AN APPLICATION OF A COX MODEL FOR LIFETIMES OF HIV PATIENTS AND A POWER ANALYSIS

AN APPLICATION OF A COX MODEL FOR LIFETIMES OF HIV PATIENTS AND A POWER ANALYSIS

By

Sanjel Sabina, M.Sc.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Master of Science

McMaster University

TO MY PARENTS

MASTER OF SCIENCE (2007) McMaster University

(Statistics) Hamilton, Ontario

TITLE: An Application of a Cox Model for Lifetimes of HIV Patients

AUTHOR: Sanjel Sabina, B.Sc (Western), BBA

(Tribhuvan University, Kathmandu, Nepal)

SUPERVISOR: Dr. N. Balakrishnan

NUMBER OF PAGES: viii, 46

Acknowledgements

I express my sincere gratitude and appreciation to my supervisor, Prof. N. Balakrishnan, for his inspiring guidance, support and encouragement throughout the preparation of this project.

It is a matter of great pleasure to thank the professors and staff in the Department of Mathematics and Statistics, McMaster University, who also made my Mac experience enjoyable and educational. In particular, I am grateful to my fellow graduate students and friends who lent a helping hand whenever I needed.

Most importantly, I would like to thank my husband Dr. D. Sanjel and my daughters for their constant love and co-operation. I couldn't have done it without them.

I am extremely thankful to my parents for the love and affection they showered on me.

Contents

List of Figures

Abstract

In this project, an application of the Cox proportional hazard model is being considered. Cox proportional hazard model is fitted to estimate the effect of the covariates, age and drugs, on the survival of the HIV positive patients. These estimates also agree with the estimates obtained by using the numerical method. Likelihood ratio, Wald test and Score test are applied to test the significance of these estimates. Power for these test are performed by Monte Carlo simulation method. Simulated powers for sample size n = 10, 20 and 30, $\beta = 0.1, 0.2, 0.4$ and 1%, 5% and 10% are tabulated.

Chapter 1

Introduction

Life tables are one of the oldest statistical techniques and are extensively used by medical statisticians and actuaries. Kaplan and Meier gave a comprehensive review of earlier work and many new results. Chiang, in a series of papers, has in particular explored the connection with birth-death processes. Cox (1972) was largely concerned with the extension of the results of Kaplan and Meier to the comparison of life tables and more generally to the incorporation of regression-like arguments into life-table analysis. The procedures are closely related to procedures for combining contingency tables by Mantel and Haenszel (1959) and Mantel for the application of life tables. There is also strong connection with a paper by R. and J. Peto.

A common problem in the analysis of survival data in medical statistics is that of obtaining treatment comparisons while adjusting for and evaluating the effects of many uncontrolled independent variables. This introduces the use of non-linear regression models which assume that independent variables affect the hazard function in a multiplicative way. The hazard function is the probability that an individual

will experience an event (for example, death) within a small time interval, given that the individual has survived up to the beginning of that interval. It can therefore be interpreted as the risk of dying at time t. If the hazard function does not depend on time and it's value is completely determined by the covariate and the unknown parameters, it means that the risk of failure is the same no matter how long the subject has been followed. The hazard function and systematic component in the regression model are inversely related. The hazard function, denoted by $h(t)$, can be estimated as follows:

$$h(t) = \frac{\text{number of individuals experiencing an event in interval beginning at t}}{\text{(number of individuals surviving at time t) x (interval width)}}$$

There are different parametric models which can be useful for analysing survival data. These models are fitted based on hazard functions. For example, exponential distribution can be used when the hazard rate is constant within a particular group of individuals. The Weibull distribution can be used when the hazard rate is not constant but smoothly increasing (may stay the same but never decreases) or decreasing with time, whereas the log-normal and the log-logistic distributions can be used when a hazard rate initially increases and then declines after reaching a peak. A graphical technique is usually applied to assess the assumptions on hazard function.

For various reasons, data resulting from these types of investigations are frequently incomplete, in the sense that observations on survival time are not known exactly for all the individuals. This may be due to the limitations on the length of the study, or to death from a cause other than that under investigation, and so on. Observations of this type are said to be censored data. It is commonly assumed that

death and censoring are determined by independent mechanisms and that the only information on the survival time t of an individual censored at t^+ is that $t > t^+$.

The proportional hazard model, introduced by Cox (1972), is the most general of the regression model because it is not based on any assumptions concerning the nature or shape of the underlying survival distribution. It is a well-recognized statistical technique for exploring the relationship between the survival of a patient and several explanatory variables. A Cox model provides an estimate of the treatment effect on survival after adjustment for other explanatory variables. It allows us to estimate the hazard (or risk) of death, or other event of interest, for individuals, given their prognostic variables. Even if the treatment groups are similar with respect to the variables known to affect survival, use of the Cox model with these prognostic variables may produce a more precise estimate of the treatment effect, for example, by resulting in a narrow the confidence interval. Cox's regression model may be considered to be a semi-parametric method.

Chapter 2

Assumptions and Definition of the model

2.1 Assumptions

While no assumptions are made about the shape of the underlying hazard function, the proportional hazard model does imply two assumptions. First, it specifies a multiplicative relationship between the underlying hazard function and the log-linear function of the covariates. This assumption is the so-called the *proportionality assumption.* In practical terms, it is assumed that, given two observations with different values for the independent variables, the ratio of the hazard functions for those two observations does not depend on time.

The second assumption, of course, is that there is a log-linear relationship between the independent variables and the underlying hazard function.

2.2 Definition of the model

The model in which the hazard function $h(t;\mathbf{x})$ of a continuous random variable T, representing the survival time for an individual with independent variable vector $\mathbf{x} = (x_1, \ldots, x_p)'$, which could be any collection of covariates: continuous covariates, design variables for nominal scale covariates, product of covariates (interactions) and other higher order terms, is given by

$$h(t, \mathbf{x}) = h_0(t)e^{\mathbf{x}'\beta},$$

where $h_0(t)$ is some baseline hazard function which may never be defined and $\beta = (\beta_1, \ldots, \beta_p)'$ is a vector of regression coefficients expressing quantitatively the effect of each of the variable in $\mathbf{x}$. Independence of the hazard function (and hence survival) on the variable x_j is implied by $\beta_j = 0$. The cumulative hazard function H(t) is given by

$$\begin{aligned} H(t;x) &= \int_0^t h_0(u)e^{\mathbf{x}'\beta}du \\ &= H_0(t)e^{\mathbf{x}'\beta}, \\ S(t;x) &= \exp\{-H(t)\} \\ &= \exp\left\{-H_0(t)e^{\mathbf{x}'\beta}\right\}, \end{aligned}$$

where $H_0(t)\text{and}S_0(t)$ are baseline cumulative hazard function and baseline survival function, respectively. We emphasize that the function $h_0(t)$ and $S_0(t)$ will remain unspecified. Only part of the model that affects the covariates is parametrized. This is why Cox's model is a *semi-parametric* model.

