2005

Survival analysis

Mohammad Alif Wardak

SURVIVAL ANALYSIS

A Thesis

Presented to the

Faculty of

California State University,

San Bernardino

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

in

Mathematics

by

Mohammad Alif Wardak

December 2005

SURVIVAL ANALYSIS

A Thesis

Presented to the

Faculty of

California State University,

San Bernardino

by

Mohammad Alif Wardak

December 2005

Approved by:

[redacted]

12/1/2005

Dr. Charles Stanton, Advisor,
Department of Mathematics

Date

[redacted]

Dr. Terry Hallett

[redacted]

Dr. Yuichiro Kakihara

[redacted]

Dr. Peter Williams, Chair
Department of Mathematics

JTHallett

Dr. Terry Hallett,
Graduate Coordinator
Department of
Mathematics

ABSTRACT

The tools used in survival analysis are the Kaplan-Meier Estimator, a non-parametric statistic, and the Cox Proportional Hazard method. The Kaplan-Meier method estimates the survival curve taking into account censored data. Cox Proportional Hazard results include total values/censored values, covariate non-parametric estimate, standard error, chi-square statistic, P-value, and hazard ratio. We used the Mayo Clinic study of 418 Primary Biliary Cirrhosis patients during a ten-year period. In using these methods we found that the Kaplan-Meier survival curves were significantly different between the groups. Kaplan-Meier results include total values/censored values.

The results indicate that drugs did not have a major difference on the outcome of the tests. Gender was the substantial determining factor.

ACKNOWLEDGMENTS

I would like to extend my sincere gratitude and appreciation to Dr. Charles Stanton for his guidance throughout this period of research. He was always there when I needed his advice, even during the summer when he had no classes. I am extremely fortunate to have this opportunity to learn and work with him. I would also like to thank Dr. Terry Hallett, and Dr. Yuichiro Kakihara, for their time in carefully reading my thesis and for the many helpful suggestions and comments that greatly improve my thesis. I would also like to extend my thanks to the Mathematics Department for the invaluable learning experience that I have received.

TABLE OF CONTENTS

CHAPTER ONE

INTRODUCTION TO SURVIVAL ANALYSIS

The term "survival analysis" pertains to a statistical approach designed to take into account the amount of time an experimental unit contributes to a study. That is, it is the study of time between entry into observation and a subsequent event. In survival analysis we observe the length of time from a starting point (such as the date of a hospital admission) until the occurrence of an endpoint event (such as death), often referred to as a "failure." A key characteristic of survival analysis is the inclusion of partially missing (so-called "censored") data. For example, if a woman is alive at study's end we do not know how long she is going to live; however if her start point occurred 180 days earlier, we do know that her survival time is at least 180 days. Loss to follow-up, and "closing the files" when a study ends are common censoring events.

There are two aspects of survival analysis that make it interesting from a data analysis perspective which are:

1. The response variable, time to failure, is usually not normally distributed.

2. Survival analysis often involves censored data.

Originally the event of interest was death, hence the term, "survival analysis." The analysis consisted of following the subject until death. The uses in the survival analysis of today vary quite a bit. Applications now include time until onset of disease, employment, equipment failure, earthquake, and so on. The best way to define such events is simply to realize that these events are a transition from one discrete state to another at an instantaneous moment in time. Of course, the term "instantaneous", which may be years, months, days, minutes, or seconds, is relative and has only the boundaries set by the researcher.

The origin of survival analysis goes back to mortality tables from centuries ago. However, it was not until World War II that a new era of survival analysis emerged (See,[8]). This new era was stimulated by interest in reliability (or failure time) of military equipment. At the end of the war these newly developed statistical methods emerging from strict mortality data research were applied to failure time research, and quickly spread through private industry as customers became more demanding of safer, more reliable products. As the uses of survival analysis grew, parametric models gave way to nonparametric and semi parametric approaches because of their appeal in dealing with the ever-growing field of clinical trials in medical research. Survival analysis was well suited for such work because medical intervention follow-up studies could start without all experimental units enrolled at the start of the observation time and could end before all experimental units had experienced an event. This is extremely important because even in the best-developed studies there will be subjects who choose to quit participating, who move too far away to follow, or who will die from some unrelated event. The researcher was no longer forced to withdraw the experimental unit and all associating data from the study; instead techniques called censoring enable researchers to analyze incomplete data due to delayed entry or withdrawal from the study. This was important in allowing each experimental unit to contribute all of the information possible to the model for the amount of time the researcher was able to observe the unit.

Current software packages and high performance computers now make applying survival analysis techniques easier to solve because of their computationally intensive algorithms.

Some of the tools used in survival analysis are the cumulative distribution function $F(t)$, the probability density function $f(t)$, the survival function $S(t)$, and the hazard function, $h(t)$. The survival function data is generally described and modeled in terms of two related functions, the survivor function and hazard function. The survivor function, $S(t)$, represents the probability that an individual survives from the time origin to some time beyond t, it is positive and ranges from 0 to 1. It is defined as $S(0)=1$ and as t approaches ∞, $S(t)$ approaches 0. The survivor function can be estimated non-parametrically from observed data, both censored and uncensored, using the Kaplan-Meier method. This method is also called the product-limit method and is based on maximum likelihood estimation. Suppose deaths occur at times $t_1 < t_2 ... < t_j ... < t_n$. The Kaplan-Meier estimator is the estimator used by most software packages because of the simplistic step idea. The Kaplan-Meier estimator incorporates information from all of the observations available, by considering any point in time as a series of steps defined by the observed survival and censored times.

$$S(t) = p(T > t) = 1 - F(t) = 1 - \int_{u=0}^{t} f(u)du$$

The above survival curve describes the relationship between the probability of survival and time.

The cumulative distribution function is very useful in describing the continuous probability distribution of a random variable, such as time, in a survival analysis. The cumulative distribution function of a random variable T, denoted by $F_t(t)$, is defined by $F_t(t) = P_T(T \leq t)$. This is interpreted as a function that will give the probability that the variable T will be less than or equal to any value t that we choose. Several properties of a distribution function $F(t)$ can be listed as a consequence of the knowledge of probabilities. Note that F(t) has the probability $0 \leq F(t) \leq 1$, and $F(t)$ is a non-decreasing function of t, and as t approaches ∞, $F(t)$ approaches 1. The resulting function is also called the survivorship or survival function. The hazard function $h(t)$ is given by the following:

$$h(t) = P\{t < T < (t + \Delta) \mid T > t\} = f(t)/(1 - F(t)) = f(t)/S(t)$$

The hazard function describes the concept of the risk of an outcome (e.g., death, failure, hospitalization) in an interval after time t, conditional on the subject having survived to time t. It is the probability that an individual dies at somewhere between t and $t+\Delta$, divided by the probability that the individual survived beyond time t. The hazard function seems to be more intuitive to use in survival analysis than the probability density function because it attempts to quantify the instantaneous risk that an event will take place at time t given that the subject survived to time t (See, [8], [9]).

The survivor function and hazard function can be estimated from observed data. If the form of $F(t)$ is not specified then non-parametric procedures can be used, otherwise parametric models can be fitted to the data. The probability density function is also very useful in describing the continuous probability distribution of a random variable. Every continuous random variable has its own density function, the probability $P(a \le T \le b)$ is the area under the curve between times a and b.

Censoring or incomplete data in survival analysis experiments are designed for a shorter period of time only, and, have to account for the lost observations. If we observe a sample only for a short period of time, we only know that some individuals were alive at the end of the survey and no information on their exact time of death is available. Similarly if observations are lost during the experiment, all we know is that these individuals were still alive at some stage and no information on their exact time of death is available.

Data are called right-censored if the current survey ends at a fixed date known in advance. If the event of interest happens after this date, the observation is censored. All we know in this case is that the event might have happened after the end of the survey. Data are called left-censored if no information on the date at which the event of interest occurred is available. All we know in this case is that a certain disease occurred before the examination. Survival in two or more groups of patients can be compared using a non-parametric test such as the log-rank test, also called the Mantel-Cox test. This is the most widely used method of comparing survival curves.

There are several reasons Cox's proportional hazards modeling was chosen to explain the effect of covariates on time until event. They are the relative risk non parametric assumptions, the use of the partial likelihood function, and the creation of survivor function estimates.

The non-parametric tests for comparing survival in the Mantel-Cox method essentially calculate at each death time, for each treatment group, the expected number of deaths under the null hypothesis of no difference between groups. These are then summed to give the total expected number of deaths in each treatment group, say E, for treatment group *i*. The log-rank test for data compares the observed number of deaths in each treatment group, say O, for treatment group *i*, to the expected number by calculating the test statistic

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

and comparing it to a chi-square distribution with g-1 degrees of freedom, where *g* is the number of treatment groups.

Nonparametric methods provide an alternative series of statistical methods that require no or very limited assumptions to be made about different circumstances. Some of the more commonly used are the nonparametric alternatives to the *t*-tests, and it is these that are covered in the present review.

EPI-Info™ version 3.3.2 is the software package used in Chapters 3 and 4, especially for Cox Proportional Hazard. EPI-Info is a public domain software package designed for the global community of public health practitioners and researchers. It provides for easy form and database construction, data entry, and analysis with epidemiologic statistics, maps, and graphs. Minitab 14 was used in Chapter 1 and 2 for Kaplan-Meier Estimator (See,[14]).

