Parametric Changepoint Survival Model with Application to Coronary Artery Bypass Graft Surgery Data

by

Suman Lata Jiwani

B.Sc., Simon Fraser University, 1995.

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in the Department

of

Statistics and Actuarial Science

SIMON FRASER UNIVERSITY

Fall 2005

APPROVAL

Name: Suman Lata Jiwani

Degree: Master of Science

Title of project: Parametric Changepoint Survival Model with Application to Coronary Artery Bypass Graft Surgery Data

Examining Committee: Dr. Richard Lockhart
Chair

Dr. Charmaine Dean
Senior Supervisor
Simon Fraser University

Dr. Rachel Altman
Simon Fraser University

Dr. John Spinelli
External Examiner
Simon Fraser University

Date Approved: _____

Abstract

Typical survival analyses treat the time to failure as a response and use parametric models, such as the Weibull or log-normal, or non-parametric methods, such as the Cox proportional analysis, to estimate survivor functions and investigate the effect of covariates. In some circumstances, for example where treatment is harsh, the empirical survivor curve appears segmented with steep initial descent followed by a plateau or less sharp decline. This is the case in the analysis of survival experience after coronary artery bypass surgery, the application which motivated this project. We employ a parametric Weibull changepoint model for the analysis of such data, and bootstrap procedures for estimation of standard errors. In addition, we consider the effect on the analyses of rounding of the data, with such rounding leading to large numbers of ties.

Dedication

To my husband, Ayaz, for his dedication and sacrifices to ensure that I could attain my goal. This endeavour would not have been possible without his help along the way, his encouragement during the difficult times, and his patience and understanding. To my family for their constant support and with whom I did not spend enough time during the past years.

Acknowledgements

I would like to thank my supervisor, Dr. Charmaine Dean, for her guidance, support, patience, and accessibility not only throughout the course of my graduate studies, but also prior to my consideration of entering the program.

I would also like to acknowledge the talented faculty members of the Statistics and Actuarial Science Department for the terrific instruction and their dedication to sharing their wealth of knowledge and experience with their students.

As well, I would like to gratefully acknowlege all the statistics graduate students without whom this journey would have been very difficult, lonely and far less enjoyable.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Many harsh medical interventions involve a substantial risk of mortality with the resulting survivor function appearing segmented in nature, with a steep initial descent followed by a less sharp decline. In some situations, if the patient survives the intervention, there are substantial gains, perhaps even a cure of the disease; after the initial rapid descent, the survivor curve declines very slowly. The estimation of the survivor curve in such instances, and particularly the changepoint of the survivor curve, marking the end of the initial steep descent, are the focus of this project.

The specific context is an understanding of the effects of Coronary Artery Bypass (CAB) grafting surgery. This is a particularly invasive procedure with some risk of mortality. With CAB, it is natural to view the distribution of the time to death (the response variable) as consisting of two or more parts. These represent operative mortality, or death within a short period after surgery, and long-term survival. In previous analyses of CAB data, operative mortality has been defined as death within 30 days of surgery (Gharamani et. al 2001; Chiu 2002), and analyses have proceeded using a logistic model for operative mortality and a proportional hazards model for long-term survival (survival time after 30 days).

The objective of this project is to explore the benefits of a parametric analysis of the CAB data using a segmented Weibull distribution to model the survivor function. One such benefit may be a data-driven approach to locating the changepoint of the survivor function and thus an empirical determination of the period which defines the intial short-term or operative mortality. The methods employed in this project are an adaptation of a model proposed by Noura and Read (1990) who outline the use of parametric modelling of the baseline hazard in terms of piecewise Weibull distributions. Bootstrap techniques are employed to obtain standard errors of the estimates.

1.1 Changepoint Models in Survival Analysis

Standard procedures for survival and event history analysis involve modelling time to death or failure, often as a function of covariates, using either parametric or semi-parametric (e.g. the Cox proportional hazards model) approaches. Various parametric families of models are used in the analysis of lifetime data, including the exponential and the Weibull, with the latter being popular due to its flexibility. In the situation we consider, the survival curve is more complex in that it appears segmented and cannot be effectively modelled with a single distribution over the entire curve.

Survival processes that involve a changepoint, a time point at which the survival experience changes, arise in both the industrial and biological contexts. In reliability analysis, changes in the failure rate can be encountered following a major overhaul or maintenance activity. In survival analysis, changepoint models arise, as discussed, in the case of harsh treatment interventions where there is substantial risk of not surviving the treatment but a much lower risk of failure if the individual survives beyond an initial short-term period after treatment. Patra and Dey (2002) describe

scenarios that arise in clinical trials where the onset of undesirable side effects may cause a different failure rate after a threshold time. They also describe other situations where such segmented models may be useful, for example, involving the introduction of a new treatment where the impact of the treatment is not immediate but affects the failure rate only after some lag time.

The study of changepoint problems in survival anlysis has mainly focussed on modelling of the hazard function. Classical approaches to modelling the hazard rate with changepoint are considered by Nguyen et al. (1984) and Loader (1991). Nguyen et al. (1984) consider a parametric approach, modelling the segmented hazard function using a mixture of truncated and delayed exponential distributions, and propose estimation techniques for obtaining consistent estimators of the changepoint and the hazard rates before and after the changepoint. Loader (1991) also considers a parametric approach and uses maximum likelihood methods for estimation of the initial hazard rate and changepoint. Approximate confidence regions for the changepoint and the size of the change are obtained through a study of the asymptotic properties of the estimators. Patra and Dey (2002) propose a Bayesian approach for studying a general class of models for hazard functions with a changepoint and, in general, for curves which are functions of survival times. Gijbels and Gurler (2003) also consider the problem of estimating hazard functions with a jump discontinuity for right-censored data; they consider not only the problem of estimating the changepoint location but also the size of the jump as well as the hazard rate before the changepoint using a comparison of three methods: a parametric maximum likelihood estimation approach, a nonparametric approach using a Nelson-Aalen type estimator, and a least squares estimation procedure which also uses the nonparametric Nelson-Aalen estimate of the cumulative hazard function. Noura and Read (1990) consider parametric modelling of the baseline hazard in terms of piecewise distributions. Their

piecewise model of the baseline hazard is adapted in this study.

1.2 Bootstrap Techniques

The bootstrap is a useful tool for obtaining standard errors and confidence intervals. Bootstrap techniques can be applied with few assumptions and minimal modeling or analysis to a variety of situations. In this project, we consider bootstrap methods specific to right-censored survival data. We experiment with different methods of resampling censored data to study the impact of such techniques on bootstrap estimates for a single changepoint model. As well, we consider a simulation study of the effects of rounding on estimation leading to tied observations as occur in this dataset.

1.3 British Columbia Cardiac Registry Database

The British Columbia Cardiac Registry database is a comprehensive, population based provincial registry that was created with the purpose of building an electronic patient record that would provide data for reporting, planning and research purposes. The database was created in 1989 by the provincial Ministry of Health in response to reported long waiting times for cardiac surgery. The data collection for the registry began in 1991.

The database captures prognostic information on all open heart surgeries performed in the province. Cardiac surgeons provide information that populates the registry by documenting patient information through the Operative Report form, which is used to approve the procedure, and clinical data.

1.4 Coronary Artery Bypass Data

The coronary arteries are the vessels that carry blood and oxygen to the heart muscle. These arteries can become clogged with fatty deposits, known as plaque, thus preventing the heart from getting enough blood and oxygen which often leads to chest pain and shortness of breath. This clogging of the arteries and the resulting heart condition is known as Coronary Artery Disease (CAD) also sometimes referred to as Coronary Heart Disease (CHD). There are three main treatment regimens for CAD: drug therapy, a surgical treatment known as angioplasty, and bypass surgery. Drugs are often prescribed as a first step to relax the arteries, lower the heart rate and blood pressure, and sometimes to thin the blood. An angioplasty procedure may be used to open and stretch a blocked artery in order to improve blood flow. For severe cases, Coronary Artery Bypass (CAB) graft surgery is recommended. CAB surgery is the most commonly performed 'open heart' operation. In CAB surgery, a blood vessel is taken from another part of the body and then attached above and below (to bypass) the narrowed part of the blocked artery thus restoring blood and oxygen flow to the heart. A bypass can be done for each blocked artery.

This study is concerned with modelling the time to death of patients that have undergone CAB surgery. The data available for the analysis are limited to CAB data from the provincial registry database from 1991 to 1994 inclusive. In order to identify death dates for patients who had died, the cardiac registry data were linked with the death file at the BC Vital Statistics Agency (VS) in Victoria, B.C. The two files were linked using the patients' unique personal health number, name, birth date, gender and place of residence at time of surgery. The method of probabilistic record linkage which calculates a weight for each pair of records and assigns a match based on the magnitude of the computed weight was used to match the data from the two sets.