Chapter 3

Estimation of parameters

3.1 Partial likelihood

Suppose that k individuals die at distinct times:

$$t_{(1)} < t_{(2)} \ldots < t_{(k)}.$$

At time $t_{(j)}$, an individual with covariate values $\mathbf{x_j}$ dies. From basic probability theory, the conditional probability that it is the specific individual j that dies out of the set R_j, given that one individual dies, is

$$\frac{h(t_{(j)}; \mathbf{x_j})}{\sum_{i \in R_j} h(t_{(j)}; \mathbf{x_i})} = \frac{e^{\mathbf{x_j'}\beta}}{\sum_{i \in R_j} e^{\mathbf{x_i'}\beta}} .$$

Since the hazard function gives the instantaneous probability of failure, the partial likelihood function expression proposed by Cox (1972), that depends only on the parameter of interest for all the failures, is

$$L(\beta) = \prod_{j=1}^{k} \left\{ \frac{e^{\mathbf{x}_j'\beta}}{\sum_{i \in R_j} e^{\mathbf{x}_i'\beta}} \right\} \quad .$$

The partial log likelihood function is then given by

$$l(\beta) = \sum_{j=1}^{k} \mathbf{x}_j'\beta - \sum_{j=1}^{k} \ln \sum_{i \in R_j} e^{\mathbf{x}_i'\beta}$$

We obtain the partial maximum likelihood estimator $\hat{\beta}$ of β by differentiating the right hand side of the partial log- likelihood function above with respect to β, setting the derivative equal to zero, and solving for the unknown parameter β. The derivative with respect to β is

$$\frac{\partial l}{\partial \beta_r} = \sum_{j=1}^{k} x_{jr} - \sum_{j=1}^{k} \left[\sum_{i \in R_j} x_{jr} e^{\mathbf{x}_i'\beta} / \sum_{i \in R_j} e^{\mathbf{x}_i'\beta} \right] \quad .$$

The variance-covariance matrix of this estimator β can be estimated by the inverse of the negative of the second derivative of the partial log-likelihood at the value of the estimator, called as the *observed information matrix,* and is given by,

$$\mathbf{I}(\beta) = -\frac{\partial^2 l}{\partial \beta_r \partial \beta_s}$$

The estimator of the variance-covariance matrix of β is then

$$\begin{aligned} \hat{\mathrm{Var}}\hat{\beta} &= \mathbf{I}(\beta)^{-1} \\ &= \sum_{j=1}^{k} \frac{\sum_{i \in R_j} e^{\mathbf{x}_i'\hat{\beta}} \sum_{i \in R_j} e^{\mathbf{x}_i'\hat{\beta}} x_{ir} x_{is} - \sum_{i \in R_j} e^{\mathbf{x}_i'\hat{\beta}} x_{ir} \sum_{i \in R_j} e^{\mathbf{x}_i'\hat{\beta}} x_{is}}{\left\{ \sum_{i \in R_j} e^{\mathbf{x}_i'\hat{\beta}} \right\}^2} \quad . \end{aligned}$$

3.2 Equal failure times

The Cox proportional hazard model assumes continuous hazard functions with no tied failure times. Since real data do often contain tied failure times, ties must be handled. There is more than one way to handle failure times that coincide. First case is when time is in principle a continuous variable. Then the $d_j > 1$ failures that occur simultaneously at time t_j cannot be exact coincidences: their failure times would have been slightly different from each other if only we had measured them with sufficient accuracy. Since these $d_j > 1$ failures are not after all simultaneous, they occur in a particular order, but we do not know which of the $d_j!$ possible orders is the correct one. The partial likelihood function therefore ought to include all of them. Infact, instead of doing this, an approximation due to Breslow is usually preferred in which the term in the partial likelihood,

$$\frac{e^{\mathbf{x}_j'\beta}}{\sum_{i\in R_j} e^{\mathbf{x}_i'\beta}}$$

that occurs for $d_j = 1$ is replaced when $d_j > 1$ by

$$\frac{e^{\mathbf{z}^*_j'\beta}}{\sum_{i\in R_j} e^{\mathbf{x}_i'\beta}} \quad ,$$

where z_j^* is the sum of the $\mathbf{x_j}$ for the d_j individuals that fail at time t_j. This approximation turns out to be accurate if the ratio d_j/n_j is small, where n_j is the number of individuals at risk (members of R_j) at time t_j. An alternative approximation has been given by Efron.

If d_j/n_j is not small, then we take another approach. We acknowledge that the time variable was in fact a discrete measurement and we use a method suitable for such data. Given that there are d_j failures at time t_j, the probability that the

Chapter 4

Tests of hypotheses

Two important steps that usually follow the fit of a regression model are the assessment of the significance of the coefficients and the formation of confidence intervals. For this purpose, it is common to use the following three different tests to assess the significance of the coefficients:

The partial lkelihood ratio test

The score test

The Wald test

4.1 The partial likelihood ratio test

The partial likelihood ratio test, denoted by G, is calculated as twice the difference between the partial log-likelihood of the model containing the covariate, and that if not containing the covariate. That is,

$$G = 2\left\{l(\hat{\beta}) - l(0)\right\}$$

G follows asymptotically a chi-square distribution with degrees of freedom determined by one for each coefficient. The significance for the test implies that at least one of the coefficients in the model explains the failure time.

4.2 The score tests

Computation of score tests for the multiple proportional hazards regression model requires matrix calculations. Specifically, we denote the vector of first partial derivatives as $\mathbf{u}(\beta)$, which is given by

$$\begin{aligned} Z^* &= \frac{\partial l/\partial \beta}{\sqrt{I(\beta)}}\Big|_{\beta} \\ &= 0 \end{aligned}$$

Under the hypothesis that all coefficients are equal to zero, and under the mathematical conditions needed for the partial likelihood tests, the vector of scores $\mathbf{u(0)} = \mathbf{u}(\beta)|_{\beta=\mathbf{0}}$ 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, $\mathbf{I(0)} = \mathbf{I}(\beta)|_{\beta=\mathbf{0}}$

The score statistic is then given by

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

which is distributed asymptotically as chi-square with p degrees of freedom.

