An Extension of Model Selection Curves Framework to Accelerated Failure Time Models

By

Md. Jamil Hasan Karami

A thesis submitted to Macquarie University

for the degree of Doctor of Philosophy

Department of Statistics

March 2017

MACQUARIE
University
SYDNEY·AUSTRALIA

Except where acknowledged in the customary manner, the material presented in this thesis is, to the best of my knowledge, original and has not been submitted in whole or part for a degree in any university.

Md. Jamil Hasan Karami

Acknowledgements

All praises are due to Allah, the only one the Almighty, the most glorified, the most merciful, the most beneficent, the most forgiving ... . Although I am unable to express thankfulness to Allah appropriately, I express my gratitude to Him from the bottom of my heart for all the blessings bestowed upon us, and I cannot but bow down my head unto Him forever.

I would like to express my deepest gratitude to my supervisor Dr kehui Luo for all the support and advices I received during my PhD study at Macquarie University. She has supported me throughout my PhD study and research with her excellent guidance, patience and understanding. I am also grateful to my co-supervisor Dr Thomas Fung who inspired me to work with zeal in the field of model selection problem. Both Kehui and Thomas have expertly guided me through this PhD project. It is my pleasure to mention that I have benefited from their hands-on teaching on many issues. They led me to solve the problems I struggled with often in an elegant manner during my PhD study. I am also thankful to professor Jonathan Carter and Dr Zongqun Ding. They helped me in understanding the ovarian cancer data and related matters. I would like to thank the Lifehouse Gynaecological Oncology Group, Sydney cancer centre, Sydney, Australia for the provision of the data and technical support.

I would also like to thank all of the academic and administrative staff of the Department of Statistics, Macquarie university, for their all out support in completing my PhD study. I have enjoyed the companionship of my fellow PhD colleagues at Macquarie University during this long period of research. I am also thankful to them for their cooperation throughout this time.

I am also grateful to my parents and brother who always inspired me for achieving

higher study. I cannot but recall the sweet memories of my father whom I lost in 2013. May his soul rest in peace. I am also thankful to my parents-in-law and brother-in-law for their never ending well wishes and encouragement.

I am grateful to my beloved wife, Shabnam Sohani, who endured many unwanted and unseen difficulties originated as I was often engrossed in my research longer time of the days and nights. Finally, special thanks go to our little kid, Shanjabir Karami, who keeps us enthusiastic and busy all the time.

Abstract

In most existing model selection criteria, a constant penalty multiplier is usually paired with a penalty function. A model selection criterion based on a single value of penalty multiplier, such as Akaike information criterion (AIC) and Bayesian information criterion (BIC), can be "unstable" as a different model may be selected if the penalty multiplier changes even by a small arbitrary amount. This thesis extends a recently developed model selection approach for (generalised) linear models, known as model selection curves (MSC), to accelerated failure time (AFT) models for survival data. In this approach, penalty multiplier in a predetermined range, instead of a single value, is considered. Model selection criteria based on this approach are thus considered more stable as the selected model is the least likely not to be selected even when the penalty multiplier changes. In addition to the two recently introduced longest cathetus criterion and longest hypotenuse criterion, a new criterion, called the triangle area criterion, is proposed in this thesis. Under some conditions, these three criteria are consistent in selecting a specified AFT model, similar to BIC. It is shown that the consistency result seems to hold even when sample size is only reasonably large using simulations. A model selection framework including these three MSC based criteria, as well as BIC and AIC, is proposed for AFT models of survival data.

The framework was investigated through simulations considering survival data of various sizes and censoring proportions from different specified models. Moreover, the performance of those model selection criteria based on the MSC was examined in comparison to AIC and BIC. The results indicate that those criteria have the potential to outperform AIC and BIC in selecting the correct model. The model selection framework has also been applied to several real world survival data. A tool in `R` program is

developed to visualise the results from applying the framework.

Contents

List of Figures

List of Tables

1

Introduction

1.1 Background

Often in medical studies, information is collected on many variables (known as risk factors, predictors and covariates) that are possibly associated with outcome variables of interest. In many of those studies, one of the aims is to obtain a parsimonious model containing only important variables, which can be used for predicting the outcome and estimating the effect of each covariate in the model. This is to identify a good subset of the covariates to be included in the model for the outcome variable, or equivalently, select a model from all possible models with different subsets of the covariates available in the data, known as model selection. Statistical model selection is an integral and generally challenging part of almost any data analysis (Claeskens and Hjort, 2008). Model selection becomes even more challenging in survival analysis where the outcome variable in the data, survival time, may be censored (i.e., incomplete) for some subjects in the study.

Most existing model selection criteria are based on two components: loss function and penalty term. Loss function measures the goodness of fit of a model, while the complexity of a model is addressed by the penalty term. Several loss functions have been considered and discussed in the literature, including residual sum of squares and log-likelihood function. A penalty term usually consists of a constant multiplier λ and a penalty function $f_n(p_\alpha)$, where p_α is number of parameters in model α and n is sample size. The constant multiplier, also known as penalty multiplier, often determines the properties of a model selection criterion. For a specific penalty function, say $f_n(p_\alpha) = p_\alpha$ that measures the model complexity in its simplest form, we can obtain several well-known model selection criteria if we restrict λ to some single value. For example, when the penalty multiplier $\lambda = 2$, it gives Akaike information criterion (AIC) (Akaike, 1973), and $\lambda = \log(n)$ leads to Bayesian information criterion (BIC) (Schwarz, 1978). Moreover, generalised information criterion (GIC) by Konishi and Kitagawa (1996) allows any single value of $\lambda \in \mathbb{R}$ to be used. These model selection criteria have been well studied in the literature and used extensively in linear and nonlinear regression models.

Model selection for models in survival analysis, on the other hand, has also gained much attention for decades. Model selection procedures for exponential survival models were developed by Krall et al. (1975) and by Greenberg et al. (1974). A stepwise selection procedure for survival models was discussed by Peduzzi et al. (1980). They proposed an algorithm for the stepwise selection procedure for nonlinear regression models. The algorithm was applied to the analysis of survival data using exponential regression model.

Several commonly used model selection criteria, such as BIC and AIC, have been extended to survival analysis. Volinsky and Raftery (2000) extended the BIC to the Cox proportional hazards model (Cox, 1972). They proposed to use the number of uncensored observations (i.e., events) as the penalty term in BIC, instead of the number of all observations in the data studied. Hurvich and Tsai (1989) obtained a bias-corrected version of AIC for nonlinear regression and autoregressive time series models. Based on their study, an improved AIC selection strategy for survival analysis was later suggested by Liang and Zou (2008). They particularly focused on model selection with accelerated failure time (AFT) model for data with a small sample size.

There was also a Bayesian variable selection method for models of survival data with censoring, based on the sufficiency and asymptotic normality of the maximum partial likelihood estimator, which was proposed by Faraggi and Simon (1998). This method is an extension of Lindley's (1968) variable selection criterion for the linear models.

Tibshirani (1997) extended his least absolute shrinkage and selection operator (lasso) technique to the Cox model. Fan and Li (2002) proposed a nonconcave penalised likelihood method for the Cox model and the Cox frailty model (Hougaard, 1995), in which a penalty known as smoothly clipped absolute deviation (SCAD) was used. Furthermore, there is no shortage of model selection studies in the literature, dealing with high dimensional data in survival analysis (e.g., Gui and Li (2005), Ma and Huang (2007), Wang et al. (2008), Huang and Ma (2010) etc.), which we are not focusing in this thesis.

1.2 Motivation

Many of the model selection criteria mentioned in Section 1.1, such as AIC and BIC, are based on only a single fixed value of the penalty multiplier. Recently, Müller and Welsh (2010), and Murray, Heritier and Müller (2013) proposed and studied model selection criteria as a function of penalty multiplier, known as the model selection curves (MSC). Two criteria based on their model selection curves approach were suggested. In the MSC approach, possible models were evaluated and ranked over a range of λ values before selecting a model. The consideration of a range of λ values in a model selection criterion function allows to examine if the model selection procedure (i.e., criterion) is considered stable. Stability can be defined in various ways (Meinshausen and Bühlmann, 2010; Müller and Welsh, 2010). According to Müller and Welsh, a model selection criterion is defined as unstable if a model selected with a particular dimension at a specific λ value is no longer selected due to a small change in λ value. They also showed that in the case of (generalised) linear regression models, criteria based on the model selection curves have the potential to outperform model selection criteria that use only a single value for the penalty multiplier. This has motivated us to consider the MSC approach for models of survival data, such as AFT models.

Here we would like to demonstrate with "`ovarian`" data from `R` package "`survival`" (Therneau, 2015; Therneau and Grambsch, 2000) that some existing model selection criteria such as AIC, based on only single value of penalty multiplier, can be unstable for selecting AFT models. The data contains survival time (outcome variable) with approximately 50% censoring and four covariates. The four covariates are age of patient in years (age), extent of residual disease (resid.ds; 1=incomplete, 2=complete), treatment (rx; 1=single and, 2=combined) and performance status (ecog.ps; 1=good and, 2=poor). The data is given in Appendix D.

Generalised information criterion (GIC), which is a linear combination of loss function and penalty term, has been calculated for each of 15 possible Weibull AFT models fitted to the "`ovarian`" data at three different values of λ (1.5, 2 and 2.5). The resulting GIC values along with the possible models are presented in Table 1.1. For $\lambda = 2$, the GIC is equivalent to AIC.

Table 1.1: GIC values at $\lambda = 1.5$, 2 and 2.5.

Variable(s)	GIC($\lambda = 1.5$)	GIC($\lambda = 2$)	GIC($\lambda = 2.5$)
age	183.00	184.00	185.00
resid.ds	194.76	195.76	196.76
rx	197.73	198.73	199.73
ecog.ps	198.22	199.22	200.22
age+resid.ds	182.54	184.04	185.54
age+rx	182.02	183.52	185.02
age+ecog.ps	184.43	185.93	187.43
resid.ds+rx	194.56	196.06	197.56
resid.ds+ecog.ps	195.36	196.86	198.36
rx+ecog.ps	198.67	200.17	201.67
age+resid.ds+rx	181.74	183.74	185.74
age+resid.ds+ecog.ps	184.03	186.03	188.03
age+rx+ecog.ps	183.49	185.49	187.49
resid.ds+rx+ecog.ps	195.46	197.46	199.46
age+resid.ds+rx+ecog.ps	183.19	185.69	188.19

From Table 1.1, one would choose the model with `age`, `resid.ds` and `rx` when using the model selection criterion at $\lambda = 1.5$ because of its lowest GIC value of 181.74. Now, suppose the value of λ increases by 0.5 to 2 (i.e., AIC), the model with `age` and `rx` (no longer including `resid.ds`) would be selected. If we increase λ further to 2.5, the model selected is different again. The researcher would end up picking a smaller model with `age` only. It is evident that the model selected has changed due to a change in λ

value. This change is considered relatively small as the λ values used in this case can potentially be between 0 and 13. This indicates that model selection criteria that based on a single value of λ, such as AIC, may lack stability. However, choosing a model based on a range of λ values may give some protection against such lack of stability in model selection.

We would also like to investigate how the extent of censoring proportion in survival data affects the performance of some existing model selection procedures, such as AIC and BIC, in identifying the true accelerated failure time (AFT) model of survival data. To do this we followed the footsteps of the studies by Braun (2015) and Tarr et al. (2015) by generating several hypothetical data sets. Similar to these two studies, we considered nine predictors x_1, x_2,...,x_9. The survival time in each of two survival data sets was generated using a single predictor x_8 and based on a Weibull AFT model. Both data sets have a sample size of 50 with 10% and 50% censoring respectively. Note that x_8 is actually a linear function of the rest eight covariates $x_1, x_2, \ldots, x_7$ and x_9 with a small random component added, and the other eight covariates in the model are also correlated.

Table 1.2: AIC and BIC values for Weibull AFT models of survival data.

Variable(s)	10% censoring		50% censoring	
	AIC	BIC	AIC	BIC
x_8	32.56	36.38	51.85	55.67
$x_4 + x_8$	33.49	39.23	49.21	54.94
$x_8 + x_9$	32.39	38.13	51.82	57.56
$x_1 + x_8 + x_9$	32.47	40.11	51.38	59.03
$x_1 + x_3 + x_8 + x_9$	33.08	42.64	52.58	62.14

AIC and BIC values for Weibull AFT models with different combinations of predictors have been computed and reported in Table 1.2. It shows that for data with 10% censoring AIC fails to pick the true model (i.e., model with only x_8), but a bigger model with both x_8 and x_9 having the lowest AIC value (32.39). On the other hand, BIC manages to select the true model. For data with 50% censoring, both AIC and BIC select the model with two predictors of x_4 and x_8, which is not the true model. Similar results were reported in Tarr et al. (2015) for the case of linear models.

This example indicates that censoring proportion may have some impact on the performance of AIC and BIC in model selection for AFT models of survival data. This

needs further investigation.

1.3 Aims and outline of this thesis

The fundamental theme of this thesis is to study several model selection criteria and develop a tool in R program, which can be used in survival analysis, particularly for accelerated failure time (AFT) models. It encompasses a few relatively new, and two commonly used model selection criteria, Akaike information criterion and Bayesian information criterion.

This thesis extends the model selection curves approach to accelerated failure time models for censored survival data. In this approach, model selection criteria are studied as a function of penalty multiplier. This means that penalty multiplier values within a predetermined range, instead of a single fixed value, are considered. Moreover, a candidate model can be assessed for how frequently it is selected since all possible models are studied and ranked over a range of λ values. Model selection via such an approach is therefore more emphasising on the stability of the model selection criterion and the model selected is least likely not to be selected even when the penalty multiplier changes considerably.

A new model selection criterion based on the MSC approach, called the triangle area criterion (TAC), is proposed for AFT model selection, in addition to the two recent criteria by Müller and Welsh (2010), namely the longest cathetus criterion (LCC) and longest hypotenuse criterion (LHC). It is shown in this thesis that, under some conditions, these three related but yet different criteria based on the MSC are all consistent in selecting a specified or true AFT model, similar to BIC.

A model selection framework is proposed for AFT models of survival data, consisting of the three MSC based model selection criteria as well as AIC and BIC. This proposed framework is investigated extensively through a comprehensive simulation study, considering survival data sets generated from different true AFT models with various sizes and censoring proportions. A stratified bootstrapping technique is proposed in this thesis for generating survival data with specified censoring proportion. In particular, the performance of those model selection criteria based on MSC is examined in comparison to AIC and BIC. The AFT model selection framework is also applied

to some published survival data sets, as well as a recent data obtained from a study of survival following ovarian cancer. Moreover, bootstrapping replications are used, when necessary, to provide additional information for and thus improve model selection.

We also intend to develop a user-friendly tool in the statistical computing project R (R Core Team, 2015) for performing our AFT model selection framework and producing tables and/or graphs with relevant results. This encourages the application of the proposed method.

The rest of the thesis is organised as follows. Chapter 2 is a brief review of some basic functions of survival analysis, accelerated failure time model and several typical distributions for modelling survival times. The materials presented in this chapter form a basis for subsequent chapters of this thesis.

Chapter 3 focuses on the methodology. The MSC approach has been discussed thoroughly in this chapter. Based on this approach, a new model selection criterion is devised. A model selection framework is proposed for AFT models.

Chapter 4 investigates the proposed model selection framework with a comprehensive simulation study. For each type of AFT models (Weibull, log-logistic and log-normal), data sets generated from several specified models with different sample sizes and censoring proportions are considered in the simulation study. The performance of the model selection criteria within the framework has been evaluated through Monte Carlo simulation. Bootstrapping technique is also considered for some cases to get more information for model selection.

Chapter 5 introduces the R tool for the AFT model selection framework. Examples based on three published data are used to illustrate the R tool.

Chapter 6 presents a case study in which the framework has been applied to a recently obtained data on patients with ovarian cancer, who were treated in Royal Prince Alfred (RPA) hospital, Sydney.

Chapter 7 gives a conclusion about the model selection framework for the AFT models. Possible directions for future research are also discussed.

2

Accelerated Failure Time Models for Survival Analysis

In this chapter we describe a class of parametric models, known as accelerated failure time (AFT) models that are specifically designed for survival data with censoring. Three typical AFT models, Weibull, log-logistic and log-normal are considered in our study of the model selection framework presented in next chapter, and are discussed here.

2.1 Introduction

Survival analysis refers to any statistical analysis of data where the outcome variable of interest is survival or failure time, i.e., time until an event (e.g., death) occurs. Such data is known as survival data. Survival analysis plays an important role in many fields, particularly medical research. Survival times for some subjects in a follow-up

study may not be fully observed, known as censoring. This makes survival analysis distinctive from other statistical analyses because of its ability to allow for or handle censored survival times. Censoring is often due to lost to follow up, withdrawal from the study or not having experienced the event before the end of the study. This kind of censoring is known as 'right censoring', a most common type of censoring (Hosmer and Lemeshow, 1999). In this case, the time to the last contact time point is used for each censored subject, which is only part of true survival time (time to event), and it is called censored survival time. Thus the true survival time for a right-censored case is always greater than the observed censored survival time. There are also other types of censoring, left censoring and interval censoring (Klein and Moeschberger, 2003), but in this thesis, only survival data with right censoring is considered.

Note that many traditional statistical and graphical procedures may not be appropriate for survival data with censoring as they depend on the data set being fully observed. Kaplan-Meier estimator of the survivor function due to Kaplan and Meier (1958) is a major step in the development of suitable procedures or methods for survival data with censoring. Cox proportional hazards (PH) model, specifically designed for modelling survival data with censoring, has been widely used for survival analysis since its introduction in 1972 (Cox, 1972), where effects of covariates, measured in hazard ratios, can be estimated. Moreover, the Cox model does not make a particular assumption for the distribution of survival time other than assumes proportional hazards. It is considered robust in the sense that it usually fits the data well no matter which parametric model is appropriate for the underlying data (Kleinbaum and Klein, 2012). However, when the proportional hazards assumption is not held by the data, the results from fitting a Cox model can be misleading and may lead to incorrect conclusions. Under some conditions (e.g., survival time follows a particular distribution), AFT model can be used as an alternative to fit the data (Orbe et al, 2002). From fitting an AFT model, the coefficient of each covariate/factor in the model can be estimated, and it gives an interpretation for the effect of the corresponding covariate on the survival time. In addition, an AFT model that assumes the Weibull (or exponential as a special case of Weibull) distribution for the survival time in the data is also a proportional hazards model, and thus gives an interpretation of covariate effect on the risk of the event using hazards ratio, similar to the Cox model.

2.2 Basic functions of survival analysis

Several basic but important functions of survival time in survival analysis, including distribution function $F(t)$, survivor function $S(t)$ and hazard function $h(t)$, are defined and described here. They are often used to characterise the distribution of survival time, denoted by T. We assume that T is a continuous, nonnegative random variable and all functions of T described below are defined over the interval $[0, \infty)$, unless otherwise specified.

Suppose the survival time T has a probability density function (pdf) $f(t)$. Then the distribution function of T is defined as

$$F(t) = \Pr(T \leq t) = \int_0^t f(x)\ dx, \tag{2.1}$$

where t $(t \geq 0)$ is a value of T for an individual. The qth quantile of the distribution of T, denoted by t_q, can be obtained from $F(t_q) = q$, or, $t_q = F^{-1}(q)$.

Survivor function, also known as survival function, is defined as the probability that an individual will survive beyond time t and is expressed as below

$$S(t) = 1 - F(t) = \Pr(T > t) = \int_t^\infty f(x)\ dx. \tag{2.2}$$

The survivor function is a monotonically nonincreasing function, which takes a value 1 at $t = 0$ and a value 0 as $t \to \infty$. The qth quantile of the distribution of T can be obtained from $S(t)$, and it is the smallest t such that $S(t_q) \leq 1 - q$, i.e., $t_q = \inf\{t : S(t) \leq (1 - q)\}$. For example, median survival time $t_{0.5}$ can be obtained by solving $S(t_{0.5}) = 0.5$. Median is considered a more reasonable summary measure than mean for survival time, and thus has often been used in practice. One reason is, as mentioned by Lee and Wang (2003), a small number of individuals with exceptionally long or short lifetimes in a survival data will cause the mean survival time to be disproportionately large or small. Mean is rather sensitive, while median is more robust, to extreme values.

Hazard function evaluates the instantaneous rate of failutre at time t given that an

individual survives up till time t. It is defined as

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t)}{\Pr(T \geq t)\ \Delta t}, \\ &= \lim_{\Delta t \to 0} \Big[\frac{F(t + \Delta t) - F(t)}{\Delta t} \Big] \frac{1}{S(t)} \\ &= \frac{f(t)}{S(t)}, \end{aligned} \tag{2.3}$$

where Δt is a small interval of time, and $\lim_{\Delta t \to 0} \left[\frac{F(t+\Delta t) - F(t)}{\Delta t} \right]$ is the first derivative of $F(t)$ with respect to t, $\frac{d}{dt} F(t)$, which is $f(t)$.

The hazard function $h(t)$ may have different shapes depending on whether the hazard rate is increasing, decreasing, constant over time or a mixture of them. More information about this can be found in Lee and Wang (2003) and Klein and Moeschberger (2003). Note that the hazard function is nonnegative, i.e., $h(t) \geq 0$, and

$$\int_0^\infty h(t)\ dt = \infty.$$

Cumulative hazard function is defined as follows

$$H(t) = \int_0^t h(x)\ dx. \tag{2.4}$$

It can be shown that these basic functions of survival time are related. Using equations (2.1) and (2.2), the pdf $f(t)$ can be expressed in terms of $S(t)$ as below:

$$f(t) = \frac{d}{dt} [1 - S(t)] = -\frac{d}{dt} S(t).$$

The hazard function $h(t)$ in equation (2.3) can be then expressed in terms of $S(t)$ as follows

$$h(t) = \frac{-\frac{d}{dt} S(t)}{S(t)} = -\frac{d}{dt} \log S(t).$$

Since $S(0) = 1$, the cumulative hazard, defined in equation (2.4), can be also expressed as

$$\int_0^t h(x)\ dx = -\log S(t).$$

Then the survivor function $S(t)$, in terms of the hazard function, is given by

$$S(t) = \exp\left(-\int_0^t h(x)\ dx\right) = \exp\left(-H(t)\right).$$

Moreover, the pdf of survival time can be also expressed in terms of hazard function as

$$f(t) = h(t)\exp\left(-\int_0^t h(x)\ dx\right).$$

It is clear that $F(t)$, $f(t)$, $S(t)$, $h(t)$ and $H(t)$ are functionally related. Therefore, based on one of these functions, other function can be obtained (see Lawless, 1982).

2.3 Distribution of survival time

There are several distributions that may be assumed for the survival time under an AFT model. Among them, three commonly used distributions, Weibull, log-logistic and log-normal, are considered in this study. Here we review their basic properties.

2.3.1 Weibull distribution

Weibull distribution was introduced by Weibull (1939). Its applications to lifetime data were illustrated and advocated by Weibull (1951) and Berretoni (1964). Since then, it has been used in many studies including biomedical applications where time to the event of interest is of prime importance. See, for example, Pike (1966), Whittemore and Altshuler (1976).

The pdf of a Weibull distribution is given by

$$f(t) = \nu\kappa(\nu t)^{\kappa-1}\exp[-(\nu t)^{\kappa}],$$

where $\nu > 0$ and $\kappa > 0$ are known as scale and shape parameters of the distribution.

Using the functional relations in equations (2.2) and (2.3), the survivor and hazard function of survival time t under the Weibull distribution can be expressed as

$$S(t) = \exp[-(\nu t)^{\kappa}] \tag{2.5}$$

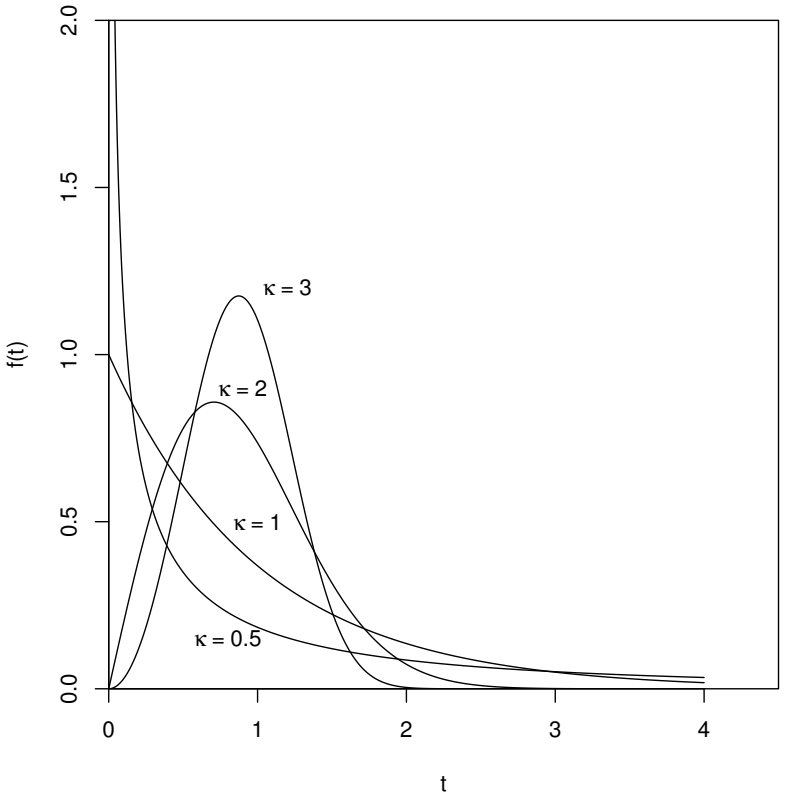


Figure 2.1: Weibull pdf's with $\nu = 1$ and various κ values.

and

$$h(t) = \nu\kappa(\nu t)^{\kappa-1}, \tag{2.6}$$

respectively. It is clear that the hazard function under the Weibull distribution is monotonically increasing if $\kappa > 1$, monotonically decreasing if $\kappa < 1$ and remained a constant if $\kappa = 1$. According to Lawless (1982), the shape parameter value κ typically varies from application to application, but in many practical situations Weibull distributions with shape parameter in the range of 1 to 3 seem appropriate. A slightly wider range of values for the shape parameter will be considered in our simulation study presented in Chapter 4. Figure 2.1 shows the pdf's of four Weibull distributions with different values of the shape parameter κ while the scale parameter ν is fixed at 1. All four distributions are right-skewed and the level of skewness of a Weibull distribution is decreasing with the increasing value in κ, as shown in the figure.

Moments of the distribution of a random variable can be used to numerically describe the variable with respect to its characteristics, such as location and variation

(Rinne, 2008). The rth raw moment, denoted by μ'_r, of a Weibull distribution is

$$\mu'_r = E(X^r) = \nu^{-r}\Gamma\left(1 + \frac{r}{\kappa}\right),$$

where $\Gamma(a) = \int_0^\infty u^{a-1}e^{-u}du$ $(a > 0)$ is a gamma function. The mean and variance of the distribution are $\nu^{-1}\Gamma\left(1 + \frac{1}{\kappa}\right)$ and $\nu^{-2}[\Gamma\left(1 + \frac{2}{\kappa}\right) - \{\Gamma\left(1 + \frac{1}{\kappa}\right)\}^2]$ respectively.

2.3.2 Log-logistic distribution

The mathematical formulation of log-logistic distribution was first studied by Fisk (1961), which is also known as the Fisk distribution in economics. This distribution has been used in survival analysis due to its simple algebraic expressions for different survival functions. See, for example, O'Quigley and Struthers (1982), Bennet (1983), Cox and Oakes (1984). Some important properties of log-logistic random variable in health care studies were discussed by Clark and El-Taha (2015).

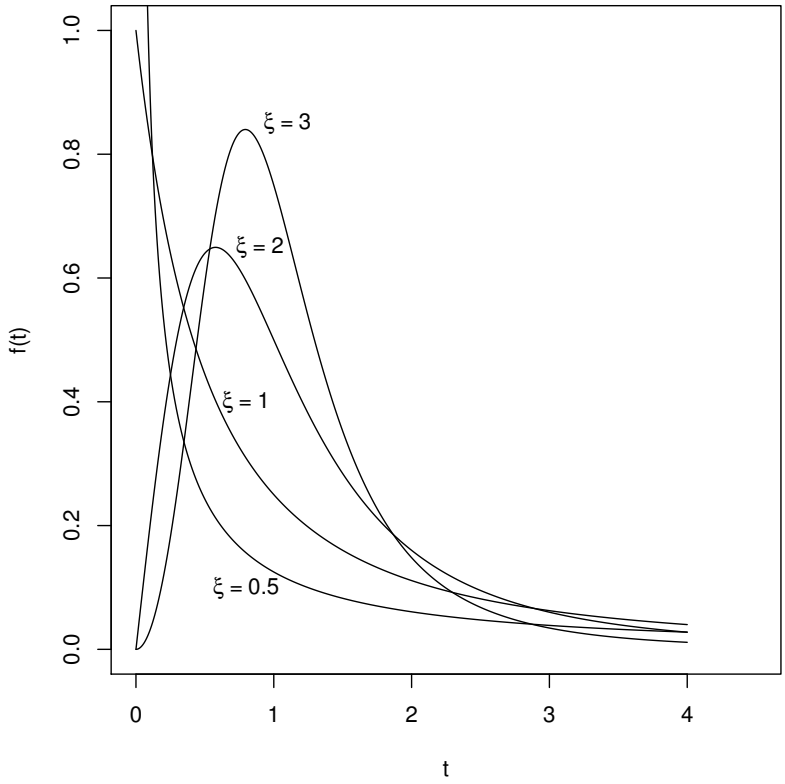


Figure 2.2: Log-logistic pdf's with $\omega = 1$ and various ξ values.