CHAPTER TWO

KAPLAN-MEIER ESTIMATOR

The Kaplan-Meier estimate is a simple way to compute the survival curve. It involves computing the number of people who died at a certain time point, divided by the number of people who were still in the study at that time. These probabilities are multiplied by any earlier computed probabilities, which is one reason this is called a "product limit estimate." The Kaplan-Meier survival curve is often illustrated graphically. It looks like a poorly designed staircase, with vertical steps downward at the time of death of each individual subject (See Appendix D).

Often we will compare curves for two different groups of subjects. For example, the survival pattern for subjects on a standard therapy may be compared to a newer therapy. We can look for gaps in these curves in a horizontal or vertical direction. A vertical gap means that at a specific time point, one group had a greater fraction of subjects surviving. A horizontal gap means that it took longer for one group to experience a certain fraction of deaths.

To compute a survival curve, we need to note the time of occurrence of events (e.g., failures, deaths) and let $t_1, t_2, t_3, \ldots$ represent the times when a death or failure occurs. It is possible for two or more events to occur at the same time, in which case the number of distinct times is less than the number of deaths or failures. We need to place the t's in order from smallest to largest, that is,

$t_1 < t_2 < t_3 < \ldots$.

We also need to define the starting point of the study, $t_0 = 0$. The basic computations for the Kaplan-Meier survival curve rely on the computation of conditional survival probabilities. In particular, the probability $P[T \geq t_i | T \geq t_{i-1}]$ which can be interpreted as the probability of a subject survival to a specific time, given that the subject survived to the previous time. This probability is easy to calculate if we know the number of deaths or failures at a specific time and if we also know the number of patients at risk at that time.

A more difficult (but more important) probability is the unconditional probability of survival, $P[T \geq t_i]$ which represents the simple probability of survival to a

Armed with this information we can now compute a Kaplan-Meier survival curve. First we need to calculate the number of patients at risk, $n_i = n_{i-1} - d_{i-1} - c_{i-1}$. In other words, the number at risk at any specific time point is just the number at risk at the previous time point, minus the number of deaths/failures and the number of censored observations. For convenience, we define n_0 to be the total number of patients in the study, c_0 to be the number of censored observations prior to the first death or failure, and $d_0 = 0$. Next we compute the estimate of the conditional probability of survival: (See, [1], [9]).

$$P[T \geq t_i \mid T \geq t_{i-1}] = 1 - \frac{d_i}{n_i}.$$

Finally, the unconditional probability of survival is simply the cumulative product of the conditional probabilities.

$$P[T \geq t_i] = \prod_{j=1}^{i} \left(1 - \frac{d_j}{n_j} \right).$$

Censoring

Censoring is a key concept for survival analysis. Censoring is a form of missing data. In an experiment in which subjects are followed over time until an event of interest (such as death or other type of failure) occurs, it is not always possible to follow every subject until the event is observed. An event is usually death (but other events used in the literature include hospital discharge, development of a disease, and relapse of a malignancy). The event is also referred to as a failure. Subjects may drop out of the study and be lost to follow-up, or be deliberately withdrawn, or the end of the data collection period may arrive before the event is observed to happen. For such a subject, all that is known is that the time to the event was at least as long as the time to when the subject was last observed. The observed time to the event under such circumstances is censored. Survival analysis methods generally allow for censored data. Censoring may occur from the right (observation stops before the event is observed) as in censorship for survival analysis, or from the left (observation does not begin until after the event has occurred).

Suppose that the following Primary Biliary Cirrhosis data are observed from 15 ($n = 15$) with Platelets. Seven patients relapse at 9.7, 10.3, 10.6, 11, 12, 12.2, 13.6, months.

The Kaplan-Meier estimates can be calculated by constructing a table with five columns following the outline below.

1. Column 1 contains all the survival time, both censored and uncensored in order from largest to smallest.

2. The second column, labeled i, consists of the corresponding rank of each observation in column 1.

3. The third column, labeled r, pertains to uncensored observations only. Let $r = i$.

4. Compute $(n-r)/(n-r+1)$, or p_i, for every uncensored observation $t_{(i)}$ in column 4 to give the proportion of patients surviving up to and then through $t_{(i)}$.

5. In column 5, $\hat{S}(t)$ is the product of all values of $(n-r)/(n-r+1)$ up to and including t. If some uncensored observations are ties, the smallest $\hat{S}(t)$ should be used.

To summarize this procedure, let n be the total number of patients whose survival times, censored or not, are available. Re-label the n survival times in order of increasing magnitude such that $t_{(1)} \leq t_{(2)} \leq \ldots \leq t_{(n)}$. Then $\hat{S}(t) = \prod_{t_{(r)}} \frac{n-r}{n-r+1}$ where r runs through those positive integers for which $t_{(r)} \leq t$ and $t_{(r)}$ is uncensored. The values of r are consecutive integers $1,2,\ldots,n$ if there are no censored observations; if there are censored observations, they are not counted. The estimated median survival time is 50 percentile, which is the value of t at $\hat{S}(t) = 0.50$. See Appendix B for an example of the calculation of a Kaplan-Meier estimate. For calculations by Minitab (see Appendix C) and for graph of Kaplan-Meier regarding survival curves of genders, (see Appendix D).

Log - Rank Test

Often it is of interest to determine whether two or more samples could have arisen from identical survivor functions. One approach would involve the use of the asymptotic results for $\hat{F}(t)$ mentioned above to devise a test for equality of the survivor functions at some pre-specified time t. Such a procedure, however, would not usually make efficient use of the available data, and attention in recent years has turned instead to test statistics that attempt to summarize differences between survivor function estimators over the whole of the study period. The log-rank test is particularly good when the ratio of hazard functions in the populations being compared is approximately constant. It can also be advocated on the basis of ease of presentation to non-statistical personnel since the test statistic is the difference between the observed number of failures in each group. It is a quantity that, for most purposes, can be thought of as the corresponding expected number of failures under the null hypothesis (See, [2], [4]).

Suppose one wishes to test the equality of the survivor functions $F_1(t),...,F_r(t)$ on the basis of samples from each of r populations. Let $t_1 < t_2 <,..,< t_k$ denote the failure times for the sample formed by pooling the r individual samples. Suppose d_j failures occur at t_j and the n_j study subjects are at risk just prior to $t_j (j = 1,...,k)$ and let d_{ij} and n_{ij} be the corresponding numbers in sample $i (i = 1,...,r)$. The data at t_j are in the form of a $2 \times r$ contingency table with d_{ij} failures and $n_{ij} - d_{ij}$ survivors in the ith row $(i = 1,..,r)$. Conditional on the failure and censoring experience up to time t_i the distribution of $d_{1j},...,d_{rj}$ is simply the product of binomial distributions

$$\prod_{i=1}^{r} \binom{n_{ij}}{d_{ij}} \lambda_j d_i (1-\lambda_j)^{n_j - d_j} \qquad (2.1)$$

where λ_j is the conditional failure probability at t_j which is common for each of the r samples under the null hypothesis. The conditional distribution for $d_{1j},...,d_{rj}$ given d_j is then the hyper-geometric distribution

$$\frac{\prod_{1}^{r} \binom{n_{ij}}{d_{ij}}}{\binom{n_j}{d_j}}.$$

The mean and variance of d_{ij} from (2.1) are, respectively,

$$w_{ij} = n_{ij} d_j n_j^{-1} \text{ and } (V_j)_{ii} = n_{ij}(n_j - n_{ij}) d_j (n_j - d_j) n_j^{-2} (n_j - 1)^{-1}$$

The covariance of d_{ij} and d_{lj} is $(V_j)_{il} = -n_{ij} n_{lj} d_j (n_j - d_j) n_j^{-2} (n_j - 1)^{-1}$. Thus the statistic $v'_j = (d_{ij} - w_{ij},...,d_{rj} - w_{rj})$ has (conditional) mean zero and variance matrix V_j, where the prime denotes vector transpose. See Appendix E for an example of the log-rank test.

CHAPTER THREE

COX PROPORTIONAL HAZARD REGRESSION MODEL

The Cox proportional Hazard model is probably the most widely used method for modeling survival data. For data with one explanatory variable, i.e. one covariate, non-parametric methods like plotting Kaplan-Meier survival probabilities may be adequate if the groups being compared are reasonably similar. Frequently however, the groups being compared differ in many respects. They may have different age distributions, different proportions of men and women, different smoking habits etc. These differences come in addition to the covariates we are really interested in, and the analysis must be adjusted to compensate for these other differences, which may otherwise confound the analysis. The Cox proportional hazards model is a semi-parametric model for fitting survival data. The basic model is as follows:

$$h(t|Z) = h_0(t) \cdot \exp(\beta' Z)$$

where $h_0(t)$ is the baseline hazard which may vary arbitrarily over time, and z is the covariate vector. The covariates may be time-dependent but are fixed at the start of the study. The vector $\beta = (\beta_1,...,\beta_n)$ is a vector of covariate coefficients. The baseline hazard is treated non-parametrically, but the individual covariate effects (β_p) are assumed to be constant throughout the study. The model is often called the proportional hazards model because of this constant covariate effect throughout the study. If two individuals are compared that have covariate values Z and Z^* the ratio of their hazard rates at any time point simplifies to

$$\frac{h(t|Z)}{h(t|Z^*)} = \frac{h_0(t)\exp[\sum_{k=1}^{p} \beta Z_k]}{h_0(t)\exp[\sum_{k=1}^{p} \beta Z_k^*]} = \exp[\sum_{k=1}^{p} \beta_k (Z_k - Z_k^*)]$$

This ratio is constant or "proportional" throughout the study. This assumption greatly facilitates the interpretation of covariate effects, as the effect of a given covariate compared to the absence of that covariate is expressed as a single constant. This does not however imply that the absolute difference between the two individuals discussed above is constant; the exponentiated covariates act multiplicatively on a baseline hazard which may vary freely (See, [3]).