4.3 The Wald test

The Wald test is obtained from equivalent theory which states that, under the null hypothesis, the estimator of the coefficient, $\hat{\beta}$, will be asymptotically nor-

mally distributed with the mean vector equal to zero and a covariance matrix that is estimated.

The Wald test statistic is then given by

$$\hat{\beta}' \mathbf{I}(\hat{\beta}) \hat{\beta},$$

which is also distributed asymptotically as chi-square with p degrees of freedom.

4.4 Asymptotic Confidence Interval

The adequacy of the asymptotic χ^2 or normal approximations used for the score test statistic, likelihood ratio test and Wald statistic can vary substantially according to the problem and the amount of information about the parameters. It is difficult to make general statements, but the distributions of likelihood ratio statistics often tend to their limiting distributions more quickly than Wald statistics or any other test statistics. Consequently, likelihood ratio methods are often preferred, especially for small to moderate sample sizes. However, confidence intervals based on these methods may require substantial computation and are not directly available from most software packages, and so the Wald test statistic is often used for this purpose and is described as follows.

For the Wald statistic, the p-value (significance level) based on the observed value $w(\theta_0)$ is approximately $Pr(\chi^2_{(k)} \geq w(\theta_0))$. A confidence region for θ with approximate confidence coefficient α consists of vectors θ_0 such that

$$w(\theta_0) \leq \chi^2_{(k),\alpha} \quad .$$

Confidence intervals for a single parametric function $\beta = g(\theta)$ is often required. The simplest approach for this is to use the normal approximation $\hat{\beta} \sim N(\beta, \hat{V}_{\beta}^{1/2})$, where $\hat{V}_{\beta}$ is given by

$$\hat{V}_{\beta} = G(\hat{\theta})\, \hat{V}_{\theta}\, G(\hat{\theta})' \quad .$$

This yields the approximate standard normal pivotal quantity

$$Z = \frac{\hat{\beta} - \beta}{\hat{V}_{\beta}^{1/2}}$$

and two-sided approximate $100(1-\alpha)\%$ confidence interval $\hat{\beta} \pm z_{\alpha/2} \hat{V}_{\beta}^{1/2}$, where z_q is the qth quantile of N(0,1). One-sided approximate $100(1-\alpha)\%$ confidence intervals are given by $\beta \leq \hat{\beta} - \hat{V}_{\beta}^{1/2}$ and $\beta \geq \hat{\beta} + \hat{V}_{\beta}^{1/2}$, respectively.

Chapter 5

Power Analysis

The power function of a test of a statistical hypothesis H_0 against an alternative hypothesis H_1 is that function, defined for all distributions under considerations, which yields the probability that the test statistic falls in the critical region C of the test; that is, it is a function that yields the probability of rejecting the null hypothesis under consideration. The value of the power function at a parameter point is called the power of the test at that point. As power increases, the chances of a Type II error decrease, and vice versa. The probability of Type II error is denoted by β_1. Therefore, power is equal to $1 - \beta_1$. Statistical power depends on the statistical significance criterion used in the test, the size of the difference or the strength of the similarity (that is, the effect size) in the population and the sensitivity of the data.

If $\theta \in \mathbf{\Theta_0}$, then $\alpha(\theta)$ is the probability of Type I error; if $\theta \in \mathbf{\Theta_1}$, then $\alpha(\theta)$ equals 1- P(Type II error). Thus, minimization of probability of Type II error is equivalent to the maximization of $\alpha(\theta)$ for $\theta \in \mathbf{\Theta_1}$ subject to the significance level requirements, namely, $\alpha(\theta) \leq \alpha$ for $\theta \in \mathbf{\Theta_0}$.

Table 5.1: Simulated power values for n = 10 for Wald test

$\beta \diagdown \alpha$	0.1	0.05	0.01
0.1	0.1167	0.0625	0.0265
0.2	0.1767	0.0987	0.0519
0.4	0.2525	0.1294	0.0782

The power comparison is done here based on for n = 10, 20 and 30 and for coefficients $\beta = 0.1, 0.2, 0.4$ and under 10%, 5% and 1% levels of significance. Tables 5.1-5.3 present the simulated power values for n = 10, 20 and 30, for Wald test.

In order to explain the power comparison procedure, we consider a particular case. For example, when n = 10, to test the hypotheses

$H_0 : \beta = 0$ Vs $H_1 : \beta \neq 0$

We generated Monte Carlo random estimates other than 0, say, for example, $\beta = 0.4$. For this we simulated the variables needed which follows an exponential distribution. From these we estimate the coefficients by fitting cox proportional hazard model and calculated the p-values for the Wald test at levels of significance 10%, 5% and 1%. This process is repeated for 400 times. Looking at these we count how many times out of 400, H_0 has been rejected. We see that 101 times it has been rejected. Hence the power of this test is 0.2525. So, when we test whether the sample is from $\beta = 0$, when it is actually from $\beta = 0.4$, we found the power to be 0.2525. For the same null hypothesis as above when the sample size is 20, the power increases to 0.7923; when the sample size is 30, the power increases to 0.9942. So as

Table 5.2: Simulated power values for n = 20 for Wald test

$\beta \diagdown \alpha$	0.1	0.05	0.01
0.1	0.2628	0.1654	0.0573
0.2	0.5467	0.2667	0.1067
0.4	0.7923	0.6533	0.36

Table 5.3: Simulated power values for n = 30 for Wald test

$\beta \diagdown \alpha$	0.1	0.05	0.01
0.1	0.4133	0.2974	0.0867
0.2	0.8124	0.6582	0.5067
0.4	0.9942	0.9857	0.9285

the sample size increases, the power increases which is to be expected. From these tables, we can conclude that as β moves away from the zero, the power increases. Notice also that the power value is substantially larger for 10% level of significance in comparision to 5% and 1%, respectively.

Tables 5.4 - 5.6 present the simulated power values for n = 10, 20 and 30 for likelihood test, while Tables 5.7 - 5.9 present the simulated power values for n = 10, 20 and 30 for Score test. We observe that the power performance of these three tests are very nearly the same. However the likelihood ratio shows the better performance.