A random variable is said to follow a log-logistic distribution if the logarithm of the random variable follows a logistic distribution. Note that logistic distribution is very similar to normal distribution but has somewhat heavier tails. A log-logistic distribution possesses similar characteristics and is a positively skewed like a log-normal distribution (Aitkin et al., 2009).

The pdf of a log-logistic distribution is

$$f(t) = \frac{\omega\xi t^{\xi-1}}{(1+\omega t^{\xi})^2},$$

which is characterised by two parameters $\omega > 0$ and $\xi > 0$. The pdf's of log-logistic distributions with $\omega = 1$ and four different values of ξ are shown in Figure 2.2. It can be seen in the figure that log-logistic distribution is positively/right skewed.

Using equations (2.2) and (2.3), the survivor and hazard functions of the log-logistic distribution can be expressed as

$$S(t) = \frac{1}{1+\omega t^{\xi}}$$

and

$$h(t) = \frac{\omega\xi t^{\xi-1}}{1+\omega t^{\xi}}$$

respectively. When $\xi > 1$, the hazard function of the log-logistic distribution takes a value 0 at $t = 0$, increases to a maximum at $t = (\frac{\xi-1}{\omega})^{\frac{1}{\xi}}$ and then decreases to 0 as $t \to \infty$. Moreover, when $\xi \leq 1$, the hazard is monotonically decreasing with t.

Variance and other characteristics of a log-logistic distribution can be obtained using its first few moments. The rth raw moment of T about zero, according to Tadikamalla and Johnson (1982), is

$$\mu'_r = \frac{r\pi}{\xi\omega^{r/\xi}} \csc\left(\frac{r\pi}{\xi}\right), \quad r < \xi.$$

The 1st moment is the mean of the log-logistic distribution, given by

$$E(T) = \frac{\pi}{\xi\omega^{1/\xi}} \csc\left(\frac{\pi}{\xi}\right), \quad \xi > 1. \tag{2.7}$$

The 2nd moment is

$$E(T^2) = \frac{2\pi}{\xi\omega^{2/\xi}} \csc\left(\frac{2\pi}{\xi}\right), \quad \xi > 2. \tag{2.8}$$

Therefore, the variance of the log-logistic distribution is given by

$$Var(T) = \frac{2\pi}{\xi\omega^{2/\xi}} \csc\left(\frac{2\pi}{\xi}\right) - \left[\frac{\pi}{\xi\omega^{1/\xi}} \csc\left(\frac{\pi}{\xi}\right)\right]^2, \quad \xi > 2.$$

2.3.3 Log-normal distribution

Log-normal distribution has been widely used in survival analysis since its application in cancer research was discussed in Boag (1949). Nelson and Hahn (1972) used this distribution in the analysis of failure times of electrical insulation, while Whittemore and Altshuler (1976) fitted the log-normal distribution to Doll and Hill's Data for British Physicians.

A random variable is said to have a log-normal distribution if its logarithm is normally distributed. That is, survival time T follows a log-normal distribution if $\log T$ is normally distributed with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. The pdf of the log-normal distribution is therefore,

$$f(t) = \frac{1}{\sigma t \sqrt{2\pi}} \exp\Big[-\frac{1}{2}\Big(\frac{\log t - \mu}{\sigma}\Big)^2\Big], \quad \text{for} \quad t > 0.$$

Note that μ is a location parameter in the normal distribution, and $\exp(\mu)$ is the scale parameter in the log-normal distribution. So change in the value of μ does not change the shape of log-normal distribution, but changes the scale on the horizontal axis. Moreover, the other scale parameter σ^2 determines the shape of the log-normal distribution. The pdf's of four log-normal distributions of survival time t, with four different values of σ (0.25, 1, 1.5 and 3) and $\mu = 0$, are shown in Figure 2.3. Note that as σ increases the level of skewness also increases. The effects of different values of μ and σ on the shape of log-normal distributions were investigated extensively by Lawless (1982) and Lee and Wang (2003).

Using equation (2.2), the survivor function of a log-normal distribution is expressed

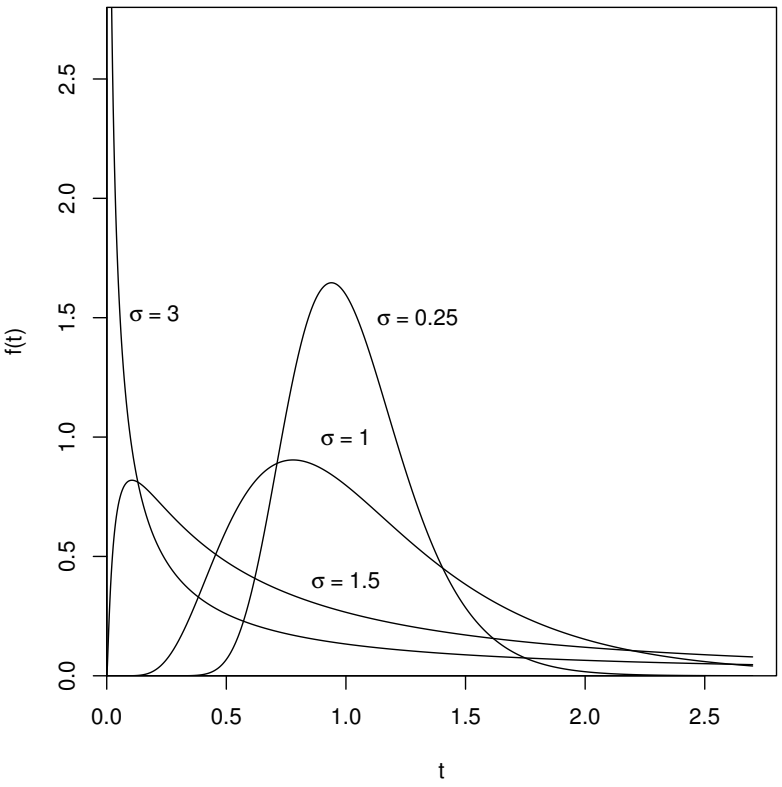


Figure 2.3: Log-normal pdf's with $\mu = 0$ and various σ values.

as

$$S(t) = 1 - \Phi\Big(\frac{\log t - \mu}{\sigma}\Big),$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. Moreover, the hazard function of the log-normal distribution, using equation (2.3), has the following form

$$h(t) = \frac{\frac{1}{\sigma t\sqrt{2\pi}} \exp\Big[-\frac{1}{2}\Big(\frac{\log t-\mu}{\sigma}\Big)^2\Big]}{1 - \Phi\Big(\frac{\log t-\mu}{\sigma}\Big)}.$$

This hazard function takes a value of 0 at $t = 0$, and increases to a maximum before decreases to 0 as $t \to \infty$. Since the hazard function is decreasing for large t, which is not realistic in many practical situations, using log-normal distribution as a survival distribution was criticised by Klein and Moeschberger (2003). However, it may be suitable for studies where large values of T are of no or little interest.

Moments of a log-normal distribution, similar to moments of Weibull and log-logistic distribution, describe the log-normal variable with respect to some characteristics (e.g., mean and variance). The rth raw moment of the log-normal distribution is given by

$$\mu'_r = E(X^r) = e^{r\mu+\frac{1}{2}r^2\sigma^2}.$$

The mean and variance of this distribution are $e^{\mu+\frac{1}{2}\sigma^2}$ and $(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$ respectively.

It is obvious from the mathematical expressions for the moments, log-logistic distribution is the only distribution considered here that may not have finite mean nor variance.

2.4 Accelerated failure time model

In this section, we would like to describe three types of accelerated failure time (AFT) models, Weibull, log-logistic and log-normal, in which the effects of covariates are multiplicative on time scale. Besides reviewing the mathematical background of the AFT models, parameter estimation of those models is also discussed.

2.4.1 The AFT model and underlying distributions

An explicit regression model for the log of survival time $\boldsymbol{T}$ can be written as:

$$\log T = \boldsymbol{\beta}^\top \boldsymbol{x} + W, \tag{2.9}$$

where $\boldsymbol{x}^\top = (1, x_1, x_2, \ldots, x_p)$ represents p covariates, $\boldsymbol{\beta}^\top = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ is a vector of coefficients and W is error term in the model. Here the logged value of $\boldsymbol{T}$ is considered because the distribution of survival time tends to be right skewed. The model in equation (2.9) can also be expressed as

$$T = \exp(\boldsymbol{\beta}^\top \boldsymbol{x})\ U,$$

where $U = \exp(W)$.

Unlike classical (log-) linear regression model where its error term is assumed to

follow a normal distribution, the error term in equation (2.9) can have other distributions. For the model above if we assume that W follows an extreme value distribution, it is equivalent to U follows an exponential distribution. More explicitly, the density function of U is

$$f(u) = \nu \exp(-\nu u);\ u \geq 0, \nu > 0.$$

Since $u = \exp(w)$ and $du = \exp(w)dw$, the density function of W is

$$f(w) = \nu \exp\left(w - \nu \exp(w)\right), \quad w \in \mathbb{R},\ \nu > 0.$$

We can write

$$W = \log T - \boldsymbol{\beta}^\top \boldsymbol{x}.$$

It implies that the differences between the observed $\log T$ and its fitted (predicted) values ($\boldsymbol{\beta}^\top \boldsymbol{x}$), i.e., residuals, follow an extreme value distribution.

The error term W can have different distributions. This has opened up opportunities for a wide variety of models to be studied. Therefore, for a given distribution of W, a different type of AFT models is considered for survival data.

A general form of an AFT model is usually expressed as

$$\begin{aligned} Y &= \log T = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \tau W \\ &= \boldsymbol{\beta}^\top \boldsymbol{x} + \tau W, \end{aligned} \tag{2.10}$$

where $\boldsymbol{x}$ and $\boldsymbol{\beta}$ are defined as before, $\tau > 0$ is a scale parameter, and W is the random error term.This model can also be expressed as

$$T = \exp(\boldsymbol{\beta}^\top \boldsymbol{x})\ \exp(\tau W).$$

The expression above is often interpreted as accelerated effect on survival time as the effect of covariate x is multiplicative on the time scale (Hosmer and Lemeshow, 1999). For this very reason, this model is called accelerated failure time model.

Here we outline the probability density function and survivor function of T under the AFT regression model as they play important roles for making inferences about the model parameters. Let $g_0(w)$ and $G_0(w)$ be the probability density function and

survivor function of W respectively. Then the survivor function of t ($t \in T$) given $\boldsymbol{x}$ can be expressed as

$$\begin{aligned} S(t;\boldsymbol{x}) &= \Pr(T \geq t) = \Pr(Y \geq \log t) \\ &= \Pr\left(w \geq \frac{\log t - \boldsymbol{\beta}^\top \boldsymbol{x}}{\tau}\right) \\ &= G_0\left(\frac{\log t - \boldsymbol{\beta}^\top \boldsymbol{x}}{\tau}\right). \end{aligned} \tag{2.11}$$

Therefore, the corresponding probability density function of t is

$$\begin{aligned} f(t;\boldsymbol{x}) &= -\frac{d}{dt} G_0\left(\frac{\log t - \boldsymbol{\beta}^\top \boldsymbol{x}}{\tau}\right) \\ &= -\frac{1}{t\tau}\, G_0'\left(\frac{\log t - \boldsymbol{\beta}^\top \boldsymbol{x}}{\tau}\right) \\ &= \frac{1}{t\tau}\, g_0\left(\frac{\log t - \boldsymbol{\beta}^\top \boldsymbol{x}}{\tau}\right). \end{aligned} \tag{2.12}$$

As mentioned above, there are various distributions that can be assumed for an AFT regression model, although here we only consider Weibull, log-logistic and log-normal distributions. Under the Weibull AFT model, the random variable W has the following density function and survivor function,

$$g_0(w) = \exp(w - \exp(w)),\ w \in \mathbb{R} \tag{2.13}$$

and

$$G_0(w) = \exp(-\exp(w)) \tag{2.14}$$

respectively. Then the Weibull AFT regression model has the following survivor and density functions of t

$$S(t;\boldsymbol{x}) = \exp\left[-\exp\left(\frac{\log t - \boldsymbol{\beta}^\top \boldsymbol{x}}{\tau}\right)\right]$$

and

$$f(t;\boldsymbol{x}) = \frac{1}{t\tau}\exp\left[\left(\frac{\log t - \boldsymbol{\beta}^\top \boldsymbol{x}}{\tau}\right) - \exp\left(\frac{\log t - \boldsymbol{\beta}^\top \boldsymbol{x}}{\tau}\right)\right]$$

respectively. Note that for a Weibull AFT model, the parameters of the Weibull distribution can be expressed using $\boldsymbol{\beta}$ and τ in the following way, as shown in Lee and Wang (2003):

$$\nu = \exp\left(-\frac{\boldsymbol{\beta}^\top \boldsymbol{x}}{\tau}\right) \quad \text{and} \quad \kappa = \frac{1}{\tau}. \tag{2.15}$$

Therefore, the hazard and survivor functions under the Weibull AFT model can be expressed in terms of covariates via ν. Moreover, once the parameters of a Weibull regression model, $\boldsymbol{\beta}$ and τ, are specified, the distribution of W is completely known. Similar to Weibull AFT model, the parameters of a distribution under other AFT models, such as log-logistic and log-normal, can be also expressed in terms of $\boldsymbol{\beta}$ and τ, and thus its hazard and survivor functions are also a function of covariates.

The random variable W under the log-logistic AFT model has density and survivor function

$$g_0(w) = \frac{\exp(w)}{[1+\exp(w)]^2} \tag{2.16}$$

and

$$G_0(w) = \frac{1}{1+\exp(w)} \tag{2.17}$$

respectively. It follows that we have the following survivor and density functions

$$S(t;\boldsymbol{x}) = \left[1+\exp\left(\frac{\log t - \boldsymbol{\beta}^\top \boldsymbol{x}}{\tau}\right)\right]^{-1}$$

and

$$f(t;\boldsymbol{x}) = \frac{1}{t\tau}\exp\left(\frac{\log t - \boldsymbol{\beta}^\top \boldsymbol{x}}{\tau}\right)\left[1+\exp\left(\frac{\log t - \boldsymbol{\beta}^\top \boldsymbol{x}}{\tau}\right)\right]^{-2}$$

for the log-logistic AFT model respectively.

The random variable W under the log-normal AFT model has the following density function

$$g_0(w) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}w^2\right). \tag{2.18}$$

Then its survivor function is

$$G_0(w) = 1 - \Phi(w), \tag{2.19}$$

where $\Phi(w)$ is the cumulative distribution function of W. For the log-normal AFT regression model, the survivor and density functions are

$$S(t;\boldsymbol{x}) = 1 - \Phi\left(\frac{\log t - \boldsymbol{\beta}^\top \boldsymbol{x}}{\tau}\right)$$

and

$$f(t;\boldsymbol{x}) = \frac{1}{t\tau\sqrt{2\pi}} \exp\left(-\frac{(\log t - \boldsymbol{\beta}^\top \boldsymbol{x})^2}{2\tau^2}\right)$$

respectively.

The probability density function and survivor function of $\boldsymbol{T}$ under the AFT regression model are important for parameter estimation via maximum likelihood. The likelihood function is obtained using the pdf's and survivor functions described above in such a way that it becomes a function of all the relevant parameters under the AFT model considered.

2.4.2 Inference about model parameters

The parameters in an AFT model can be estimated based on a likelihood function. However, constructing the likelihood function for survival data with censoring need to be handled differently for complete and censored observations in the data.

Suppose we have n observations $t_1, t_2, \ldots, t_n$, which may come from same or different distributions and all of them have complete information. Let $\boldsymbol{\theta} = \left(\theta_1, \theta_2, \ldots, \theta_p\right)^\top$ denote the vector of unknown parameters associated with the distribution(s) of $t_1, t_2, \ldots, t_n$. The likelihood function of $\boldsymbol{\theta}$, $L(\boldsymbol{\theta})$, based on a set of observed data is the probability of observing the data given $\boldsymbol{\theta}$, i.e.,

$$L(\boldsymbol{\theta}) = \Pr(data;\ \boldsymbol{\theta}).$$

Assuming that all the observations are independent, the log-likelihood function can be

then written in the following form

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log L_i(\boldsymbol{\theta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}),$$

where $L_i(\boldsymbol{\theta})$ is the probability density function for the continuous data or the probability mass function for the discrete data.

In a survival data, we may have censored observations. This means we only know incomplete or partial information of survival time for some subjects in the data, which is a special feature of survival data. Such censoring issue needs to be taken into consideration in the construction of a likelihood function in survival analysis.

Let us consider data $\{(y_i, \delta_i), i = 1, 2, \ldots, n\}$ containing right censoring. Here

$$y_i = \min(T_i, C_i), \ \delta_i = I(T_i \leq C_i), \tag{2.20}$$

where the survival time of ith subject is T_i if fully observed (i.e., $\delta_i = 1$) or C_i if censored (i.e., $\delta_i = 0$), and δ_i is the censoring indicator. Assume that censored survival time C_i is fixed. In this case $y_i \in (0, C_i)$. When $\delta_i = 0$ then $y_i = C_i$ and their joint distribution is

$$\begin{aligned} \Pr(y_i = C_i, \delta_i = 0) &= \Pr(T_i > C_i) \\ &= S(C_i). \end{aligned}$$

The cumulative distribution function of y_i jointly with $\delta_i = 1$ is

$$\begin{aligned} \Pr(y_i \leq y, \delta_i = 1) &= \Pr(T_i \leq y) \\ &= F(y), \ y \leq C_i. \end{aligned}$$

So the density function of y_i jointly with $\delta_i = 1$ is

$$\frac{d}{dy} \Pr(y_i \leq y, \delta_i = 1) = f(y).$$

Note that uncensored observations ($\delta_i = 1$) give information on both the hazard of the event and the survival of individuals prior to that event, while censored observations

($\delta_i = 0$) only give information on the survival of individuals no further than C_i. Thus uncensored observations contribute to the likelihood function via their density function, but censored observations contribute to it through their survivor function. It follows that the likelihood for ith observation can be expressed as

$$\begin{aligned} L_i &= f(y_i)^{\delta_i} S(C_i)^{1-\delta_i} \\ &= f(y_i)^{\delta_i} S(y_i)^{1-\delta_i}, \end{aligned}$$

as $y_i = C_i$ when $\delta_i = 0$. If the pairs (y_i, δ_i) are independent, the likelihood function of the whole data of size n is given by

$$L = \prod_{i=1}^{n} f(y_i)^{\delta_i} S(y_i)^{1-\delta_i}. \tag{2.21}$$

It is natural to choose likelihood approach for the parameter estimation and inference for AFT regression models. Consider a sample of n independent subjects with p explanatory variables $\boldsymbol{x}_i$, $i = 1,2,\ldots,n$. Following equation (2.21), its log-likelihood function is

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \Big(\delta_i \log f(y_i; \boldsymbol{x}_i) + (1-\delta_i) \log S(y_i; \boldsymbol{x}_i)\Big),$$

where $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top, \tau\right)^\top$ and $\boldsymbol{x}_i$ ($i = 1,2,\ldots,n$) is the vector of values on the p covariates for ith subject. In terms of $g_0(w)$ and $G_0(w)$, using equations (2.11) and (2.12), the log-likelihood function then takes the following form

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \left(\delta_i \log\left(\frac{1}{y_i \tau} g_0(w_i)\right) + (1-\delta_i) \log G_0(w_i)\right) \\ &= -\log \tau \sum_{i=1}^{n} \delta_i + \sum_{i=1}^{n} \Big(\delta_i \log g_0(w_i) + (1-\delta_i) \log G_0(w_i)\Big) - \sum_{i=1}^{n} \delta_i \log y_i, \end{aligned} \tag{2.22}$$

where

$$w_i = \frac{\log y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i}{\tau}, \quad i = 1,2,\ldots,n.$$

Now we can find the score vectors by considering the derivative of the log-likelihood

function above. We have

$$\frac{\partial w_i}{\partial \beta_j} = -\frac{1}{\tau} x_{ij} \quad \text{for} \quad j = 0, 1, 2, \ldots, p$$

and

$$\frac{\partial w_i}{\partial \tau} = -\frac{1}{\tau} w_i.$$

Note that $x_{ij} = 1$ when $j = 0 \; \forall \; i$. Therefore, the score vector $U(\boldsymbol{\theta})$ has the following components

$$U_j(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \beta_j} = -\frac{1}{\tau} \sum_{i=1}^{n} b_i x_{ij}, \quad j = 0, 1, 2, \ldots, p,$$

$$U_{p+1}(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \tau} = -\frac{1}{\tau} \sum_{i=1}^{n} (\delta_i + b_i w_i),$$

where

$$b_i = \frac{d}{dw_i} \Big(\delta_i \log g_0(w_i) + (1 - \delta_i) \log G_0(w_i) \Big), \quad i = 1, 2, \ldots, n.$$

An iterative procedure such as Newton-Raphson method can be used to solve the equation $U(\boldsymbol{\theta}) = \mathbf{0}$, and thus the maximum likelihood estimates (MLE) of the parameters of interest $(\boldsymbol{\theta})$ can be obtained for an AFT regression model.

The observed Fisher information matrix, denoted by $\boldsymbol{I}_0$, can be determined by an evaluation of the following derivatives:

$$-\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \beta_j \partial \beta_k} = -\frac{1}{\tau^2} \sum_{i=1}^{n} B_i x_{ij} x_{ik}, \quad j, k = 0, 1, 2, \ldots, p;$$

$$-\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \beta_j \partial \tau} = -\frac{1}{\tau^2} \sum_{i=1}^{n} (b_i + B_i w_i) x_{ij} = \frac{1}{\tau} U_j(\boldsymbol{\theta}) - \frac{1}{\tau^2} \sum_{i=1}^{n} B_i w_i x_{ij};$$

$$-\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \tau^2} = \frac{2}{\tau} U_{p+1}(\boldsymbol{\theta}) + \frac{1}{\tau^2} \sum_{i=1}^{n} (\delta_i - B_i w_i^2);$$

where

$$B_i = \frac{\partial b_i}{\partial w_i} = \frac{\partial^2}{\partial w_i^2} \Big(\delta_i \log g_0(w_i) + (1 - \delta_i) \log G_0(w_i) \Big).$$

Since the MLEs satisfy $U(\boldsymbol{\theta}) = \mathbf{0}$, the expression of the observed information matrix

becomes a bit simpler. The observed information matrix is

$$\boldsymbol{I}_0 = \begin{pmatrix} -\frac{1}{\tau^2}\sum_{i=1}^{n} B_i & -\frac{1}{\tau^2}\sum_{i=1}^{n} B_i x_{ik} & -\frac{1}{\tau^2}\sum_{i=1}^{n} B_i w_i \\ & -\frac{1}{\tau^2}\sum_{i=1}^{n} B_i x_{ij}^2 & -\frac{1}{\tau^2}\sum_{i=1}^{n} B_i w_i x_{ij} \\ & & \frac{1}{\tau^2}\sum_{i=1}^{n}(\delta_i - B_i w_i^2) \end{pmatrix}.$$

Now using the entries from the information matrix above, standard errors (SE) of estimates and confidence intervals (CI) of relevant parameters can be obtained.

In next chapter, the methodology devised and used in this thesis is presented and explained.

3

Model Selection Framework

In this chapter, a recently developed model selection approach by Müller and Welsh (2010), and also by Murray, Heritier and Müller (2013) for (generalised) linear regression models, known as model selection curves (MSC), is extended to AFT models for censored survival data. In addition to two existing criteria based on the MSC, a new criterion is derived and proposed for AFT model selection. Utilising these MSC based model selection criteria, a model selection framework is proposed for the AFT models.

Almost all commonly known model selection approaches are based on a linear combination of loss function and penalty term. Often an expression in the form "Loss function + Penalty term" is minimised over a set of models for the construction of a model selection criterion. These two components for AFT models are considered in Section 3.1.

3.1 Loss function and penalty term for AFT model

For many models, including accelerated failure time model in survival analysis, log-likelihood based loss function is often a choice, and it has been used to compare models ever since the original AIC introduced by Akaike (1973). Hurvich and Tsai (1989) also used log-likelihood as part of their derivation of a bias-corrected version of AIC for nonlinear regression and autoregressive time series models. Later, Liang and Zou (2008) used this bias-corrected version of AIC to compare AFT models for survival analysis. Murray et al. (2012) developed a number of graphical tools for model selection in generalised linear models, where a log-likelihood based loss function was also used. Müller and Welsh (2010), however, used residual sum of squares as the loss function when illustrating model selection curves approach for linear regression model. In this case, it can be shown that the residual sum of squares is proportional to the log-likelihood function and thus using residual sum of squares based loss function is equivalent to using log-likelihood based loss function. In our study of model selection framework for AFT models, a log-likelihood based loss function is chosen.

Loss functions alone may be used in the evaluation of model performance. However, comparing the values of loss function only may not be sufficient for model comparison. For example, if only maximisation of log-likelihood (l) is considered, incorporating more parameters in a model will almost result in a larger value of l, and thus model with the largest number of parameters will always be chosen. This does not seem to be consistent with the idea of model selection to obtain a parsimonious and optimal model. Therefore, an extra term that can penalise models with respect to the number of parameters, known as penalty term, is considered in many existing model selection approaches for comparing models. There are also other forms of penalty term. For example, the lasso penalty, $f_n(\beta) = \sum_{k=1}^{p} |\beta_k|$ and the ridge penalty $f_n(\beta) = \sum_{k=1}^{p} \beta_k^2$, where β is a vector of parameters and p is the number of the parameters in a model.

In our investigation of model selection framework for AFT models, a penalty term in the form of $\lambda_n f_n(p_\alpha)$ is considered. Here, λ_n, known as penalty multiplier, is a non-stochastic sequence. A range between 0 and $4\log(n)$ for λ_n, i.e., $\lambda_n \in [0,\ 4\log(n)]$, is considered in our study. Though the upper bound of this interval is arbitrary, it has been chosen in such a way that values of penalty multipliers in many existing model selection criteria are covered by the interval. The other component in the penalty term,

$f_n(\cdot)$, known as penalty function, is a non-stochastic sequence of functions of number of parameters (i.e., p_α) in model α. Different forms of $f_n(\cdot)$, such as $f_n(p) = p$ and $f_n(p) = \frac{p+1}{n-p-2}$ have been investigated in the literature. We have chosen $f_n(p_\alpha) = p_\alpha$, same as Müller and Welsh's study. In the formulation of model selection framework in this thesis, generalised information criterion that is expressed as a function of the penalty multiplier λ_n is considered.

3.2 A brief introduction to model selection curves

Suppose we wish to choose one model from the set of all possible models $\mathcal{A}$. To do this, an expression of generalised information criterion (GIC) for model α is expressed as follows:

$$M(\lambda;\alpha) = -2l + \lambda p_\alpha,\ \alpha \in \mathcal{A}, \tag{3.1}$$

where l is log-likelihood function. Note that residual sum of squares was used in Müller and Welsh's paper (2010) where model selection curves approach was applied to classical linear regression model. For each specified $\lambda > 0$, a model is chosen by minimising $M(\lambda;\alpha)$ over $\alpha \in \mathcal{A}$ in equation (3.1). The function $M(\lambda;\alpha)$ is computed for each $\alpha \in \mathcal{A}$ at each λ over a range from 0 to $4\log(n)$. The models are then ranked at each λ in an increasing order of $M(\lambda;\alpha)$ values.