Cox Model with Several Covariates

Fitting of the multivariate Cox proportional hazards model would be conducted by starting with a model with all variables listed above. One by one, the least significant variable would be removed until only significant variables remained in the model. Data for overall survival was modeled in the same way. (See,[5]).

The Assumption of Proportional Hazards

Since the Cox proportional hazards model relies on the hazards to be proportional, i.e. that the effect of a given covariate does not change over time, it is very important to verify that the covariates satisfy the assumption of proportionality. If the assumption is violated, the simple Cox model is invalid, and more sophisticated analyses are required. If the interest centers upon a binary covariate, Z_1 whose relative risk changes over time, one approach is to introduce a time-dependent covariate as follows. Let

$$Z_2(t) = \begin{cases} Z_1 & \text{if the covariate } Z_1 \text{ takes on the value 1} \\ 0 & \text{if the covariate } Z_1 \text{ takes on the value 0,} \end{cases}$$

where $g(t)$ is a known function of time. In such cases, it may be preferable to use a procedure that would allow the function $g(t)$ to be estimated from the data. One approach to this problem is to fit a model with an indicator function for $g(t)$. In the simplest approach, define a time-dependent covariate $Z_2(t) = \begin{cases} Z_1 & \text{if } t > \tau \\ 0 & \text{if } t \leq \tau \end{cases}$.

To determine the optimal value of τ, the model including the new covariate $z_2(t)$ is fitted for a set of values for τ, and the value of the maximized log partial likelihood is the optimal value to use. Proportional hazards can, then, be tested for each region and if it fails, for t on either side of τ then this process can be repeated in that region.

The assessment of the proportional hazards assumption can be done numerically or graphically. A great number of procedures have been proposed over the years. Some of the procedures require partitioning of failure time, some require categorization of covariates, some include a spline function, and some can be applied to the untransformed dataset. None of the methods, either numerical or graphical, are today known to be better than the others in finding out whether the hazards are proportional or not. Some authors recommend using numerical tests and others recommend graphical procedures since they believe that the proportional hazards assumption only approximates the correct model for a covariate and that any formal test, based on a large enough sample, will reject the null hypothesis of proportionality.

Maximum Likelihood

The likelihood and log-likelihood functions are the basis for deriving estimators for parameters, given data. While the shapes of these two functions are different, they have their maximum point at the same value. In fact, the value of θ that corresponds to this maximum point is defined as the Maximum Likelihood Estimate and that value is denoted as $\hat{\theta}$. This is the value that is "most likely" relative to the other values. This is a simple, concept and it has a host of good statistical properties. Thus, in general, we seek $\hat{\theta}$ such that this value maximizes the log-likelihood function (See, [4], [7]).

Generally, the calculus is used to find the maximum point of the log-likelihood function and obtain Maximum Likelihood Estimations in closed form. This is tedious and often not useful in real problems (where closed form estimator may often not even exist). The log-likelihood functions we will see have a single mode or maximum point and no local optima. These conditions make the use of numerical methods appealing and efficient. (See,[6]).

Consider first, the binomial model with a single unknown parameter, θ. Using calculus one could take the first partial derivative of the log-likelihood function with respect to the θ, set it to zero and solve for θ. This solution will give $\hat{\theta}$, the Maximum Likelihood Estimation. This value of $\hat{\theta}$, is the one that maximizes the likelihood function. It is the value of the parameter that is most likely, given the data.

The likelihood function provides information on the relative likelihood of various parameter values, given the data and the model (here, a binomial). Think of 10 of your friends, 9 of which have one raffle ticket, while the 10th friend who has 4 tickets, has a higher likelihood of winning relative to the other 9 friends. If you were to try to select the most likely winner of the raffle, which person would you pick? Most would select the person with 4 tickets. Now, what if 8 people had a single ticket, one had 4 tickets, but the last had 80 tickets. Surely the person with 80 tickets is most likely to win (but not with certainty). In this simple example you have a feeling about the "strength of evidence" about the likely winner. In the first case, one person has an edge, but not much more. In the second case, the person with the 80 tickets is relatively very likely to win.

The shape of the log-likelihood function is important in a conceptual way to the raffle ticket example. If the log-likelihood function is relatively flat, one can make the interpretation that several (perhaps many) values of p are nearly equally likely. They are somewhat alike; this is quantified as the sampling variance or standard error. If the log-likelihood function is fairly flat, this implies considerable uncertainty and this is reflected in large sampling variances and standard errors, and wide confidence intervals. On the other hand, if the log-likelihood function is fairly peaked near its maximum point, this indicates some values of p are relatively very likely compared to others (like the person with 80 raffle tickets). There is some considerable degree of certainty implied and this is reflected in small sampling variances and standard errors, and narrow confidence intervals. So, the log-likelihood function at its maximum point is important as well as the shape of the function near this maximum point.

The shape of the likelihood function near the maximum point can be measured by the analytical second partial derivatives and these can be closely approximated numerically by a computer. Such numerical derivatives are important in complicated problems where the log-likelihood exists in 20-60 dimensions. This method's advantage is that maximum likelihood provides a consistent approach to parameter estimation problems. This means that maximum likelihood estimates can be developed for a large variety of estimation situations. For example, they can be applied in reliability analysis to censored data under various censoring models (See, [10], [11]).

Maximum likelihood methods have desirable mathematical and optimality properties. Specifically,

1. They become minimum variance unbiased estimators as the sample size increases. By unbiased, we mean that if we take (infinitely many number of) random samples with replacement from a population, the average value of the parameter estimates will be theoretically exactly equal to the population, the average value of the parameter estimates will be theoretically exactly equal to the population value. By minimum variance, we mean that the estimator has a smallest variance, and thus the narrowest confidence interval, of all estimators of that type.
2. They have approximate normal distributions and approximate sample variances that can be used to generate confidence bounds and hypothesis tests for the parameters.

Several popular statistical software packages provide excellent algorithms for maximum likelihood estimates for many of the commonly used distributions. This helps mitigate the computational complexity of maximum likelihood estimation. This method's disadvantage is that, the likelihood equations need to be specifically worked out for a given distribution and estimation problem. The mathematics is often non-trivial, particularly if confidence intervals for the parameters are desired.

The numerical estimation is usually non-trivial. Except for a few cases where the maximum likelihood formulas are in fact simple, it is generally best to rely on high quality statistical software to obtain maximum likelihood estimates. Fortunately, high quality maximum likelihood software is becoming increasingly common.

Maximum likelihood estimates can be heavily biased for small samples. The optimality properties may not apply for small samples. Maximum likelihood can be sensitive to the choice of starting values.

Partial Likelihood

To obtain estimates of the covariate parameters, Cox developed a nonparametric method he called partial likelihood. Estimation of the parameter values is then obtained by use of maximum partial likelihood estimation. The partial likelihood method based on this assumption is related to h_0 being undetermined. The intervals between successive duration times (or failure times) contribute no information regarding the relationship between the covariates and the hazard rate.

This is in contrast to the parametric methods, where the actual survival times are used in the construction of the likelihood function. Because the Cox model only uses "part" of the available data ($h_0(t)$is not estimated), the likelihood function for the Cox model is a "partial" likelihood function, hence the name. To get a sense for how this works, look at the logic underlying the partial likelihood method. Consider the data in Appendix F. Here are the survival times for fifteen cases. Of these fifteen cases four of them are right-censored and coded 1. All the tables in the Appendix, 0 represents male, 1 represents female.

In the Appendix F table, the first case for t_1 occurs at 51 follow up days, t_2 occurs at 264 follow up days, t_3 occurs at 611 follow up days, t_4 occurs at 762, t_6 occurs at 1012 follow up days, t_7 occurs at 1217 follow up days, t_8 occurs at 1427 follow up days, t_9 occurs at 2466 follow up days, t_{11} occurs at 2689 follow up days, t_{14} occurs at

4079 follow up days, t_{15} occurs at 4191 follow up days, t_5, t_{10}, t_{12}, and t_{13} are censored (See Appendix G).

- Events can be ordered.
- At t_0 all cases are at risk of failing.
- After the first failure, the risk set decreases by 1.
- The risk set successively dwindles as events occur.