Table 5.4: Simulated power values for n = 10 for likelihood test

$\beta \diagdown \alpha$	0.1	0.05	0.01
0.1	0.2248	0.0833	0.0565
0.2	0.2667	0.1682	0.0733
0.4	0.3781	0.1867	0.0925

Table 5.5: Simulated power values for n = 20 for likelihood test

$\beta \diagdown \alpha$	0.1	0.05	0.01
0.1	0.2933	0.2167	0.0879
0.2	0.5733	0.3067	0.1333
0.4	0.8123	0.7467	0.5033

Table 5.6: Simulated power values for n = 30 for likelihood test

$\beta \diagdown \alpha$	0.1	0.05	0.01
0.1	0.4667	0.3281	0.1453
0.2	0.8425	0.7033	0.5729
0.4	1	0.9957	0.9895

Table 5.7: Simulated power values for n = 10 for score test

$\beta \backslash \alpha$	0.1	0.05	0.01
0.1	0.1667	0.0751	0.0338
0.2	0.2267	0.1333	0.0614
0.4	0.3562	0.2122	0.1673

Table 5.8: Simulated Power values for n = 20 for score test

$\beta \backslash \alpha$	0.1	0.05	0.01
0.1	0.2725	0.1842	0.0773
0.2	0.5682	0.3200	0.1467
0.4	0.8025	0.7386	0.4933

Type I Errors values are simulated for different test. We observe that the wald test is more conservative than score and likelihood test. If we look at the values of likelihood ratio compared to wald and score the error is high. Hence the claim we made that the likelihood ratio test is better is not true as the type I error is affecting the power.

Table 5.9: Simulated Power values for n = 30 for score test

$\beta \backslash \alpha$	0.1	0.05	0.01
0.1	0.4582	0.3067	0.1228
0.2	0.8124	0.6882	0.5533
0.4	1	0.9867	0.9547

Table 5.10: Simulated Type I Error values for Wald test

$n \backslash \alpha$	0.1	0.05	0.01
10	0.0631	0.0396	0.0286
20	0.0825	0.0448	0.0184
30	0.0893	0.0489	0.0154

Table 5.11: Simulated Type I Error values for Score test

$n \backslash \alpha$	0.1	0.05	0.01
10	0.1665	0.0842	0.0439
20	0.1467	0.0833	0.0162
30	0.1072	0.0527	0.0124

Table 5.12: Simulated Type I Error values for Likelihood test

$n \backslash \alpha$	0.1	0.05	0.01
10	0.1682	0.0892	0.0743
20	0.1667	0.1167	0.0316
30	0.1593	0.0618	0.0245

Chapter 6

Data Description and Analysis

6.1 Description of Data

These data were extracted from the website

"http://people.umass.edu/statdata/statdata/data/hmohiv.xls ".

Only the part of the data is shown here.

ID	time	age	drug	censor	entdate	enddate
1	5	46	0	1	5/15/1990	10/14/1990
2	6	35	1	0	9/19/1989	3/20/1990
3	8	30	1	1	4/21/1991	12/20/1991
4	3	30	1	1	1/3/1991	4/4/1991
5	22	36	0	1	9/18/1989	7/19/1991
6	1	32	1	0	3/18/1991	4/17/1991
7	7	36	1	1	11/11/1989	6/11/1990

A large health maintenance organization (HMO) wishes to evaluate the survival time of its HIV+ members using a follow up study. Subjects were enrolled in the study from January 1, 1989 to December 31, 1991. The study ended on December 31, 1995. After a confirmed diagnosis of HIV, members were followed until death due to AIDS to AIDS-related complications, until the end of the study, or until the subject was lost to follow up. We assume that there were no deaths due to other causes (e.g, auto accident). The primary outcome variable of interest is the survival time after a confirmed diagnosis of HIV. Since subjects entered the study at different times over a 3 year period, the maximum possible follow up time is different for each study participant. Possible predictors of survival time were collected at enrollment into the study. Data for 100 subjects were observed as follows:

TIME: the follow-up time is the number of months between the entry date (ENT DATE) and the end date (END DATE),
AGE: the age of the subject at the start of the follow-up (in years),
DRUG: history of prior IV drug use (1 = YES, 0 = NO), and
CENSOR: vital status at the end of the study (1 = Death due to AIDS, 0 = Lost to follow-up or alive). Of many possible covariates, age and prior drug use were chosen for their clinical relevance.

The variable TIME actually records two different things: for those subjects who died, it is the outcome variable of interest, the actual survival time. However, for subjects who were alive at the end of the study, or for subjects who were lost, TIME indicates the length of follow-up (which is partial or incomplete observation

of survival time). These incomplete observations are referred to as being *censored*. For example, subject 1 died from AIDS 5 months after being seen in the HMO clinic (CENSOR = 1) while subject 2 was not known to have died from AIDS at the conclusion of the study and had been followed for 6 months (CENSOR = 0). It is possible for a subject to have entered the study much earlier, eventually becoming lost to follow-up as a result of moving, failing to return to the clinic or some other reason.

The main goal for a statistical analysis of these data is to fit a model that will yield biologically plausible and interpretable estimates of the effect of age and drug use on survival time for HIV+ patients.

6.2 Analysis of Data

In any applied setting, a statistical analysis should begin with a thoughtful description of the data. Sample mean, variance, median, etc. will not yield estimates of the desired parameters when the data include censored observations as in the case of our data. Hence, we must obtain an estimate of the cumulative distribution function. However, we are more interested in describing how long the study subjects live, than how quickly they die. So, the estimation focuses on the survival function. The survival function is estimated by Kaplan-Meier estimator. This estimator incorporates information from all of the observations available, both censored and uncensored, by considering survival to any point in a time series of steps defined by the observed survival and censored times.

Table 6.1: Estimate of the survival function

time	n.risk	n.event	survival	std.err
1	100	15	0.8500	0.0357
2	83	5	0.7988	0.0402
3	73	10	0.6894	0.0473
4	61	4	0.6442	0.0493
5	56	7	0.5636	0.0517
6	49	2	0.5406	0.0521
7	46	6	0.4701	0.0526
8	39	4	0.4219	0.0525
9	35	3	0.3857	0.0520
10	32	3	0.3496	0.0511
11	28	3	0.3121	0.0500
12	25	2	0.2872	0.0490
13	21	1	0.2735	0.0486
14	20	1	0.2598	0.0480
15	19	2	0.2325	0.0467
22	16	1	0.2179	0.0460
30	14	1	0.2024	0.0453
31	13	1	0.1868	0.0444
32	12	1	0.1712	0.0433
34	11	1	0.1557	0.0421
35	10	1	0.1401	0.0407

36	9	1	0.1245	0.0390
43	8	1	0.1090	0.0371
53	7	1	0.0934	0.0349
54	6	1	0.0778	0.0324
57	4	1	0.0584	0.0296
58	3	1	0.0389	0.0253

From the above table we see that at the probability of surviving for at least one month is 0.85. The conditional probability of surviving the second month after having survived the first month is calculated as (83-5)/83 = 0.94 and the overall probability of surviving the second month is 0.83*0.94 = 0.7988 and so forth. The standard errors related to its survival estimates are listed in the last column. It is to be noted that the standard error is increasing until n is 46 and starts to decrease. This is intuitive as the standard error of proportion gets larger when the probability is around 0.5. This is true as our survival estimate 0.4701.