Let us define a rank function as below:

$$r(\lambda;\alpha) = rank(M(\lambda;\alpha)). \tag{3.2}$$

Note that rank functions are step functions, pairs of which have jumps at the values of λ where the ranks of models change. Now, assuming no ties in $M(\lambda;\alpha)$'s for $\alpha \in \mathcal{A}$, the k rank model selection curve can be defined by

$$\gamma_{(k)}(\lambda;\mathcal{A}) = \max\Big(M(\lambda;\alpha);\ \alpha \in \mathcal{A} \wedge r(\lambda;\alpha) \leq k\Big),$$

where $1 \leq k \leq m$, and m is the number of models in $\mathcal{A}$. This definition can be extended to $M(\lambda;\alpha)$'s with ties for $\alpha \in \mathcal{A}$ by considering continuous locus of $\gamma_{(k)}(\lambda;\mathcal{A})$ at each $\lambda > 0$. The 1 rank ($k = 1$) model selection curve is then the lower enveloping curve

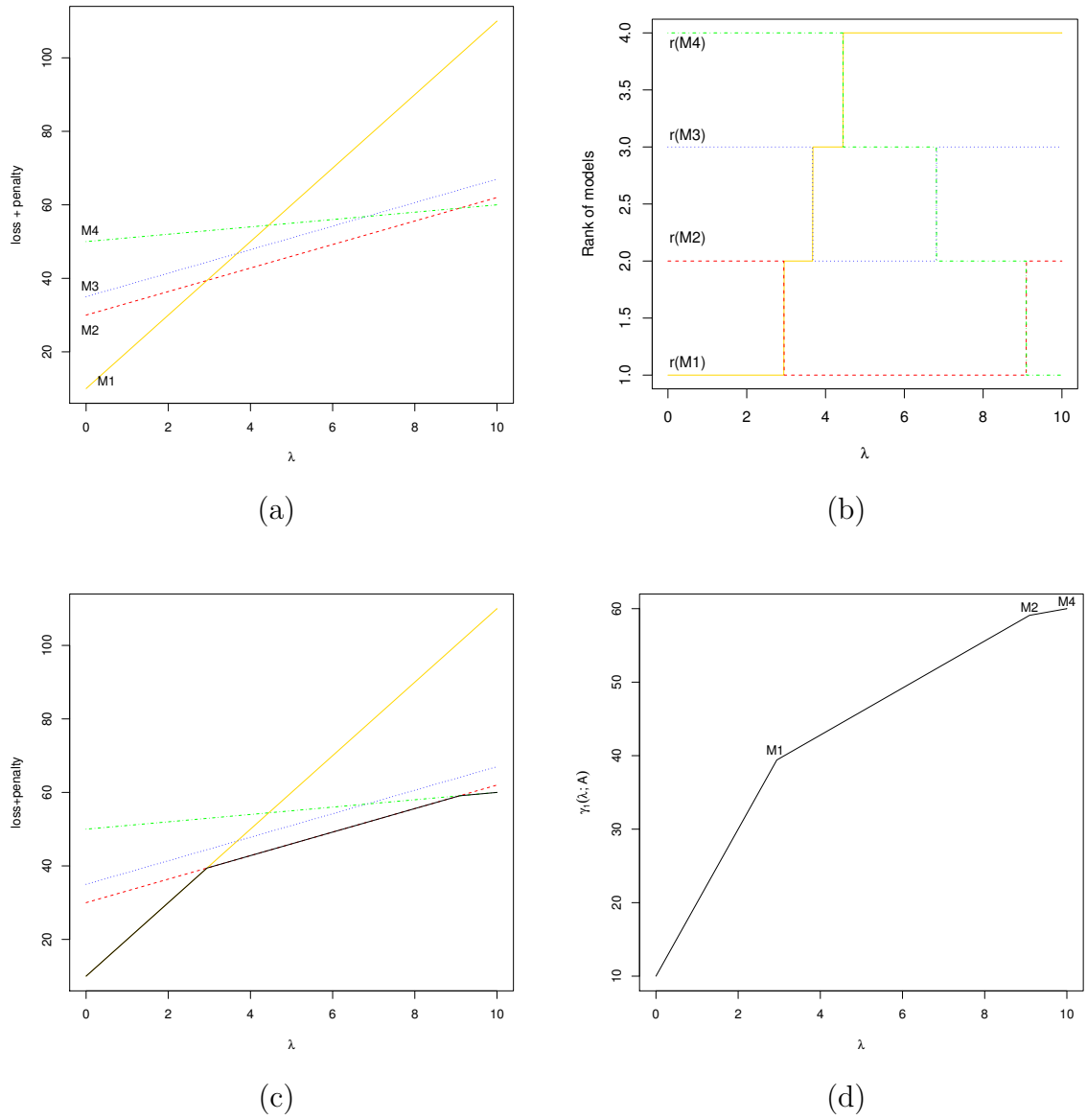


Figure 3.1: An illustration of the construction of model selection curves: (a) plot of $M(\lambda;\alpha)$ against λ; (b) plot of $r(\lambda;\alpha)$ against λ; (c) lower enveloping curve; (d) the 1 rank model selection curve.

defined by the expression below:

$$\gamma_{(1)}(\lambda;\mathcal{A}) = \min\left(M(\lambda;\alpha);\ \alpha \in \mathcal{A}\right). \tag{3.3}$$

Let us illustrate how to construct model selection curves with a simple example where only four models ($m = 4$) are considered and denoted by $M1$ to $M4$ respectively. The $M(\lambda;\alpha)$'s for all these four models are computed at each λ as in equation (3.1)

over $\lambda \in [0,\ 4\log(n)]$, and then plotted against λ as shown in Figure 3.1(a). The four models are ranked at each λ and then the resulting ranks of the four models ($r(\lambda,\alpha)$ in equation (3.2)) are plotted against λ. This plot is known as rank plot and shown in Figure 3.1(b). Note that $M1$, $M2$ and $M4$ have achieved rank 1 over low, middle and high λ values respectively and each corresponds to one of the three sections of the lower enveloping curve presented in bold line on Figure 3.1(c). This lower enveloping curve as shown in Figure 3.1(d) is known as model selection curve. It is the nucleus of all the MSC based model selection criteria considered in this thesis. We will return to this hypothetical example later.

3.3 Model selection criteria based on the MSC

Two model selection criteria based on the MSC approach, longest cathetus criterion and longest hypotenuse criterion, were suggested, and the former was also applied to linear models in Müller and Welsh (2010). The longest cathetus criterion was investigated for AFT models in survival analysis by Karami, Luo and Fung (2015). In this thesis, a new model selection criterion based on the MSC approach is proposed. It is named triangle area criterion. To describe and illustrate each of these three model selection criteria, let us consider another hypothetical example. Figure 3.2 shows an artificially constructed 1 rank model selection curve according to equation (3.3). The lower enveloping curve above the truncated polygons shown in the figure is often referred as model selection curve. There are three models appeared on this 1 rank model selection curve, the full model α_f and submodels of α_1 and α_2.

3.3.1 The longest cathetus criterion

The longest cathetus criterion (LCC) was first studied based on the MSC by Müller and Welsh (2010) for linear regression model. The basics of the LCC is given below.

The 1 rank model selection curve in Figure 3.2 may be regarded as a convex polygon with the maximum number of knot points $\mathcal{N}(p_\alpha) - 1$, where $\mathcal{N}(p_\alpha)$ is the number of distinct values of p_α for $\alpha \in \mathcal{A}$. The other model selection curves (rank k; $k \neq 1$) might be considered as piecewise convex polygon, at least on the consecutive points of $\gamma_{(i)} \cap \gamma_{(j)}$ for $i \neq j$ and $\gamma_{(i)} \cap \gamma_{(j)} \neq \varnothing$, an empty set. Therefore, they are all bounded

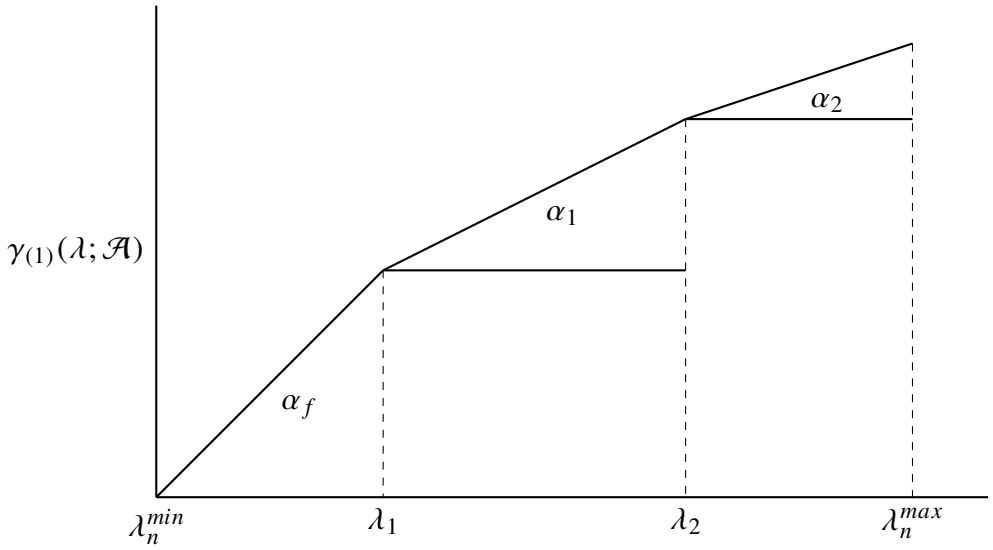


Figure 3.2: The truncated polygon on the 1 rank model selection curve.

from below by $\gamma_{(1)}(\lambda;\mathcal{A})$, the 1 rank model selection curve.

Model α appears on the 1 rank model selection curve if $r(\lambda;\alpha) = 1$ for some λ values. The model being selected based on this criterion may be obtained by minimising

$$\int \psi_\alpha(r(\lambda;\alpha))d\Delta_n, \tag{3.4}$$

where $\psi_\alpha(x) = 1 - \mathbf{1}\{x = 1\}$ so that $\mathbf{1}\{x = 1\}$ takes value 1 if $x = 1$, and 0 otherwise, and Δ_n is the Dirac measure that puts mass 1 on the point $\lambda = \lambda_n$, a specific value of λ. The penalty multiplier λ can be used instead of Dirac measure. This leads to a consideration of using a measure that might contain a range of λ values $[\lambda_n^{min}, \lambda_n^{max}]$ at the same time. In this study, we consider uniform distribution on an interval $[0, 4\log(n)]$. Minimising the expression (3.4) means that model α achieves rank 1 over the largest range of λ. Geometrically this measure (i.e., using uniform distribution over the λ interval) would select a model α, for which the length of its cathetus is the longest. That is why, this criterion is called longest cathetus criterion. Cathetus here is defined as the horizontal edge of a right-angled triangle (truncated polygon as shown in Figure 3.2) that has its hypotenuse being a segment on 1 rank model selection curve $\gamma_{(1)}(\lambda;\mathcal{A})$. Note that the cathetus length of model α, denoted by C_{L_α}, is determined by the difference between the $x-$coordinates of the upper and lower end points of the cathetus.

As shown in Figure 3.2, α_f achieves rank 1 from λ_n^{min} and λ_1, α_1 from λ_1 and λ_2, α_2 from λ_2 and λ_n^{max}. The cathetus lengths corresponding to models α_f, α_1 and α_2 are

$(\lambda_1 - \lambda_n^{min})$, $(\lambda_2 - \lambda_1)$ and $(\lambda_n^{max} - \lambda_2)$, respectively. Note that the length of cathetus of a model indicates the stability of that model, i.e., how long the model preserves its 1 rank position throughout the entire range of the penalty multiplier.

3.3.2 The longest hypotenuse criterion

Besides the cathetus, the hypotenuse of the triangle can also be utilised as a model selection criterion. A model may be selected on the basis of having the longest hypotenuse, which is the longest edge of all truncated polygon on the 1 rank model selection curve where catheti are obtained (Müller and Welsh, 2010). This criterion is referred to as the longest hypotenuse criterion (LHC). Mathematically, a model α with the longest hypotenuse is selected by minimising the following expression:

$$\int_{\lambda_n^{min}}^{\lambda_n^{max}} \sqrt{1 + p_{\alpha_f}^2} - \sqrt{1 + p_{\alpha}^2} \cdot \mathbf{1}\{r(\lambda; \alpha) = 1\}\, d\lambda, \tag{3.5}$$

where p_{α_f} is the column rank (number of predictors + 1) under full model α_f.

In order to minimise expression (3.5) we have to evaluate the integral for all models appearing on the 1 rank model selection curve. As an example, suppose there are only three models, α_f, α_1 and α_2 that achieve rank 1 over the interval $[\lambda_n^{min},\ \lambda_n^{max}\ ]$ as shown in Figure 3.2. For the full model α_f, the expression (3.5) can be written as

$$(\lambda_n^{max} - \lambda_n^{min}) \sqrt{1 + p_{\alpha_f}^2} - (\lambda_1 - \lambda_n^{min}) \sqrt{1 + p_{\alpha_f}^2}, \tag{3.6}$$

where λ_1 is the abscissa of the point at which legs of the right-angled triangle associated with model α_f intersect. Clearly, $(\lambda_1 - \lambda_n^{min})$ is the length of cathetus and $(\lambda_1 - \lambda_n^{min}) \sqrt{1 + p_{\alpha_f}^2}$ is the length of hypotenuse of the right-angled triangle.

For the model α_1, the legs of the right-angled triangle intersect at a point whose abscissa is λ_2. Then the expression (3.5) becomes

$$(\lambda_n^{max} - \lambda_n^{min}) \sqrt{1 + p_{\alpha_f}^2} - (\lambda_2 - \lambda_1) \sqrt{1 + p_{\alpha_1}^2}. \tag{3.7}$$

Here $(\lambda_2 - \lambda_1) \sqrt{1 + p_{\alpha_1}^2}$ is the length of hypotenuse of the right-angled triangle, and $(\lambda_2 - \lambda_1)$ is the corresponding cathetus length.

For the model α_2 that achieves rank 1, the expression (3.5) integrates to

$$(\lambda_n^{max} - \lambda_n^{min}) \sqrt{1 + p_{\alpha_f}^2} - (\lambda_n^{max} - \lambda_2) \sqrt{1 + p_{\alpha_2}^2}. \tag{3.8}$$

In this case, the length of hypotenuse is $(\lambda_n^{max} - \lambda_2) \sqrt{1 + p_{\alpha_2}^2}$, and $(\lambda_n^{max} - \lambda_2)$ is the corresponding cathetus length. Hypotenuse lengths for models in other scenarios on the 1 rank model selection curve can be determined similarly. Note that if the full model α_f is the only model that achieves rank 1, the expression (3.5) goes down to its minimum value zero. For a model never achieving rank 1, the expression (3.5) reduces to its maximum value $(\lambda_n^{max} - \lambda_n^{min}) \sqrt{1 + p_{\alpha_f}^2}$.

We can see that first terms in expressions (3.6), (3.7) and (3.8) are exactly the same, but second terms vary. So minimisation of expression (3.5) can be done by maximisation of these second terms because of its negative sign. Note that each second term in expressions (3.6), (3.7) and (3.8) corresponds to the hypotenuse length of the right-angled triangle produced by the 1 rank model α_f, α_1 and α_2 respectively. Clearly, model with the longest hypotenuse has the lowest value in expression (3.5) and thus may be selected accordingly.

There are some philosophical differences between models selected by the longest hypotenuse criterion and the longest cathetus criterion, as the minimisation of expression (3.5) for the longest hypotenuse selection criterion generally favours a larger model than the longest cathetus criterion. It can be seen that larger p_α or slope of generalised information criterion, leads to a larger angle or steeper hypotenuse for a right-angled triangle. However, a right-angled triangle with a steeper hypotenuse or even the longest hypotenuse does not necessarily mean that the cathetus is the longest among a set of right-angled triangles. On the other hand, the right-angled triangle with the longest cathetus must have a relatively long hypotenuse. This hypotenuse is very likely to be the longest among right-angled triangles of all models having ever achieved rank 1 over the range of λ. When the longest hypotenuse and cathetus criteria select different models, the longest hypotenuse criterion usually picks a larger model. In other words, the longest cathetus criterion tends to select more parsimonious model than the longest hypotenuse criterion.

3.3.3 The triangle area criterion

Besides the cathetus and hypotenuse, we believe the area of the triangle can also be used to construct a new model selection criterion as it utilises most of the information on the 1 rank model selection curve. This new criterion should have a good ability to identify an appropriate model. Since the criterion is based on the area of a triangle, we name this triangle area criterion (TAC).

Suppose the cathetus length in the truncated polygon of model α, as one of the models in Figure 3.2, is denoted by C_{L_α}. The hypotenuse length is determined by $C_{L_\alpha}\sqrt{1+p_\alpha^2}$. Clearly, these two criteria depend on the λ values for a specific model α with dimension p_α, which is also a slope parameter of generalised information criterion as mentioned earlier. Now let θ $(0 \le \theta \le \frac{\pi}{2})$ be the angle between the cathetus and the hypotenuse of the right-angled triangle for model α.

A model α can be selected according to the following expression

$$\max_{\alpha\in\mathcal{A}} \iint\limits_{(\lambda,\ M(\lambda;\alpha))\in\mathbb{R}} \mathbf{1}\{r(\lambda;\alpha)=1\}\ dM(\lambda;\alpha)\ d\lambda,$$

or equivalently,

$$\max_{\alpha\in\mathcal{A}}\ \frac{1}{2}\cdot C_{L_\alpha}\cdot[C_{L_\alpha}\sqrt{1+p_\alpha^2}]\ \sin\theta\cdot\mathbf{1}\{\varrho(C_{L_\alpha};\alpha)=1\}, \tag{3.9}$$

where $\mathbf{1}\{\varrho(C_{L_\alpha};\alpha)=1\}$ takes a value 1 if model α with cathetus length C_{L_α} achieves rank 1 and 0 otherwise. Note that the part to be maximised in the expression (3.9) is the area of a right-angled triangle associated with model α in the truncated polygon. Since the model $\alpha \in \mathcal{A}$ appears in $\gamma_{(1)}(\lambda;\mathcal{A})$, the area of the triangle related with model α in expression (3.9), denoted by AT_α, can be written as

$$\begin{aligned}
AT_\alpha &= \frac{1}{2}C_{L_\alpha}^2\sqrt{1+p_\alpha^2}\ \sin\theta \\
&= \frac{1}{2}C_{L_\alpha}^2\sqrt{1+\tan^2\theta}\ \sin\theta \\
&= \frac{1}{2}C_{L_\alpha}^2\sqrt{\frac{\cos^2\theta+\sin^2\theta}{\cos^2\theta}}\ \sin\theta \\
&= \frac{1}{2}C_{L_\alpha}^2\ \frac{1}{\cos\theta}\ \sin\theta,
\end{aligned}$$

so, the area becomes

$$\begin{aligned} AT_\alpha &= \frac{1}{2} C_{L_\alpha}^2 \tan\theta \\ &= \frac{1}{2} C_{L_\alpha}^2 \; p_\alpha . \end{aligned}$$

The measurement unit of an area is naturally in square of the original unit. For easy comparison with the other two MSC based criteria in this study, the square root of the triangle area may be considered, and it takes the following form:

$$\sqrt{TAC} = \max_{\alpha \in \mathcal{A}} \; C_{L_\alpha} \sqrt{\frac{p_\alpha}{2}}.$$

Therefore, a model α is chosen for which TAC or $\sqrt{TAC}$ value is the maximum among all models $\alpha \in \mathcal{A}$ having ever achieved rank 1.

Note that $\sqrt{TAC}$ is a product of cathetus length and a simple function of model dimension, p_α; it is reduced to LCC when a model with $p_\alpha = 2$ attains the longest cathetus. Furthermore, TAC is less likely to pick an intercept-only model, which is obvious from the expression of this criterion. TAC can thus shed extra light in selecting model when several candidate models have similar cathetus length, becasue of its relation with the model dimension illustrated. For example, if two or more models with different dimensions have similar cathetus length that is the longest among all, LCC cannot distinguish those models because this criterion is solely based on cathetus length. So a researcher will have difficulty in selecting a model if LCC is used alone. However, in this situation, TAC may be able to distinguish those models and is more likely to pick the model with larger dimension. Although LHC also has model dimension in its expression, it may pick a model with even larger dimension than TAC. The advantage of TAC over LCC and LHC in this circumstances is quite obvious. Moreover, in the special case where those models with exact the same dimension (when $p_\alpha \geq 2$), based on the 1 rank model selection curve, all three criteria will perform similarly. In this case we recommend to use bootstrap replications (see Section 4.2) of the data to get more information and thus reach more concrete decision in choosing an appropriate model.

Under some conditions for AFT models in survival analysis, it is shown below

that LCC is consistent for AFT model selection. Therefore, LHC and TAC are also consistent since both are functions of the cathetus length on which LCC is based.

Lemma 1 *Given a true AFT model α_0 with dimension p_{α_0}, assume the generalised information criterion (GIC) is consistent when $\lambda_n \to \infty$ and $\lambda_n/n \to 0$ as $n \to \infty$. Let $\xi_n < \lambda_n^{max}$ be a nonnegative quantity that satisfies $\xi_n \to \infty$, $\lambda_n^{max}/n \to 0$ as $n \to \infty$ such that the penalty measure $\Lambda_n([\lambda_n^{min} + \xi_n,\ \lambda_n^{max}]) \to 1$. Then the longest cathetus criterion is consistent for the true AFT model α_0.*

Proof. *Since the GIC, defined for AFT model with fixed p_{α_0}, is consistent when $\lambda_n \to \infty$ and $\lambda_n/n \to 0$ as $n \to \infty$, it is consistent for $\lambda_n = \lambda_n^{min} + \xi_n$ and also for $\lambda_n = \lambda_n^{max}$. Now we can write*

$$
\begin{aligned}
&\int_{\lambda_n^{min}}^{\lambda_n^{max}} (1 - \mathbf{1}\{r(\lambda;\alpha_0) = 1\}) d\Lambda_n \\
&= \int_{\lambda_n^{min}}^{\lambda_n^{min}+\xi_n} (1 - \mathbf{1}\{r(\lambda;\alpha_0) = 1\}) d\Lambda_n + \int_{\lambda_n^{min}+\xi_n}^{\lambda_n^{max}} (1 - \mathbf{1}\{r(\lambda;\alpha_0) = 1\}) d\Lambda_n \\
&\leq \int_{\lambda_n^{min}}^{\lambda_n^{min}+\xi_n} 1 d\Lambda_n + \int_{\lambda_n^{min}+\xi_n}^{\lambda_n^{max}} (1 - \mathbf{1}\{r(\lambda;\alpha_0) = 1\}) d\Lambda_n.
\end{aligned}
$$

The first term in the above relation is $o_p(1)$ since $\Lambda_n([\lambda_n^{min}+\xi_n,\ \lambda_n^{max}]) \to 1$, $\xi_n < \lambda_n^{max}$ satisfy $\xi_n \to \infty$ and $\lambda_n^{max}/n \to 0$ as $n \to \infty$ for the true AFT model α_0. The second term is bounded from above and can be written as

$$
\begin{aligned}
&\int_{\lambda_n^{min}+\xi_n}^{\lambda_n^{max}} (1 - \mathbf{1}\{r(\lambda;\alpha_0) = 1\}) d\Lambda_n \\
&\leq \int_{\lambda_n^{min}+\xi_n}^{\lambda_n^{max}} (1 - \mathbf{1}\{r(\lambda_n^{min} + \xi_n;\alpha_0) = 1\}) d\Lambda_n + \int_{\lambda_n^{min}+\xi_n}^{\lambda_n^{max}} (1 - \mathbf{1}\{r(\lambda_n^{max};\alpha_0) = 1\}) d\Lambda_n \\
&= o_p(1) + o_p(1) \\
&= o_p(1),
\end{aligned}
$$

since the GIC for the true AFT model α_0 is consistent over the limits and also by dint of $\Lambda_n([\lambda_n^{min}+\xi_n,\ \lambda_n^{max}]) \to 1$, the penalty measure puts mass 1 both at $\lambda_n = \lambda_n^{min}+\xi_n$ and at $\lambda_n = \lambda_n^{max}$, and because if $r(\lambda_n^{min} + \xi_n;\alpha_0) = 1$, it is not possible that $r(\lambda^*;\alpha_0) \neq 1$, where $\lambda^* \in [\lambda_n^{min} + \xi_n,\ \lambda_n^{max}]$ is any specific value of λ (e.g., $\lambda^* = \frac{\lambda_n^{min}+\xi_n+\lambda_n^{max}}{2}$) and $r(\lambda^*;\alpha_0)$ converges to rank 1 faster than $r(\lambda_n^{min} + \xi_n;\alpha_0)$ as $n \to \infty$.

Now consider a model $\alpha \neq \alpha_0$. Under the assumptions as stated, we have the

following:

$$\int_{\lambda_n^{min}}^{\lambda_n^{max}} (1 - \mathbf{1}\{r(\lambda;\alpha) = 1\}) d\Lambda_n$$
$$= \int_{\lambda_n^{min}}^{\lambda_n^{min}+\xi_n} (1 - \mathbf{1}\{r(\lambda;\alpha) = 1\}) d\Lambda_n + \int_{\lambda_n^{min}+\xi_n}^{\lambda_n^{max}} (1 - \mathbf{1}\{r(\lambda;\alpha) = 1\}) d\Lambda_n. \quad (3.10)$$

The first term of the right hand side of equation (3.10) is bounded from below by zero. This is because the integrand is non-negative, and so the integrated value cannot be negative. Thus it can be written as

$$\int_{\lambda_n^{min}}^{\lambda_n^{min}+\xi_n} (1 - \mathbf{1}\{r(\lambda;\alpha) = 1\}) d\Lambda_n \geq 0.$$

The second term of the right hand side of equation (3.10) is

$$\int_{\lambda_n^{min}+\xi_n}^{\lambda_n^{max}} (1 - \mathbf{1}\{r(\lambda;\alpha) = 1\}) d\Lambda_n$$
$$= \int_{\lambda_n^{min}+\xi_n}^{\lambda_n^{max}} 1 d\Lambda_n - \int_{\lambda_n^{min}+\xi_n}^{\lambda_n^{max}} \mathbf{1}\{r(\lambda;\alpha) = 1\} d\Lambda_n$$
$$\geq 1 - o_p(1).$$

Therefore, equation (3.10) becomes

$$\int_{\lambda_n^{min}}^{\lambda_n^{max}} (1 - \mathbf{1}\{r(\lambda;\alpha) = 1\}) d\Lambda_n$$
$$= \int_{\lambda_n^{min}}^{\lambda_n^{min}+\xi_n} (1 - \mathbf{1}\{r(\lambda;\alpha) = 1\}) d\Lambda_n + \int_{\lambda_n^{min}+\xi_n}^{\lambda_n^{max}} (1 - \mathbf{1}\{r(\lambda;\alpha) = 1\}) d\Lambda_n$$
$$\geq 1 - o_p(1),$$

when a model $\alpha \neq \alpha_0$. Hence the longest cathetus criterion is consistent for selecting a true model α_0. ■

3.4 Model selection framework for AFT models

Here we propose a model selection framework for AFT models in survival analysis. It consists of all three MSC based model selection criteria, discussed and derived in the

previous section, as well as two commonly used criteria AIC and BIC. Let us continue using the example in Section 3.2, in particular Figure 3.1, to illustrate how each model selection criterion under this framework is displayed and interpreted. In Figure 3.1(d), if perpendiculars are drawn from the vertices of the lower enveloping curve to the horizontal axis, the horizontal yellow, red and green dotted line segments correspond to the lengths of cathetus for models $M1$, $M2$ and $M4$ respectively. This is illustrated in Figure 3.3.

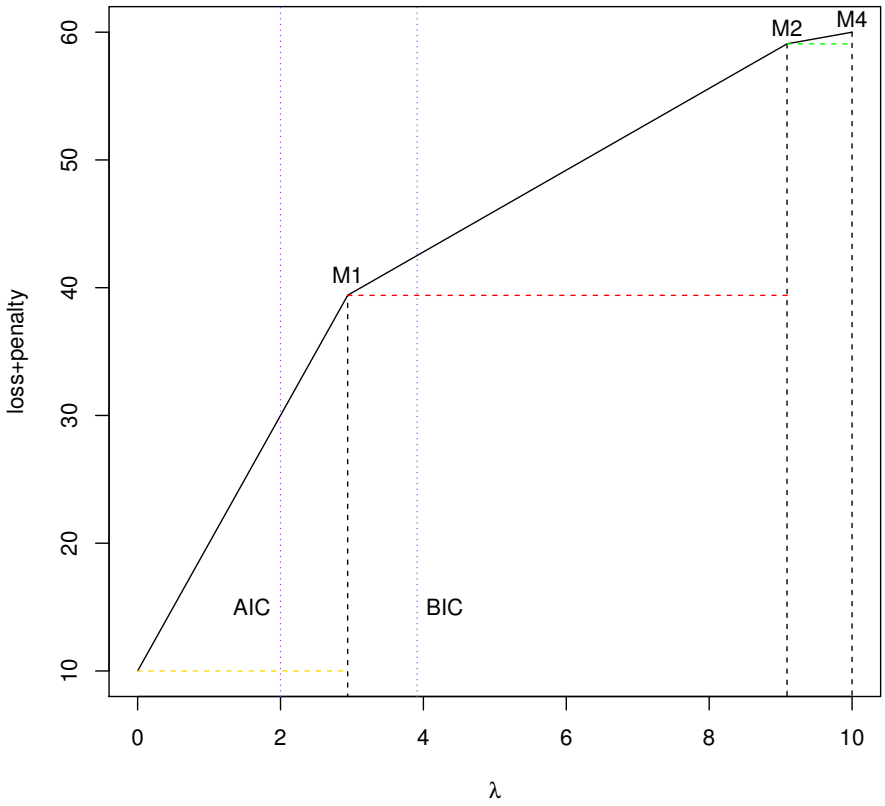


Figure 3.3: Model selection criteria under study.