For this study, the registry data were further limited to a subset consisting of the first isolated CAB surgery of all individuals who received at least one CAB surgery in this period. The term isolated refers to the scenario that no other procedure (such as a valve replacement) could be done at the same time as the CAB surgery. Here, we focus on one-year survival experience. Preliminary analyses of the five-year study data indicate that a model with a single changepoint would isolate one at about two years after surgery and the intention here is to consider whether an earlier changepoint exists, specifically one shortly after surgery. The total number of patients in this subset is 6060. The ages of the patients in the study ranged from 27 to 92 with the median age being 65. The breakdown of the 6060 CAB surgery cases by year and number of deaths in a particular year is given in Table 1.1.

Year	Number of Cases	Number of Deaths
1991	1372	53
1992	1571	76
1993	1546	53
1994	1571	72
Total	6060	254

Table 1.1: Coronary artery bypass data summary

Figure 1.1 illustrates the Kaplan-Meier survivor function for the CAB patients for the 1-year period of follow up. The scale of the y-axis was narrowed to begin at 0.95 to show more clearly the shape of the survivor curve, especially within the first 30 days. The estimated 30 day and 1-year survival probabilities are 98% ($97.6\% \pm 0.2\%$) and 96% ($95.8\% \pm 0.3\%$) respectively. The steep initial descent in the Kaplan-Meier curve defines the period of operative mortality and is followed by the less rapid decline in survival probabilities.

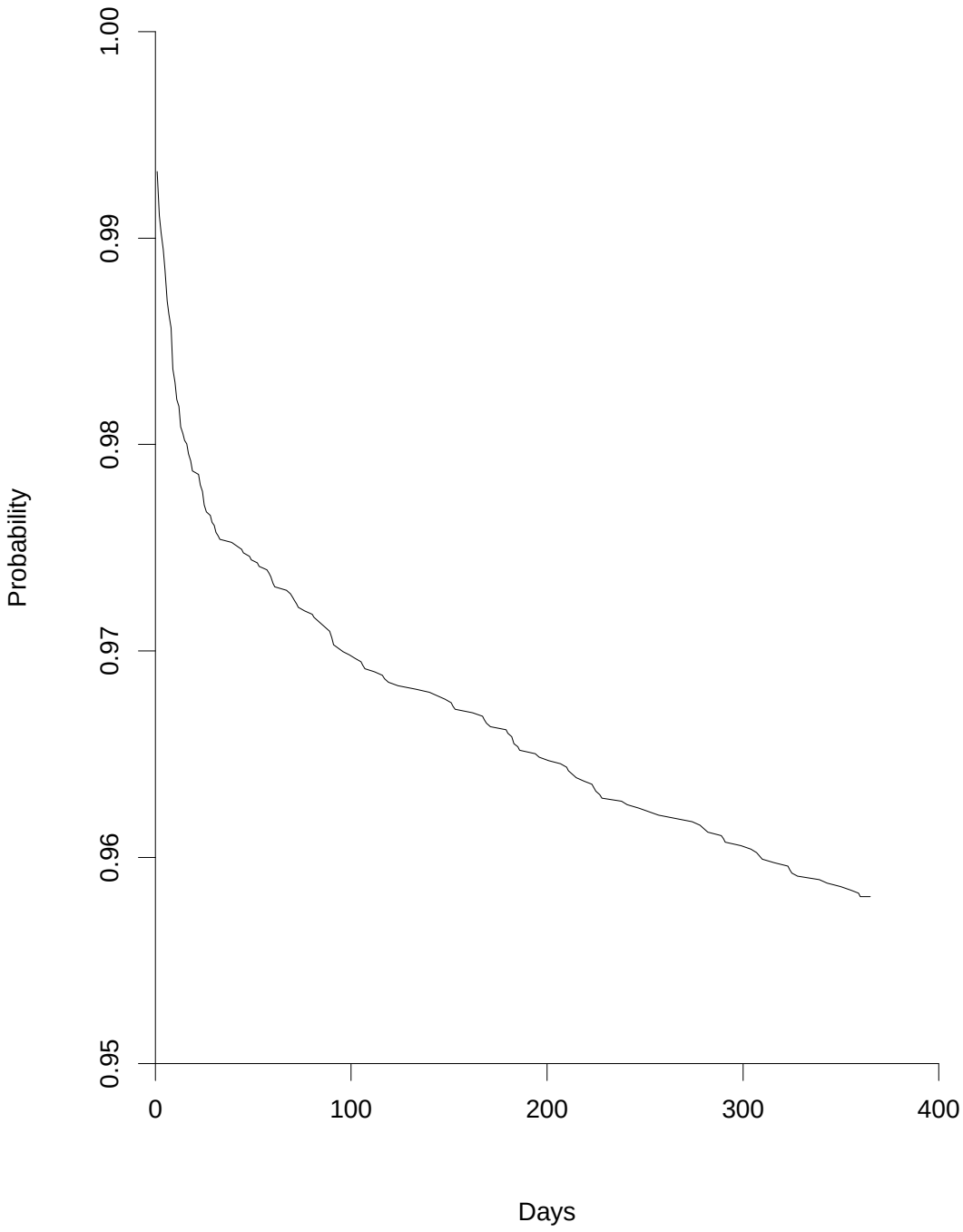

Figure 1.1: Estimated survivor function: one year follow up data

1.5 Plan of the Project

The plan of the project is as follows.

In Chapter 2 we consider a parametric analysis using a segmented Weibull distribution to model a survivor function with a single changepoint. Bootstrap methods for estimating variability of estimators are discussed.

In Chapter 3 the model is fitted to the British Columbia cardiac registry data and compared to the fit from a Weibull model without a changepoint.

Chapter 4 presents a simulation study to investigate the effect of rounding on parameter estimation.

Chapter 5 provides an overview of the project and a discussion of future work.

Chapter 2

Modelling with Piecewise Weibulls

2.1 Introduction and Model Assumptions

Traditional survival analysis involves fitting a model to a single response, survival time, which is measured relative to a relevant time-origin (for example, the start of a treatment). Both parametric and nonparametric approaches can be considered for this purpose. Within the group of fully parametric statistical models, the Weibull model is very widely used. The model is flexible enough to describe many different types of lifetime data. It is often applied to lifetimes of a variety of manufactured items, as well as in biological and medical applications. This flexibility and the fact that the model has simple expressions for the probability density and survivor and hazard functions partly account for its popularity (Lawless 2003).

Under the assumption of a Weibull distribution, the probability density of lifetime, T, is

$$\frac{\alpha}{\delta}\left(\frac{t}{\delta}\right)^{\alpha-1}\exp\left[-\left(\frac{t}{\delta}\right)^{\alpha}\right] \qquad t \geq 0; \tag{2.1}$$

Here, α ($\alpha > 0$) is the shape parameter and δ ($\delta > 0$) is the scale parameter.

Incorporating covariates only into the scale parameter, δ, implies proportional hazards for lifetimes. We focus here on the development of a two-stage Weibull model with one changepoint.

Let a represent the single changepoint considered. Let T_i denote the i-th lifetime, L_i denote the i-th censoring time and $t_i = min\{T_i, L_i\}$. Here, lifetime is defined as the interval between date of surgery and date of death. Though written here in a broader context, note that for the CAB data censoring time is defined as 365 days since we are considering only a one year follow up for all patients and all patients were followed for this period. Then for $i = 1, \ldots, n$ let

$$w_i = \begin{cases} 0 & \text{if the } i^{th} \text{ individual is censored} \\ 1 & \text{otherwise} \end{cases} \qquad c_i = \begin{cases} 1 & \text{if } 0 < t_i \leq a \\ 0 & \text{otherwise} \end{cases}$$

For a Weibull distribution the cumulative hazard function is $(t/\delta)^{\alpha}$ and its logarithm is $\alpha \log t + \lambda^*$ where $\lambda^* = -\alpha \log \delta$. Let $g(t)$ denote the logarithm of the cumulative hazard for a piecewise Weibull distribution with one changepoint. Then for the i^{th} individual:

$$g(t_i) = c_i(\lambda_1^* + \alpha_1 \log t_i) + (1 - c_i)(\lambda_2^* + \alpha_2 \log t_i), \tag{2.2}$$

where λ_1^* and α_1 refer to the parameters of the Weibull segment before the changepoint, a, and λ_2^* and α_2 refer to the parameters of the Weibull segment after a. In order to have continuity of the survivor function and hence $g(t)$ at the changepoint a, we require that

$$\alpha_1 \log a + \lambda_1^* = \alpha_2 \log a + \lambda_2^* \tag{2.3}$$

so that

$$\lambda_2^* = \lambda_1^* + (\alpha_1 - \alpha_2) \log a \tag{2.4}$$

Note that the restriction (2.3) that imposes continuity of $g(t)$ ensures continuity at the changepoint of the survivor function $S(t)$ or equivalently, the cumulative hazard

function $H(t)$. However, this is not the case for the hazard function $h(t)$. Denoting $\lambda = \lambda_1^*$, we write $g(t_i) = \log H(t_i)$ in terms of the three model parameters λ ($\lambda \in \Re$), α_1 ($\alpha_1 > 0$), and α_2 ($\alpha_2 > 0$), as

$$g(t_i) = \lambda + c_i(\alpha_1 \log t_i) + (1 - c_i)[\alpha_2 \log t_i + (\alpha_1 - \alpha_2)\log a] \tag{2.5}$$