The Figure 6.1 shows the Kaplan-Meier estimate of the survival function using all subjects in the study. The estimate demonstrates conventions for handling tied survival times as well as tied survival and censored times. The columns in Table 6.1 presents the time, the number at risk of dying, the number of deaths, the estimate of survivorship function, and its standard error.

Further,we wish to compare groups of subjects. The graph of Kaplan-Meier estimator of survival function for each of the groups in Figure 6.1 help us to visualize it. A variable related to the survival of the subjects is the history of IV drug use.

Figure 6.1: Kaplan-Meier estimate of the survival function

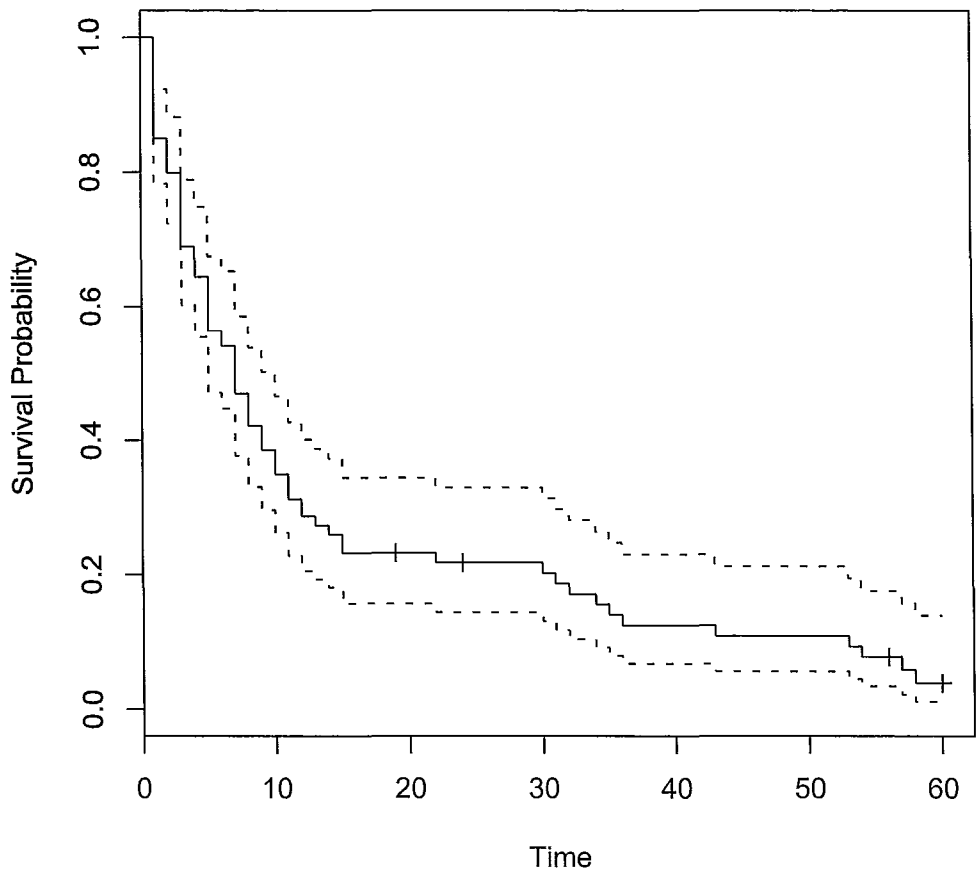
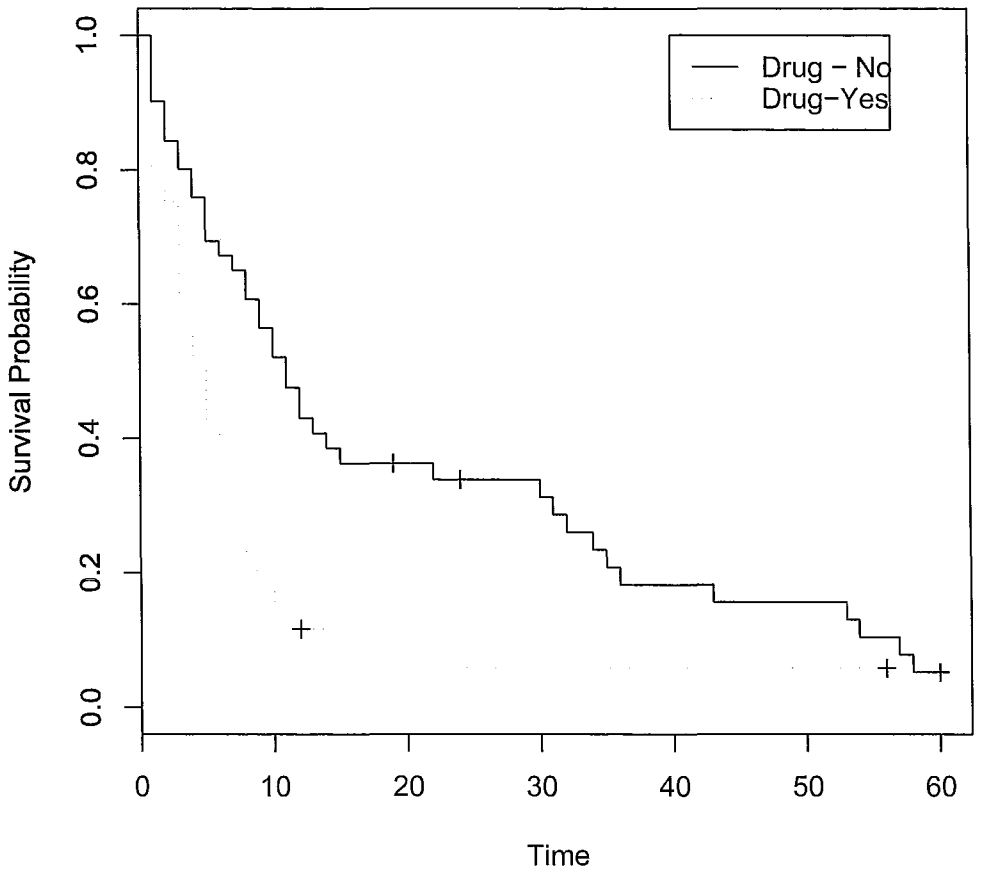


