cancer. This program was generated in SAS analyst.

title; footnote; *** Life Table Analysis *** ; options pageno=1; proc lifetest data=proj.one2 method=pl OUTSURV=work.surv; time T * X9 (0); strata X4; test X1 X2 X3 X5 X6 X7 X8; run; quit; goptions reset=all device=WIN; data work.surv; set work.surv; if survival ¿ 0 then lsurv = -log(survival); if lsurv ¿ 0 then llsurv = log(lsurv); run; ** Survival plots **; title; footnote; goptions reset=symbol; goptions ftext=SWISS ctext=BLACK htext=1 cells; proc gplot data=work.surv ; label t = 'Survival Time'; axis2 minor=none major=(number=6) label=(angle=90 'Survival Distribution Function'); symbol1 i=stepj l=1 width=1; symbol2 i=stepj l=2 width=1; symbol3 i=stepj l=3 width=1; plot survival * t = x4 / description="SDF of t by x4" frame cframe=CXF7E1C2 caxis=BLACK vaxis=axis2 hminor=0 name='SDF'; run; symbol1 i=join l=1 width=1; symbol2 i=join l=2 width=1; symbol3 i=join l=3 width=1; quit; goptions ftext= ctext= htext= reset=symbol;

Vita

Lindsay Smith was born on October 16, 1980. She grew up in the small town of Corning, New York. After graduation from high school in June 1998, she entered College of Charleston in Charleston, South Carolina, with no particular major. After two years there, she transferred to Virginia Polytechnic Institute and State University with mathematics major. In May 2002, she graduated with a bachelor's of science degree in applied and computational mathematics with a concentration and minor in statistics. She then moved to Baton Rouge, Louisiana, to pursue a master's degree in mathematics. Upon completion of her master's degree, Lindsay plans on securing a position teaching in Baton Rouge.