The hazard function $h(t_i)$ for the i^{th} individual is

$$h(t_i) = H'(t_i) = \exp[g(t_i)]\, g'(t_i) = \frac{H(t_i)}{t_i}[c_i\alpha_1 + (1 - c_i)\alpha_2] = \frac{H(t_i)}{t_i}\left(\alpha_1^{c_i} \cdot \alpha_2^{(1-c_i)}\right) \tag{2.6}$$

and the survivor function is $S(t_i) = \exp(-\exp[g(t_i)])$, or

$$S(t_i) = \exp(-\exp\{\lambda + c_i(\alpha_1 \log t_i) + (1 - c_i)[\alpha_2 \log t_i + (\alpha_1 - \alpha_2)\log a]\}). \tag{2.7}$$

The probability density function is

$$f(t_i) = h(t_i)\exp\{-H(t_i)\} \tag{2.8}$$

2.2 Likelihood Development and Maximum Likelihood Estimation

We build the likelihood function for the segmented model using (2.7) and (2.8) by considering the contribution of each individual to the likelihood. Suppose that a sample of n individuals yields observed lifetimes $T_1 \ldots, T_n$. For each individual we have $t_i = min(T_i, L_i)$ and a censoring indicator w_i. Thus, the data arise in pairs (t_i, w_i), and assuming independence among the data pairs for the n individuals we can build the likelihood for the i^{th} individual as

$$L_i = [f(t_i)]^{w_i}[S(t_i)]^{1-w_i} = h(t_i)^{w_i}\left[e^{-H(t_i)}\right].$$

The logarithm of the likelihood becomes

$$\log L = \sum_{i=1}^{n} \{w_i [c_i \log \alpha_1 + (1 - c_i) \log \alpha_2] - w_i \log t_i + w_i \log H(t_i) - H(t_i)\} \quad (2.9)$$

where $\log H(t_i)$ is defined in (2.5).

To maximize the logarithm of the likelihood with respect to the parameters, we employ a grid search or likelihood profile approach: maximum likelihood estimates of λ, α_1, and α_2 are obtained for a fixed value of the changepoint parameter a and the search covers a range of values of a to locate the overall joint maximum likelihood estimates.

The first derivatives of the logarithm of the likelihood with respect to the parameters α_1, α_2, and λ, are required for the grid search and they are:

$$\frac{\partial logL}{\partial \alpha_1} = \sum_{i=1}^{n} \frac{w_i c_i}{\alpha_1} + \{w_i - H(t_i)\} [c_i \log t_i + (1 - c_i) \log a_1]$$

$$\frac{\partial logL}{\partial \alpha_2} = \sum_{i=1}^{n} \frac{w_i (1 - c_i)}{\alpha_2} + \{w_i - H(t_i)\} [(1 - c_i) \log t_i - (1 - c_i) \log a_1]$$

$$\frac{\partial logL}{\partial \lambda} = \sum_{i=1}^{n} w_i - H(t_i)$$

For fixed a, the maximum likelihood estimates of α_1, α_2, and λ may be found using a Newton-Raphson updating algorithm. Experience shows that there are no problems in implementing this algorithm in this scenario. An alternative updating algorithm may be constructed as follows. Let a^p, λ^p, ${\alpha_1}^p$, and ${\alpha_2}^p$ denote current values of the parameters a, λ, α_1, and α_2 respectively. As well, let ${c_i}^p$ and $H^p(t_i)$ denote c_i and $H(t_i)$ evaluated at current values of the parameters. Then, the likelihood equations for α_1 and α_2 may be arranged to provide updates using:

$${\alpha_1}^{p+1} = \frac{-\sum_{i=1}^{n} w_i {c_i}^p}{\sum_{i=1}^{n} [w_i - H^p(t_i)] [{c_i}^p \log t_i + (1 - {c_i}^p) \log a^p]} \quad (2.10)$$

$${\alpha_2}^{p+1} = \frac{-\sum_{i=1}^{n} w_i (1 - {c_i}^p)}{\sum_{i=1}^{n} [w_i - H^p(t_i)] [(1 - {c_i}^p) \log t_i - (1 - {c_i}^p) \log a^p]} \quad (2.11)$$

An algorithm for finding the mle of the parameters λ, α_1, and α_2 for fixed a may then be obtained as follows. Given current values a^p, λ^p, ${\alpha_1}^p$, and ${\alpha_2}^p$:

Step 1. Compute $H^p(t_i)$. Solve $\frac{\partial logL}{\partial \lambda} = 0$ iteratively updating λ to convergence with all other parameters fixed at their current values. Set λ^{p+1} to be the value of λ at such convergence.

Step 2. Compute $H(t_i)$ evaluated at a^p, ${\alpha_1}^p$, ${\alpha_2}^p$, and using λ at λ^{p+1} from step 1. Denote this to be $H^p(t_i)$ for this step 2. and then obtain a one-step update of α_1 and α_2 using (2.10) and (2.11).

Repeat steps 1 and 2 to convergence; either the score vector is suitably close to zero or updates of λ, α_1, and α_2 using steps 1 and 2 above do not change substantially from the previous iteration.

2.3 Bootstrap Methods for Confidence Interval Estimation

Bootstrap methods are based on simulations or resampling of the data and are very useful for assigning measures of accuracy to statistical estimates. The advantage of the bootstrap is that it requires few assumptions and little modelling and can be applied in a systematic way to a large number of scenarios.

One can best describe the distinction between bootstrap methods and traditional parametric statistical inference through the concept of the sampling distribution of a statistic. Consider a population probability distribution F which has a parameter, θ, that is estimated by means of a statistic, say, T_n, whose value for the sample is $\hat{\theta}$ computed from a sample of size n drawn from the population under consideration. The sampling distribution of T_n is the relative frequency distribution of all possible

values of T_n computed from an infinite number of random samples of size n drawn from the population. It is of interest to estimate this sampling distribution in order to make inferences about the population parameter, θ. Traditional parametric inference involves making assumptions about the shape of the sampling distribution of T_n, however, the nonparametric bootstrap is distribution-free relying instead on the fact that the sample's distribution is a good estimate of the population distribution.

A brief description based on the work of Efron and Tibshirani (1993) of the essential concepts involved in the nonparametric bootstrap method follows. Let $x_1, \ldots, x_n$ be a random data sample of size n which are independent and identically distributed (i.i.d) outcomes of random variables $X_1, \ldots, X_n$ from a population with cumulative density function (CDF) denoted by F. An estimate of the CDF, say, $\hat{F}$, can be constructed from this sample. The empirical distribution function (EDF), or $\hat{F}$, is defined such that there is probability $1/n$ on each observed value $x_i, i = 1, 2, \ldots, n$. The notion of the plug-in principle is also important in understanding the bootstrap. This principle states that if a parameter of a probability distribution F is to be estimated from a random sample drawn from F, and the EDF $\hat{F}$ is used to estimate F, then any function $\theta = t(F)$ can be estimated by applying the same function to $\hat{F}$, $\hat{\theta} = t\left(\hat{F}\right)$. The bootstrap is advantageous in that it allows the study of the bias and standard error of $\hat{\theta} = t\left(\hat{F}\right)$ regardless of how complicated the functional mapping $\theta = t(F)$ is. Having defined the EDF, a random i.i.d. sample of size n is drawn from $\hat{F}$ *with replacement*. The bootstrap sample is denoted as $x^* = (x_1{}^*, x_2{}^*, \ldots, x_n{}^*)$ where the asterisk indicates that the components of x^* are not the actual data set but a randomized or resampled version of the original data set $x_1, \ldots, x_n$. The parameter estimate from the bth bootstrap sample, $b = 1, \ldots, B$, is denoted $\hat{\theta}^*(b)$. Having obtained parameter estimates from B independent bootstrap samples, the bootstrap estimate of the standard error, $se_F\left(\hat{\theta}\right)$, is found through an application of

the plug-in principle that uses the empirical distribution $\hat{F}$ in place of the unknown distribution F. Specifically, the bootstrap estimate of $se_F\left(\hat{\theta}\right)$ is defined by $se_{\hat{F}}\left(\hat{\theta}^*\right)$ and is known as the ideal bootstrap estimate of standard error of $\hat{\theta}$. A computational way of approximating the numerical value of $se_{\hat{F}}\left(\hat{\theta}^*\right)$ is by computing the sample standard deviation of the B replications:

$$\hat{se}_B = \left\{ \sum_{b=1}^{B} \left[\hat{\theta}^*(b) - \bar{\hat{\theta}}^* \right]^2 / (B-1) \right\}^{\frac{1}{2}} \tag{2.12}$$

where

$$\bar{\hat{\theta}}^* = \sum_{b=1}^{B} \hat{\theta}^*(b) / B.$$

Note that

$$\lim_{B\to\infty} \hat{se}_B = se_{\hat{F}} = se_{\hat{F}}\left(\hat{\theta}^*\right).$$