To motivate the partial likelihood estimator, let $\psi = \exp(\beta' x_i)$. The partial likelihood function for these data would be equivalent to:

$$L_P = \left(\frac{\Psi(1)}{\sum_{j=1}^{15}\Psi(j)}\right) \cdot \left(\frac{\Psi(2)}{\sum_{j=2}^{15}\Psi(j)}\right) \cdot \left(\frac{\Psi(3)}{\sum_{j=3}^{15}\Psi(j)}\right) \left(\frac{\Psi(4)}{\sum_{j=4}^{15}\Psi(j)}\right) \cdot \left(\frac{\Psi(6)}{\sum_{j=6}^{15}\Psi(j)}\right) \cdot \left(\frac{\Psi(7)}{\sum_{j=7}^{15}\Psi(j)}\right) \cdot$$

$$\left(\frac{\Psi(8)}{\sum_{j=8}^{15}\Psi(j)}\right) \cdot \left(\frac{\Psi(9)}{\sum_{j=9}^{15}\Psi(j)}\right) \cdot \left(\frac{\Psi(11)}{\sum_{j=11}^{15}\Psi(j)}\right) \cdot \left(\frac{\Psi(14)}{\sum_{j=14}^{15}\Psi(j)}\right) \cdot \left(\frac{\Psi(15)}{\sum_{j=15}^{15}\Psi(j)}\right)$$

For a similar illustrative calculation see Appendix G. In words, this tells us that each of the fifteen cases is at risk of experiencing an event up to the first failure time, t_5. After the first failure in the data set, the risk set decreases in size by 1; thus, the risk set up to the second failure time, t_{10}, includes all cases. By the fourth failure time in the data, t_{13}, the risk set includes only cases 5, 10, 12, and 13. By the last failure time, only case 13 remains in the risk set. This exercise shows that the partial likelihood function is solely based on the ordered duration times, and not on the length of the interval between duration times. Also, censored observations contribute information to the "risk set," that is, cases that are surviving to time t_i, but contribute no information regarding failure times. To be more formal, suppose we have a data set with n observations and k distinct failure (event) times. Cox estimation first proceeds by sorting the ordered failure times, such that

$$t_1 < t_2 < \ldots < t_k,$$

where t_i denotes the failure time for the ith individual. For censored cases, we define δ_i to be 1 if the case is right-censored, and 0 if the case is uncensored. Finally, the ordered event times are modeled as a function of covariates, x.

The partial likelihood function is derived by taking the product of the conditional probability of a failure at time t_i, given the number of cases that are at risk of failing at time t_i. That is to say, given that some event has occurred, what is the probability the event occurred to the ith individual from a risk set of size n? More formally, if we define $R(t_i)$ to denote the number of cases that are at risk of experiencing an event as time t_i, that is, the "risk set," then the probability that the jth case will fail at time T_i is given by

$$\Pr(t_j = T_i | R(t_i)) = \frac{e^{\beta' x_i}}{\sum_{j \in R(t_i)} e^{\beta' x_j}}, \qquad (3.1)$$

where the summation operator in the denominator is summing over all individuals in the risk set. Taking the product of the conditional probabilities in (3.1) yields the partial likelihood function (a similar example can be found in [5], [6]),

$$L_p = \prod_{i=1}^{k} \left[\frac{e^{\beta' x_i}}{\sum_{j \in R(t_i)} e^{\beta' x_j}} \right]^{\delta_i}$$

with corresponding log-likelihood function,

$$\log L_p = \sum_{i=1}^{k} \delta_i \left[\beta' x_i - \log \sum_{j \in R(t_i)} e^{\beta' x_j} \right] \qquad (3.2)$$

By maximizing the log-likelihood in (3.2), estimates of the β may be obtained. What is the importance of this result?

- Specifying the baseline hazard, $h_0(t)$ is unnecessary.
- The interval between events does not inform the partial likelihood function.
- Censored cases contribute information only pertinent to the risk set (i.e. the denominator, not the numerator)

The critical thing here is to note that no assumptions about the shape of the baseline hazard need to be made. Another way to see this is to think about the heuristic partial likelihood function above. All we need to know to compute a probability is ψ (or $\exp(\beta' x_i)$).

Cox demonstrated that maximum partial likelihood estimation produces parameter estimates that have the same properties as maximum likelihood estimates. This is convenient because under the same set of regularity conditions as maximum likelihood estimation the parameter estimates from partial likelihood are asymptotically normal, asymptotically efficient, consistent, and invariant. So the usual kinds of hypothesis tests discussed in the context of parametric models are directly extended to the Cox model. The first step in applying the results of the Application to the primary biliary cirrhosis data is to order the survival times from smallest to largest. Appendix G shows an example of this data. The partial likelihood for β is now formed by taking the product over all failure points to give

$$L(\beta) = \prod_{i=1}^{k} \left(\frac{\exp(z_{(i)}\beta)}{\sum_{l \in R(t_i)} \exp(z_l \beta)} \right)$$

The partial likelihood is not a likelihood in the usual sense in that the general construction does not give a result that is proportional to the conditional of marginal probability of any observed event. This is an example of a partial likelihood to be found in Appendix G.

$$PL \;:=\; \frac{e^{4\beta}}{\left(6e^{\beta}+9\right)\cdot\left(5e^{\beta}+9\right)\cdot\left(4e^{\beta}+9\right)\cdot\left(4e^{\beta}+8\right)\cdot\left(4e^{\beta}+6\right)} \cdot \frac{1}{\left(4e^{\beta}+5\right)\cdot\left(4e^{\beta}+4\right)\cdot\left(3e^{\beta}+4\right)\cdot\left(1e^{\beta}+4\right)\cdot\left(0e^{\beta}+2\right)}$$

$$
\begin{aligned}
\frac{d}{d(PL)} := {} & \frac{2e^{4\beta}}{\left(6e^{\beta}+9\right)\left(5e^{\beta}+9\right)\left(4e^{\beta}+9\right)\left(4e^{\beta}+8\right)\left(4e^{\beta}+6\right)} \cdot \\
& \frac{1}{\left(4e^{\beta}+5\right)\left(4e^{\beta}+4\right)\left(3e^{\beta}+4\right)\left(e^{\beta}+4\right)} - \\
& \frac{3e^{4\beta}e^{\beta}}{\left(6e^{\beta}+9\right)^{2}\left(5e^{\beta}+9\right)\left(4e^{\beta}+9\right)\left(4e^{\beta}+8\right)\left(4e^{\beta}+6\right)} \cdot \\
& \frac{1}{\left(4e^{\beta}+5\right)\left(4e^{\beta}+4\right)\left(3e^{\beta}+4\right)\left(e^{\beta}+4\right)} - \\
& \frac{5e^{4\beta}e^{\beta}}{2\left(6e^{\beta}+9\right)\left(5e^{\beta}+9\right)^{2}\left(4e^{\beta}+9\right)\left(4e^{\beta}+8\right)\left(4e^{\beta}+6\right)} \cdot \\
& \frac{1}{\left(4e^{\beta}+5\right)\left(4e^{\beta}+4\right)\left(3e^{\beta}+4\right)\left(e^{\beta}+4\right)} - \\
& \frac{2e^{4\beta}e^{\beta}}{\left(6e^{\beta}+9\right)\left(5e^{\beta}+9\right)\left(4e^{\beta}+9\right)^{2}\left(4e^{\beta}+8\right)\left(4e^{\beta}+6\right)} \cdot \\
& \frac{1}{\left(4e^{\beta}+5\right)\left(4e^{\beta}+4\right)\left(3e^{\beta}+4\right)\left(e^{\beta}+4\right)} - \\
& \frac{2e^{4\beta}e^{\beta}}{\left(6e^{\beta}+9\right)\left(5e^{\beta}+9\right)\left(4e^{\beta}+9\right)\left(4e^{\beta}+8\right)^{2}\left(4e^{\beta}+6\right)} \\
& \frac{1}{\left(4e^{\beta}+5\right)\left(4e^{\beta}+4\right)\left(3e^{\beta}+4\right)\left(e^{\beta}+4\right)} - \\
& \frac{2e^{4\beta}e^{\beta}}{\left(6e^{\beta}+9\right)\left(5e^{\beta}+9\right)\left(4e^{\beta}+9\right)\left(4e^{\beta}+8\right)\left(4e^{\beta}+6\right)^{2}} - \\
& \frac{1}{\left(4e^{\beta}+5\right)\left(4e^{\beta}+4\right)\left(3e^{\beta}+4\right)\left(e^{\beta}+4\right)} - \\
& \frac{2e^{4\beta}e^{\beta}}{\left(6e^{\beta}+9\right)\left(5e^{\beta}+9\right)\left(4e^{\beta}+9\right)\left(4e^{\beta}+8\right)\left(4e^{\beta}+6\right)} \cdot \\
& \frac{1}{\left(4e^{\beta}+5\right)^{2}\left(4e^{\beta}+4\right)\left(3e^{\beta}+4\right)\left(e^{\beta}+4\right)} - \\
& \frac{2e^{4\beta}e^{\beta}}{\left(6e^{\beta}+9\right)\left(5e^{\beta}+9\right)\left(4e^{\beta}+9\right)\left(4e^{\beta}+8\right)\left(4e^{\beta}+6\right)} \cdot \\
& \frac{1}{\left(4e^{\beta}+5\right)\left(4e^{\beta}+4\right)^{2}\left(3e^{\beta}+4\right)\left(e^{\beta}+4\right)} - \\
& \frac{3e^{4\beta}e^{\beta}}{2\left(6e^{\beta}+9\right)\left(5e^{\beta}+9\right)\left(4e^{\beta}+9\right)\left(4e^{\beta}+8\right)\left(4e^{\beta}+6\right)} \cdot \\
& \frac{1}{\left(4e^{\beta}+5\right)\left(4e^{\beta}+4\right)\left(3e^{\beta}+4\right)^{2}\left(e^{\beta}+4\right)} - \\
& \frac{e^{4\beta}e^{\beta}}{\left(6e^{\beta}+9\right)\left(5e^{\beta}+9\right)\left(4e^{\beta}+9\right)\left(4e^{\beta}+8\right)\left(4e^{\beta}+6\right)} \cdot \\
& \frac{1}{\left(4e^{\beta}+5\right)\left(4e^{\beta}+4\right)\left(3e^{\beta}+4\right)\left(e^{\beta}+4\right)^{2}} = 0.2875 \qquad (3.3)
\end{aligned}
$$

$PL = 0.2875$ if $\beta = 0$ The first step in applying the results of the data is to order the survival times from smallest to largest with the additional convention that failure times precede censored times. An efficient computer solution to the problem would require essentially the same organization of the data set. In general, there is advantage to begin the calculation at the last failure time since the risk set can then be formed by adding the labels of items failing or censored.