The length of a cathetus for model α is $C_{L_\alpha} = c_{u_\alpha} - c_{l_\alpha}$ where c_{u_α} and c_{l_α} are respectively the $x-$coordinates of the upper and lower end points of the cathetus. So, the relative length of the cathetus over $\lambda \in [0, 4\log(n)]$ is $\frac{C_{L_\alpha}}{4\log(n)}$. The cathetus length of a model may indicate the stability of the model as explained earlier in this thesis. Based on Figure 3.3, model $M2$ (in red dotted line) has the longest cathetus and is thus selected by LCC from the four models considered. Note that model $M2$ is also selected by BIC because the vertical line at $\lambda = \log(n)$ for $n = 50$ in the example crosses over

the model $M2$ segment of the 1 rank model selection curve. However, using AIC in this instance leads to selection of model $M1$ (in yellow dotted line) instead. It is because the vertical line at $\lambda = 2$ for AIC in the figure crosses over with the part of the 1 rank model selection curve that corresponds to $M1$.

It is obvious that the lengths of catheti in Figure 3.3 can also be observed in the rank plot in Figure 3.1(b), where they are presented as intervals of λ on horizontal axis for models achieved rank 1. So, the 1 rank plot that only contains models that have achieved rank 1 is considered a sufficient display for model selection using LCC.

The segments on the lower enveloping curve are the hypotenuses of the right-angled triangle as shown in Figure 3.3. Here model $M2$ clearly has the longest hypotenuse and thus is selected by LHC. Note that the model selection via LHC is simply based on segments of the 1 rank model selection curve as shown in Figure 3.1(d).

It can be seen in Figure 3.3 that the triangle corresponding to model $M2$ seems to have the maximum value (i.e., largest area) and is thus selected according to TAC. The TAC for a model can be presented graphically by the triangle area corresponding to the model. For the example shown in Figure 3.3, all three criteria LCC, LHC and TAC select model $M2$ same as BIC, while AIC picks a different model $M1$. For this example, it is not difficult to decide which model should be selected using each of the three criteria based on the MSC. It is possible in some cases that two or more of the models considered have similar length of cathetus and may thus be difficult to decide which one to choose using LCC. In this case, we may use TAC or LHC. Similarly, when TAC or LHC falls into similar situation, the other MSC based criteria may be used. Bootstrap technique can be incorporated to provide additional information, such as the chance of selecting a particular model, a variable and/or models with a particular dimension. This helps with the exploration on the stability of a candidate model through multiple perspectives, which can enhance model selection.

In this thesis, we adopted similar bootstrap technique used in Müller and Welsh (2010). In this technique, the empirical bootstrap estimate of the probability of selecting model α, denoted by $\pi^*(\lambda;\alpha)$, is obtained by computing proportion of this model being selected by LCC across all bootstrap replications for each λ values within its range. The $\pi^*(\lambda;\alpha)$'s of all models can be plotted against λ, and used as an additional model detection plot. Note that to avoid having too many curves in the plot, we can

include only models that appear frequently in 1 rank model selection curve. The frequency can be measured by the marginal probability of selecting a model α over the range of λ denoted by $\pi^*(\alpha)$. It is defined by

$$\pi^*(\alpha) = \int_\Lambda \pi^*(\lambda;\alpha)\, d\Lambda,$$

where Λ is assumed to be uniformly distributed over the interval $[0,\ 4\log(n)]$. Here we consider a model α appeared frequently if $\pi^*(\alpha) > 4\%$, same as Müller and Welsh's study. If $\pi^*(\alpha)$ for model α is the largest among all possible models and also very large, say greater than 0.50, this is considered as the best model on the specified range of λ values.

Bootstrapping can also be used to quantify the importance of each variable considered in a study. To do this, the probability of including a variable in a model at each λ, denoted by $\pi^*\left(\lambda, x_j\right),\ j = 1,2,\ldots,p$, is computed. It is simply the proportion of times when the variable is included in a model across all bootstrap replications for each λ. Then $\pi^*\left(\lambda, x_j\right)$'s can be plotted against λ to indicate the order of importance of covariates over the range of λ. This plot is known as variable inclusion plot. A diagnostic measure for the inclusion or exclusion of variable x_j can be obtained by summing $\pi^*\left(\lambda; x_j\right)$ over λ, i.e., $\pi^*_{x_j} = \sum_\lambda \pi^*\left(\lambda, x_j\right)$.

The whole model selection framework, including LCC, LHC, TAC, AIC and BIC will be investigated through simulations and real world data examples in subsequent chapters.

3.5 Construction of the model selection curves for AFT models

Depending on the distribution function of the error term W in equation (2.10), the type of AFT model is assumed. Here Weibull AFT model is used as an example to illustrate how to construct model selection curves. Note that the error term W in a Weibull AFT model expressed in equation (2.10) follows an extreme value distribution. The density and survivor functions are given in equations (2.13) and (2.14) respectively. Substituting these two functions into equation (2.22), the log-likelihood for the Weibull

AFT model is thus

$$
\begin{aligned}
l\left(\boldsymbol{\theta}\right) &= -\log\tau\sum_{i=1}^{n}\delta_i + \sum_{i=1}^{n}\left(\delta_i(w_i - \exp(w_i)) + (1-\delta_i)(-\exp(w_i))\right) - \sum_{i=1}^{n}\delta_i \log y_i \\
&= -\log\tau\sum_{i=1}^{n}\delta_i + \sum_{i=1}^{n}\delta_i\left(\frac{\log y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i}{\tau}\right) - \sum_{i=1}^{n}\exp\left(\frac{\log y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i}{\tau}\right) - \sum_{i=1}^{n}\delta_i \log y_i \\
&= \left(\frac{1-\tau}{\tau}\right)\sum_{i=1}^{n}\delta_i \log y_i - \sum_{i=1}^{n}\exp\left(\frac{\log y_i - \boldsymbol{\beta}^\top \mathbf{x}_i}{\tau}\right) - \sum_{i=1}^{n}\delta_i\left(\log\tau + \frac{\boldsymbol{\beta}^\top \mathbf{x}_i}{\tau}\right). \qquad (3.11)
\end{aligned}
$$

The MLEs ($\hat{\boldsymbol{\theta}}$) of the parameters in equation (3.11) can be obtained using an iterative procedure (e.g., Newton-Raphson). Using the MLEs the $loss + penalty$ form for the Weibull AFT model is

$$
M\left(\lambda;\alpha\right) = -2l\left(\hat{\boldsymbol{\theta}}\right) + \lambda p_\alpha,\ \lambda > 0. \qquad (3.12)
$$

This provides a basis for the construction of model selection curves under the Weibull AFT model. This approach works similarly for log-logistic and log-normal AFT models considered in our study. The only difference is they have different log-likelihood functions.

The log-likelihood function for the log-logistic AFT model, after substituting equations (2.16) and (2.17) into equation (2.22), is

$$
\begin{aligned}
l(\boldsymbol{\theta}) &= -\log\tau\sum_{i=1}^{n}\delta_i + \sum_{i=1}^{n}\left(\delta_i(w_i - 2\log(1+\exp(w_i))) + (1-\delta_i)(-\log(1+\exp(w_i)))\right) \\
&\quad - \sum_{i=1}^{n}\delta_i \log y_i, \\
&= \sum_{i=1}^{n}\delta_i\left(\frac{\log y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i}{\tau} - \log y_i - \log\tau\right) \\
&\quad - \sum_{i=1}^{n}(1+\delta_i)\log\left(1+\exp\left(\frac{\log y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i}{\tau}\right)\right).
\end{aligned}
$$

Similarly, the log-likelihood function for the log-normal AFT model can be obtained by substituting equations (2.18) and (2.19) into equation (2.22), and is expressed as

$$
\begin{aligned}
l(\boldsymbol{\theta}) &= -\log\tau \sum_{i=1}^{n}\delta_i + \sum_{i=1}^{n}\left(\delta_i(\log(\frac{1}{\sqrt{2\pi}}) - \frac{1}{2}w_i^2) + (1-\delta_i)\log\left(1-\Phi(w_i)\right)\right) \\
&\quad - \sum_{i=1}^{n}\delta_i \log y_i \\
&= \sum_{i=1}^{n}\delta_i\left(\log(\frac{1}{\sqrt{2\pi}}) - \log\tau\right) - \frac{1}{2}\sum_{i=1}^{n}\delta_i\left(\frac{\log y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i}{\tau}\right)^2 \\
&\quad + \sum_{i=1}^{n}(1-\delta_i)\log\left(1-\Phi\left(\frac{\log y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i}{\tau}\right)\right) - \sum_{i=1}^{n}\delta_i \log y_i.
\end{aligned}
$$

For constructing model selection curves, we have chosen $\lambda \in [0,\ 4\log(n)]$ in equation (3.12), same as Müller and Welsh's study (2010). As mentioned earlier, such range of λ would cover most of existing model selection criteria that are based on single values of λ, such as AIC ($\lambda = 2$) and BIC ($\lambda = \log(n)$). Both AIC and BIC would appear as single points on the model selection curves.

The model selection framework for AFT models proposed in this chapter is studied extensively through simulations in next chapter.

4

Simulation Study

In this chapter, the model selection framework, proposed in Section 3.4, is investigated via a comprehensive simulation study. The mechanism of generating survival data with censoring is also discussed. A number of survival data sets are generated from three types of distributions, Weibull, log-logistic and log-normal. Different sample sizes ranging from 30 to 300, various censoring proportions such as 10% and 50%, and different sets of AFT regression coefficients are considered in the simulations. The performance of all the model selection criteria included in the AFT model selection framework is examined for Weibull, log-logistic and log-normal AFT models using the data sets simulated. The Weibull AFT models are chosen to illustrate how to interpret the graphs and the tables of the results that are produced under the model selection framework for AFT models.

4.1 Parameterisation of distributions under study

For simulating survival data from three different distributions of Weibull, log-logistic and log-normal, our emphasis was placed on generating data sets that are comparable across these three distributions to allow for easy comparison. Each of the three distributions considered here can exhibit four different shapes as explained in Chapter 2. In our simulation study, four Weibull distributions are chosen to cover those four different shapes. These four distributions with specified parameters $\nu = \exp(0.1)$ and $\kappa = (1,\ 2,\ 3,\ 4)$ are shown in Figure 4.1, and the four shapes shown on the figure will be referred as severely right-skewed, moderately right-skewed, nearly symmetric and moderately left-skewed, respectively. Note that the horizontal axis there is based on the 99th percentile of the respective Weibull distribution. For each of these four Weibull distributions, a data set is generated, and a comparable data set with log-logistic and one with log-normal distribution are then generated.

Suppose the Weibull distribution shown in Figure 4.1(a) was chosen first, and then a data set was generated. Our initial scheme was to find the parameters for each of the other two distributions by trial and error so that data sets generated from these three distributions have comparable mean and standard deviation (SD). This approach did not work well for this particular case as the resulting data sets have exhibited vastly different range across these three heavily right-skewed distribution. This can be explained by the fact that, when those three distributions are all right skewed with extreme observations, the right side tail of the log-logistic can be much heavier than other two distributions as log-logistic distribution is the only distribution here that may not have finite first and/or second moments (see equations 2.7 and 2.8 in Section 2.3). This means neither sample mean nor sample SD is an appropriate summary measure for comparisons across those skewed distributions, and more robust summary measures such as median, inter-quartile range (IQR) and median absolute deviation (MAD) should be used instead. Keeping this in mind, we have come up with a modified scheme for generating data sets from the three distributions.

Here is our modified scheme. At first a large number of observations (say, $N = 500{,}000$) is drawn from a Weibull distribution parameterised by one of the set-ups shown in Table 4.2. Note that main reason for simulating a very large data is to ensure the data is close to its population distribution specified. To work out the parameters

to be used for generating data from another distribution, say a log-logistic distribution, an intercept only log-logistic AFT model is fitted by maximum likelihood to the large number of observations drawn from the specified Weibull distribution. Based on the MLEs, observations are generated from the log-logistic distribution. Similar procedure can be followed to generate data from log-normal distribution. We found out that the data generated through this scheme would have comparable robust summary measures.

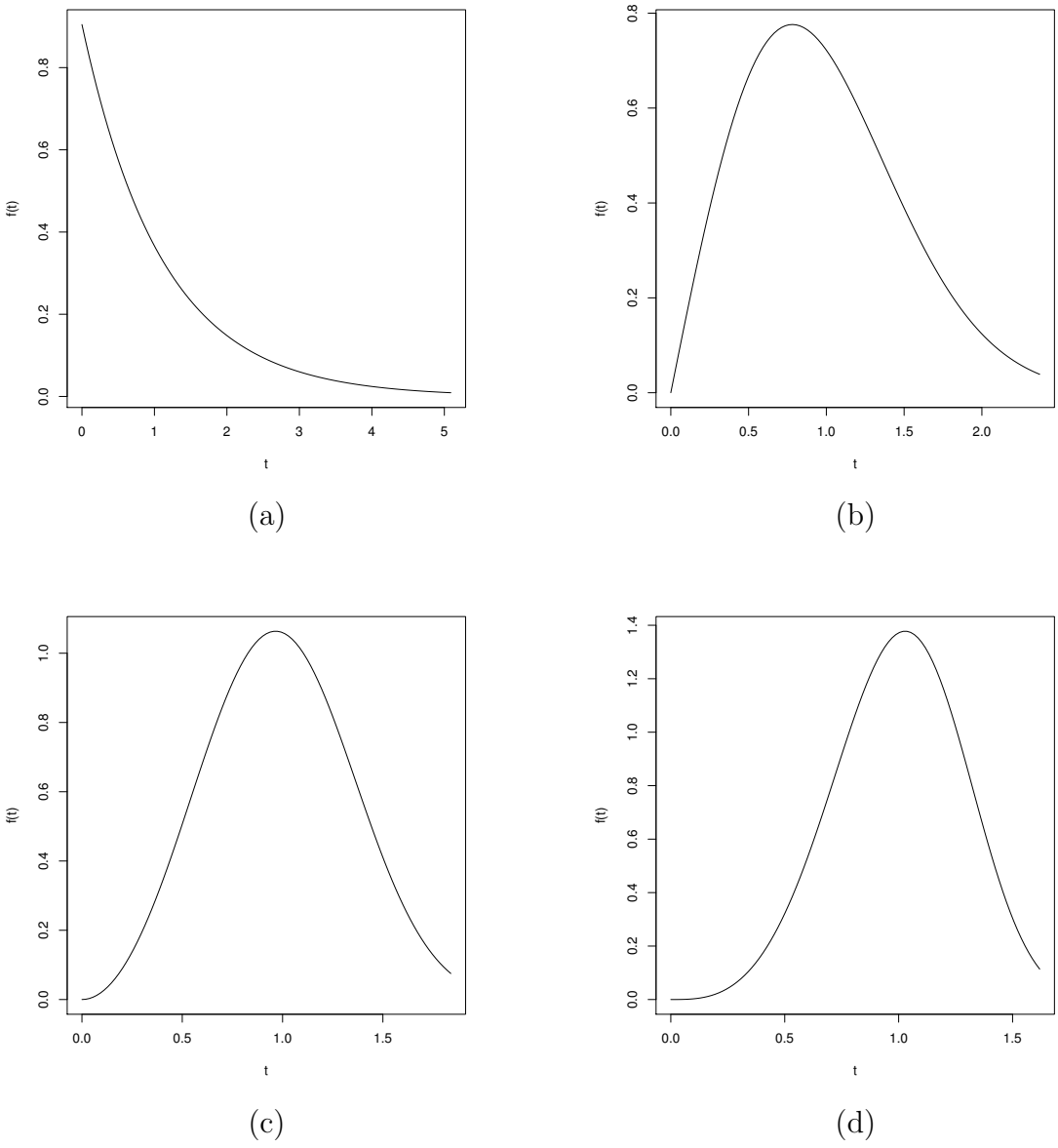


Figure 4.1: Weibull density curves (plots) for fixed $\nu = \exp(0.1)$: (a) $\kappa = 1$; (b) $\kappa = 2$; (c) $\kappa = 3$; (d) $\kappa = 4$.

Table 4.1: Summary measures of simulated data.

Statistic	Weibull sample	Log-logistic sample	Log-normal sample
Mean	0.979	1.083	1.013
Median	0.920	0.877	0.827
SD	0.511	0.941	0.720
IQR	0.707	0.689	0.736
MAD	0.519	0.481	0.504

To demonstrate that the data sets generated via our modified scheme are comparable across the three distributions of interest, $500,000$ observations were simulated from the Weibull distribution corresponding to Figure 4.1(a). Then the data sets with the same size were generated from the matching log-logistic and log-normal distributions. Summary measures of the three data sets are presented in Table 4.1. As shown in this table, median, MAD and IQR values are very close to each other across those three distributions. However, SD for the generated log-logistic data is almost double of, and the log-normal data is about 40% higher than the SD of the Weibull data. As mentioned earlier, mean and SD are not appropriate summary measures for survival data if skewed and/or censored. However, to generate comparable survival data with censored observations, median and MAD are thus more suitable summary measures to be used.

Table 4.2: Parameterisation of distributions.

Set-up	Weibull (κ,ν)	Log-logistic (ξ,ω)	Log-normal (μ,σ)
1	$(2,\exp(0.1))$	$(\frac{1}{0.35},\exp(-0.13))$	$(-0.19,0.64)$
2	$(1,\exp(0.1))$	$(\frac{1}{0.69},\exp(-0.36))$	$(-0.48,1.28)$
3	$(3,\exp(0.1))$	$(\frac{1}{0.23},\exp(-0.05))$	$(-0.09,0.43)$
4	$(4,\exp(0.1))$	$(\frac{1}{0.17},\exp(-0.02))$	$(-0.04,0.32)$

Using this modified scheme, the parameter settings of the four Weibull and its matching log-logistic and log-normal distributions considered are summarised in Table 4.2. These four settings are used throughout our study. The set-up 1 is for moderately right-skewed distribution, the set-up 2 corresponds to severely right-skewed distribution, the set-up 3 is for nearly symmetric distribution and the last set-up corresponds to moderately left-skewed distributions. For example, Figure 4.1 shows different levels of skewness for four different Weibull distributions considered in the simulation study.

4.2 Method of generating survival data with censoring

In the previous section, simulation of identically and independently distributed complete data (without censoring) was discussed. In this section, the method and process for generating samples of survival data with specified censoring proportion while incorporating covariates for simulation study are described. The method for obtaining bootstrap samples is also explained.

4.2.1 Generating survival data with specified censoring proportion

Suppose we want to simulate data of size n containing survival time with censoring and several covariates. In order to control the properties of the sample, we will go through a two stage process by first simulating a much larger number of observations (say N) and then using stratified sampling via proportional allocation.

General speaking, two survival distributions are required to generate survival times in the data. One corresponds to the uncensored (complete) survival times ($\boldsymbol{T}$) and the other corresponds to the censored (incomplete) survival times ($\boldsymbol{C}$) as described in the paper by Moriña and Navarro (2014). Since both censored and uncensored survival time distributions considered under the AFT model are functionally related with the linear predictors as explained in Section 2.4, not only we need to assign values for the coefficients of the assumed AFT model, but also need to simulate a set of covariates for each survival time. To simulate the covariates, we first specify its mean vector $\boldsymbol{\mu}$ and dispersion matrix $\boldsymbol{\Sigma}$ and then the covariates are drawn from a multivariate normal distribution with the specified $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Once the linear predictors are generated one of the parameters (e.g., scale) of the distribution is computed using the functional relation. On the other hand, to generate a censored survival time a constant term is added to each of the linear predictors of the uncensored survival time and the functional relation with the linear predictor is used once again to evaluate the specific values of scale and shape parameters of the distribution of the censored survival time. The survival times with censoring are then generated and defined using equation (2.20) and

censoring status determined accordingly. Note that the sample censoring proportion is controlled via the size of the constant term mentioned above. As such relationship is implicit, for a specific censoring proportion the size of the constant term is determined by trial and error such that the sample censoring proportion over a large sample is as close as possible to the desired population censoring proportion. This is the basis for further sampling to get data sets of different sizes n.

Notice that even if we can control the censoring proportion in the much larger sample with size N, simply drawing random samples of a particular size n from there, the censoring proportion across these samples can vary. In some instances, the censoring proportions can be far off the desired proportion, especially when the sample size is small. For example, when the desired censoring proportion is 50%, the censoring proportion in the sample drawn can be as low as 36%, according to our simulations.

In order to control the censoring proportion even in a sample of size n, we consider stratified random sampling technique. Here is how to get a sample of survival data of size n with a desired censoring proportion p, $0 < p < 1$. Suppose we have a very large number of N observations such that the censoring proportion is very close to the specified censoring proportion p. In those N observations, there are two subgroups, defined by censoring status, which can be used as strata in our proposed stratified sampling. Stratum one consists of all the uncensored observations, say N_1, and stratum two consists of all the censored observations with size $N_2 = N - N_1$. A simple random sample of size $n_1 = \lfloor n * (1 - p) \rfloor$, where $\lfloor \cdot \rfloor$ is the integer part of the number, is drawn from the N_1 uncensored observations, and another simple random sample of size $n_2 = n - n_1$ is drawn from the N_2 censored observations. These two samples obtained are then merged to form one data set of size n, which will have the censoring proportion p as desired for any given sample size n.

4.2.2 Bootstrap sampling schemes

As discussed in Section 3.4, bootstrap replications can be used in some cases to provide additional information to enhance model selection. Two bootstrapping schemes, ordinary and stratified, are used in this study. Ordinary bootstrap sampling is simply resampling from the original data (Davison and Hinkley, 1997). One limitation of this sampling for survival data is that its resulting bootstrap samples may not have the same

censoring proportion as the original data. To overcome this problem, i.e., obtain all bootstrap samples with the same censoring proportion as the original sample, we have used stratified sampling by censoring status. In this approach, a stratified bootstrap sample is drawn from censored and uncensored portion of the original data, separately, with respect to the censoring proportion in the data. For our simulation study in next section, 1,000 bootstrap replications are used whenever bootstrap technique is considered.

4.3 An Illustration of the model selection framework

In this section, we describe the details of specific Weibull model for simulating data based on set-up 1 (see Table 4.2). Although covariates in most of the data sets generated for our simulations in this chapter are almost uncorrelated, we have also considered an example where few of the covariates are highly correlated. Our proposed model selection framework is applied to all these data sets generated as illustration. In addition, how MSC based criteria are used to handle AFT models with a combination of continuous and categorical covariates with more than two levels have been addressed in this section.

4.3.1 Simulating data for Weibull AFT models

Suppose we want to simulate some data from the Weibull AFT model below with four continuous covariates/predictors,

$$\log T_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \tau W_i,\ i = 1,\ 2,\ \ldots,\ n. \tag{4.1}$$

Firstly, observations for the almost uncorrelated predictors ($\rho = 0.001$) x_1, x_2, x_3 and x_4 are drawn from a multivariate normal distribution with mean vector $\boldsymbol{\mu}_0$ and

dispersion matrix $\boldsymbol{\Sigma}_0$ as given below

$$\boldsymbol{\mu}_0^\top = (0,\ 0,\ 0,\ 0) \text{ and } \boldsymbol{\Sigma}_0 = \begin{pmatrix} 1 & 0.001 & 0.001 & 0.001 \\ 0.001 & 1 & 0.001 & 0.001 \\ 0.001 & 0.001 & 1 & 0.001 \\ 0.001 & 0.001 & 0.001 & 1 \end{pmatrix}.$$

We consider three different sets of coefficients (0.1, 1, 0.7, 0.9, 0.8), (0.1, 0, 0, 0.9, 0.8) and (0.1, 0, 0, 0.9, 0) corresponding to specified Weibull AFT models for generating survival times in different data sets. These coefficients correspond to models {1, 2, 3, 4, 5} (full model with all predictors), {1, 4, 5} (model with x_3 and x_4) and {1, 4} (model with x_3) respectively. The "1" inside the curly brackets represents the intercept term that is included in all models considered in our simulations. Note that the scale parameter of a Weibull distribution can be computed in equation (2.15) after specifying the shape parameter.

One particular important aspect in simulating survival data with right censoring is to simulate a complete survival time vector and, independently, a censored survival time vector (Moriña and Navarro, 2014). Therefore, we draw separately two sets of observations (i.e., T's and C's) from two different Weibull distributions with different scale but same shape parameter $\kappa = 2$. This is achieved by creating the censored time vector C through adding some constant value to the intercept term of the AFT model under consideration as explained earlier. Then survival times with censoring status are defined according to equation (2.20). This together with the covariates drawn from the multivariate normal distribution as mentioned in previous section constitute a survival data to be used in this section.

It is possible that two or more predictors considered in a model may be correlated. In this case, a subset of these may be sufficient to represent the rest of them in the model. Suppose that two covariates x_3 and x_4 in equation (4.1) are highly correlated. Let the correlation coefficient between x_3 and x_4 is $\rho = 0.90$. Then entries of cell (3, 4) and (4, 3) in $\boldsymbol{\Sigma}_0$ are 0.9. Based on specified model coefficients (0.1, 1, 0, 0.8, 0.9), survival data with some correlated predictors can be generated in a similar manner as described above. Note that the strength of correlation may have impact on model selection. This can be investigated via bootstrapping, where a variable inclusion plot

as discussed in Section 3.4, can be constructed and used to assist model selection.

In survival analysis, it is very common that some of the covariates in the data are categorical with more than two levels. In this section, we have considered a case where covariates x_1 and x_2 in the model given by equation (4.1) are categorical variables with three levels, coded 1, 2 and 3, while x_3 and x_4 are continuous or binary. In this case, the Weibull AFT model in equation (4.1) can be expressed as

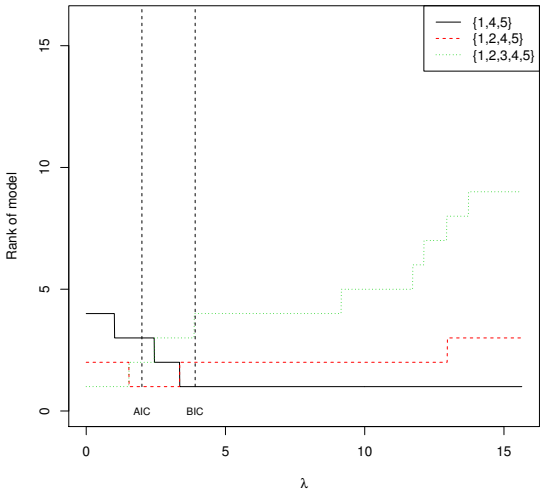
$$\log T_i = \beta_0+\beta_{12}x_{i12}+\beta_{13}x_{i13}+\beta_{22}x_{i22}+\beta_{23}x_{i23}+\beta_3x_{i3}+\beta_4x_{i4}+\tau W_i,\ i = 1, 2, \cdots, n, \quad (4.2)$$

where the 1st level of the categorical variables x_{11} and x_{21} serves as a reference group, respectively. Then based on specified model coefficients, (0.1, 0.7, 0.5, 0, 0, 1, 0), survival times can be generated similarly as discussed before. Note that the `R` tool developed for the model selection framework in Chapter 5 can handle this kind of data. Unlike the model with continuous or binary covariates only where the number of parameters is simply the number of covariates plus 1, the number of parameters in the model with categorical covariate(s) having more than two levels is not easily determined based on the number of covariates. If a categorical covariate with more than two levels is included in a model, the number of parameters for the covariate presented in this model is one less than the number of levels (i.e., $s - 1$, where s is the number of categories) in this covariate. If this covariate is then removed from the model, the number of parameters in the model goes down by $s - 1$.

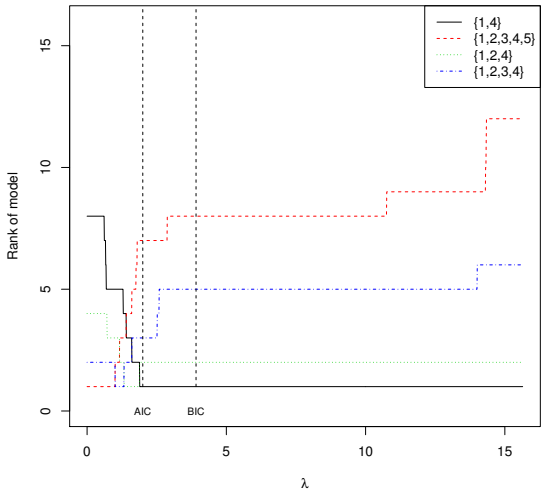
4.3.2 Study of the criteria under the model selection framework

Here we intend to examine the performance of the criteria under the model selection framework for AFT models, using several survival data sets generated based on Weibull AFT models as described in previous section.