The bootstrap estimate of standard error usually has relatively little bias; the smallest possible standard deviation among nearly unbiased estimates of $se_F\left(\hat{\theta}\right)$ occurs with $B = \infty$ in the asymptotic ($n \to \infty$) sense. Since we must stop after a finite number of replications, $\hat{se}_B$ always has greater standard deviation than $\hat{se}_\infty$, and the magnitude of the discrepancy can be illustrated in terms of the coefficient of variation of $\hat{se}_B$, the ratio of the standard deviation of $\hat{se}_B$ to its expectation (see Efron and Tibshirani 1993). The coefficient of variation reflects variation both at the resampling level (due to stopping after B bootstrap replications) and at the population sampling level, as the ideal estimate $\hat{se}_\infty$ can still have considerable variability as an estimate of $se_F\left(\hat{\theta}\right)$ due to the variability of using $\hat{F}$ as an estimate of F. Thus reliable results are best obtained by using many bootstrap replications.

For this project, standard errors of parameter estimates were obtained using the nonparametric bootstrap that resampled the CAB data survival times with replacement and imposed fixed time censoring at 365 days for each of the bootstrap samples. Standard errors for each of the parameter estimates were obtained by applying

equation (2.10). One additional bootstrap technique was employed using the Boot library, developed by Angelo Canty, which includes special algorithms for resampling of right-censored data. Specifically, the other method considered was the so-called *weird* bootstrap.

The weird bootstrap method for resampling censored data was introduced by Andersen et al. (1993). This method of resampling works by simulating from the Nelson-Aalen estimate of the cumulative hazard function. At each of the observed event times (lifetimes or failure times), the risk-sets as given by the original sample are kept fixed. In this way, the censored observations are held as fixed. For each of the bootstrap samples, new events are randomly drawn within each risk set. Let $Y(t)$ represent the number of observations in the risk set at time t. Then, the number of deaths at time t is simulated from a $Binomial\left(Y(t), \frac{dN(t)}{Y(t)}\right)$ distribution where $dN(t)$ is the observed number of events at time t. Hence the weird bootstrap (i) fixes the censored data and (ii) generates the number of deaths from the binomial distribution each time a death was recorded. Since the events are drawn independently among the fixed risk sets, the strangeness of this bootstrap is that the resampling strategy can result in data sets with either fewer or more observations than the original data although the observed number of censored observations will remain the same.

Chapter 3

Application to the CAB Data

3.1 Preliminary Data Exploration

The CAB data consist of 6,060 cases, with 254 deaths. Of the 254 deaths, 145 of them, or 57%, occurred on or before 30 days. This large percentage of deaths early on is reflected in the Kaplan-Meier survivor function presented in Figure 1.2, which shows a steep initial descent. The average of the 254 lifetimes for this data set was 77 days. An important point to note is that lifetimes are rounded here to the nearest day. In Chapter 4 we explore the effect of such rounding on our analysis. Table 3.1 summarizes the average lifetimes of those individuals who died within two groups: on or before 30 days and after 30 days.

Survival Time	Avg of Lifetimes	No. of Deaths
$\leq$30 days	8 days	145
$>$30 days	169 days	109

Table 3.1: CAB data average lifetimes before and after 30 Days

3.2 Model Fitting

Graphical inspection of the Kaplan-Meier survivor function estimate is often useful in assessing the appropriateness of a parametric model. If the piecewise model is appropriate, a diagnostic plot should show well-defined sections meeting at the changepoint value. Visual inspection to locate the changepoint is also useful in providing good initial estimates for the maximum likelihood grid search procedure. Figure 3.1 shows a plot of $\ln \hat{S}(t)$ against t where $\hat{S}(t)$ is the Kaplan-Meier estimate of the survivor function. The plot does appear to reveal distinct segments. As well, it is somewhat suggestive of a changepoint at 30 days which supports the initial intuition about the changepoint location.

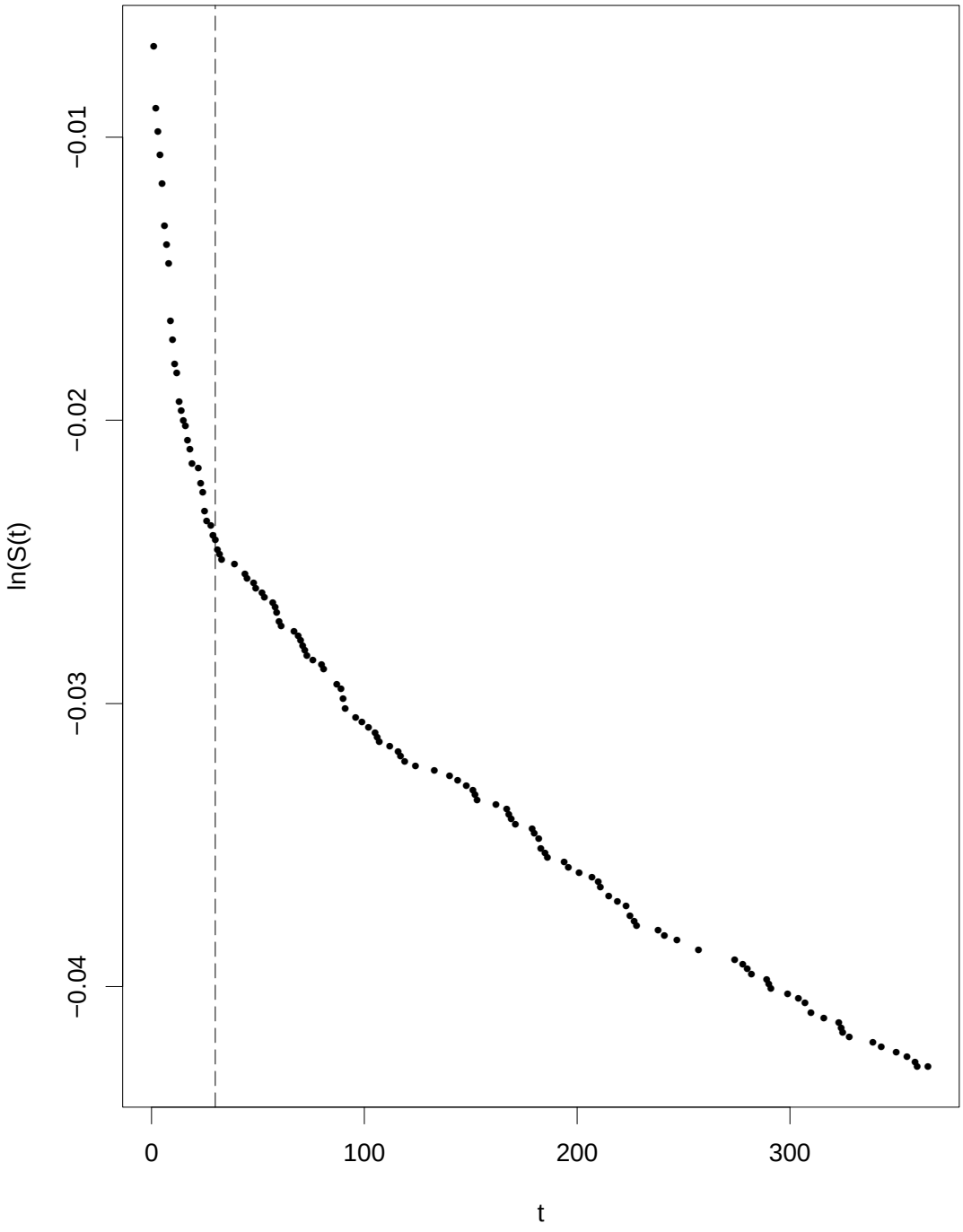

Figure 3.1: Diagnostic plot of CAB data: the logarithm of the Kaplan-Meier estimate of the survivor function versus time

The parameter estimates of the postulated segmented model (2.2) are provided in Table 3.2.

Parameter	Maximum Likelihood Estimate
a	9.0 days
α_1	0.78
α_2	0.26
λ	-5.81

Table 3.2: Parameter estimates for segmented Weibull model applied to CAB data

Note that there were 99 patients who died before 9 days which represented approximately 39% of all deaths, and 68% of all deaths before 30 days. The changepoint estimate is much lower than initially postulated.