Information Matrix

Fisher information is a key concept in the theory of statistical inference and is defined in the following manner: Let $X = (X_1, \ldots, X_n)$ be a random sample, and let $f(X|\theta)$ denote the probability density function for some model of the data, which has parameter vector $\theta = (\theta_1, \ldots, \theta_k)$. Then the Fisher information matrix $I_n(\theta)$ of sample size n is given by the $k \times k$ symmetric matrix whose ij-th element is given by the covariance between first partial derivatives of the log-likelihood,

$$I_n(\theta)_{i,j} = Cov\left[\frac{\partial \ln f(X|\theta)}{\partial \theta_i}, \frac{\partial \ln f(X|\theta)}{\partial \theta_j}\right].$$

An alternative, but equivalent, definition for the Fisher information matrix is based on the expected values of the second partial derivatives, and is given by

$$I_n(\theta)_{i,j} = -E\left[\frac{\partial^2 \ln f(X|\theta)}{\partial\theta_i \partial\theta_j}\right].$$

Strictly, this definition corresponds to the expected Fisher information. If no expectation is taken we obtain a data-dependent quantity that is called the observed Fisher information. As a simple example, consider a normal distribution with mean μ and variance σ^2, where $\theta = (\mu, \sigma^2)$.. The Fisher information matrix for this situation is given by:

$$I_n(\theta) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$

It is worth noting two useful properties of the Fisher information matrix. Firstly, $I_n(\theta) = nI_1(\theta)$, meaning that the expected Fisher information for a sample of n independent observations is equivalent to n times the Fisher information for a single observation. Secondly, it is dependent on the choice of parameterization. Suppose the parameter θ is changed into another parameter $\eta = (\eta_1, ..., \eta_k)$ with $\eta_i = g_i(\theta)$ where g_i is one-to-one so its inverse $g^{-1}(\eta_i) = \theta_i$ exists. The Fisher information $I_n^*(\eta)$ for the new parameterization is obtained using the chain rule $I_n^*(\eta) = J(\eta)^T I_n(\theta(\eta))J(\eta)$, where $J(\eta)$ is the Jacobian matrix with elements $J(\eta)_{ij} = \partial g^{-1}(\eta_i)/\partial\eta_j (i,j=1,...,k)$.

Let $T(X)$ be any statistic and let $\psi(\theta)$ be its expectation such that $\psi(\theta) = E[T(X)]$. Under some regularity conditions, it follows that for all θ,

$$\mathrm{var}(T(X)) \geq \frac{\left(\frac{d\psi(\theta)}{d\theta}\right)^2}{I_n(\theta)} \qquad (3.4)$$

The value of the right hand side of (3.4) is known as the Information inequality lower bound. In particular, if $T(X)$ is an unbiased estimator for θ, then the numerator becomes 1, and the lower bound is simply $\frac{1}{I_n(\theta)}$. Note that this explains why $I_n(\theta)$ is called the "information" matrix: The larger the value of $I_n(\theta)$ is, the smaller the variance becomes, and therefore, we would be more certain about the location of the unknown parameter value. The information inequality generalizes to the multi-parameter case, where $\theta = (\theta_1,...,\theta_k)$. Let the statistic $W(X)$ be an estimator for some function $g(\theta)$. Then the inequality states that

$Var(W(X)) \geq \gamma(\theta)^T I_n(\theta)^{-1} \gamma(\theta)$ where $\gamma(\theta)$ is a $k \times 1$ column vector with elements $\gamma(\theta)_i = \partial g(\theta) / \partial \theta_i$. The Asymptotic Theory involves the maximum likelihood estimator that has many useful properties, including re-parametrization-invariance, consistency, and sufficiency. Further, it follows under some regularity conditions that the sampling distribution of a maximum likelihood estimator $\hat{\theta}_{ML}$ is asymptotically unbiased and also asymptotically normal with its variance-covariance matrix obtained from the inverse Fisher information matrix of sample size 1, that is $\hat{\theta}_{ML} \to N(\theta, I_1(\theta)^{-1} / n)$ as n goes to infinity. The Fisher information matrix also arises in Bayesian inference.

The log partial likelihood ratio test is not only the easiest test to compute, but is also the best of the three tests for assessing the significance of the fitted model. The computation of information matrix tests for the multiple proportional hazards regression model requires matrix calculations. Specifically, we denote the vector of first partial derivatives whose elements are given as $u(\beta)$. Under the hypothesis that all coefficients are equal to zero, and under the mathematical conditions needed for the partial likelihood ratio test, the vector of scores

$u(0) = u(\beta)|_{\beta=0}$ will be distributed as multivariate normal with mean vector equal to zero and covariance matrix given by the information matrix evaluated at the coefficient vector equal to zero, $I(0) = I(\beta)|_{\beta=0}$. The elements in this matrix are obtained by evaluating the expressions with the coefficient vector equal to zero. The score test statistic is

$$u'(0)[I(0)]^{-1}u(0),$$

which is distributed asymptotically as chi-square with n degrees-of-freedom. This statistic can be used to test the null hypothesis $\beta = 0$ by using a chi-square test.

CHAPTER FOUR

PRIMARY BILIARY CIRRHOSIS DATA

Primary biliary cirrhosis is a disease characterized by inflammatory destruction of the small bile ducts within the liver. Primary biliary cirrhosis eventually leads to cirrhosis of the liver. The cause of primary biliary cirrhosis is unknown, but because of the presence of auto-antibodies, it is generally thought to be an auto-immune disease. Other etiologies, such as infectious agents, have not been completely excluded. Primary biliary cirrhosis has a worldwide prevalence of approximately 5/100,000 and an annual incidence of approximately 6/1,000,000. The prevalence and incidence appear to be similar in different regions of the world. About 90% of patients with primary biliary cirrhosis are women. Most commonly, the disease is diagnosed in patients between the ages of 40 and 60 years. (See, [13]).

This data set is a follow-up to the original primary biliary cirrhosis data set. "Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits." The data from the Mayo Clinic trial in primary biliary cirrhosis of the liver conducted between

1974 and 1984 contains a description of the clinical background for the trial and the covariates. A total of 418 primary biliary cirrhosis patients, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. The first 310 cases in the data set participated in the randomized trial and contain largely complete data. The additional 108 cases did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival. Six of those cases were lost to follow-up shortly after diagnosis, so the data here are on an additional 102 cases as well as the 310 randomized participants.

The data contains only baseline measurements of the laboratory parameters. This data contains multiple laboratory results, but only on the first 310 patients. Some baseline data values in this file differ from the original primary biliary cirrhosis file, for instance, the data errors in prothrombin time and age which were discovered after the original analysis, during research work on dfbeta residuals. Another major difference is that there was significantly more follow-up for many of the patients at the time this data was assembled.

One "feature" of the data deserves special comment. The last observation before death or liver transplant often has many more missing covariates than other data rows. The original clinical protocol for these patients specified visits at 6 months, 1 year, and annually thereafter. At these protocol visits lab values were obtained for a large pre-specified battery of tests. "Extra" visits, often undertaken because of worsening medical condition, did not necessarily have all this lab work. The missing values are thus potentially informative, and violate the usual "missing at random" assumptions that are assumed in analyses. Because of the earlier published results on the Mayo primary biliary cirrhosis risk score, however, the 5 variables involved in that computation were usually obtained, i.e. age, bilirubin, albumin, prothrombin time, and edema score. The variables used were: Case number; Number of days between registration and the earlier of death, trans-plantation, or study analysis time; Status: 0=alive, 1=transplanted, 2=dead; Drug: 1= D-penicillamine, 0=placebo; Age in days, at registration; Sex: 0=male, 1=female; Day: number of days between enrollment and this visit date, remaining values on the line of data refer to this visit; Serum bilirubin n mg/dl; Serum cholesterol in mg/dl; Albumin in gm/dl; Alkaline phosphates in u/liter; SGOT in u/ml (serum glutamic-oxaloacetic transaminase, the enzyme name has subsequently changed to "ALT" in the medial literature); Platelets per cubic ml / 1000; Prothrombin time in seconds; Histologic stage of disease. We used EPI-Info to calculate Kaplan-Meier for 312 patients Primary Biliary Cirrhosis by Gender for Age, Albumin, Alkaline, Bili, Platelets, and Spiders (See Appendix J). The smaller the p-value is, more changes can be seen affecting the outcomes. The larger the p-value is, covariates are not significant. In the outcome for 312 patients by Gender there was a noticeable change. In the outcome for 312 patients by Drug in there was not a noticeable change (See Appendix L). The smallest p-value was shown for Age, Albumin, Alkaline, and Bili. The Coefficient (β) for Gender was $\beta = -.0804$, based on $h(t) = h_0(t)e^{\beta x}$ indicating that the hazard function for female is smaller (sex=1), and the male (sex=0) hazard ratio could be lower with 95 percent confidence (See Appendix H). The nonparametric survival plot for follow up days by

Placebo and Penicillamine for 312 patients is illustrated in Appendix K on the Kaplan-Meier curve for all patients $n = 312$.