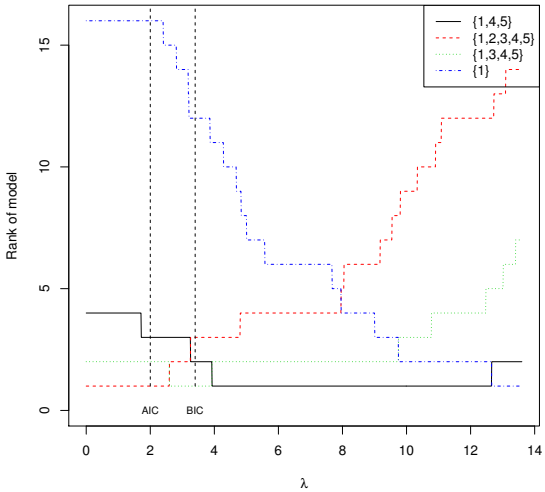
Suppose that a sample survival data of size 50 with 10% censoring is generated from a Weibull AFT model {1, 4, 5}, that is, the Weibull AFT model that contains two covariates x_3 and x_4 out of the four covariates in the data. Based on this sample, GIC values for all possible models are computed and ordered to determine ranks of these models. Based on the GIC values, ranks for the models that have ever achieved rank 1



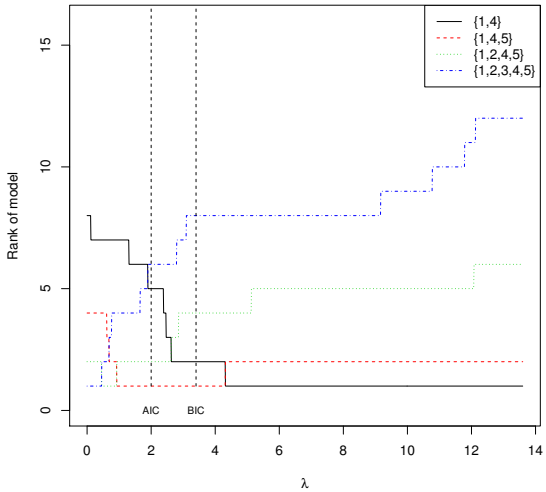
(a) $n = 50$, True model: $\{1, 4, 5\}$



(b) $n = 50$, True model: $\{1, 4\}$



(c) $n = 30$, True model: $\{1, 4, 5\}$



(d) $n = 30$, True model: $\{1, 4\}$

Figure 4.2: Rank plots for data with low censoring (10%) and small sample sizes.

over range of λ are plotted against λ to get rank plot (see Figure 4.2(a)). Similarly, for another even smaller sample of size 30 with 10% censoring the rank plot is obtained as shown in Figure 4.2(c). Similar samples of size 50 and 30 are also simulated based on Weibull AFT model $\{1, 4\}$. The corresponding rank plots are obtained as shown in Figure 4.2(b) and Figure 4.2(d). It was mentioned earlier that rank plot is a sufficient display for model selection via LCC (see Section 3.4). On these four rank plots, AIC and BIC are marked using the dotted vertical lines.

As shown in Figure 4.2(b), all the three criteria LCC, AIC and BIC, have identified the specified or true model $\{1, 4\}$. Both LCC and BIC identify the specified model

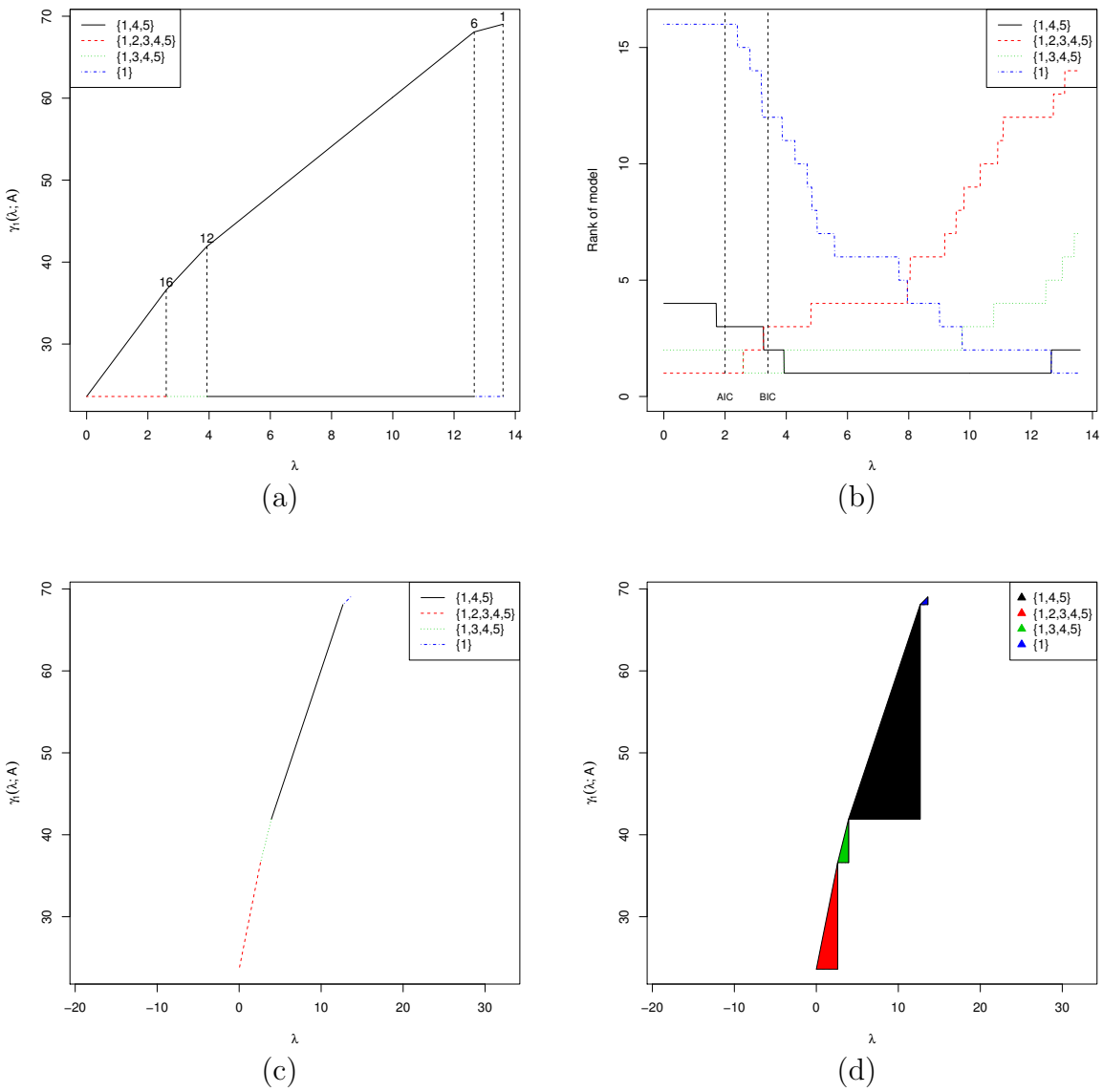


Figure 4.3: Plots from the model selection framework for data with low censoring (10%) when true model is {1, 4, 5} and $n = 30$: (a) model selection curve and cathetus lengths; (b) rank plot; (c) hypotenuse plot; (d) TAC plot.

{1, 4, 5} in Figure 4.2(a), but AIC does not. It is seen from Figure 4.2(c) and Figure 4.2(d) that the specified (true) model is selected if using LCC. However, in these two cases, both AIC and BIC have failed to identify the specified model but selected model {1, 2, 3, 4, 5} and model {1, 3, 4, 5} respectively.

It is interesting to note that different criteria have selected different models as shown in Figure 4.2(c). So, we are eager to look into this case again in more details by also considering two additional criteria under the model selection framework. The plots are generated, including hypotenuse plot and TAC plot, and shown in Figure

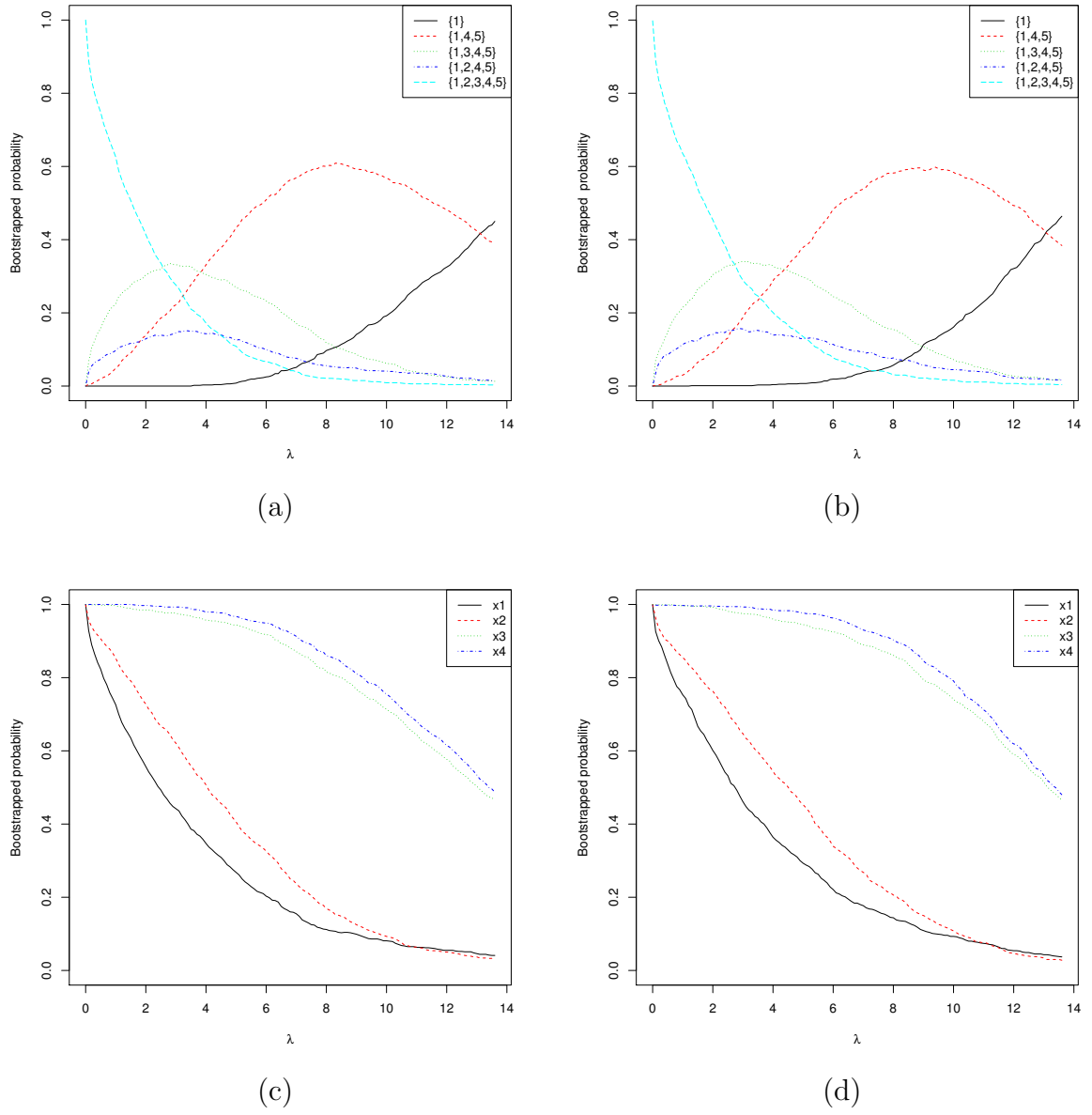
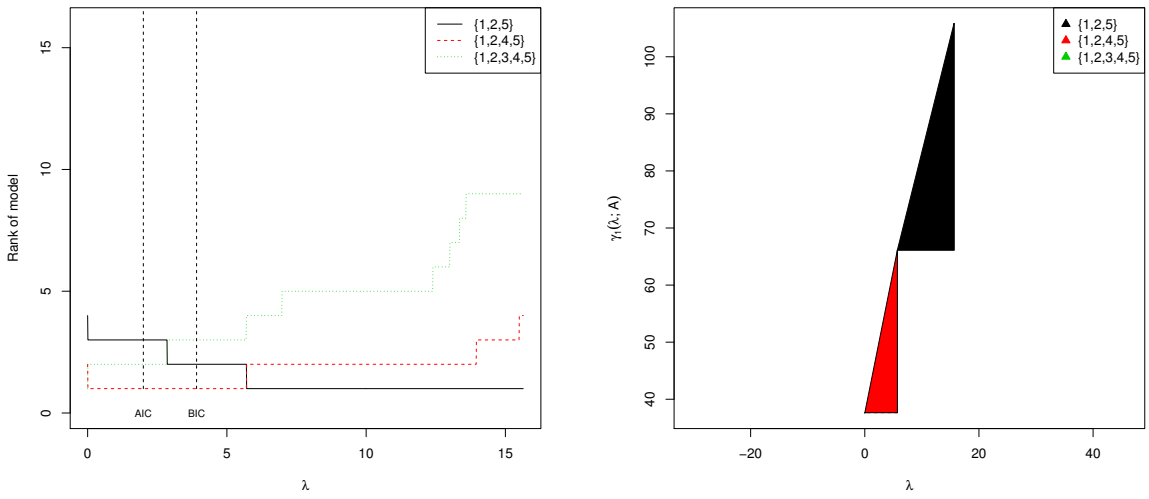


Figure 4.4: Plots from ordinary and stratified bootstrapping for data with low censoring (10%) when true Weibull AFT model is {1, 4, 5} and $n = 30$: (a) model detection plot from ordinary bootstrapping; (b) model detection plot from stratified bootstrapping; (c) variable inclusion plot from ordinary bootstrapping; (d) variable inclusion plot from stratified bootstrapping.

4.3. Note that all the five model selection criteria under the framework are presented in the figure. The model selection curve in Figure 4.3(a) shows all the models that have achieved rank 1 along with their catheti. Clearly, the specified model {1, 4, 5} has the longest cathetus and thus is selected by LCC. This is also evident in Figure 4.3(b). LHC and TAC also select the specified model {1, 4, 5} because of its longest



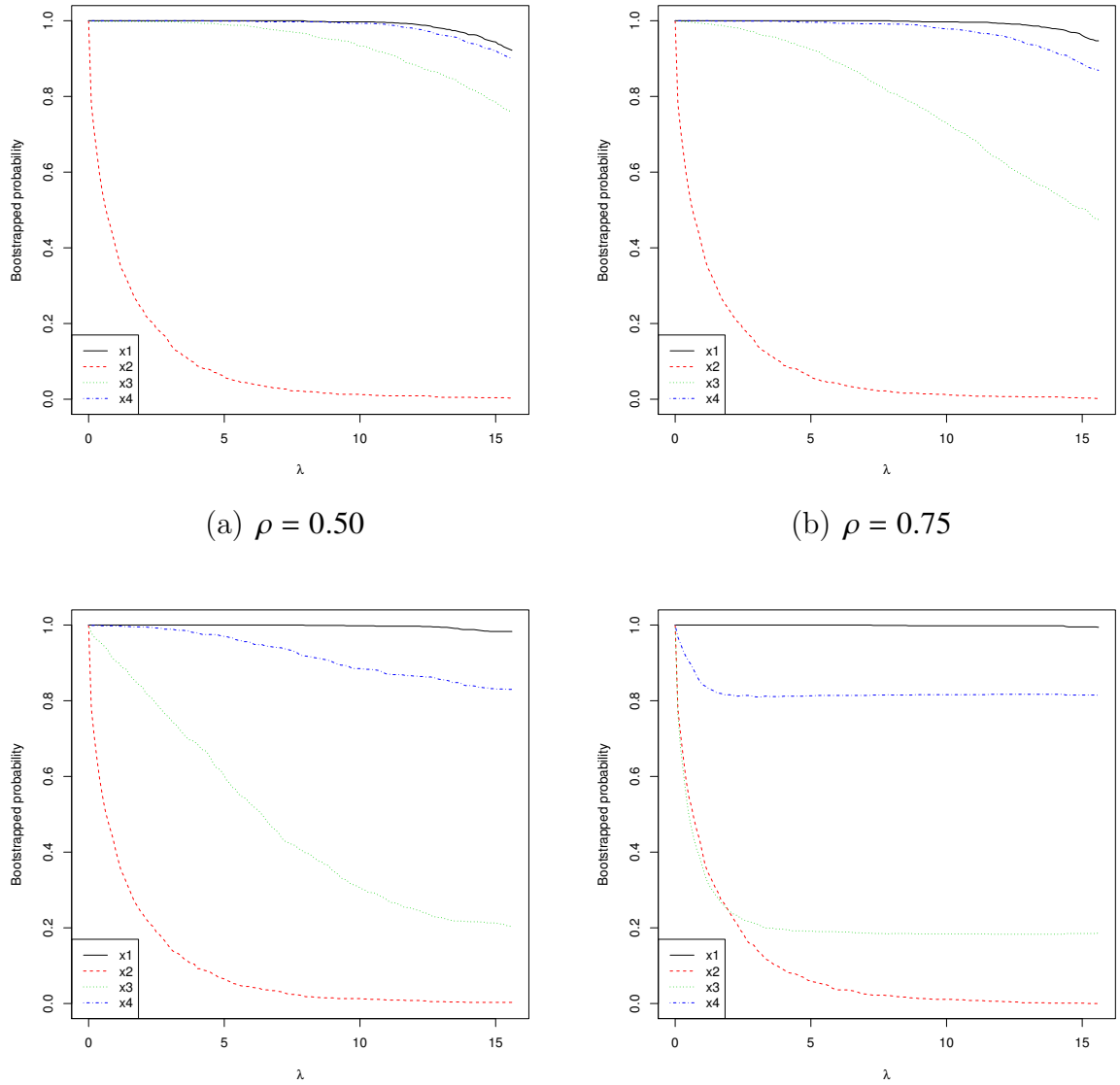
(a) (b)

Figure 4.5: Rank and TAC plots for data ($n = 50$ and 50% censoring) with correlated covariates (x_3 and x_4) when the true Weibull AFT model is {1, 2, 4, 5}: (a) rank plot; (b) TAC plot.

hypotenuse and largest triangle area (see Figure 4.3(c) and 4.3(d)).

We have also considered using bootstrap replications for the case presented in Figure 4.2(c). Both ordinary and stratified bootstrap samples, 1,000 each, are generated to investigate the likelihood that LCC correctly identifies the true model. Based on bootstrapping both the model detection plots and variable inclusion plots are obtained and shown in Figure 4.4. It is evident that ordinary and stratified bootstrapping give similar model detection plots and variable inclusion plots. The true model {1, 4, 5} is highly likely to be selected as the area under the curve corresponding to this model is the largest among candidate models as shown in Figure 4.4(a) and 4.4(b). Moreover, the variable inclusion plots in Figure 4.4(c) and 4.4(d) show that covariates x_3 and x_4 are highly likely to be included in the model. This agrees with the fact that covariates x_1 and x_2 are not included in the specified model.

Now let us consider a survival data with strongly correlated covariates. Suppose x_3 and x_4 in equation (4.1) are strongly correlated ($\rho = 0.90$). The data with sample size of 50 was generated from the AFT model {1, 2, 4, 5}. The results from applying the model selection framework to this data is summarised graphically in Figure 4.5. It is seen in this figure that TAC, LCC and LHC select the same model, model {1, 2, 5} that



(a) $\rho = 0.50$ (b) $\rho = 0.75$

(c) $\rho = 0.90$ (d) $\rho = 0.99$

Figure 4.6: Variable inclusion plot for data ($n = 50$ and 50% censoring) with correlated covariates at various levels ($\rho = 0.5,\ 0.75,\ 0.9$ and 0.99).

includes only x_1 and x_4. This model does not include the covariate x_3 that is in the specified model and also highly correlated with x_4. However, both AIC and BIC select the specified model $\{1,\ 2,\ 4,\ 5\}$ containing the two highly correlated covariates. In order to understand the effects of the correlation between covariates, $1{,}000$ bootstrap samples have been generated to construct the variable inclusion plot shown in Figure 4.6 for each of four samples with two correlated covariates of various correlation coefficients. It is clear from the figure that, as the strength of correlation increases, one of the two

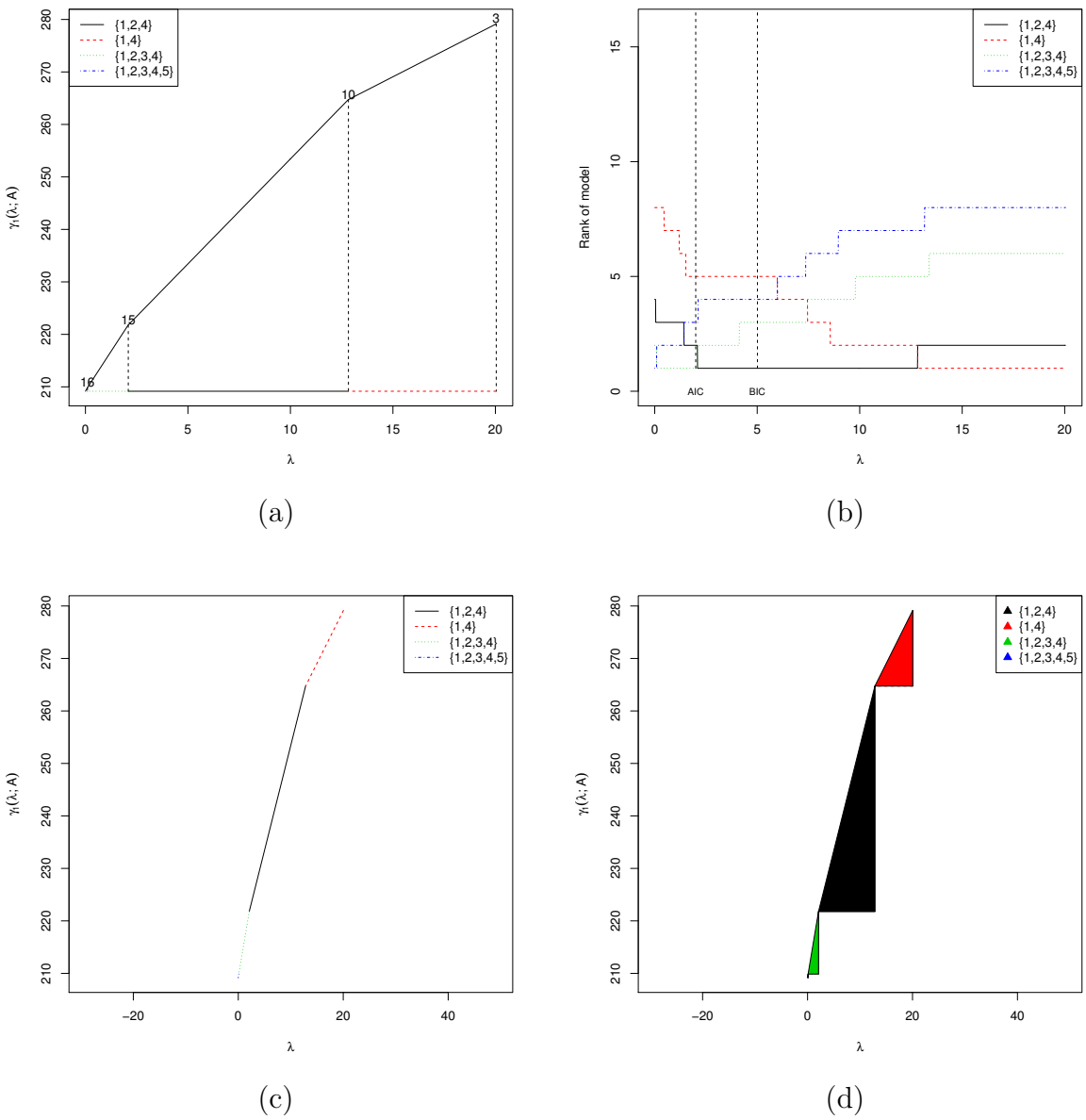


Figure 4.7: Plots from applying model selection framework for Weibull AFT model of data with continuous and categorical covariates at high censoring (50%) and $n = 150$ when true model is {1, 2, 4}: (a) model selection curve and cathetus lengths; (b) rank plot; (c) hypotenuse plot; (d) TAC plot.

highly correlated predictors is becoming less likely to be included in the model. This means that model to be selected may only include one of the two highly correlated covariates.

Now we consider Weibull AFT model with both continuous (or binary) and categorical variables with more than two levels as shown in the model given by equation (4.2). The true model is specified using the coefficients (0.1, 0.7, 0.5, 0, 0, 1, 0), i.e.,

model {1, 2, 4}. The results from applying model selection framework for this data is graphed in Figure 4.7. It is seen that models {1, 2, 4}, {1, 4}, {1, 2, 3, 4} and {1, 2, 3, 4, 5} have appeared in the model selection curve due to achieving rank 1 with respect to GIC values (Figure 4.7(a)). Although the true model is selected if using LCC and BIC, it is not selected if using AIC, as shown in the rank plot (Figure 4.7(b)). The longest hypotenuse in Figure 4.7(c) and triangle with the largest area in Figure 4.7(d) correspond to the specified model. Accordingly, the specified Weibull AFT model is selected if using LHC and TAC.

Note that if data are generated based on a specified full model, all model selection criteria under the AFT model selection framework can identify it irrespective of sample size and censoring proportion. These results are not shown here. It should be pointed out that the graphical results, shown and discussed in this section, are based on only one sample simulated under the Weibull AFT model. Therefore, it is possible that if a different sample is used, the results may change considerably. To address this, similar samples are drawn repeatedly for hundred or more times, known as Monte Carlo simulation. This would provide more convincing results. This is done in the following section.

4.4 Performance of model selection criteria under the framework for AFT models

The performance of TAC, LHC and LCC in selecting AFT models for survival data has been investigated in this section using Monte Carlo simulation, which considers the data simulated repeatedly, in comparison to two commonly used model selection criteria, BIC and AIC. Simulated data sets have been generated and studied under many settings: three types of AFT models (Weibull, log-logistic and log-normal); several sets of model parameters (one for each true model considered); low (10%) and high (50%) censoring proportions; different sample sizes (30, 50, 150 and 300). Each simulated data here contains six columns: a column of survival times that are generated under a specified distribution (Weibull, log-logistic or log-normal), a column of censoring status, and four columns for almost uncorrelated covariates, unless otherwise stated.

For each combination of those AFT model types, true models, censoring proportions and sample sizes under each set-up listed in Table 4.2, one hundred simulation runs have been carried out and studied. To evaluate the performance of TAC, LHC and LCC in comparison to BIC and AIC, percentage of selecting the specified (true) model using each of these five model selection criteria under the framework proposed in Chapter 3 has been computed for each set of 100 simulation runs. These percentages are presented in Tables 4.3–4.14.

Set-up 1: Data based on moderately right-skewed distribution

The results in Table 4.3 are based on simulated data sets from each of two Weibull AFT models {1, 4, 5} and {1, 4}. The survival times are generated under these two models and from Weibull distribution with shape parameter value of 2 and scale parameter values that are computed based on the covariates. In Table 4.3, when the Weibull AFT model {1, 4, 5} is the true model, TAC identifies it highly frequently irrespective of sample size or censoring proportion. Only in few instances, the true model is not detected by TAC. Similar results are observed when LHC or LCC is used. For BIC, proportions of identifying the true model are also high across different sample sizes, but it has more instances of not detecting the true model than TAC, LHC and LCC. On the other hand, proportions of correctly identifying the true model are much lower if using AIC than any of the other criteria under the model selection framework.

Table 4.3: Proportions of identifying true Weibull ($\kappa = 2$) AFT model.

Censoring	n	{1, 4, 5}					{1, 4}				
		LCC	LHC	TAC	BIC	AIC	LCC	LHC	TAC	BIC	AIC
10%	30	0.90	0.93	0.93	0.88	0.60	0.98	0.97	0.97	0.86	0.56
	50	0.99	1.00	1.00	0.96	0.78	1.00	1.00	1.00	0.90	0.64
	150	1.00	0.99	1.00	0.93	0.72	1.00	0.98	1.00	0.93	0.65
	300	1.00	1.00	1.00	0.96	0.74	1.00	1.00	1.00	0.92	0.59
50%	30	0.91	0.89	0.91	0.81	0.66	0.95	0.93	0.95	0.73	0.52
	50	0.97	0.93	0.95	0.82	0.60	0.98	0.95	0.97	0.76	0.47
	150	1.00	1.00	1.00	0.94	0.71	1.00	1.00	1.00	0.93	0.61
	300	1.00	1.00	1.00	0.96	0.73	1.00	1.00	1.00	0.95	0.60

At both 10% and 50% censoring proportions, if the true model is {1, 4}, proportions of selecting the true model obtained using TAC, LHC or LCC are similar and close to

100% across all sample sizes. Although BIC has a higher proportion of selecting the true model than AIC, as expected, its proportion is generally much lower than that via TAC, LHC or LCC as shown in Table 4.3.