The usual nonparametric bootstrap and the so-called *weird* bootstrap provided approximations of the standard error and bias for the parameter estimates. For both methods, 1,000 replications were obtained and for each replication, a grid search of 1 day increments was employed for the maximum likelihood estimation. Tables 3.3 and 3.4 summarize the results of the bootstrap replications including the average value of parameter estimates, standard deviation, bias, the absolute value of the bias divided by the standard error and 95% confidence intervals based on percentiles of the bootstrap distributions.

1,000 Nonparametric Bootstrap Replicates						
	Estimate	Mean	Std.Dev	Bias	Abs(bias)/Std.Dev	95% C.I.
a	9.0	10.14	3.52	1.14	0.32	$(4, 17)$
α_1	0.78	0.80	0.09	0.02	0.19	$(0.65, 1.01)$
α_2	0.26	0.25	0.02	-0.01	0.26	$(0.21, 0.30)$
λ	-5.81	-5.83	0.15	-0.02	0.16	$(-6.15, -5.56)$

Table 3.3: Standard errors and bias - nonparametric bootstrap

1,000 Weird Bootstrap Replicates						
	Estimate	Mean	Std.Dev	Bias	Abs(bias)/Std.Dev	95% C.I.
a	9.0	10.35	3.41	1.35	0.40	$(4, 19)$
α_1	0.78	0.79	0.09	0.01	0.13	$(0.66, 1.00)$
α_2	0.26	0.25	0.02	-0.01	0.28	$(0.20, 0.30)$
λ	-5.81	-5.83	0.15	-0.02	0.14	$(-6.14, -5.57)$

Table 3.4: Standard errors and bias - weird bootstrap

The bootstrap results show very little variation between the methods. Efron and Tibshirani (1993) state that values of the ratio of the bias to standard error less than about 0.25 indicate that the small sample bias observed can be ignored. This ratio is presented in Tables 3.3 and 3.4 and the results obtained from these bootstraps indicate that there may be some small sample bias in the estimate of the changepoint parameter.

The histograms of the 1,000 bootstrap replicates from the nonparametric bootstrap for each of the parameters appear in Figure 3.2. Corresponding plots from the weird bootstrap method were very similar and are not provided here. The distribution of the changepoint parameter is bimodal, with the first mode at 9 days and the second at 13 days. Figures 3.3 and 3.4 are boxplots and qqplots for the bootstrap replicates for each parameter. The qqplots for all parameters except α_2 demonstrate non-normal distributions.

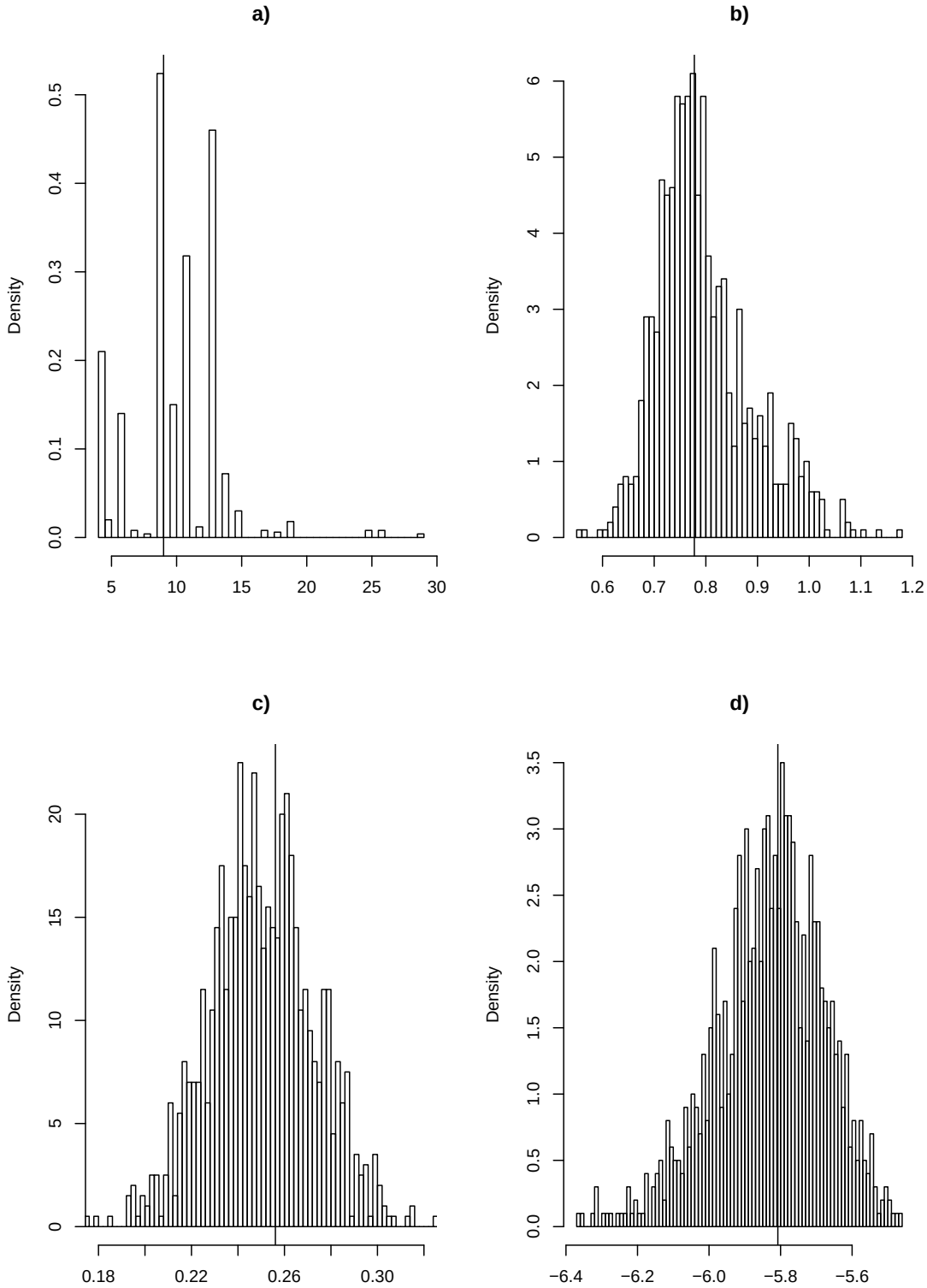

Figure 3.2: Histograms for 1,000 bootstrap replicates a) Changepoint parameter b) α_1 c) α_2 d) λ