Figure 6.2: Estimates of the survival function of the subjects with and without IV drug use



The estimates of the survival function of the subjects in the non-IV drug use and IV drug use have been plotted in Figure 6.2. Both groups show a similar pattern of survival. There is a rapidly descending survival function with a long right tail. This is a result of a number of early deaths and a few subjects with survival near the maximum follow-up time. The figure also shows a separation of the function for the two groups. The estimated survivorship function for the non-IV drug users lies completely above that of IV drug users, which means the group defined by the upper curve lived lot longer than the group defined by the lower curve.

Now the obvious statistical question that arises is whether the observed difference here is significant. There are several ways of performing this test. The log-rank test has been applied here.

```
survdiff(Surv(time, censor)~drug, data=hmohiv)
survdiff(formula = Surv(time, censor) ~ drug, data = hmohiv,rho = 0)
        N Observed Expected (O-E)^2/E (O-E)^2/V
drug=0 51       42     54.9      3.02      11.9
drug=1 49       38     25.1      6.60      11.9
Chisq= 11.9  on 1 degrees of freedom, p= 0.000575
```

Since the p-value is very small(0.000575), it indicates that the test is highly significant and supports that those with a prior history of IV drug use tended to die sooner than those who did not have a history of IV drug use.

The semi-parametric Cox proportional hazards model is the most commonly used model in hazard regression. Exact partial likelihood method is usually used for this model under the assumptions of no tied data. However, our data have some

Table 6.2: Estimated coefficients for age

covariate	$\hat{\beta}$	$\hat{SE}$
age	0.0814	0.0174

tied observations. Slight modification is therefore needed. Breslow and Efron have provided expressions that are more easy to compute than the exact partial likelihood to deal with tied observations. Setting method as Breslow or as Efron in a built-in function coxph in R yields the appropriate analysis.

We have fitted the model with single covariate each, both covariates and both covariates with interaction by making use of built in functions in R. The coefficients that were obtained with this method have been verified by numerically maximizing the partial log-likelihood function as well. We have done this by applying the function nlm. This function carries out a minimization of the function using a Newton-type algorithm.

When this model is fitted with only one covariate (age) in our study, the value of the estimated coefficient and the corresponding estimated standard error is given in Table 6.2 which also agrees with the estimates obtained by using the numerical method.

```
mle<-nlm(ll1,c(0.06),hessian = T)
$minimum [1] 288.5180
$estimate
[1]0.08140366
$hessian
```

Table 6.3: Estimated coefficients for drug

covariate	$\hat{\beta}$	$\hat{SE}$
drug	0.78	0.242

```
         [,1]
[1,] 3289.773
varcov1<-solve(mle$hessian)
varcov1
            [,1]
[1,] 0.05868916
se<-sqrt(varcov1)
           [,1]
[1,]  0.01743480
```

When this model is fitted with only one covariate (drug) in our study, the value of the estimated coefficient and the corresponding estimated standard error is given in Table 6.3 which also agrees with the estimates obtained by using the numerical method.

```
mle<-nlm(ll1,c(0.06),hessian = T)
mle
$minimum
[1] 294.0964
```

Table 6.4: Estimated coefficients for age and drug

covariate	$\hat{\beta}$	$\hat{SE}$
age	0.0915	0.0185
drug	0.9414	0.2555

```
$estimate
[1] 0.77923
$hessian
         [,1]
[1,] 17.03892
varcov1<-solve(mle$hessian)
se<-sqrt(varcov1)
          [,1]
[1,] 0.2422584
```

When this model is fitted with two covariates (age and drug) in our study, the value of the estimated coefficients and the corresponding estimated standard errors of the estimated coefficients are given in Table 6.4 which also agrees with the estimates obtained by using the numerical method.

```
mle<-nlm(ll1,c(0.06),hessian = T)
mle
$minimum
[1] 281.7040
```

Table 6.5: Estimated coefficients for age,drug and interaction

covariate	$\hat{\beta}$	$\hat{SE}$
age	0.0942	0.0229
drug	1.1859	1.2565
age : *drug*	-0.0067	0.0337

```
$estimate
[1] 0.09153135 0.94138352
$hessian
           [,1]       [,2]
[1,] 3026.91177 -40.07171
[2,]  -40.07171  15.84739
varcov1<-solve(mle$hessian)
se<-sqrt(varcov1)
         [,1]       [,2]
[1,] 0.01848815 0.02939907
[2,] 0.02939907 0.25551392
```

When this model is fitted with two covariates (age and drug) and their interactions in our study, the value of the estimated coefficients and the corresponding estimated standard errors are given in Table 6.5 which also agrees with the estimates obtained by using the numerical method.

```
nlm(ll1,c(0.06,0.1,0.1),hessian = T)
```

```
$minimum
[1] 281.6844
$estimate
[1]  0.094234795  1.186308762 -0.006713934
$hessian
           [,1]       [,2]       [,3]
[1,] 3062.49362 -44.63811  -391.2862
[2,]  -44.63811  15.97741   564.4688
[3,] -391.28615 564.46880 21299.3127
varcov<-solve(mle$hessian)
varcov
        [,1]         [,2]          [,3]
[1,]  0.0005258306  0.01770068 -0.0004594388
[2,]  0.0177006797  1.57816125 -0.0414988393
[3,] -0.0004594388 -0.04149884 0.0011383010
se<-sqrt(varcov)
sqrt(varcov)
           [,1]       [,2]       [,3]
[1,] 0.02293100 0.1330439        NaN
[2,] 0.13304390 1.2562489        NaN
[3,]        NaN       NaN 0.03373872
```

Chapter 7

Conclusions

Based on this analysis, we can do several tests. The model may be used to determine whether the association of age with survival time is different for subjects with and without a history of IV drug use. The likelihood ratio test, Wald test and score test are performed in order to draw conclusions.