Another interesting observation is that the performance of AIC and BIC decreases when a smaller model is specified. The proportions have gone down remarkably when true model is {1, 4}, instead of the larger true model {1, 4, 5}, particularly for AIC. Such reduction in proportion of selecting the true model ranges from 7% to 22% approximately for AIC, and can be up to 10% for BIC across all sample sizes and censoring proportions considered. Also note that from the true model {1, 4, 5} to {1, 4}, the proportions for TAC, LHC and LCC tend to increase or remain 100% in all simulated instances.

It is also noted that as sample size gets larger the proportion of detecting the true model using TAC, LHC or LCC gets higher. This empirical result agrees with what we have mentioned in the Lemma 1 of chapter 3, where it is shown that these three model selection criteria are consistent.

Table 4.4: Proportions of identifying true log-logistic ($\xi = \frac{1}{0.35}$) AFT model.

Censoring	n	{1, 4, 5}					{1, 4}				
		LCC	LHC	TAC	BIC	AIC	LCC	LHC	TAC	BIC	AIC
10%	30	0.93	0.90	0.91	0.77	0.63	0.97	0.95	0.95	0.75	0.56
	50	0.97	0.96	0.97	0.86	0.68	0.98	0.95	0.97	0.79	0.56
	150	1.00	1.00	1.00	0.95	0.76	1.00	1.00	1.00	0.94	0.66
	300	1.00	1.00	1.00	0.97	0.71	1.00	1.00	1.00	0.95	0.62
50%	30	0.90	0.93	0.94	0.78	0.64	0.96	0.94	0.94	0.76	0.61
	50	0.97	0.97	0.97	0.86	0.63	0.96	0.96	0.96	0.85	0.52
	150	1.00	1.00	1.00	0.94	0.73	1.00	0.99	1.00	0.89	0.59
	300	1.00	1.00	1.00	0.98	0.73	1.00	1.00	1.00	0.95	0.66

The proportions of selecting true model for log-logistic AFT model are presented in Table 4.4. The survival times are again generated under two different models {1, 4, 5} and {1, 4} from log-logistic distribution with shape parameter value of $\frac{1}{0.35}$ and varying scale parameter values. At both 10% and 50% censoring proportions, when the true model is {1, 4, 5}, the proportions of selecting the true model using TAC, LHC and LCC are very high across all sample sizes. The proportions for these three criteria vary between 90% and 100%. BIC also has high proportions of selecting the true model {1,

4, 5} although lower than those by LCC, LHC or TAC. These proportions obtained via AIC are the smallest among all the five criteria. Similar pattern is upheld when the true model is {1, 4}.

Note that the proportions of identifying the true model using BIC and AIC at both 10% and 50% censoring have gone down when the true model is smaller. For example, if the true model is reduced from {1, 4, 5} to {1, 4}, the proportions for BIC decrease between 1% and 8% approximately; and for AIC they decrease even more, between 5% and 19% approximately. However, there is almost no reduction in performance for TAC, LHC and LCC at both censoring proportions when the smaller model is specified.

The proportions of identifying the true model for each of the criteria studied do not differ much across censoring proportions at a specific sample size. Except for AIC, the proportions get larger as sample size increases, particularly for BIC. Also as n increases, the proportions of identifying the true model for TAC and LCC approach to unity.

Table 4.5: Proportions of identifying true log-normal ($\sigma = 0.64$) AFT model.

Censoring	n	{1, 4, 5}					{1, 4}				
		LCC	LHC	TAC	BIC	AIC	LCC	LHC	TAC	BIC	AIC
10%	30	0.97	0.91	0.94	0.76	0.62	0.98	0.92	0.93	0.74	0.54
	50	0.99	0.97	0.98	0.82	0.61	0.99	0.97	0.98	0.76	0.49
	150	1.00	1.00	1.00	0.96	0.70	1.00	1.00	1.00	0.93	0.55
	300	1.00	1.00	1.00	0.97	0.66	1.00	1.00	1.00	0.96	0.57
50%	30	0.89	0.88	0.91	0.81	0.63	0.93	0.92	0.93	0.75	0.54
	50	0.99	0.96	0.99	0.89	0.65	0.99	0.97	0.98	0.84	0.62
	150	1.00	1.00	1.00	0.93	0.72	1.00	1.00	1.00	0.91	0.57
	300	0.99	0.99	0.99	0.97	0.71	0.99	0.99	0.99	0.96	0.62

The proportions of selecting a true model in the case of log-normal AFT model are presented in Table 4.5. The survival times are generated from log-normal distribution with scale of 0.64 (σ) and varying mean values. At both 10% and 50% censoring, when the true model is {1, 4, 5} or {1, 4}, TAC, LHC and LCC can identify the true model in 90% or higher proportions of all runs for all sample sizes (small to large). BIC also has high proportions of selecting the true model at these censoring proportions. However, the proportions of identifying the true model are the smallest if using AIC.

When the true model is {1, 4} or {1, 4, 5}, the proportions of selecting the true model are reasonably similar for TAC, LHC and LCC but are much lower for BIC

and AIC. This difference in the proportions ranges between 1% and 7% approximately for BIC, and between 5% and 21% approximately for AIC across the two censoring proportions.

Based on the above discussion of results in Tables 4.3–4.5, the proportion of selecting the true model for moderately skewed survival data using TAC, LHC or LCC is generally higher or at least same as BIC; often much better than BIC and far exceeds AIC. This indicates these three MSC based model selection criteria TAC, LHC and LCC have outperformed both AIC and BIC in selecting the true model across all sample sizes at the low or high censoring proportions considered.

Set-up 2: Data based on severely right-skewed distribution

From Figure 4.1(a), we can see that the Weibull density is severely right-skewed for this set-up. The results in Table 4.6 are based on simulated data from two Weibull AFT models {1, 4, 5} and {1, 4}. The survival times are generated under these two models from Weibull distribution with shape parameter value of 1 and varying scale parameter values. We see in the table that at both 10% and 50% censoring when true model is {1, 4, 5} or {1, 4} and sample size is large ($n = 150$ or $n = 300$), proportions of selecting the true model obtained using TAC, LHC, LCC or BIC are all very high (> 90%). However, for small sample size such as $n = 30$, the proportions for all these criteria are considerably low, disregarding the censoring proportion in the data.

For small sample size, the proportions of selecting the true model {1, 4, 5} obtained using LCC is the lowest among all criteria (Table 4.6). It seems that the severely right-skewed distribution of the data is likely the reason for LCC not to perform as well as it usually does. For these severely skewed cases, LHC or TAC may substantially improve the performance of selecting the true model over LCC. The proportions obtained using LHC are closer to those of obtained using BIC.

Table 4.6: Proportions of identifying true Weibull ($\kappa = 1$) AFT model.

Censoring	n	{1, 4, 5}					{1, 4}				
		LCC	LHC	TAC	BIC	AIC	LCC	LHC	TAC	BIC	AIC
10%	30	0.24	0.40	0.35	0.55	0.49	0.53	0.60	0.61	0.78	0.54
	50	0.65	0.74	0.73	0.85	0.75	0.85	0.88	0.87	0.89	0.64
	150	1.00	1.00	1.00	0.93	0.66	1.00	1.00	1.00	0.92	0.60
	300	1.00	1.00	1.00	0.90	0.62	0.99	0.98	0.99	0.86	0.50
50%	30	0.35	0.49	0.44	0.57	0.57	0.52	0.58	0.59	0.60	0.49
	50	0.57	0.70	0.64	0.72	0.60	0.79	0.80	0.81	0.72	0.47
	150	0.99	0.99	0.99	0.94	0.71	1.00	1.00	1.00	0.93	0.61
	300	1.00	1.00	1.00	0.96	0.73	1.00	1.00	1.00	0.95	0.60

The proportions of selecting the true log-logistic AFT model, based on data generated under log-logistic distribution with shape parameter value of $\frac{1}{0.69}$ and varying scale parameter values are presented in Table 4.7. At both 10% and 50% censoring, when the sample size is large ($n = 150,\ 300$), the proportions of identifying the true model using TAC, LHC, LCC or BIC are high, disregarding if the true model is {1, 4, 5} or {1, 4}. Moreover, the proportions for data with low censoring and small sample

size (n = 50 or 30) are higher than those at high censoring with small samples. The proportions obtained using AIC are relatively low. Proportions for TAC, LHC and LCC are relatively high and close to those if using BIC for small sample size and when the smaller model {1, 4} is true. At 50% censoring and when sample size is very small (n = 30), the proportion of selecting true model {1, 4, 5} using LCC is the lowest among all model selection criteria considered (Table 4.7). However, this proportion is higher if LHC or TAC is used. This scenario is also observed if the true model is {1, 4}.

Table 4.7: Proportions of identifying true log-logistic ($\xi = \frac{1}{0.69}$) AFT model.

Censoring	n	{1, 4, 5}					{1, 4}				
		LCC	LHC	TAC	BIC	AIC	LCC	LHC	TAC	BIC	AIC
10%	30	0.61	0.67	0.65	0.67	0.57	0.75	0.76	0.77	0.71	0.49
	50	0.83	0.86	0.86	0.83	0.71	0.91	0.94	0.94	0.80	0.59
	150	0.99	0.99	0.99	0.95	0.74	1.00	0.99	1.00	0.92	0.66
	300	1.00	1.00	1.00	0.97	0.76	1.00	1.00	1.00	0.97	0.67
50%	30	0.37	0.48	0.44	0.60	0.58	0.60	0.65	0.64	0.72	0.61
	50	0.66	0.78	0.74	0.82	0.62	0.79	0.84	0.83	0.84	0.52
	150	1.00	1.00	1.00	0.94	0.73	1.00	0.99	1.00	0.89	0.59
	300	1.00	1.00	1.00	0.98	0.73	1.00	1.00	1.00	0.95	0.66

For the log-normal AFT models, data are generated from the log-normal distribution with scale σ = 1.28 and varying mean values. Proportions for identifying the true model are computed and reported in Table 4.8. The proportions for small sample size (n = 30, 50) have similar patterns to those reported above for log-logistic models (see Table 4.7). Although the proportions of selecting the true model {1, 4, 5} or {1, 4} obtained using TAC, LHC or LCC are very high (100% or close) at 10% or 50% censoring for data with large sample size (n = 150 or 300), and higher than those for BIC. However, those proportions at high censoring with small sample size are considerably low (Table 4.8). The lowest proportions are obtained if LCC is used in identifying the true model. Note that higher proportions are obtained if using other MSC based criteria, LHC and TAC.

From the results in Table 4.6–4.8, we see that LCC may not perform very well when dealing with data containing extreme observations, especially when the sample size is

Table 4.8: Proportions of identifying true log-normal (σ = 1.28) AFT model.

Censoring	n	{1, 4, 5}					{1, 4}				
		LCC	LHC	TAC	BIC	AIC	LCC	LHC	TAC	BIC	AIC
10%	30	0.42	0.60	0.52	0.72	0.55	0.56	0.67	0.64	0.73	0.53
	50	0.80	0.87	0.85	0.84	0.61	0.88	0.89	0.88	0.83	0.52
	150	1.00	0.99	1.00	0.95	0.67	1.00	0.99	1.00	0.91	0.58
	300	0.99	0.99	0.99	0.96	0.64	0.99	0.99	0.99	0.95	0.46
50%	30	0.29	0.42	0.38	0.50	0.52	0.51	0.56	0.54	0.61	0.50
	50	0.47	0.60	0.58	0.76	0.64	0.78	0.82	0.81	0.81	0.61
	150	1.00	1.00	1.00	0.93	0.72	1.00	1.00	1.00	0.91	0.57
	300	0.99	0.99	0.99	0.97	0.71	0.99	0.99	0.99	0.96	0.62

small. However, using LHC or TAC may improve, to some extent, this situation. The performance of identifying the true model using LHC or TAC in this special case is thus advisable.

Set-up 3: Data based on nearly symmetric distribution

A Weibull density with this set-up is close to symmetric as shown in Figure 4.1(c). The proportions in Table 4.9 are based on simulated data from two Weibull AFT models {1, 4, 5} and {1, 4}. The survival times are generated under these two models from Weibull distribution with shape parameter value of 3 and varying scale parameter values. From Table 4.9, one can see that the proportions of selecting the true model {1, 4, 5} or {1, 4} using TAC, LHC and LCC are high in all combinations of censoring proportions and sample sizes. BIC, on the other hand, is not performing as well particularly for small samples, but increasing censoring proportion does not seem to impair its performance. However, when smaller model {1, 4} instead of bigger model {1, 4, 5} is true, proportions obtained using BIC are relatively low. These proportions for AIC are much lower than those obtained using other model selection criteria.

We can see that as the sample sizes get larger, the proportions of identifying the true model using TAC, LHC and LCC also get larger, close to unity (Table 4.9). This is because these criteria are consistent in selecting the true model.

The proportions of selecting the true model in Table 4.10 are based on simulated data from two log-logistic AFT models {1, 4, 5} and {1, 4}. The survival times are generated under these two different models from log-logistic distribution with shape

Table 4.9: Proportions of identifying true Weibull ($\kappa = 3$) AFT model.

Censoring	n	{1, 4, 5}					{1, 4}				
		LCC	LHC	TAC	BIC	AIC	LCC	LHC	TAC	BIC	AIC
10%	30	0.95	0.95	0.95	0.83	0.66	0.94	0.92	0.93	0.77	0.61
	50	0.99	0.99	0.99	0.92	0.64	0.99	0.98	0.99	0.90	0.51
	150	1.00	1.00	1.00	0.94	0.77	1.00	0.99	1.00	0.93	0.68
	300	1.00	1.00	1.00	0.98	0.77	1.00	1.00	1.00	0.97	0.63
50%	30	0.96	0.95	0.95	0.83	0.66	0.97	0.94	0.97	0.73	0.52
	50	0.99	0.95	0.97	0.82	0.60	0.99	0.96	0.98	0.76	0.47
	150	1.00	1.00	1.00	0.94	0.71	1.00	1.00	1.00	0.93	0.61
	300	1.00	1.00	1.00	0.96	0.73	1.00	1.00	1.00	0.95	0.60

parameter value of $\frac{1}{0.23}$ and varying scale parameter values. At both **10%** and **50%** censoring, when the true model is {1, 4, 5} or {1, 4}, the proportions of selecting the true model using TAC, LHC and LCC are very close to each other at each sample size. The proportions obtained using BIC are also high, but the proportions are considerably low if using AIC.

Table 4.10: Proportions of identifying true log-logistic ($\xi = \frac{1}{0.23}$) AFT model.

Censoring	n	{1, 4, 5}					{1, 4}				
		LCC	LHC	TAC	BIC	AIC	LCC	LHC	TAC	BIC	AIC
10%	30	0.94	0.94	0.94	0.83	0.64	0.95	0.90	0.94	0.74	0.53
	50	1.00	0.99	0.99	0.91	0.69	0.99	0.94	0.97	0.83	0.55
	150	1.00	1.00	1.00	0.96	0.73	1.00	0.99	0.99	0.94	0.56
	300	1.00	1.00	1.00	0.97	0.73	1.00	1.00	1.00	0.95	0.61
50%	30	0.97	0.96	0.97	0.78	0.64	0.97	0.95	0.95	0.76	0.61
	50	0.97	0.97	0.97	0.86	0.63	0.96	0.96	0.96	0.85	0.52
	150	1.00	1.00	1.00	0.94	0.73	1.00	0.99	1.00	0.89	0.59
	300	1.00	1.00	1.00	0.98	0.73	1.00	1.00	1.00	0.95	0.66

The proportions of identifying the true model in Table 4.11 are based on simulated data from two log-normal AFT models {1, 4, 5} and {1, 4}. The survival times are generated under these models from log-normal distribution with scale of **0.43** (σ) and varying μ values. In Table 4.11, the pattern of proportions of selecting the true model based on the model selection criteria under the framework are similar to the pattern described for Tables 4.9–4.10.

Based on the results for Weibull, log-normal and log-logistic AFT models as shown

Table 4.11: Proportions of identifying true log-normal ($\sigma = 0.43$) AFT model.

Censoring	n	{1, 4, 5}					{1, 4}				
		LCC	LHC	TAC	BIC	AIC	LCC	LHC	TAC	BIC	AIC
10%	30	0.97	0.97	0.97	0.74	0.54	0.97	0.89	0.96	0.73	0.44
	50	0.98	0.97	0.97	0.85	0.63	0.99	0.98	0.98	0.79	0.48
	150	1.00	1.00	1.00	0.97	0.70	1.00	1.00	1.00	0.94	0.54
	300	1.00	1.00	1.00	0.95	0.71	1.00	1.00	1.00	0.94	0.63
50%	30	0.95	0.95	0.95	0.81	0.63	0.95	0.93	0.95	0.75	0.54
	50	1.00	0.96	0.99	0.89	0.65	0.99	0.97	0.98	0.84	0.62
	150	1.00	1.00	1.00	0.93	0.72	1.00	1.00	1.00	0.91	0.57
	300	0.99	0.99	0.99	0.97	0.71	0.99	0.99	0.99	0.96	0.62

in Tables 4.9–4.11, we see that TAC, LHC and LCC have clearly outperformed BIC and AIC in most of the cases, irrespective of sample size and censoring proportions, when the data is less skewed or close to symmetric. Since TAC, LHC and LCC perform fairly similarly, especially when the data is less chaotic such as the current set-up (set-up 3), any of them can be used for choosing a suitable model.

Set-up 4: Data based on moderately left-skewed distribution

The results in Table 4.12 are based on simulated data from two Weibull AFT models {1, 4, 5} and {1, 4}. The survival times are generated under these two models from Weibull distribution with shape parameter value of 4 and varying scale parameter values. In Table 4.12, when the Weibull AFT model {1, 4, 5} or {1, 4} is true, TAC, LCC and LHC can identify it with very high proportions (> 94%) irrespective of sample size and censoring proportion. Although the proportions of identifying the true model using BIC are also high across different sample sizes, it cannot detect the true model as often as TAC, LHC and LCC can. On the other hand, the proportions of correctly identifying the true model using AIC are the lowest of all, irrespective of sample size and censoring proportion.

Each of the MSC based criteria TAC, LHC and LCC, for moderately left-skewed data has outperformed both AIC and BIC in identifying the true model irrespective of sample size and censoring proportion as observed from Table 4.12. Moreover, we can see as the sample size gets larger the proportions of detecting the true model using TAC, LHC and LCC get larger too. This empirical result agrees with what we have

Table 4.12: Proportions of identifying true Weibull ($\kappa = 4$) AFT model.

Censoring	n	{1, 4, 5}					{1, 4}				
		LCC	LHC	TAC	BIC	AIC	LCC	LHC	TAC	BIC	AIC
10%	30	0.96	0.94	0.94	0.84	0.60	0.97	0.90	0.94	0.80	0.48
	50	0.97	0.97	0.97	0.85	0.70	0.99	0.96	0.98	0.81	0.60
	150	0.99	0.98	0.99	0.93	0.66	1.00	0.99	0.99	0.91	0.62
	300	1.00	1.00	1.00	0.92	0.68	1.00	1.00	1.00	0.92	0.56
50%	30	0.97	0.95	0.96	0.82	0.65	0.97	0.94	0.97	0.73	0.52
	50	0.98	0.94	0.95	0.81	0.59	0.99	0.96	0.98	0.76	0.47
	150	1.00	1.00	1.00	0.94	0.71	1.00	1.00	1.00	0.93	0.61
	300	1.00	1.00	1.00	0.96	0.73	1.00	1.00	1.00	0.95	0.60

mentioned in the Lemma 1 in Chapter 3, where it is shown that these criteria are consistent.

Table 4.13: Proportions of identifying true log-logistic ($\xi = \frac{1}{0.17}$) AFT model.

Censoring	n	{1, 4, 5}					{1, 4}				
		LCC	LHC	TAC	BIC	AIC	LCC	LHC	TAC	BIC	AIC
10%	30	0.97	0.95	0.96	0.85	0.60	0.96	0.94	0.94	0.78	0.51
	50	0.99	0.98	0.98	0.87	0.63	0.99	0.96	0.98	0.85	0.56
	150	0.99	0.98	0.99	0.97	0.71	0.99	0.98	0.99	0.96	0.65
	300	1.00	1.00	1.00	0.98	0.69	1.00	1.00	1.00	0.98	0.58
50%	30	0.97	0.96	0.97	0.78	0.64	0.97	0.95	0.95	0.76	0.61
	50	0.97	0.97	0.97	0.86	0.63	0.96	0.96	0.96	0.85	0.52
	150	1.00	1.00	1.00	0.94	0.73	1.00	0.99	1.00	0.89	0.59
	300	1.00	1.00	1.00	0.98	0.73	1.00	1.00	1.00	0.95	0.66

The proportions of selecting true log-logistic AFT model are presented in Table 4.13. The survival times are generated under two different models {1, 4, 5} and {1, 4} from log-logistic distribution with $\xi = \frac{1}{0.17}$ and varying scale parameter values. At both 10% and 50% censoring, when the true model is {1, 4, 5}, the proportions of selecting the true model using TAC, LHC or LCC are very high, close to 1, across all sample sizes. BIC also has high proportions of selecting the true model {1, 4, 5} although much lower than those by TAC, LHC or LCC. The proportions obtained using AIC are the smallest among all as usual. Similar pattern is seen when {1, 4} is the true model.

Note that the proportions of identifying the true model using BIC and AIC at both 10% and 50% censoring are lower if the true model is smaller. For example, if the

true model is reduced from {1, 4, 5} to {1, 4}, the proportions for BIC decrease up to 8% approximately; and for AIC they decrease even more, between 5% and 19% approximately. However, changes in proportions, when smaller model is true, are very little for TAC, LHC or LCC across all sample sizes at both censoring proportions.

Table 4.14: Proportions of identifying true log-normal ($\sigma = 0.32$) AFT model.

Censoring	n	{1, 4, 5}					{1, 4}				
		LCC	LHC	TAC	BIC	AIC	LCC	LHC	TAC	BIC	AIC
10%	30	0.99	0.98	0.99	0.83	0.70	0.98	0.97	0.97	0.81	0.61
	50	1.00	0.99	1.00	0.92	0.72	1.00	0.99	1.00	0.89	0.64
	150	1.00	1.00	1.00	0.92	0.73	1.00	1.00	1.00	0.92	0.65
	300	1.00	0.99	1.00	0.99	0.72	1.00	0.99	0.99	0.97	0.61
50%	30	0.96	0.95	0.95	0.81	0.63	0.94	0.93	0.94	0.75	0.54
	50	1.00	0.96	0.99	0.89	0.65	0.99	0.97	0.98	0.84	0.62
	150	1.00	1.00	1.00	0.93	0.72	1.00	1.00	1.00	0.91	0.57
	300	0.99	0.99	0.99	0.97	0.71	0.99	0.99	0.99	0.96	0.62

A large number of survival times are also generated from log-normal distribution with scale of 0.32 (σ) and varying mean values under log-normal AFT models {1, 4, 5} and {1, 4}, similarly as discussed before. The proportions of selecting a true model based on samples from the simulated data are computed and presented in Table 4.14. When the true model is {1, 4, 5} or {1, 4}, TAC, LHC and LCC can identify the true model almost in all cases irrespective of sample size and censoring proportion. BIC has also moderately high proportions of selecting the true model. However, the proportions of identifying the true model are the smallest if using AIC. Another observation from the Table 4.14 is that when the true model is {1, 4} instead of {1, 4, 5}, the proportions of selecting the true smaller model are similar for TAC, LHC, LCC and BIC, but are much lower for AIC.

From the above discussion and based on the results in Tables 4.12–4.14, each of the criteria, TAC, LHC and LCC outperforms both AIC and BIC in selecting the true model across all sample sizes and censoring proportions considered in this study. Note that TAC always lies in between LCC and LHC as expected. Moreover, it upholds its conservative nature among these three MSC based criteria and is, therefore, advisable for the AFT model selection under the proposed framework.

Overall, the MSC based criteria, TAC, LHC and LCC, are consistent in selecting

the true model irrespective of censoring proportion and sample size. They can also outperform other criteria if the data is not severely chaotic. In the next chapter, the model selection framework has been implemented through constructing a user-friendly tool in R.

5

A Tool in R for AFT Model Selection

A model selection framework for the AFT models of survival data with right censoring has been established in the previous chapters of this thesis, based on the MSC approach (Müller and Welsh, 2010; Murray, Heritier and Müller, 2013). To visualise the results from applying the framework, a user-friendly tool has been developed in the statistical computing project R. This tool consists of several functions, and it is explained and illustrated with three published data sets: the ovarian cancer data (Edmunson et al., 1979), the lung cancer data (Lawless, 1982) and the Stanford heart transplant data (Crowley and Hu, 1977). Note that R packages, including `survival` and `mplot` have been utilised in writing up the R codes for implementing the model selection framework for AFT models.

5.1 Functions and programs developed in R

In this section, the functions and programs are described, which are developed to construct a tool in R program for the AFT model selection framework proposed in Chapter 3. The constructed R tool is then trialled and illustrated using several examples.

5.1.1 Functions in the R tool

Several main functions developed for the R tool are discussed in details in this section. Among these functions, `aftmsc` is the most important one as it executes all the required calculations of the model selection, provides arguments to be used in the plot functions for visualisation. The description of a few important sub-functions under the main functions can be found in the Appendix B.

aftmsc: `aftmsc` is the principal function as it executes all the required calculations. A call to this function is

```
aftmsc(formula, data, distn='weibull', lambda).
```

The arguments in this function are summarised as follows:

- `formula`: an expression like in other regression models, which consists of response variable and predictors. The response lies on the left of the ~ operator and the predictors lie on the right of that operator, separated by + operators. The response must be a `survival` object obtained from the `Surv` function in the `survival` package. See the documentation for `Surv`, `lm` and `formula` in R for more details.

- `data`: a data frame for full or subset of data to be specified. The name of the variables in the data frame should be consistent with variables named in the `formula` argument. If all the variables in `formula` are defined individually, `data` in `aftmsc` can be left unspecified.

- `distn`: assumed distribution for log of survival time. The default distribution is set to '`Weibull`'. However, other distributions such as '`exponential`' (special case of '`Weibull`'), '`log-normal`', '`log-logistic`', '`Gaussian`' and '`logistic`' can be

assumed. Any distribution accepted by the `survreg.distributions` function in the `survival` package can be utilised here.

- `lambda`: a sequence of values or a single value can be assigned to `lambda`, which is the penalty multiplier in a model selection procedure. The default value for `lambda` is a sequence between 0 and $4\log(n)$ with a step size of 0.01, where n is the number of observations (or rows) in a data frame. A step size of 0.01 is used to balance the computational time and accuracy of the process.

The `aftmsc` function returns an object of class "`msc`", which includes several items, such as a table of 1 rank models with their frequencies of appearing on the model selection curve, a table of models according to LHC and TAC, a table containing numerical values of cathetus lengths, hypotenuse lengths and $\sqrt{TAC}$, etc.

There are four main functions for visualising the results of the model selection process: `plot_cat`, `plot_msc`, `plot_hyp` and `plot_tac`.

plot_cat: `plot_cat` gives the catheus plot, which is equivalent to the plot of ranks of models that achieve rank 1. It can be used as follows:

`plot_cat(object, location="topright", inset=0, defaults_number)`.

The arguments in the function are described as below:

- `object`: an object of class "`msc`", obtained from a call to the `aftmsc` function.

- `location`: a position to be specified for the legend to appear on the cathetus plot. The default position is "`topright`". However, it can be changed to other positions such as "`topleft`", "`right`", "`center`", "`left`" etc. See the document style of `legend` in `R` for more details.

- `inset`: a numerical value used to adjust the position of the legend box inside a graphical frame. See the document style of `legend` in `R` for more details.

- `defaults_number`: a numerical value that indicates how many models to be shown on the plot from the set of models appearing on the 1 rank model selection curve. If nothing is specified, the plot will by default show all models that have appeared on the model selection curve.

The plot_msc, plot_hyp and plot_tac are used almost identically to plot_cat except that the default position of the legend in the plot_msc is "topleft". The plot_msc function returns the plot of model selection curve and shows the model number on the curve and models on the legend.

The plot_hyp function returns the plot of hypotenuses, which are segments on the model selection curve. On this plot, a model corresponding to each segment is shown on the legend in the order of the highest to lowest length of hypotenuse.

The plot_tac function returns the TAC plot, i.e., plot of the areas of triangles. Each triangle represents a model that achieves rank 1 and appears on the model selection curve. The legend of this plot shows the models ordered from largest to smallest area of triangles (TAC values).