We used the nonparametric survival plot for follow up days by gender for 312 patients. The plot shows the survival curves for all categories for Primary Biliary Cirrhosis (See Appendix H). The result of Minitab calculations are in Appendix I.

We used EPI-Info to calculate Primary Biliary Cirrhosis by drug for Age, Albumin, Alkaline, Bili, Platelets, and Spiders for 312 patients (See Appendix L). In the outcome for 312 patients by Drug there was no noticeable change. The result in Appendix L show that the hazard ratio for drugs is not noticeably different. The hazard rate for the drug was 0.9775. This difference could be due to the non-linear effect of the drug itself.

APPENDIX A

DATA OF PRIMARY BILIARY CIRRHOSIS I

ID	FU Days	Status	Censor	Age	Sex	Asictes	Trig	Platelets
1	321	2	0	15116	1	0	158	124
2	552	2	0	18799	0	0	122	119
3	691	0	1	21185	1	*	*	269
4	769	2	0	19060	1	0	128	224
5	877	1	1	12912	0	0	194	306
6	890	2	0	24622	0	0	91	360
7	939	0	1	22767	1	0	100	234
8	1487	2	0	22977	1	0	188	178
9	1746	2	0	19724	0	*	*	325
10	2033	1	1	12839	0	0	210	344
11	2386	2	0	18460	0	0	93	362
12	2400	2	0	15526	1	0	88	251
13	2576	0	1	17323	1	0	71	356
14	2689	2	0	12227	0	0	155	337
15	2812	2	0	18628	1	*	*	*
16	3069	0	1	19318	0	0	107	182

APPENDIX B

CALCULATION OF THE KAPLAN-MEIER ESTIMATE

Time (t)	Rank i	r	$(n-r)/(n-r+1)$	$\hat{S}(t)$
321	1	1	15/16	0.938
552	2	2	14/15	(0.938)(0.933) = 0.875
691	3	-	-	
769	4	4	12/13	(0.875)(0.923) = 0.808
877	5	-	-	
890	6	6	10/11	(0.808)(0.909) = 0.734
939	7	-	-	
1487	8	8	8/9	(0.734)(0.889) = 0.653
1746	9	9	7/8	(0.653)(0.875) = 0.571
2033	10	-	-	
2386	11	11	5/6	(0.571)(0.833) = 0.476
2400	12	12	4/5	(0.476)(0.800) = 0.381
2576	13	-	-	
2689	14	14	2/3	(0.381)(0.667) = 0.254
2812	15	15	1/2	(0.254)(0.500) = 0.127
3069	16	-	0	0

APPENDIX C

CALCULATION OF THE MINITAB FOR 15 PATIENTS

Kaplan-Meier Estimates

```
                       Number
 at  Number     Survival  Standard   95.0% Normal CI
Time      Risk  Failed  Probability     Error      Lower
                        Upper
 321         8       1     0.875000  0.116927  0.645828
                       1.00000
 769         6       1     0.729167  0.164976  0.405819
                       1.00000
1487          4       1     0.546875  0.200580  0.153745
                       0.94000
2400          3       1     0.364583  0.200086  0.000000
                       0.75675
2812          1       1     0.000000  0.000000  0.000000
                       0.00000
```

APPENDIX D

GRAPH OF NONPARAMETRIC SURVIVAL PLOT FOR FOLLOW UP DAYS FOR 15 PATIENTS

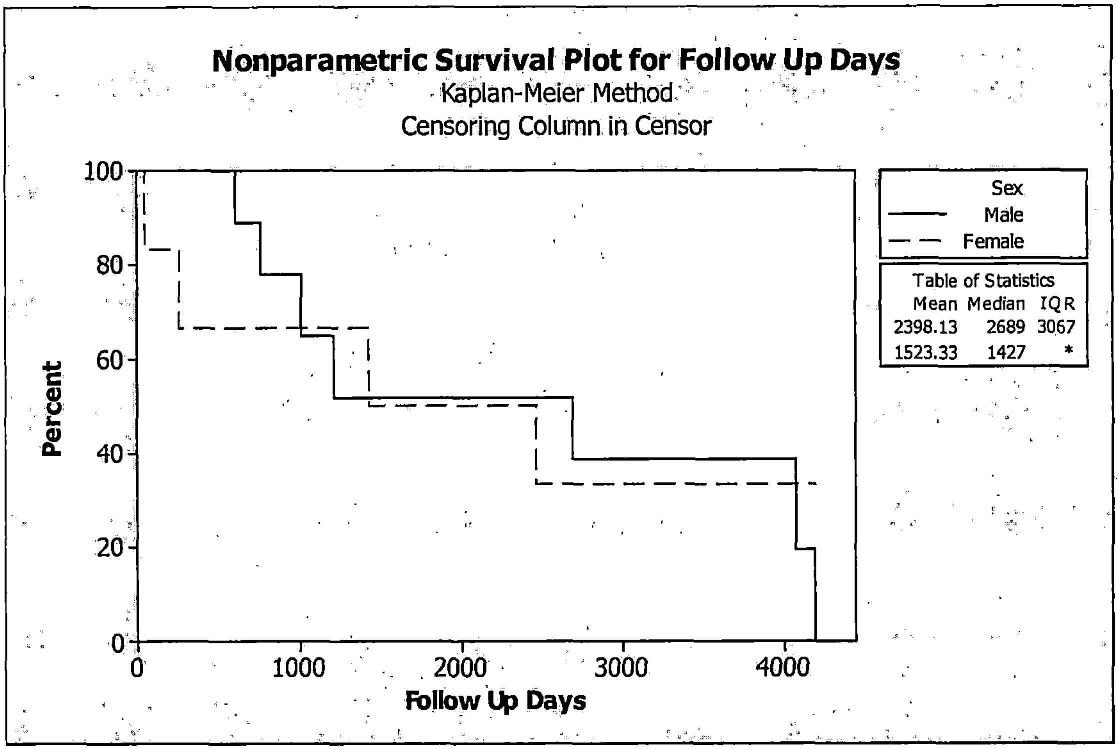

Table of Statistics

Mean	Median	IQR
2398.13	2689	3067
1523.33	1427	*

APPENDIX E

CALCULATION OF THE LOG-RANK TEST FOR 16 PATIENTS

t_i	m_i	r_i	$m_{(i)} / r_{(i)}$	$e(t_{(i)})$	w_i
321	1	16	0.625	0.625	0.375
552	1	15	0.067	0.692	0.308
691	-	-	-	-	-
769	1	13	0.077	0.769	0.231
877	-	-	-	-	-
890	1	11	0.909	1.678	- 0.678
939	-	-	-	-	-
1487	1	9	0.111	1.789	- 0.789
1746	1	8	0.125	1.914	- 0.914
2033	-	-	-	-	-
2386	1	6	0.167	2.081	- 1.081
2400	1	5	0.200	2.281	- 1.281
2576	-	-	-	-	-
2689	1	3	0.333	2.614	- 0.614
2812	1	-	-	-	-
3069	-	-	-	-	-

APPENDIX F

PRIMARY BILIARY CIRRHOSIS FOR 15 PATIENTS

Patient	ID	FU Days	Status	Censor	Drug	Age	Sex	Asictes	Platelets
1	10	51	2	0	2	25772	1	1	302
2	23	264	2	0	2	20442	1	1	214
3	97	611	2	0	2	26259	0	0	344
4	149	762	2	0	1	22574	0	0	140
5	295	877	1	1	1	12912	0	0	306
6	3	1012	2	0	1	25594	0	0	151
7	14	1217	2	0	2	20535	0	1	156
8	148	1427	2	0	2	11273	1	0	330
9	8	2466	2	0	2	19379	1	0	373
10	190	2504	0	1	1	19916	1	0	327
11	90	2689	2	0	1	12227	0	0	337
12	21	3445	0	1	2	23445	0	0	336
13	16	3672	0	1	2	14772	1	0	198
14	24	4079	2	0	1	16261	0	0	70
15	66	4191	2	0	1	16967	0	0	123

APPENDIX G

PROPORTIONAL HAZARDS MODEL
APPLIED TO PRIMARY
BILIARY CIRRHOSIS

Patient	FU Days	Sex	Censored	Contribution to Likelihood
1	51	1		$\frac{e^{1\beta}}{6e^{\beta}+9}$
2	264	1		$\frac{e^{\beta}}{5e^{\beta}+9}$
3	611	0		$\frac{1}{4e^{\beta}+9}$
4	762	0	0(877)	$\frac{1}{4e^{\beta}+8}$
6	1012	0		$\frac{1}{4e^{\beta}+6}$
7	1217	0		$\frac{1}{4e^{\beta}+5}$
8	1427	1		$\frac{e^{\beta}}{4e^{\beta}+4}$
9	2466	1	1(2504)	$\frac{e^{\beta}}{3e^{\beta}+4}$
11	2689	0	0(3445), 1(3672)	$\frac{1}{1e^{\beta}+4}$
14	4079	0		$\frac{1}{0e^{\beta}+2}$
15	4191	0		$\frac{1}{0e^{\beta}+1}$