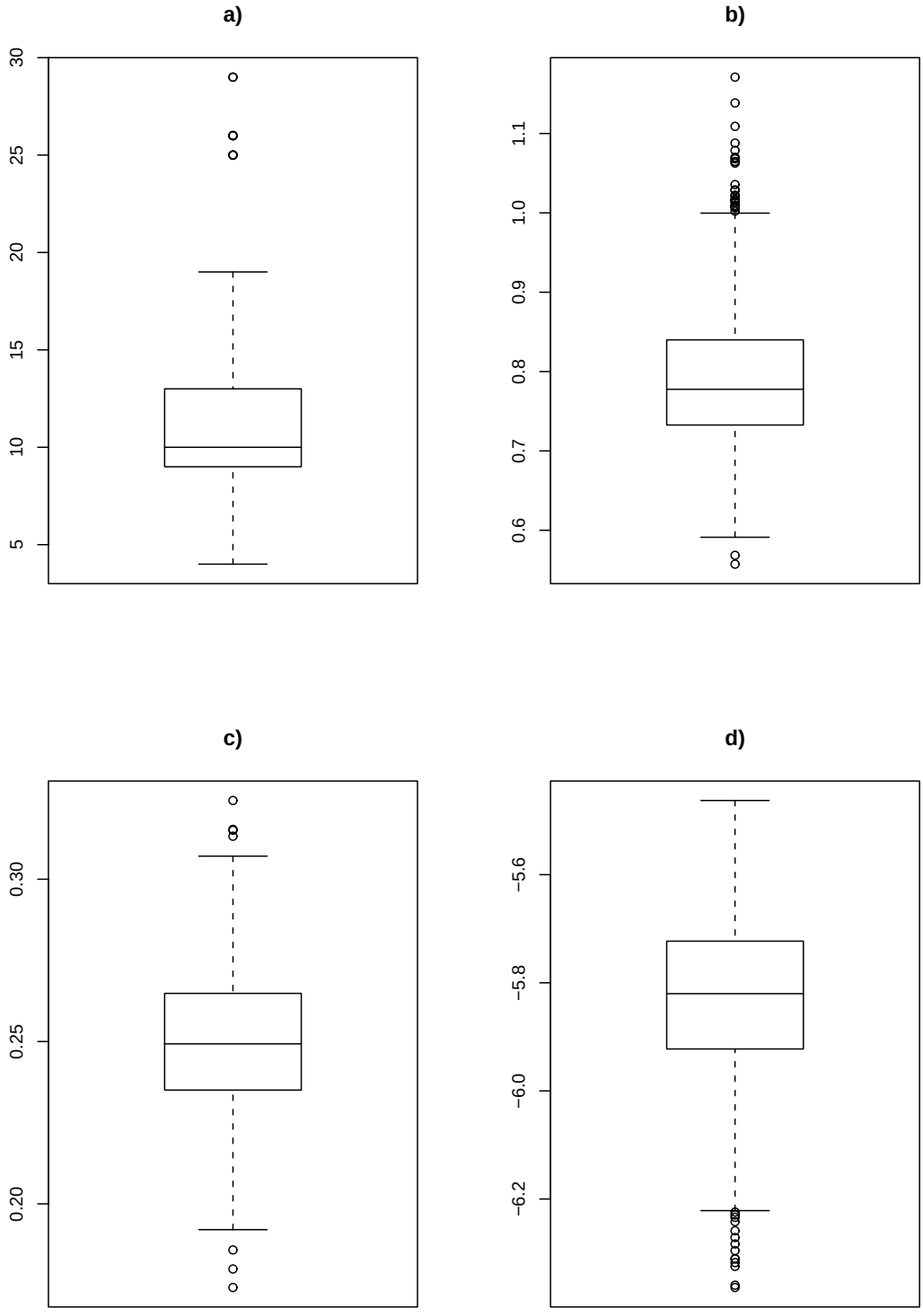

Figure 3.3: Boxplots for the 1,000 bootstrap replicates a) Changepoint parameter b) α_1 c) α_2 d) λ

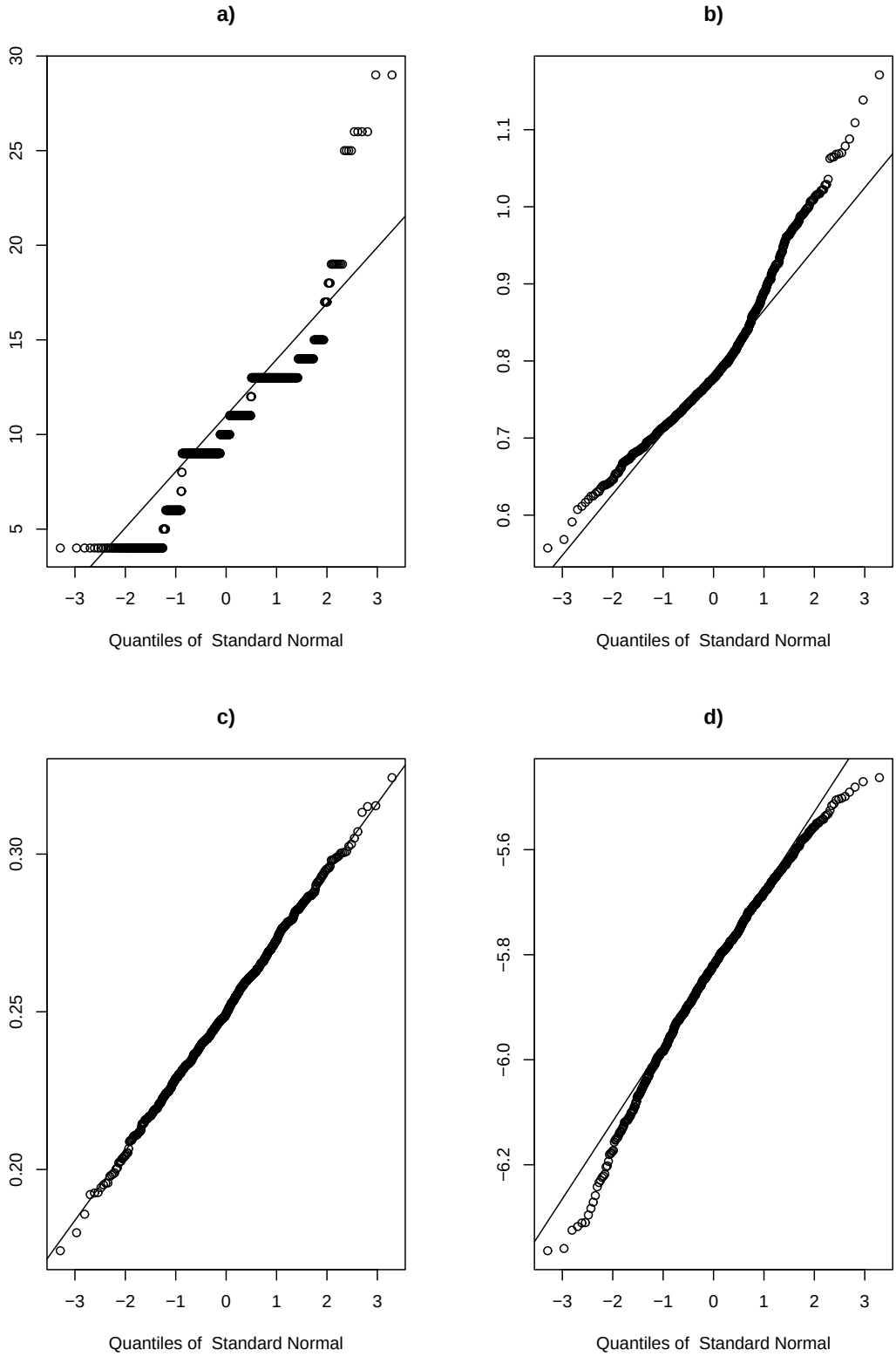

Figure 3.4: QQplots for the 1,000 bootstrap replicates a) Changepoint parameter b) α_1 c) α_2 d) λ

3.3 Comparison with Single Weibull Model

For the single Weibull model, the shape parameter estimate is 0.35 with $\hat{\lambda}$ estimated as -5.19. Figure 3.5 compares the fit of the piecewise and single Weibull models with the Kaplan-Meier estimate of the survivor function. Based only on this visual inspection, the piecewise Weibull model seems to give a better overall fit to the data.

When $\alpha_1 = \alpha_2$, or equivalently, when $\alpha_1 - \alpha_2 = 0$, the piecewise Weibull model reduces to the single Weibull model. The minimum value of the bootstrap estimate of $\alpha_1 - \alpha_2$ from the previous section is 0.348, providing further evidence that the single Weibull model does not give a good fit. A 95% confidence interval for $\hat{\alpha}_1 - \hat{\alpha}_2$ based on the bootstrap distribution is (0.414, 0.737).