Finally, print.msc and plot.msc are two generic functions for object of class "msc". The print.msc is the print method for class "msc". This function summarises important information from the object of class "msc". It can be called by

```
print(object).
```

On the other hand, plot.msc displays rank plot, model selection curve plot, hypotenuse plot and TAC plot from the model selection framework, numbered 1 to 4 respectively. By default, all four plots are provided. A call to this function can be

```
plot(object, defaults_number, which=c(1L,2L,3L,4L), ask =
  prod(par("mfcol")<length(which)) && dev.interactive()).
```

The first two arguments are described earlier. The which in the plot function is a numerical argument, and is used if a subset of the plots is required. The ask is a logical argument such that if it is TRUE, the user is asked before producing each plot (see the documentation of par(ask=·) in R for more details).

There are also a few sub-functions in this tool, to help with some computations, such as DSM and surv.gic. These sub-functions are considered very important in the Monte Carlo simulation and bootstrapping technique. The descriptions for these and sample R codes that we wrote and used in this thesis can be found in Appendix B and Appendix C.

5.1.2 Programs in the R tool

The functions described in this chapter are designed for graphical presentation of the model selection framework for AFT models. Besides this, several programs were written to carry out Monte Carlo simulation, which were mainly used for comparing the performance of model selection criteria considered in the framework (see Section 4.4). Moreover, programing codes were also written for carrying out bootstrap replications to help to get additional information and assist with the model selection process as suggested in Section 3.4 and Section 4.2. These codes can also be found in Appendix C.

5.2 Examples

In this section, our R tool for AFT model selection framework is trialled and illustrated using several published data sets mentioned at the beginning of this chapter.

5.2.1 Ovarian cancer data

A set of ovarian cancer data from a study by Edmunson et al. (1979) is considered here. In this study, anti-tumour effects of two different forms of chemotherapy following the surgical treatment of ovarian cancer were compared. Only 26 patients were included in the study. The response variable was survival time in days (T), and the covariates considered were age in years (x_1), residual disease (incomplete or complete) (x_2), treatment (single or combined) (x_3) and performance status (good or poor) (x_4). This data is available in an R package "survival" under the name "ovarian". Using Weibull AFT model, Collett (2003) analysed the data and suggested a final model for the data. Here, we have also used Weibull AFT model for this data to illustrate the model selection framework.

The output from applying the functions in R tool to the ovarian cancer data is presented below. Any line starting with the prompt sign (>) indicates R code and the relevant output is beneath the prompt sign.

```
> out=aftmsc(Surv(futime,fustat)~.,data=ovarian)
> out
```

```
   Model.num cat.length hyp.length    TAC.sqrt Intercept age ... ecog.ps
1          5      10.55  23.590517 10.55000000         1   1  ... 0
2         15       1.74   7.174204  2.46073160         1   1  ... 0
3         10       0.69   2.181972  0.84507396         1   1  ... 0
4         16       0.05   0.254951  0.07905694         1   1  ... 1
```

```
> plot_msc(out)
```

```
   Model.num       Model cat.length Intercept age resid.ds rx ecog.ps
1          5       {1,2}      10.55         1   1        0  0       0
2         15   {1,2,3,4}       1.74         1   1        1  1       0
3         10     {1,2,4}       0.69         1   1        0  1       0
4         16 {1,2,3,4,5}       0.05         1   1        1  1       1
```

```
> plot_cat(out)
```

```
 Model.num       Model cat.length Intercept age resid.ds rx ecog.ps
1          5       {1,2}      10.55         1   1        0  0       0
2         15   {1,2,3,4}       1.74         1   1        1  1       0
3         10     {1,2,4}       0.69         1   1        0  1       0
4         16 {1,2,3,4,5}       0.05         1   1        1  1       1
```

```
> plot_hyp(out)
```

```
    Model.num       Model hyp.length Intercept age resid.ds rx ecog.ps
1          5       {1,2}  23.590517         1   1        0  0       0
2         15   {1,2,3,4}   7.174204         1   1        1  1       0
3         10     {1,2,4}   2.181972         1   1        0  1       0
4         16 {1,2,3,4,5}   0.254951         1   1        1  1       1
```

```
> plot_tac(out)
```

```
   Model.num       Model    TAC.sqrt Intercept age resid.ds rx ecog.ps
1          5       {1,2} 10.55000000         1   1        0  0       0
2         15   {1,2,3,4}  2.46073160         1   1        1  1       0
3         10     {1,2,4}  0.84507396         1   1        0  1       0
4         16 {1,2,3,4,5}  0.07905694         1   1        1  1       1
```

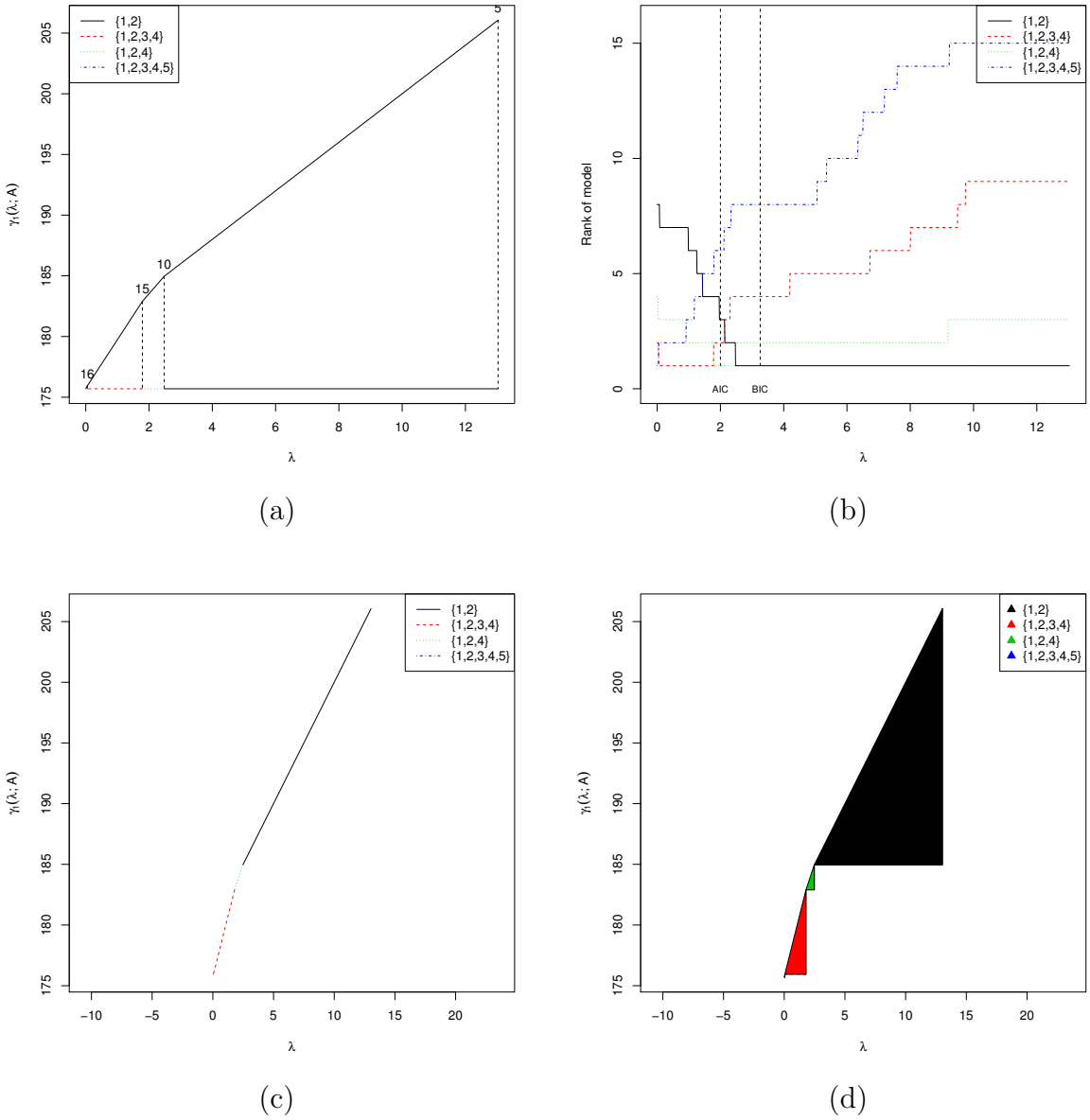


Figure 5.1: Model selection framework for the AFT model with ovarian data: (a) model selection curve with the model number; (b) rank plot based on 1 rank models; (c) hypotenuse plot; (d) TAC plot.

After executing the code mentioned, numerical summaries of important measures of model selection framework are obtained. Moreover, the plots that are generated by running the four plot functions are given in Figure 5.1. It is obvious from the output above that the model {1, 2} is selected by all the three model selection criteria LCC, LHC and TAC based on the MSC. This model was also the final model suggested by Collett (2003). BIC selected the same model, but AIC did not.

It was mentioned before that bootstrapping could be used to provide additional

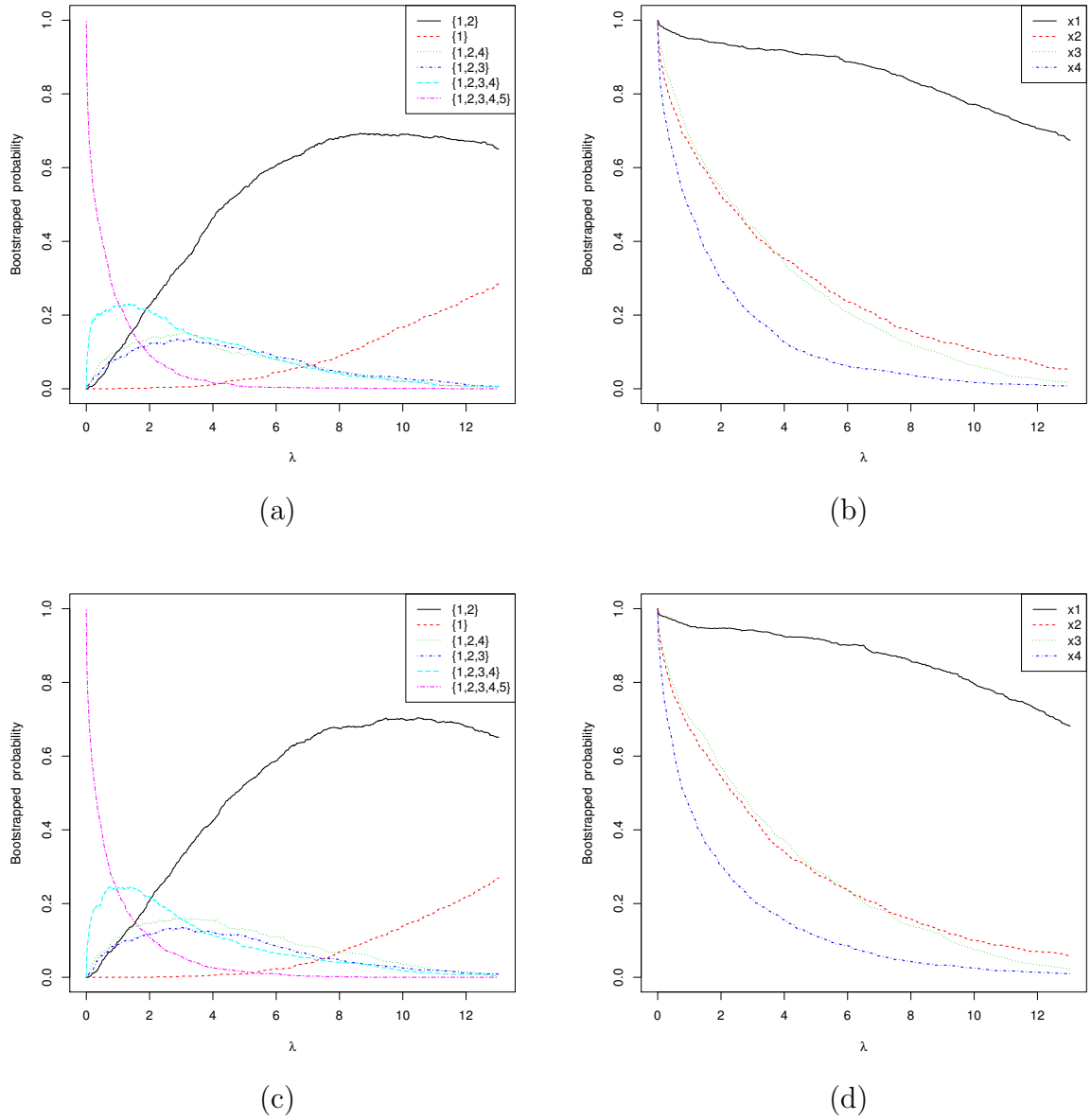


Figure 5.2: Plots based on bootstrap of 1,000 replications: (a) model detection plot (ordinary bootstrap); (b) variable inclusion plot (ordinary bootstrap); (c) model detection plot (stratified bootstrap); (d) variable inclusion plot (stratified bootstrap).

information to enhance the model selection process. As an illustration, based on the ordinary and stratified bootstrapping technique respectively, corresponding model detection and variable inclusion plots are presented in Figure 5.2. The plots that based on ordinary and stratified bootstrap replications look very similar for this data. As mentioned before, the bootstrap procedure has not been included in the R tool developed.

5.2.2 Lung cancer survival data

Survival data on 137 advanced lung cancer patients were initially reported and discussed by Prentice (1973). The data include survival time and censoring status, as well as several covariates. A subset of 40 patients from this data was presented and analysed in Lawless (1982). The subset consisted of all patients who received prior therapy and were then randomly assigned to one of the two chemotherapeutic agents, i.e., treatments, labelled as 'standard' and 'test'. In this subset of data, there were 21 patients who received 'standard' treatment and 19 patients received the other treatment 'test'. The goal was to compare the effects of chemotherapeutic treatment on prolonging survival time, while accounting for possible effects due to other covariates. In other words, a suitable model was sought to represent the relations between survival time and the covariates under the study.

One of the covariates in the data is tumour cell type with four types, squamous, small, adeno, and large. Another covariate is performance status at the time of diagnosis, which is a measure of general medical status on a scale of 10, 20, 30, ..., 90. Here the scores 10, 20, and 30 indicate that the patient has been completely hospitalised; 40, 50, and 60 indicate that the patient has been partially confined to hospital; and 70, 80, and 90 indicate that the patient is able to care for self. Other covariates are patient's age and time in months from the diagnosis of lung cancer to the entry into the study. Lawless (1982) used both exponential and Weibull AFT models to analyse this subset data of 40 patients. Similar conclusions were drawn for both models in his analysis, suggesting that performance status was very important to be considered in the model.

Here we consider fitting Weibull AFT model for the same subset of lung cancer data used by Lawless, which includes performance status (x_1), age (x_2), time in months from diagnosis of lung cancer to entry into the study (x_3), treatment (test vs standard) (x_4), and tumour cell type (x_5). Note that covariates considered here not only include continuous and binary variables but also a variable (x_5) with more than two categories. We have applied the model selection framework for AFT models to this data, to choose a final model. Since five covariates are considered, there are 32 possible models including intercept only model and the model with all five covariates. Note that we are not considering interactions terms in the models being considered in this thesis.

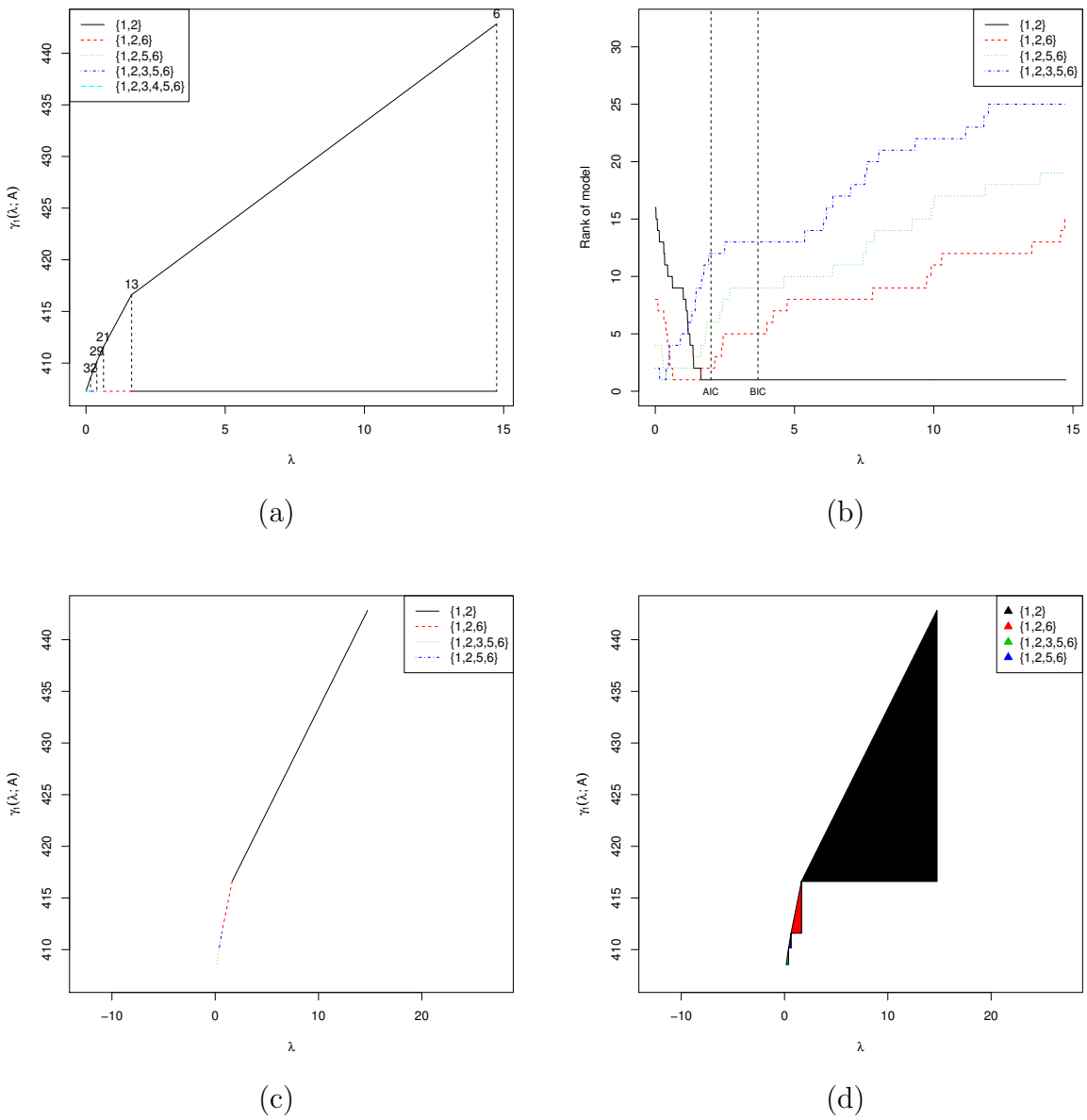


Figure 5.3: Model selection framework for the AFT model with lung cancer data: (a) model selection curve with the model number; (b) rank plot based on 1 rank models; (c) hypotenuse plot; (d) TAC plot.

Main graphical outputs obtained from applying the model selection framework to the lung cancer survival data are presented in Figure 5.3. We can see that 5 out of 32 possible models have achieved rank 1 and appeared on the model selection curve in Figure 5.3(a). However, only the top four models are shown in the other plots in Figure 5.3 to limit the clutter on these graphs. The model with performance status only, model {1, 2}, has the longest cathetus as shown in Figure 5.3(b), and is thus selected according to LCC. If LHC is used, same model {1, 2} is selected as shown in

Figure 5.3(c). Also the same model {1, 2} is identified by TAC (see Figure 5.3(d)), since the triangle for model {1,2} has the largest area among all models. Note that AIC and BIC, each is represented by a single point on the model selection curves (see Figure 5.3(b)), select the same model as the three MSC based criteria in this instance.

Table 5.1: Summary statistics extracted from model selection framework for AFT model with lung cancer survival data.

Model α	p_α	C_{L_α}	H_{L_α}	$\sqrt{TAC}$	$\pi^*(\alpha)$
{1,2,3,4,5,6}	6	0.16	0.97	0.28	0.03
{1,2,3,5,6}	5	0.23	1.72	0.36	0.01
{1,2,5,6}	4	0.24	0.99	0.34	0.03
{1,2,6}	3	1.01	3.19	1.24	0.04
{1,2}	2	13.11	29.31	13.11	0.69

Summary statistics, from applying the AFT model selection framework to the lung cancer survival data, are presented in Table 5.1. The first five columns of this table are obtained from the analysis of the subset lung cancer survival data via the `R` tool. Obviously, the same conclusion can be drawn as each of the model selection criteria, LCC, LHC, and TAC, achieves the maximum value corresponding to model {1, 2} and thus this model is chosen as the best model. Note that the cathetus length C_{L_α} and $\sqrt{TAC}$ value for the chosen model are identical as expected. This is because this model, selected by both LCC and TAC, has dimension two, i.e., $p_\alpha = 2$ (see the explanation in Section 3.3). Moreover, $\sqrt{TAC}$ values shown in Table 5.1 lie in between LCC and LHC values $\left(\sqrt{TAC} \in (C_{L_\alpha}, H_{L_\alpha})\right)$. This supports the conservative nature of TAC as a model selection criterion, which was also reported in our simulation study in Section 4.4.

Now let's run 1,000 ordinary bootstrap replications of the lung cancer survival data ($n = 40$). From this, the marginal probabilities of selecting model α, $\pi^*(\alpha)$, obtained using LCC are reported in the last column of Table 5.1. Clearly, model {1, 2} is confirmed as the best model for this data set since it occurs most frequently in the 1,000 bootstrap replications with a high $\pi^*(\alpha)$ value (see Table 5.1). Besides the five models presented in column 1 of Table 5.1, other models also appeared on the model selection curves for some of the bootstrapped samples, but we did not report here as its $\pi^*(\alpha)$ values are generally very small. This explains why $\sum \pi^*(\alpha) \neq 1$ in Table 5.1.

5.2.3 Stanford heart transplant data

The Stanford heart transplant data on 103 patients was reported by Crowley and Hu (1977) who analysed the data using piecewise exponential models. Later, Aitkin, Laird and Francis (1983) reanalysed this survival data to explore parametric models (e.g., Weibull AFT). We obtained a subset of the Stanford heart transplant data from R package SMIR (Aitkin et al., 2012), which only contains data on 65 post-transplant patients, studied by Crowley and Hu (1977). In this subset, the survival time was measured in days after transplant (surv) and censoring status (an indicator variable) is given in the variable 'died' with 1 for 'yes' and 0 for 'no'. Other variables in the data are: age of the patient at time of transplantation (age); whether a patient has prior open-heart surgery (surg); the number of mismatches alleles between donor and recipient (nmm); hla that is a dichotomous variable with 1 if the donor has the antigen HLA-A2 and the recipient has neither HLA-A2 nor the similar HLA-A28, and 0 otherwise; mismatch score (mm) representing mismatch between the patient's and donor's tissue type; time of acceptance into the program (acc); and an indicator for death by rejection (rej). There were also other variables in the original data, however, they are not considered in our analysis because of its limited importance on the prognosis of survival for the transplanted patients.

Following the discussion by Aitkin, Laird and Francis (1983) in which the same data was analysed, we have applied Weibull AFT model to the same data. Here we consider four covariates age (x_1), surg (x_2), acc (x_3) and hla (x_4). In this data on 65 patients, there are 41 deaths and 24 censored survivals. The data has a censoring proportion of almost 37%. The Weibull model for this data showed a declining hazard, also known as monotonically decreasing hazard, as the estimated shape parameter of the Weibull distribution is below 1 (Aitkin, Laird and Francis, 1983). In this circumstance (i.e., declining hazard), the Weibull distribution assumed for this data is considered heavily right-skewed. For data with such a chaotic nature, our simulation study in Section 4.4 showed that LCC could not perform as good as other criteria, and thus we suggested to use LHC and TAC instead, which performs very closely to BIC in identifying a specified (or true) model (see Section 4.4).

Given four covariates considered in the Weibull AFT model, there are 16 possible models, including from the intercept only model to the model with all four covariates.

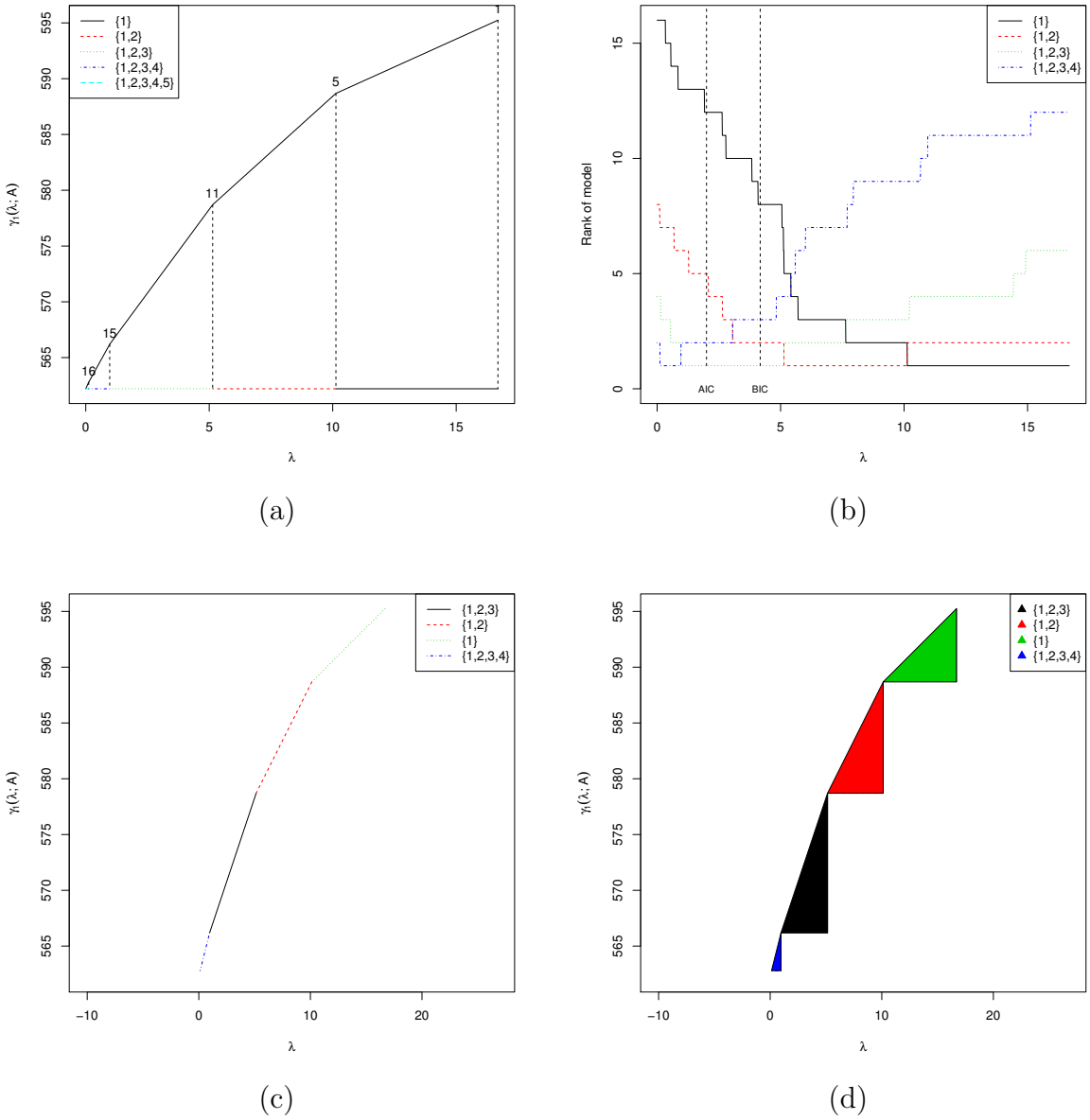


Figure 5.4: Model selection framework for the AFT model with Stanford heart transplant data: (a) model selection curve with the model number; (b) rank plot based on 1 rank models; (c) hypotenuse plot; (d) TAC plot.

Important outputs from our analysis of the data are shown in Figure 5.4 and summarised in Table 5.2. Out of the **16** possible models, **5** models have achieved rank 1 and thus appeared in the model selection curve (Figure 5.4(a)), however, only the top four models are shown in the Figure 5.4(b)–5.4(d) to limit the clutter on these graphs. From Figure 5.4(b), we can see that the intercept only model {1} having the longest cathetus is selected by LCC (see also $\boldsymbol{C}_{L_\alpha}$ column in Table 5.2). According to LHC and TAC, the model with covariates age and surgery, model {**1**, **2**, **3**}, is selected as

shown in Figure 5.4(c) and Figure 5.4(d). Note that the same model is also selected if the model selection criteria, AIC and BIC, are used.

Table 5.2: Summary statistics extracted from MSC for heart transplant data.