As we see from the full model, the p-value of covariate, age, is very small and hence may be significant whereas the p-value associated with drug and the interaction between age and drug terms are larger than 0.05 and hence it may not be significant. We need to explore this further. However, the p-values of likelihood ratio test, Wald test and the score test are very small. It provides evidence against the null hypothesis that all three coefficients are simultaneously equal to zero and indeed provides evidence towards the alternative hypothesis that at least one of the coefficients in the model is significantly associated with survival time.

Now, the reduced model is fitted. The model contains age and drug as covariates, but their interaction is excluded. The p-value for both age and drug coefficients

are small. Hence, we can conclude that both age and drug are significantly associated with survival time. We can even test this further by fitting the model with only one covariate, age and drug, separately.

We can also conclude in power analysis that the power increases as the sample size increases, the power increases as β moves away from the zero, the power value is substantially larger for 10% level of significance in comparison to 5% and 1%, respectively.

Appendix A

R Code and Results for Data Analysis

Kaplan-Meier survival function and it's graph.

```
hmohiv.surv <- survfit( Surv(time, censor)~ 1 ,data =
hmohiv)
summary(hmohiv.surv)
plot (hmohiv.surv, xlab="Time",ylab="Survival Probability" )
```

The graph of Kaplan-Meier estimator of survival function for treatment and control.

```
timestrata.surv <- survfit( Surv(time, censor)~ strata(drug),
hmohiv, conf.type="log-log")
plot(timestrata.surv, lty=c(1,3), xlab="Time", ylab="Survival
```

```
Probability")
legend(40, 1.0, c("Drug - No", "Drug - Yes") , lty=c(1,3) )
```

To estimate the coefficients fitting Cox model and comparing with estimates obtained by using numerical method.

```
age.coxph <- coxph( Surv(time,censor)~age, data =
hmohiv,method="breslow")
summary(age.coxph)
coxph(formula=Surv(time, censor) ~ age, data = hmohiv, method = "breslow")
  n= 100
      coef exp(coef) se(coef)    z     p
age 0.0814      1.08   0.0174 4.67 3e-06
    exp(coef) exp(-coef) lower .95 upper .95
age      1.08      0.922      1.05      1.12
Rsquare= 0.192(maxpossible= 0.997 )
Likelihood ratio test= 21.4 on 1 df, p=3.82e-06
Wald test = 21.8  on 1 df, p=3.03e-06
Score (logrank)test= 22  on 1 df, p=2.72e-06

drug.coxph <- coxph(Surv(time,censor)~drug, data
= hmohiv,method="breslow")
summary(drug.coxph)
coxph(formula =Surv(time,censor) ~ drug, data = hmohiv, method = "breslow")
  n= 100
```

```
     coef exp(coef) se(coef)    z      p
drug 0.78      2.18    0.242 3.22 0.0013
     exp(coef) exp(-coef) lower .95 upper .95
drug      2.18      0.459      1.36      3.50 Rsquare= 0.097
(maxpossible= 0.997 ) Likelihood ratio test= 10.2 on 1 df,
p=0.00141 Wald test            = 10.3 on 1 df,  p=0.00130 Score
(logrank) test = 10.7 on 1 df,  p=0.00105

main.coxph <- coxph( Surv(time,censor)~age+drug,data =
hmohiv,method="breslow")
summary(main.coxph)
Call: coxph(formula =
Surv(time,censor) ~ age + drug, data = hmohiv, method = "breslow")
  n= 100
       coef exp(coef) se(coef)    z       p
age  0.0915      1.10   0.0185 4.95 7.4e-07 drug 0.9414      2.56
0.2555 3.68 2.3e-04
     exp(coef) exp(-coef) lower .95 upper .95
age       1.10      0.913      1.06      1.14
drug      2.56      0.390      1.55      4.23
Rsquare= 0.295   (max possible= 0.997 )
Likelihood ratio test= 35   on 2 df, p=2.53e-08
Wald test=              32.5 on 2 df, p=8.76e-08
Score (logrank) test = 34.3 on 2 df, p=3.56e-08
```

```
inter.coxph <- coxph( Surv(time,censor)~age+drug+age:drug, data =
hmohiv,method="breslow")
summary(inter.coxph)
coxph(formula =Surv(time, censor) ~ age + drug + age:drug, data = hmohiv,
    method = "breslow")
  n= 100
             coef exp(coef) se(coef)      z       p
age        0.0942     1.099   0.0229  4.110 0.00004
drug       1.1859     3.274   1.2565  0.944 0.35000
age:drug -0.0067     0.993   0.0337 -0.199 0.84000
          exp(coef) exp(-coef) lower .95 upper .95
age           1.099      0.910     1.051      1.15
drug          3.274      0.305     0.279     38.42
age:drug      0.993      1.007     0.930      1.06
Rsquare= 0.295   (max possible= 0.997 )
Likelihood ratio test=35    on 3 df,   p=1.21e-07
Wald test            = 32.2  on 3 df,  p=4.83e-07
Score (logrank) test = 35.1  on 3 df,  p=1.13e-07

lll<- function(beta){
-(sum(p41%*%beta)sum(15*log(sum(exp(xi11%*%beta))),5*log(sum(exp
(xi21 %*%beta))),10*log(sum(exp(xi31%*%beta))),4*log(sum(exp(xi41
%*%beta))),7*log(sum(exp(xi51%*%beta))),2*log(sum(exp(xi61%*%beta
```

```
))),6*log(sum(exp(xi71%*%beta))),4*log(sum(exp(xi81%*%beta))),3*
log(sum(exp(xi91%*%beta))),3*log(sum(exp(xi101%*%beta))),3*log(
sum(exp(xi111%*%beta))),2*log(sum(exp(xi121%*%beta))),log(sum(exp
(xi131%*%beta))),log(sum(exp(xi141%*%beta))),2*log(sum(exp(xi151%
*%beta))),log(sum(exp(xi221%*%beta))),log(sum(exp(xi301%*%beta))),
log(sum(exp(xi311%*%beta))),log(sum(exp(xi321%*%beta))),log(sum(
exp(xi341%*%beta))),log(sum(exp(xi351%*%beta))),log(sum(exp(xi361
%*%beta))),log(sum(exp(xi431%*%beta))),log(sum(exp(xi531%*%
beta))),log(sum(exp(xi541%*%beta))),log(sum(exp(xi571%*%beta))),
log(sum(exp(xi581%*%beta)))))}

nlm(ll1,c(0.06),hessian = T)
```