APPENDIX H

NONPARAMETRIC SURVIVAL PLOT
FOR FOLLOW UP DAYS BY GENDER
FOR 312 PATIENTS

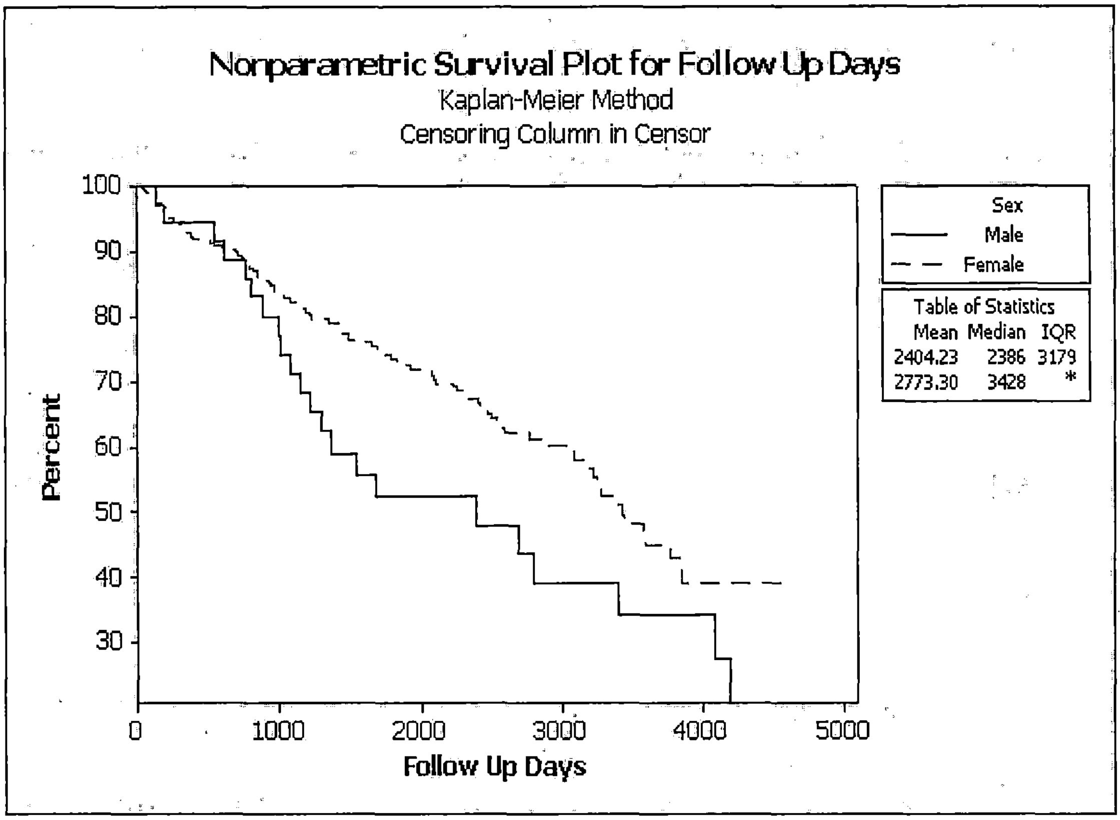

Table of Statistics

Mean	Median	IQR
2404.23	2386	3179
2773.30	3428	*

APPENDIX I

MINITAB CALCULATIONS: PRIMARY BILIARY CIRRHOSIS BY GENDER FOR 312 PATIENTS

```
                    Kaplan-Meier Estimates

       Number
           at  Number     Survival   Standard    95.0%   Normal CI
  Time   Risk   Failed Probability  Error       Lower    Upper
    41    275        1    0.996364  0.0036297  0.989249  1.00000
    51    274        1    0.992727  0.0051239  0.982685  1.00000
    71    273        1    0.989091  0.0062639  0.976814  1.00000
    77    272        1    0.985455  0.0072196  0.971304  0.99960
   110    271        1    0.981818  0.0080569  0.966027  0.99761
   130    270        1    0.978182  0.0088095  0.960915  0.99545
   131    269        1    0.974545  0.0094977  0.955930  0.99316
   179    268        1    0.970909  0.0101345  0.951046  0.99077
   186    267        1    0.967273  0.0107291  0.946244  0.98830
   198    266        1    0.963636  0.0112882  0.941512  0.98576
   207    265        1    0.960000  0.0118168  0.936840  0.98316
   216    264        1    0.956364  0.0123188  0.932219  0.98051
   223    263        1    0.952727  0.0127974  0.927645  0.97781
   264    262        2    0.945455  0.0136941  0.918615  0.97229
   304    260        1    0.941818  0.0141160  0.914151  0.96948
   321    259        1    0.938182  0.0145223  0.909719  0.96664
   326    258        1    0.934545  0.0149143  0.905314  0.96378
   334    257        1    0.930909  0.0152932  0.900935  0.96088
   348    256        1    0.927273  0.0156598  0.896580  0.95797
   388    255        1    0.923636  0.0160150  0.892248  0.95503
   400    254        1    0.920000  0.0163596  0.887936  0.95206
   460    253        1    0.916364  0.0166942  0.883644  0.94908
   515    252        1    0.912727  0.0170194  0.879370  0.94608
   549    251        1    0.909091  0.0173357  0.875114  0.94307
   597    250        1    0.905455  0.0176436  0.870874  0.94004
   673    249        1    0.901818  0.0179436  0.866649  0.93699
   694    248        1    0.898182  0.0182360  0.862440  0.93392
   708    247        1    0.894545  0.0185211  0.858245  0.93085
   733    245        1    0.890894  0.0188020  0.854043  0.92775
   750    243        1    0.887228  0.0190787  0.849834  0.92462
   769    242        1    0.883562  0.0193489  0.845639  0.92148
   786    241        1    0.879896  0.0196129  0.841455  0.91834
   790    240        1    0.876229  0.0198709  0.837283  0.91518
   797    239        1    0.872563  0.0201231  0.833123  0.91200
   824    238        1    0.868897  0.0203698  0.828973  0.90882
   850    235        1    0.865199  0.0206160  0.824793  0.90561
   853    234        1    0.861502  0.0208568  0.820623  0.90238
   859    233        1    0.857805  0.0210925  0.816464  0.89915
   904    231        1    0.854091  0.0213255  0.812294  0.89589
   930    230        1    0.850378  0.0215537  0.808133  0.89262
   943    228        1    0.846648  0.0217795  0.803961  0.88933
   971    227        1    0.842918  0.0220006  0.799798  0.88604
   974    226        1    0.839189  0.0222171  0.795644  0.88273
   980    225        1    0.835459  0.0224293  0.791498  0.87942
  1000    223        1    0.831712  0.0226394  0.787340  0.87608
  1037    221        1    0.827949  0.0228476  0.783168  0.87273
  1080    219        1    0.824168  0.0230540  0.778983  0.86935
  1083    218        1    0.820388  0.0232561  0.774807  0.86597
  1165    214        1    0.816554  0.0234613  0.770571  0.86254
  1170    213        1    0.812721  0.0236623  0.766343  0.85910
  1191    212        2    0.805053  0.0240521  0.757912  0.85219
  1212    210        1    0.801220  0.0242412  0.753708  0.84873
  1235    205        1    0.797311  0.0244360  0.749418  0.84521
  1350    195        1    0.793223  0.0246504  0.744909  0.84154
  1356    194        1    0.789134  0.0248601  0.740409  0.83786
```