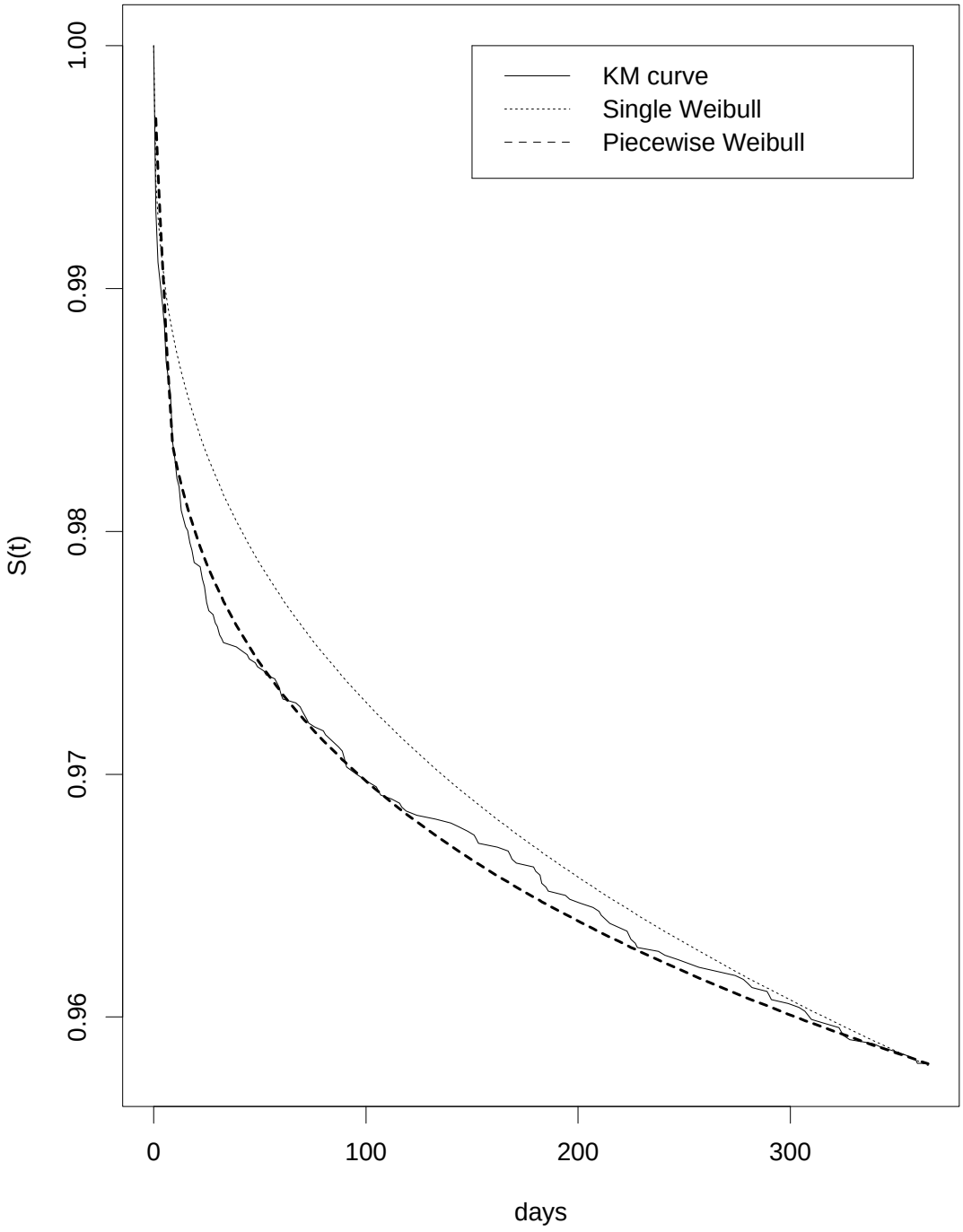

Figure 3.5: Comparison of single Weibull and piecewise Weibull fit

3.4 Residual Analysis

A primary tool for model validation is graphical residual analysis. Graphical methods have the advantage that they readily illustrate a broad range of complex aspects of the relationship between the model and the data. Specifically, we consider the modified Cox-Snell residual in determing lack-of-fit. The residual in this case is defined as follows:

$$\hat{e}_i = \begin{cases} \hat{H}(t_i) & \text{if } i^{th} \text{ observation is a death} \\ \hat{H}(t_i) + 1 & \text{if } i^{th} \text{ observation is censored} \end{cases}$$

The definition above follows from the fact that if a continuous random variable T has survivor function $S(t)$, then $S(T) \sim U(0,1)$, so that the cumulative hazard function, $H(T) = -\log S(T)$ has a standard exponential distribution. That is, the full set of residuals should look roughly like a sample from the standard exponential distribution. Kalbfleisch and Prentice (2002) recommend plotting these residauls against the expected order statistics of the standard exponential distribution when there are few censored observations. If the fit of the model is adequate, the plot should be a straight line with slope 1. Alternatively, having computed the residuals, one could calculate the product-limit estimate of the survivor function of $\hat{e}_i$ ($S^{PL}(\hat{e}_i)$) and then plot $-\log S^{PL}(\hat{e}_i)$ versus $\hat{e}_i$. Again this should be roughly linear. Figure 3.6 illustrates plots of the modified Cox-Snell residuals for both the piecewise and single Weibull models. In the plot of the piecewise Weibull model residuals, a roughly linear shape is seen and no glaring discrepancies surface. However, the plot of the residuals from the single Weibull model does not demonstrate the same linearity.

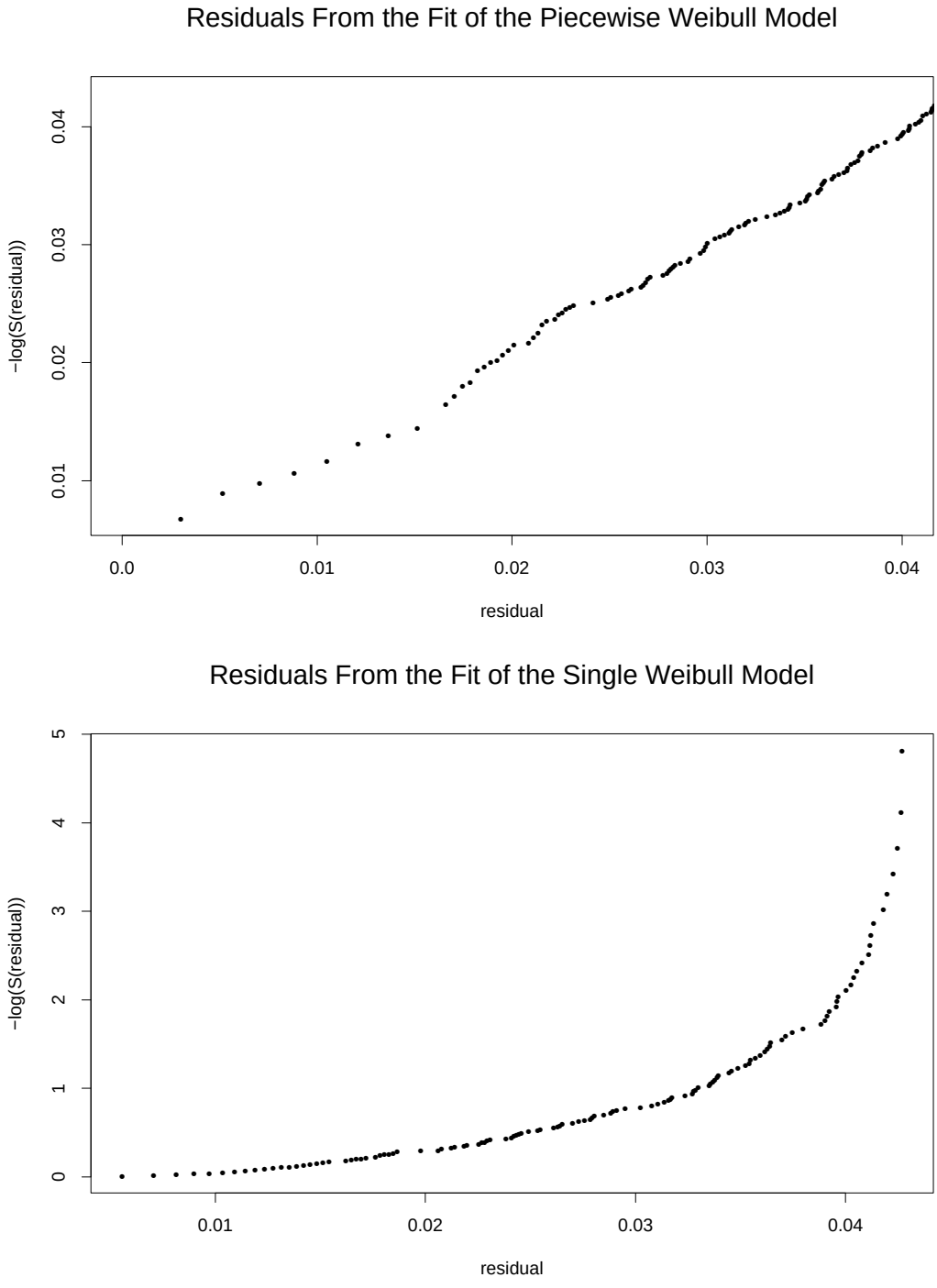

Figure 3.6: Modified Cox-Snell residuals for piecewise and single Weibull models

Chapter 4

Simulation Study on Rounding Effects

4.1 Introduction and Simulating Data

A simulation study was performed to investigate the effect of rounding on parameter estimation and on bootstrap estimation of standard errors.

Using the parameter estimates from the CAB data presented in Table 3.2, lifetimes were generated from a piecewise Weibull model using the inverse transform algorithm. Equation 2.7 gives the survivor function of the i^{th} individual under the segmented Weibull model, which can be written as:

$$S(t_i) = \begin{cases} \exp\{-\exp[\lambda + \alpha_1 \log t_i]\} & \text{if } 0 < t_i \leq a \\ \exp\{-\exp[\lambda + \alpha_2 \log t_i + \log a_1 (\alpha_1 - \alpha_2)]\} & \text{if } a < t_i < \infty \end{cases}$$

The CDF is then given by:

$$F(t_i) = \begin{cases} 1 - \exp\{-\exp[\lambda + \alpha_1 \log t_i]\} & \text{if } 0 < t_i \leq a \\ 1 - \exp\{-\exp[\lambda + \alpha_2 \log t_i + \log a (\alpha_1 - \alpha_2)]\} & \text{if } a < t_i < \infty \end{cases}$$

We have that $0 \leq F(t_i) \leq 1$ for all t_i. At the changepoint, a, $F(a) = 1 - \exp\{-\exp[\lambda + \alpha_1 \log a]\}$. The simulated data set is created by first generating random numbers, u, from the uniform distribution $U[0,1]$ and then transforming these to the survival times of interest using the CDF as given above. We then have:

$$t = \begin{cases} \exp\left(\frac{\log[-\log(1-u)]-\lambda}{\alpha_1}\right) & \text{if } u \leq F(a) \\ \exp\left(\frac{\log[-\log(1-u)]-\lambda-\alpha_1 \log a+\alpha_2 \log a}{\alpha_2}\right) & \text{if } u > F(a) \end{cases}$$

For this study, censoring was imposed through a fixed time censoring mechanism to mimic the CAB data; all individuals with survival times of greater than 365 days were censored.