Power Analysis

```
x<-(1:10)/2-3
myrates<-exp(0.1*x+1)
Sim2reg <- function(x1,inputrates){
y <- rexp(length(inputrates), rate = inputrates)
cen<-rexp(length(inputrates), rate = 0.1) + ycen <- pmin(y,cen)
di<-as.numeric(y<=cen)
temp1<-coxph(Surv(ycen, di)~x1)
return(temp1$coef) }
result <- rep(NA, 400)
for(i in 1:400) result[,i] <- Sim2reg(x1,myrates)
```

```
x<-(1:20)/2-3
myrates<-exp(0.1*x+1)
Sim2reg <- function(x1,inputrates){
y <- rexp(length(inputrates), rate = inputrates)
cen<-rexp(length(inputrates), rate = 0.1) + ycen <- pmin(y,cen)
di<-as.numeric(y<=cen)
temp1<-coxph(Surv(ycen, di)~x1)
return(temp1$coef) }
result <- rep(NA, 400)
for(i in 1:400) result[,i] <- Sim2reg(x1,myrates)

x<-(1:30)/2-3
myrates<-exp(0.1*x+1)
Sim2reg <- function(x1,inputrates){
y <- rexp(length(inputrates), rate = inputrates)
cen<-rexp(length(inputrates), rate = 0.1) + ycen <- pmin(y,cen)
di<-as.numeric(y<=cen)
temp1<-coxph(Surv(ycen, di)~x1)
return(temp1$coef) }
result <- rep(NA, 400)
for(i in 1:400) result[,i] <- Sim2reg(x1,myrates)
```

```
x<-(1:10)/2-3
myrates<-exp(0.2*x+1)
Sim2reg <- function(x1,inputrates){
y <- rexp(length(inputrates), rate = inputrates)
cen<-rexp(length(inputrates), rate = 0.2) + ycen <- pmin(y,cen)
di<-as.numeric(y<=cen)
temp1<-coxph(Surv(ycen, di)~x1)
return(temp1$coef) }
result <- rep(NA, 400)
for(i in 1:400) result[,i] <- Sim2reg(x1,myrates)
```

```
x<-(1:20)/2-3
myrates<-exp(0.2*x+1)
Sim2reg <- function(x1,inputrates){
y <- rexp(length(inputrates), rate = inputrates)
cen<-rexp(length(inputrates), rate = 0.2) + ycen <- pmin(y,cen)
di<-as.numeric(y<=cen)
temp1<-coxph(Surv(ycen, di)~x1)
return(temp1$coef) }
result <- rep(NA, 400)
for(i in 1:400) result[,i] <- Sim2reg(x1,myrates)
```

```
x<-(1:30)/2-3
myrates<-exp(0.2*x+1)
Sim2reg <- function(x1,inputrates){
y <- rexp(length(inputrates), rate = inputrates)
cen<-rexp(length(inputrates), rate = 0.2) + ycen <- pmin(y,cen)
di<-as.numeric(y<=cen)
temp1<-coxph(Surv(ycen, di)~x1)
return(temp1$coef) }
result <- rep(NA, 400)
for(i in 1:400) result[,i] <- Sim2reg(x1,myrates)
```

```
x<-(1:10)/2-3
myrates<-exp(0.4*x+1)
Sim2reg <- function(x1,inputrates){
y <- rexp(length(inputrates), rate = inputrates)
cen<-rexp(length(inputrates), rate = 0.4) + ycen <- pmin(y,cen)
di<-as.numeric(y<=cen)
temp1<-coxph(Surv(ycen, di)~x1)
return(temp1$coef) }
result <- rep(NA, 400)
for(i in 1:400) result[,i] <- Sim2reg(x1,myrates)
```

```
x<-(1:20)/2-3
myrates<-exp(0.4*x+1)
Sim2reg <- function(x1,inputrates){
y <- rexp(length(inputrates), rate = inputrates)
cen<-rexp(length(inputrates), rate = 0.4) + ycen <- pmin(y,cen)
di<-as.numeric(y<=cen)
temp1<-coxph(Surv(ycen, di)~x1)
return(temp1$coef) }
result <- rep(NA, 400)
for(i in 1:400) result[,i] <- Sim2reg(x1,myrates)

x<-(1:30)/2-3
myrates<-exp(0.4*x+1)
Sim2reg <- function(x1,inputrates){
y <- rexp(length(inputrates), rate = inputrates)
cen<-rexp(length(inputrates), rate = 0.4) + ycen <- pmin(y,cen)
di<-as.numeric(y<=cen)
temp1<-coxph(Surv(ycen, di)~x1)
return(temp1$coef) }
result <- rep(NA, 400)
for(i in 1:400) result[,i] <- Sim2reg(x1,myrates)
```

Bibliography

[1] N. Mantel & W. Haenszel, *Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease,* Journal of National Cancer Institute, 22, 719-748 [1959].

[2] D.R.Cox, *Regression Models and Life Tables(with discussion),* Journal of the Royal Statistical Society, Series, B 34, 187-220 [1972].

[3] D.A.Schoenfeld *Sample Size Formula for the Proportion Hazard Model,* Biometrics, Vol. 39, 499-503 [1983].

[4] A.A.Tsiatis, *A Large Sample Study of Cox's Regression Model,* The Annals of Statistics, Vol.9, No. 1, 93-108 [1993].

[5] P.D.Allison, *Survival Analysis Using the SAS System: A Practical Guide,* Cary NC: SAS Institute. [1995].

[6] D.R.Cox & D. Oakes, *Analysis of Survival Data,* London: Chapman and Hall [1995].

[7] T.M. Therneau & P.M. Grambsch *Modeling Survival Data: Extending the Cox Model,* Newyork: Springer [2000].

[8] B.Efron, *The Efficiency of Cox's Likelihood Function for Censored Data,* Journal of American Statistical Association, 72, 557-565 [1977].

[9] N.E.Breslow, *Covariance Analysis of Censored Survival Data,* Biometrics, 30, 89-99 [1974].

[10] J.D.Kalbfleisch & R.L. Prentice, *Marginal Likelihoods Based on Cox's Regression and Life Model,* Biometrika, 60, 267-278 [1973].

[11] R. Kay, *Proportional Hazard Regression Models and the Analysis of Censored Survival Data,* Applied Statistics, 26, 227-237 [1977].

[12] JE. Lawless *Statistical Models and Methods for Lifetime Data,* New York: Wiley [1982].

[13] P. Hu, M. Davidian & A.A.Tsiatis, *Estimating the Parameters in the Cox Model when Covariate Variables are Measured with Error,* Biometrics, 54, 1407-1419 [1998].