1413	188	1	0.784936	0.0250797	0.735781	0.83409
1427	185	1	0.780693	0.0253005	0.731105	0.83028
1434	183	1	0.776427	0.0255194	0.726410	0.82644
1444	180	1	0.772114	0.0257396	0.721665	0.82256
1487	175	1	0.767702	0.0259679	0.716806	0.81860
1492	174	1	0.763290	0.0261908	0.711957	0.81462
1576	167	1	0.758719	0.0264298	0.706918	0.81052
1657	162	1	0.754036	0.0266783	0.701747	0.80632
1690	159	2	0.744551	0.0271727	0.691293	0.79781
1741	154	1	0.739716	0.0274230	0.685968	0.79346
1786	148	1	0.734718	0.0276894	0.680448	0.78899
1827	145	1	0.729651	0.0279582	0.674854	0.78445
1847	142	1	0.724513	0.0282296	0.669184	0.77984
1925	137	1	0.719224	0.0285146	0.663337	0.77511
2055	128	1	0.713605	0.0288401	0.657080	0.77013
2081	127	1	0.707986	0.0291553	0.650843	0.76513
2090	126	1	0.702367	0.0294604	0.644626	0.76011
2105	125	1	0.696749	0.0297557	0.638428	0.75507
2224	114	1	0.690637	0.0301158	0.631611	0.74966
2256	111	1	0.684415	0.0304805	0.624674	0.74416
2288	109	1	0.678136	0.0308408	0.617689	0.73858
2297	107	1	0.671798	0.0311970	0.610653	0.73294
2400	98	1	0.664943	0.0316228	0.602963	0.72692
2419	97	1	0.658088	0.0320312	0.595308	0.72087
2466	92	1	0.650935	0.0324719	0.587291	0.71458
2503	89	1	0.643621	0.0329204	0.579098	0.70814
2540	85	1	0.636049	0.0333926	0.570601	0.70150
2583	77	1	0.627788	0.0339653	0.561218	0.69436
2598	76	1	0.619528	0.0345082	0.551893	0.68716
2769	66	1	0.610141	0.0352389	0.541074	0.67921
2847	62	1	0.600300	0.0360185	0.529705	0.67090
3086	52	1	0.588756	0.0371298	0.515983	0.66153
3090	51	1	0.577212	0.0381542	0.502431	0.65199
3170	45	1	0.564385	0.0394035	0.487155	0.64161
3222	44	1	0.551558	0.0405420	0.472097	0.63102
3244	42	1	0.538426	0.0416493	0.456794	0.62006
3282	40	1	0.524965	0.0427279	0.441220	0.60871
3358	37	1	0.510777	0.0438656	0.424802	0.59675
3428	34	1	0.495754	0.0450746	0.407409	0.58410
3445	33	1	0.480731	0.0461443	0.390290	0.57117
3574	31	1	0.465224	0.0471896	0.372734	0.55771
3584	28	1	0.448608	0.0483409	0.353862	0.54335
3762	24	1	0.429916	0.0498096	0.332291	0.52754
3839	22	1	0.410375	0.0512357	0.309955	0.51079
3853	20	1	0.389856	0.0526224	0.286718	0.49299

APPENDIX J

EPI-INFO CALCULATIONS:
PRIMARY BILIARY CIRRHOSIS BY GENDER
FOR 312 PATIENTS

Term	Hazard Ratio	95%	C.I.	Coefficient	S. E.	Z-Statistic	P-Value
Sex(Yes/No)	0.9228	0.4903	1.7367	-0.0804	0.3227	-0.2492	0.8032
Age	1.0001	1.0001	1.0002	0.0001	0.0	4.2321	0.0
Albumin	0.3024	0.1798	0.5086	-1.196	0.2653	-4.5083	0.0
Alkaline	1.0051	1.002	1.0081	0.0051	0.0016	3.2592	0.0011
Bili	1.4099	1.3181	1.508	0.3435	0.0343	10.0035	0.0
Platelets	0.9982	0.9963	1.0001	-0.0018	0.001	-1.8612	0.0627
Spiders	1.4712	0.9356	2.3135	0.3861	0.231	1.6717	0.0946

Convergence:	Diverged
Iterations:	2
-2 * Log-Likelihood:	1357.5497

Test	Statistic	D.F.	P-Value
Score	261.9749	7	0.0
Likelihood Ratio	-104.5501	7	1.0

APPENDIX K

NONPARAMETRIC SURVIVAL PLOT FOR FOLLOW UP DAYS BY DRUG FOR 312 PATIENTS

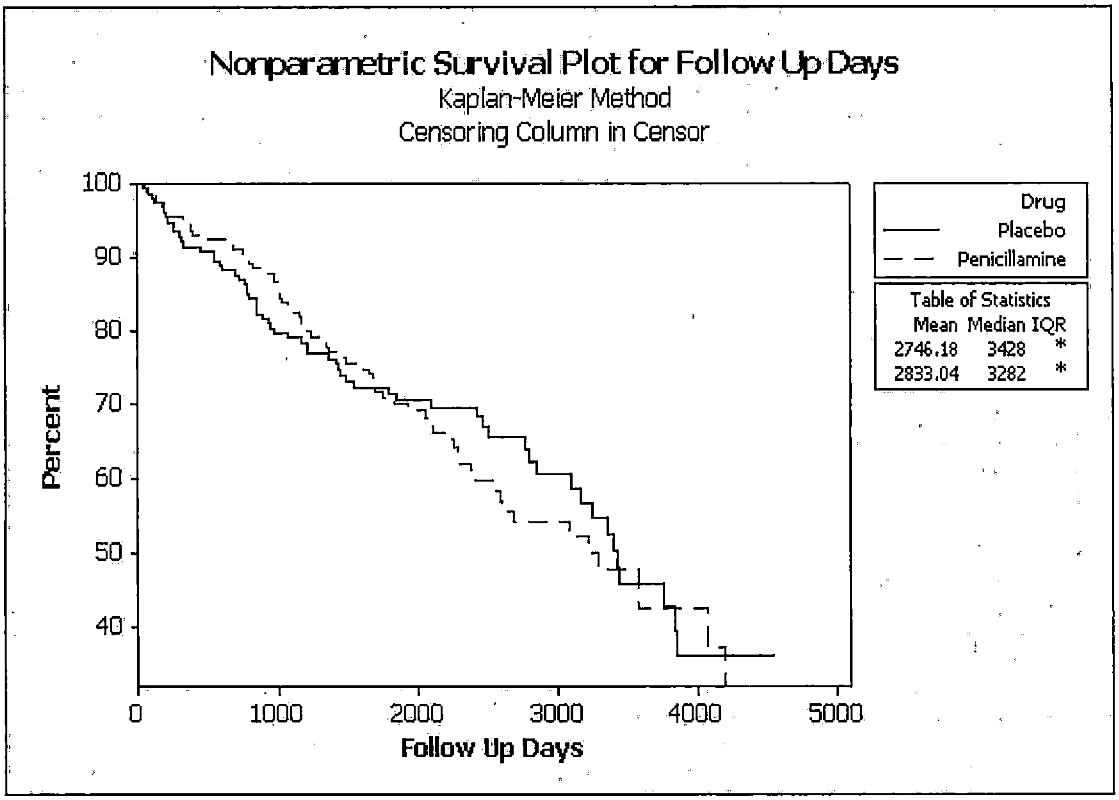

Table of Statistics

Mean	Median	IQR
2746.18	3428	*
2833.04	3282	*

APPENDIX L

EPI-INFO CALCULATIONS:
PRIMARY BILIARY CIRRHOSIS BY DRUG
FOR 312 PATIENTS

Term	Hazard Ratio	95%	C.I.	Coefficient	S. E.	Z-Statistic	P-Value
Drug(Yes/No)	0.9775	0.683	1.3991	-0.0227	0.1829	-0.1242	0.9011
Age	1.0001	1.0001	1.0002	0.0001	0.0	4.2723	0.0
Albumin	0.3059	0.1831	0.5111	-1.1844	0.2619	-4.5231	0.0
Alkaline	1.0052	1.0024	1.0081	0.0052	0.0014	3.6178	0.0003
Bili	1.4092	1.3176	1.5073	0.343	0.0343	9.9969	0.0
Platelets	0.9981	0.9963	1.0	-0.0019	0.001	-1.9534	0.0508
Spiders	1.4579	0.9318	2.2813	0.377	0.2284	1.6505	0.0988

Convergence: Diverged
Iterations: 2
-2 * Log-Likelihood: 1358.8139

Test	Statistic	D.F.	P-Value
Score	261.9283	7	0.0
Likelihood Ratio	-105.8143	7	1.0

REFERENCES

[1] Bernardo, J.M., Journal of the Royal Statistical Society, Series B, 4. Reference posterior distributions for Bayesian inference (with discussion).

[2] Conover, W.J., Practical Nonparametric Statistics, John Wiley and Sons, New York c.1999.

[3] Fitzmaurice, Garrett M., Laird, Nan M., and Ware, James H., Applied Longitudinal Analysis, John Wiley and Sons, New Jersey c.2004.

[4] Hogg, Robert V., and Tanis, Elliot A., Probability and Statistical Inference, Prentice Hall, New Jersey c.2001

[5] Hollander, Myles, and Wolfe, Douglas A, Nonparametric Statistical Methods, John Wiley and Sons, New York c.1999.

[6] Hosmer, David W. Jr., and Lemeshow, Stanley, Applied Survival Analysis, John Wiley and Sons, New Jersey c.1999.

[7] Jeffreys, H., Theory of Probability, Third Edition, Oxford University Press, London, UK.

[8] Kalbfleisch, John D., and Prentice, Ross L, The statistical Analysis of Failure Time Data, John Wiley and Sons, New York c.1980.

[9] Lee, Elisa T., and Wang, John Wenyu, Statistical Methods for Survival Data Analysis, Third Edition, John Wiley and Sons, New Jersey c.2003.

[10] Meier P., Karison T., Chappel R., Xie, H. The Price of Kaplan-Meier Journal of American Statistical Assn. (2004) 99,467.

[11] Nelson, Wayne B., APPLIED LIFE DATA ANALYSIS, John Wiley and Sons, New Jersey c.2004.

[12] Wackerly, Dennis D., Mendenhall III, William, and Scheaffer, Richard L., Mathematical Statistics with Applications, Wadsworth Publishing Company, California c.1996

[13] Worman, Howard J., M.D, What is Primary Biliary Cirrhosis (PBC)? C. 1999
http://cpmcnet.columbia.edu/ dept/gi/PBC.html

[14] EPI-INFO (TM) Computer Program, 2005
http://www.cdcc.gov/epo/dphsi/contact.htm