4.2 Rounding of Simulated Data

We consider the effects on estimation given that lifetime data are rounded to the nearest day or to the nearest hour. The grid search increment size for obtaining the maximum likelihood estimates was dictated by the rounding scheme. For the unrounded data, a grid size of 0.5 days was used. For data rounded to the nearest day, the grid size was 1 day. A grid size of 0.5 days was also used for the data rounded to the nearest hour. Note that when imposing rounding on the data, there is the possibility that some very short survival times will round to zero values. In addition to consideration of the rounding scheme, it is important to also determine how best to deal with rounded zeros. For the purposes of this project, when rounding to the nearest day, those values that round to zero were set to a nominal survival time of 0.05 days, and when rounding to the nearest hour, rounded zeroes were set to 0.005 days.

For the simulation study, 1000 data sets were generated from a piecewise Weibull model with parameter values set to be the maximum likelihood estimates as given

in 3.2, and the three different approaches to rounding were applied to each data set. Table 4.1 shows the number of distinct and tied lifetimes in the CAB data as well as corresponding averages for the 1000 simulated data sets where generated data are rounded to the nearest day. There is fair agreement in the number of ties in the simulated data sets with those in the CAB data. Although details are not presented here, note that the sorts of extreme number of tied observations in the CAB data, however, are not replicated in the simulations.

Level of Ties	CAB Data	Simulated Data Averages
distinct lifetimes	86	82.50
2-5 ties	30	29.47
>5 ties	6	9.14

Table 4.1: Ties in CAB data and simulated data sets with rounding to nearest day

Table 4.2 summarizes the mean values of the 1000 parameter estimates obtained under each of the three rounding methods. The parameter estimates are very close to the generating model parameters. Surprisingly, even data rounded to the nearest day seem to provide good estimates of parameters.

Parameter	True Value	Maximum Likelihood Estimates		
		Unrounded Data	Data Rounded To Nearest Hour	Data Rounded To Nearest Day
a	9.0	9.1	9.0	8.7
α_1	0.7777	0.79	0.79	0.76
α_2	0.2561	0.26	0.26	0.26
λ	-5.8076	-5.84	-5.84	-5.74

Table 4.2: Mean value of simulation estimates

Table 4.3 summarizes the standard deviations of the parameter estimates of the 1000 data sets. Here again standard errors are quite similar for the three rounding schemes. There is good agreement between the standard deviations presented below and the standard errors presented in Tables 3.3 and 3.4 for the parameters α_1, α_2, and λ. However, this is not the case for the changepoint a where larger standard error

estimates are obtained from the nonparametric bootstrap approaches. In addition, the distribution of $\hat{a}$ based on the parametric bootstrap is closer to normality than that obtained from the non-parametric bootstrap procedures of the previous chapter. However, the parametric simulation discussed here has been somewhat helpful in providing reassurance that rounding does not drastically affect estimators.

	Maximum Likelihood Estimates		
Parameter	Unrounded Data	Data Rounded To Nearest Hour	Data Rounded To Nearest Day
a	1.11	1.14	1.41
α_1	0.08	0.08	0.09
α_2	0.02	0.02	0.02
λ	0.21	0.20	0.21

Table 4.3: Standard deviations of parameter estimates from simulated data sets

Chapter 5

Discussion

In this project, we have proposed a parametric piecewise Weibull model with a single changepoint for analysing CAB data to reflect two distinct outcomes: operative mortality and long-term survival. A nonparametric bootstrap method provides the standard errors of parameter estimates. A simulation study of the effects of rounding of the data on parameter estimation found that even with the rounding of survival times to the nearest day, good estimates can be obtained.

In examining the diagnostic plot presented in Figure 3.1, it seems natural to attempt to locate a changepoint by looking for changes in linear segments which define sharp changes in slope. Visually then it would appear that a changepoint at approximately 30 days meets this criterion. The question is thus raised, in the Weibull changepoint model, how informative is the changepoint in determining important features of the data. It may be that the changes in slope are more important, in which case an approach using linear splines could be considered. In addition, Figure 3.1 seems to suggest a multi-changepoint scenario. Expanding the proposed model to include more than one changepoint, as per the model outlined by Noura and Read (1990), would be useful, especially as preliminary analyses of data from a five-year

followup suggest another changepoint at about 2 years.

As the goal of a typical analysis of lifetime data is not only to model the survivor function but also to investigate the relationship between the response (survival time) and covariates, a natural extension of the work presented in this project is to include covariates into the modelling process. Primarily it is of interest to determine the covariate effects which can predict operative mortality in order to be better able to distinguish those individuals who should pursue a less severe treatment regimen. It is important that covariate effects be allowed to be different over parts of the segments of the survival curve as previous work by Ghahramani et. al (2001), and Chiu (2002), has shown that certain prognostic factors for operative mortality and long-term survival do in fact differ. In their segmented model, Noura and Read (1990) do include covariate effects. However, their formulation assumes that both segments of the survivor function are influenced by the same set of covariates.

Bibliography

[1] Andersen, P.K., Borgun, O., Gill, R.D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

[2] Canty, A.J. (2002). Resampling Methods in R: The Boot Package. *RNews*, Vol 2/3.

[3] Chen, Y.Q., Ronde, C.A., Wang, M.C. (2002). Models with Latent Treatment Effectiveness Lag Time. *Biometrika*, 89(4):917-931.

[4] Chiu, M. (2002). *Nonparametric Simultaneous Modelling of Operative Mortality and Long-Term Survival after Coronary Artery ByPass Surgery*. M.Sc. Project, Simon Fraser University.

[5] Ebrahimi, N. (1991). On estimating Change Point in a Mean Residual Life Function *Sankhya: The Indian Journal of Statistics*, 53(A), 206-219.

[6] Efron, B. (1981) Censored Data and The Bootstrap. *Journal of the American Statistical Association*, 76(374):312-319.

[7] Efron, B., Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

[8] Ghahramani, M. (1998) *Simultaneous Modelling of Long and Short Term Survival after Coronary Artery Bypass Graft Surgery* M.Sc. Project, Simon Fraser University.

[9] Ghahramani, M. Dean, C.B., Spinelli, J.J. (2001), Simultaneous Modelling of Operative Mortality and Long-Term Survival after Coronary Artery Bypass Surgery. *Statistics in Medicine*, 20:1931-1945.

[10] Gijbels, I., Gurler, U. (2003). Estimation of a Change Point in a Hazard Function Based on Censored Data. *Lifetime Data Analysis*, 9, 395-411.

[11] Hjort, N.L. (1985). Bootstrapping Cox's Regression Model. *Technical Report NSF-241*, Department of Statistics, Stanford University.

[12] Kalbfleisch, J.D., Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edn. Wiley, New Jersey.

[13] Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data*, 2nd edn. Wiley, New Jersey.

[14] Levy, A.R., Sobolev, B.G., Hayden, R., et. al. (2005). Time on Wait Lists for Coronary Bypass Surgery in British Columbia Canada, 1991-2000. *BMC Health Services Research*, 5:22.

[15] Liang, K.Y., Self, S.G., Liu, X. (1990). The Cox Proportional Hazards Model with Change Point: An Epidemiologic Application. *Biometrics*, 46,783-793.

[16] Lim, H., Sun, J., Mathews, D.E. (2002). Maximum Likelihood Estimation of a Survival Function with a Change Point for Truncated and Interval-Censored Data. *Statistics in Medicine*, 21:743-752.

[17] Loader, C.R. (1991). Inference for a Hazard Rate Change Point *Biometrika*, 78(4):749-757.

[18] Nguyen, H.T., Rogers, G.S., and Walker, E.A. (1984) Estimation in Change-Point Hazard Rate Models. *Biometrika*, 71(2):299-304.

[19] Noura, A.A., Read, K.L.Q. (1990). Proportional Hazards Changepoint Models in Survival Analysis. *Applied Statistics*, 239, No. 2, 241-253.

[20] Patra, K., Dey, D.K. (2002). A General Class of Change Point and Change Curve Modeling for Life Time Data. *Annals of the Institute of Statistical Mathematics*, 54, No. 3, 517-530.

[21] Ross, S.M. (1997). *Simulation*, 2nd edn. Academic Press, San Diego

[22] Wu, C.Q., Zhao, L.C., Wu, Y.H. (2003). Estimation in Change-Point Hazard Function Models. *Elsevier Science statistics and Probability Letters*, 63, 41-48.