Model α	p_α	C_{L_α}	H_{L_α}	$\sqrt{TAC}$	$\pi^*(\alpha)$
{1,2,3,4,5}	5	0.12	0.61	0.19	0.03
{1,2,3,4}	4	0.85	3.50	1.20	0.06
{1,2,3}	3	4.17	13.19	5.11	0.18
{1,2}	2	4.99	11.16	4.99	0.30
{1}	1	6.56	9.28	4.64	0.29

The results of the model selection procedure can also be summarised numerically. Numerical results from the AFT model selection for the heart transplant data are presented in Table 5.2. The first five columns of this table are obtained from the application of the R tool to the data. We see that $\sqrt{TAC}$ and LHC reach maximum value by the model {1, 2, 3}. It is notable that the variation in the $\sqrt{TAC}$ value is the smallest among the three MSC based criteria within the framework.

We have also run 1,000 bootstrap replications of the heart transplant data. The marginal probabilities of selecting model α, $\pi^*(\alpha)$, obtained using LCC are reported in the last column of Table 5.2. As we mentioned earlier that for heavily skewed data, LCC may not perform well, we are thus expecting these marginal probabilities provide some additional insight. According to these, models {1, 2}, {1} or {1, 2, 3} may be considered.

6

A Case Study Using the Model Selection Framework

The model selection framework for the AFT model was illustrated and investigated in simulations as well as several published examples of survival data in the previous chapters. In this chapter, we apply this framework to some survival data on ovarian cancer, which was recently obtained from the Royal Prince Alfred (RPA) hospital, Sydney, Australia, aiming to identify and select important prognostic factors from a number of variables available in the data. This is part of a collaborative project with researchers at the hospital.

6.1 Background

Ovarian cancer is the leading cause of gynaecological cancer related deaths in many countries of the world. In general, women with ovarian cancer have poor prognosis

with short survival. Winter et al. (2008) estimated an overall median survival of 29 months among patients with advanced ovarian cancer.

The prognosis is worse when the cancer is diagnosed at a late stage, which is quite common for ovarian cancer as there are not many clear symptoms for the cancer until tumour has metastasised. Therefore, it is very important to investigate and identify key prognostic factors of survival following a diagnosis of ovarian cancer. This may provide doctors with a better understanding of their patients' prognosis and potentially enable doctors to understand patients' survival prospects and thus treat them appropriately.

Ovarian cancer is not a homogeneous disease but rather a group of diseases, each with different morphology and biologic behaviour. There have been a large number of studies on ovarian cancer over the last few decades. Many of them aimed to predict or estimate survival rate and identify the prognostic factors associated with the survival for women with ovarian cancer, using various methods of survival analysis such as the Cox proportional hazards model. However, the results from those studies are not all that consistent. According to Colombo et al. (2009), the survival of patients with ovarian cancer at advanced stage is mainly influenced by a few factors, such as the biology and chemosensitivity of the tumour, the size of the residual disease as well as the extent of the disease at the time of diagnosis.

Ovarian cancer can be broadly divided into two main types, epithelial and non-epithelial ovarian cancer. Any ovarian cancer that started in the surface layer covering the ovary is called epithelial ovarian cancer and the rest is non-epithelial.

There are many factors that may be associated with the prognosis of ovarian cancer, including residual disease, histology, stage of cancer, grade of tumour, CA125 and age. Detailed definition for each of these variables/factors is given in Section 6.2. It is also of interest to identify possible change in survival prospect over time. Note that a set of covariates with a specified relationship with survival represents the true prognostic importance of each covariate, according to Burton et al. (2006).

A number of previous studies (for example, Polterauer et al., 2012; Tingulstad et al., 2003 etc) showed that residual disease is one of the most important prognostic factors. In a study of patients with advanced ovarian cancer, Colombo et al. (2009) found that patients with no residual disease after the debulking procedure, i.e., surgery, had experienced a better prognosis with a greater 5-year survival rate.

There are a number of different histologic types of tumour, including serous, mucinous, endometrioid, clear cell and other epithelial type. Women with mucinous and endometriod tumours had better 5-year survival than the other histologic types (Barnholtz-Sloan et al., 2003).

Ovarian cancer can also be classified by stages, known as FIGO stages. FIGO stage is another important prognostic factor reported in the previous studies (e.g., Obermair et al., 2013; Kotsopoulos et al., 2012; and Chan et al., 2006, etc). Different FIGO stages may indicate differing prognoses.

Grade of ovarian cancer was considered as an important prognostic variable for ovarian cancer by Eisenhauer et al. (1999) and also by Seiden (2001). Also, it was reported that older women (age > 60) experienced worse prognosis with a shorter survival following the diagnosis of ovarian cancer (Barnholtz-Sloan et al., 2003; Colombo et al., 2009).

CA125 (Cancer Antigen 125) was discovered in early 1980s (Bast et al., 1981). It has been considered as a biomarker to monitor epithelial ovarian cancer (Felder et al., 2014). It is usually found in a greater concentration in tumour cells than other part of the body. The plausibility of CA125 as a prognostic factor for ovarian cancer was evaluated in a number of previous studies. Markmann et al. (2007) found that the level of CA125 was correlated with overall survival. According to this study, level around 100 U/l is an indication of a bad prognosis. Nowadays, CA125 is routinely used to monitor the response to chemotherapy for patients with ovarian cancer (Schmidt, 2011).

A number of other prognostic factors of ovarian cancer were also investigated in the literature. Among them, the following prognostic factors are worth noting: performance status, physician's speciality, race, oestrogen receptor, progesterone receptor, HER-2/neu (c-erb-B_2), epidermal growth factor receptor (EGFR), p53, mitotic activity index, volume percentage epithelium, morphometric groups, marital status, height, weight, BMI, alkaline phosphatise (ALP), albumin, presence or absence of ascites, GST-pi etc.

When there are many potential prognostic factors of survival following ovarian cancer studied under a statistical method (model), it is likely that only a small subset of them is sufficient in describing or predicting the survival of ovarian cancer. A natural

query is which combination of prognostic factors better describes or predicts the survival sufficiently and thus should be included in the model, i.e., the final model. Model selection is thus an important part in the study of ovarian cancer data or any other data.

The Cox proportional hazards model has often been used in the literature for ovarian cancer survival, or cancer survival in general. Under certain distributions of the survival time, such as Weibull, the AFT model is also a proportional hazards model and thus may be used as a parametric alternative to the Cox model. An AFT model is used to analyse the ovarian cancer survival data using our model selection framework for AFT models.

6.2 The RPA data and variable definitions

As mentioned earlier, data used in this chapter was obtained from the RPA hospital, Sydney. An ethical application was submitted and approved by the relevant committee of the hospital. Then a formal request was made to get the information on clinical and other prognostic factors of the ovarian cancer for patients being treated in the hospital. Unfortunately, not all information requested were available in its hospital database of ovarian cancer. Cases and variables from a number of sources were merged and matched to get a reasonably complete data on those ovarian cancer patients.

6.2.1 Data

Data used here are confined to patients with epithelial ovarian cancer, which is the common and serene type of ovarian cancer. It counts for 85%–95% of patients with ovarian cancer (Roett and Evans, 2009). Considering possible reporting delay and problems in data acquisition and entry in some years, only patients who were diagnosed with epithelial ovarian cancer during the two decades between 1990 and 2009 were considered in this study. After remove patients without sufficient information, only 347 epithelial ovarian cancer patients were included. Although this only represents a subset of all patients treated in the RPA hospital over the study period, there is no reason to suggest that this subset is biased. Furthermore, the main focus in our analysis of this data is to show how well the AFT model selection works.

6.2.2 Variable definitions

Variables considered in the data include survival time, residual disease, histology, FIGO stages, grade, CA125, age and year of diagnosis, and they are defined as follows.

The survival time ($\boldsymbol{T}$) is defined as time from the diagnosis of ovarian cancer to death or last medical contact, whichever comes first. Note that a survival time is considered censored in this study if a patient did not have date of death recorded or did not die from ovarian cancer or was lost to follow up. As the record of one of the patients showed the same diagnosis date and date of death, one day was added to the survival time to all patients to get rid of zero survival time, and thus allow for logarithmic transformation when applying the AFT model.

Residual disease (RD) is defined as size of the residual tumour after surgery (i.e., debulking procedure). It is categorised into the following three groups in our analysis: microscopic (i.e., not visible to naked eyes), ≤ 1 cm and > 1 cm.

According to the histology of ovarian cancer tumour, there are several subtypes of epithelial ovarian cancer. In this study, it is classified into the following six groups: high-grade serous, low-grade serous, mucinous, endometrioid, clear cell and other epithelial type.

FIGO stage describes how advanced the cancer is and how far it has spread at the time of diagnosis for all types of cancer. For ovarian cancer, the staging is based on the location of cancer in the ovary and in other parts of body. There are four main FIGO stages, stage I–stage IV, where FIGO stages III and IV are known as advanced stage of ovarian cancer.

Grade of ovarian cancer is defined according to how similar (or different) the cancer cells are to normal cells. In this study, three broad grades, grade 1–grade 3 of epithelial ovarian cancer, are considered.

CA125 is a biomarker used for monitoring epithelial ovarian cancer. It is classified into two groups, normal and elevated. A CA125 level below 35 U/l is considered normal (see Markmann et al., 2007), otherwise elevated in our study.

Age is patient's age (in years) at diagnosis of ovarian cancer. It is a continuous variable. Like most of previous studies (e.g., Chan et al., 2006; Zhang et al., 2005), it is also considered in this study as a categorical variable of three groups: < 50, ≥ 50 but < 70, and ≥ 70.

Year of diagnosis in our data is between 1990 and 2009. In our study, it is grouped into two broad periods of 1990 to 1997 and 1998 to 2009. This grouping was made based on a preliminary analysis of the data.

6.3 Preliminary analysis of the RPA data

In this part, each variable considered in our analysis is explored, followed by an examination of survival patterns by each variable. The inter-relationships between possible prognostic factors of the survival are also investigated to indicate possible confounding effects.

6.3.1 Brief summary of survival times

More than 50% of survival times in the RPA data studied are censored, which is considered very high. Note that survival data with relatively high censoring was investigated in our simulation study in Section 4.4, where it was shown that our AFT model selection framework would work well even for survival data like this.

The Kaplan-Meier (K-M) survival curve for all 347 patients with a 95% confidence interval (illustrated with dotted lines) is presented in Figure 6.1, where short vertical bar indicates one or more observations censored. The estimated survival curve drops down sharply till about four years of survival, and then falls gradually. Overall, the chance of surviving 5 years or more is about 60%.

Let us assume that the survival times in the RPA ovarian cancer data follow a Weibull distribution with a pdf as expressed in Section 2.3. In order to justify the parametric Weibull assumption, we will take a closer look at its survivor and hazard functions given in equations (2.5) and (2.6) respectively. The $\log(-\log)$ of the estimated survivor function for Weibull distribution can be expressed as a linear function of log time:

$$\log\{-\log \hat{S}(t)\} = \hat{\kappa}\log\hat{\lambda} + \hat{\kappa}\log t,$$

where, $\hat{S}(t)$ is the Kaplan-Meier estimator of $S(t)$. A plot of $\log\{-\log \hat{S}(t)\}$ (log of cumulative hazard) against $\log(t)$ (log of survival times) for the RPA ovarian cancer data is presented in Figure 6.2. Since the plot appears approximately a straight line,

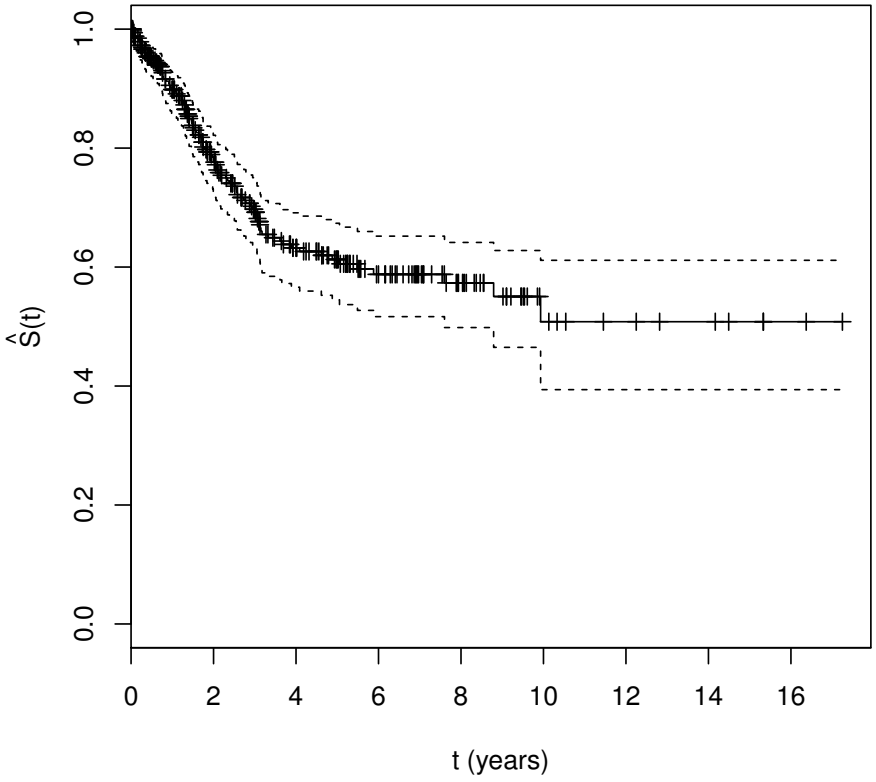


Figure 6.1: K-M survival curves for all 347 patients.

the Weibull AFT model would be a sensible choice for the data. Note that among the three commonly used distributions in AFT model, Weibull, log-normal and log-logistic, only under Weibull distribution, $\log\{-\log \hat{S}(t)\}$ is a linear function of $\log(t)$. Moreover, Weibull AFT is the only parametric survival model that has both a proportional hazards representation and an accelerated failure time representation as explained earlier. Parameters of assumed Weibull distribution for the RPA data can be estimated via fitting an intercept only Weibull AFT model to the data. The estimated shape and scale parameters from the RPA ovarian cancer data are about 0.812 ($\hat{\kappa}$) and 0.118 ($\hat{\nu}$) respectively. According to the results from our simulation study, LHC and TAC are expected to perform better than other criteria for data like this. As a comparison, we have still considered all three MSC based model selection criteria, as well as AIC and BIC, for this data by applying our constructed `R` tool for the AFT model selection framework.

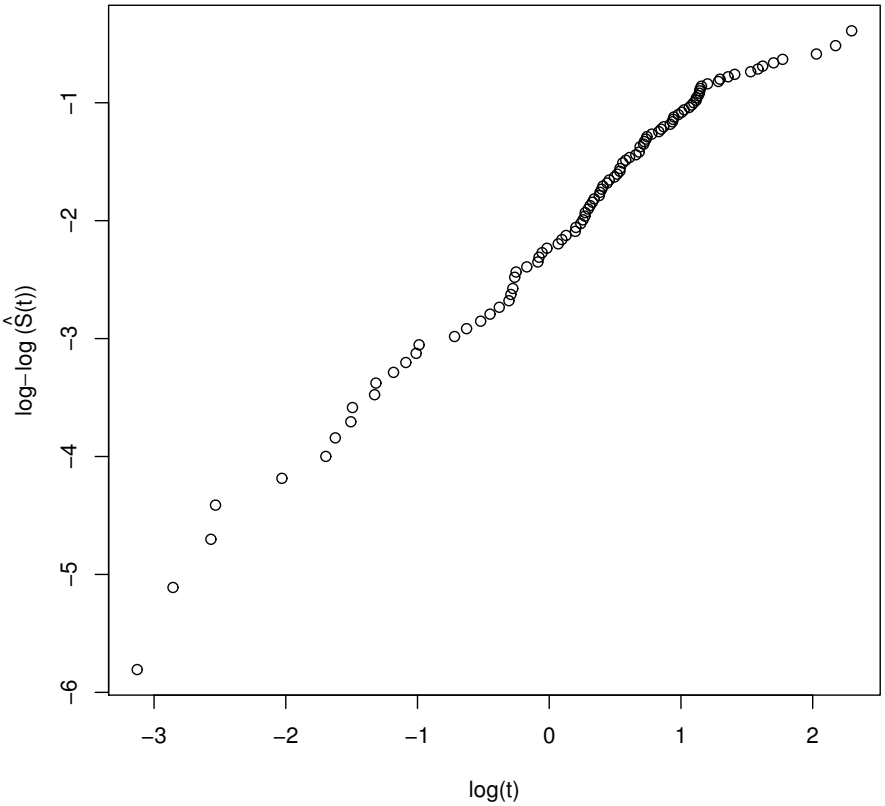


Figure 6.2: Log-cumulative hazard plot for the ovarian cancer patients studied.

6.3.2 Exploration of potential prognostic factors

As noticed, the RPA ovarian cancer data contains both continuous and categorical prognostic factors (variables). Most of the categorical variables in the data have more than two categories, so this data would be a nice example of survival data for us to demonstrate the features of our model selection framework as well as the `R` tool developed.

Summary statistics of the prognostic factors studied and group mean age are presented in Table 6.1. It can be seen that proportion of patients with residual disease (RD) $>$ 1 cm is three times of that who have RD $\leq$ 1 cm in the RPA ovarian cancer data although more than 50% patients belong to the microscopic group. Note that patients in group (RD $>$ 1 cm) are older on average than the other two residual disease groups.

Across the six histology groups studied, over 50% of patients belong to high-grade serous group, who are also oldest. With respect to FIGO stage, over 50% are classified

Table 6.1: Summary of prognostic factors for RPA ovarian cancer patients.

Factor	Levels	No. of patients	% distribution	Mean Age
Residual disease	Microscopic	190	54.76	53.10
	≤ 1 cm	38	10.95	55.75
	> 1 cm	119	34.29	60.45
Histology	High-grade serous	183	52.74	58.77
	Low-grade serous	28	8.07	49.84
	Mucinous	47	13.54	51.25
	Endometrioid	25	7.20	50.96
	Clear cell	23	6.63	54.91
	Other epithelial type	41	11.82	56.32
FIGO stage	Stage I	108	31.12	51.77
	Stage II	38	10.95	54.87
	Stage III	167	48.13	57.40
	Stage IV	34	9.80	63.06
Grade	Grade 1	71	20.46	51.47
	Grade 2	28	8.07	51.00
	Grade 3	248	71.47	57.76
CA125	Normal	58	16.71	52.92
	Elevated	289	83.29	56.53
Age group	Below 50	116	33.43	40.83
	50–69.99	169	48.70	59.00
	70 or above	62	17.87	75.78
Diagnosis year	1990–1997	228	65.71	56.47
	1998–2009	119	34.29	54.87

as advanced stages, i.e., Stage III (48.13%) and Stage IV (9.80%). The proportion of patients who belong to Stage II and Stage IV are similar around 10%. The mean age for patients within Stage III or IV is higher.

Majority of patients (71.47%) were classified as Grade 3 with a higher average age. Patients in Grade 1 or Grade 2 have similar age on average. With respect to the level of CA125, more than 80% of patients have CA125 elevated at the time of diagnosis of ovarian cancer. These patients are older on average when compared with the patients with normal level of CA125.

Almost 50% of the patients belong to the age group 50 to 69 in the RPA ovarian cancer data, about 33% were diagnosed with ovarian cancer before 50 years old and about 18% were diagnosed with the disease at 70 years or older. The proportion of younger ovarian cancer patients (below 50 years) is almost twice the proportion of older patients (70 years or above). With respect to diagnosis year, 65.71% were diagnosed in earlier years (1990–1997) and the rest of patients (34.29%) were diagnosed in later

years (1998–2009) in this data. Note that mean ages do not differ much between these two groups of patients.

6.3.3 Survival by each prognostic factor

Since the distribution of survival times is usually skewed, an appropriate summary measure of survival times is median survival time, which is often of interest in survival analysis. Median survival times can be approximated from Kaplan-Meier survival curves. The Kaplan-Meier survival curves for various prognostic factors are shown in Figure 6.3–Figure 6.6.

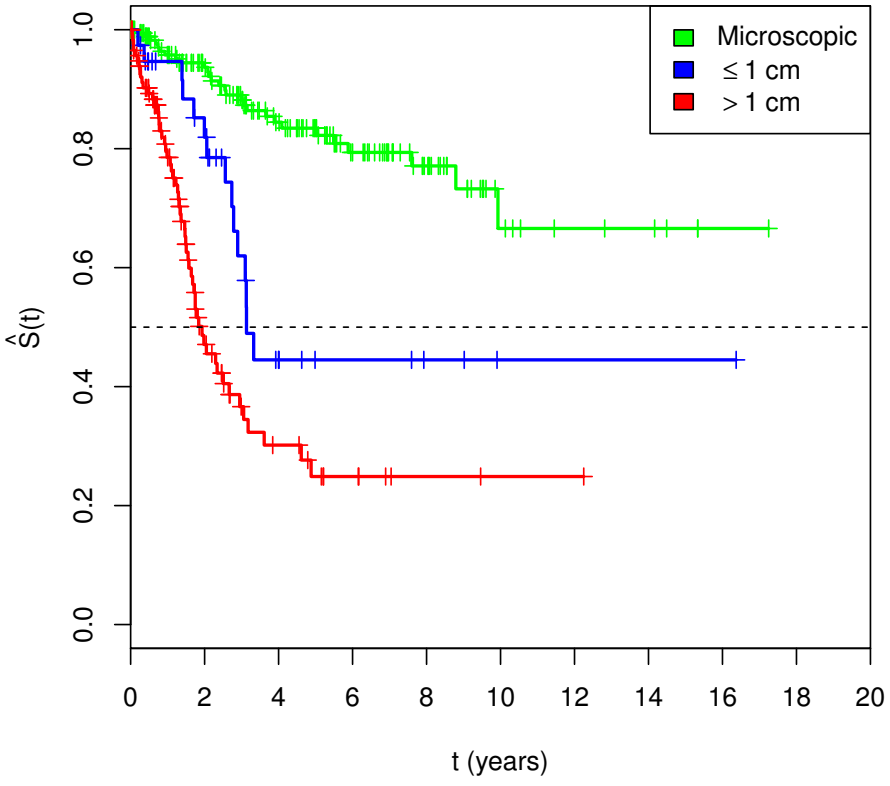
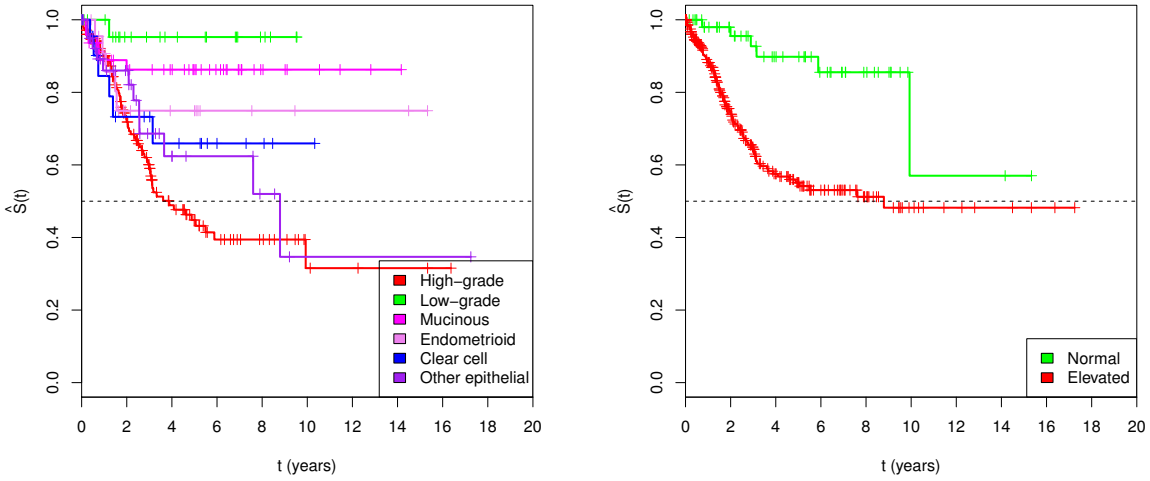


Figure 6.3: K-M curves by residual disease.

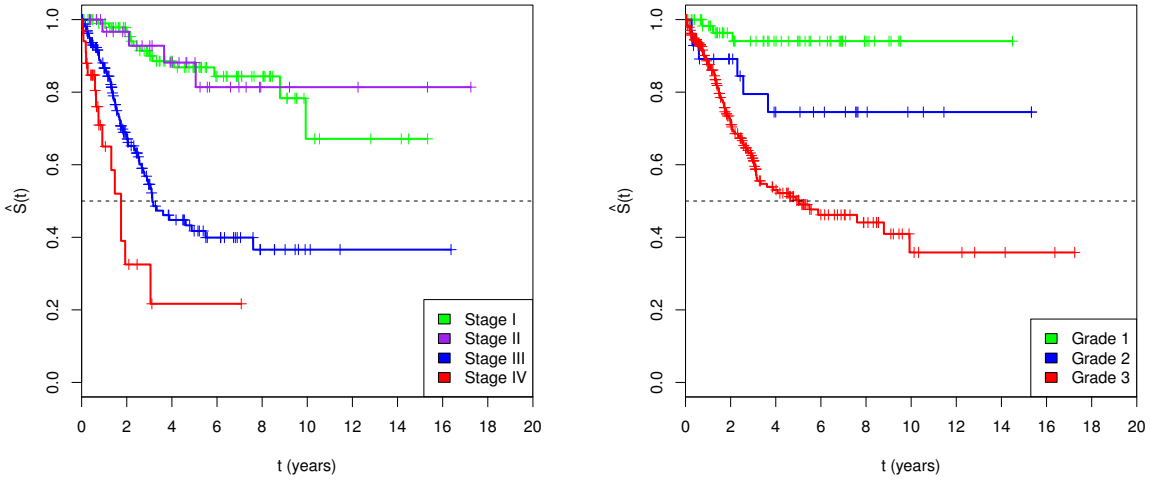
In Figure 6.3, the survival curve for patients with residual tumour greater than 1 cm falls much sharply compared to the survival curves for patients with smaller sizes of residual tumour. This group of patients has experienced relatively poorer survival with a median survival time of 1.9 years, while the median survival time for patients with residual tumour size less than or equal to 1 cm is around 2.5 to 3 years. The

microscopic group has a much better survival as its survival curve is well above the other two residual disease groups.



(a) (b)

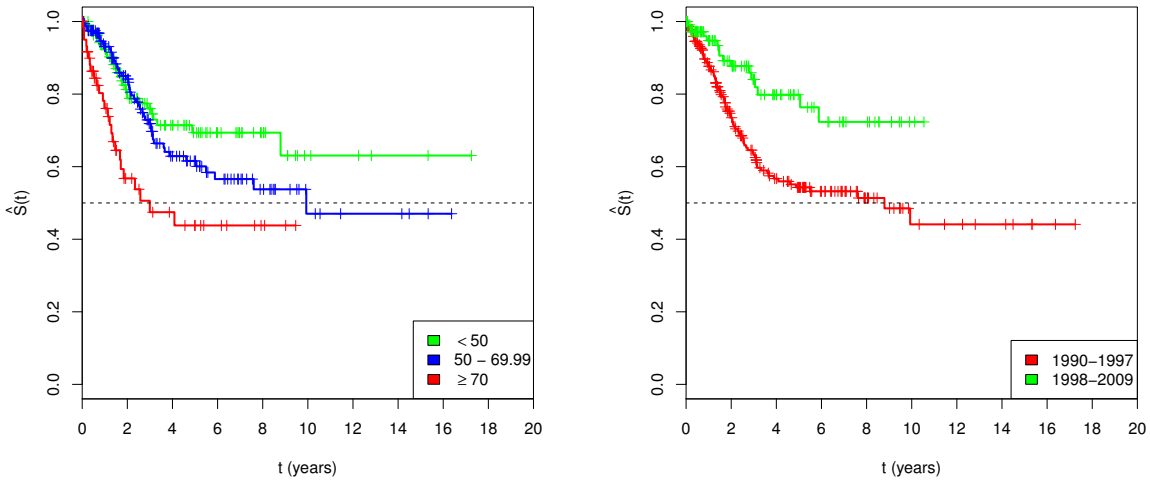
Figure 6.4: (a) K-M curves by histology; (b) K-M curves by CA125.



(a) (b)

Figure 6.5: (a) K-M curves by FIGO stage; (b) K-M curves by grade.

From Figure 6.4(a), patients with high-grade serous had experienced the worst survival among all the histology groups considered. The median survival time for this group is 3.9 years approximately. Patients with normal CA125 level had better survival



(a) (b)

Figure 6.6: (a) K-M curves for three age groups; (b) K-M curves for two diagnosis periods.

than those whose CA125 level were elevated (see Figure 6.4(b)) as expected.

The survival prospect of patients with FIGO stage I is similar to that for those with Stage II, and both these two groups have a much better survival than those patients at advanced stages (stages III and IV) as shown in Figure 6.5(a). Patients with FIGO stage IV had experienced the worst survival across the four stages with a median survival time of only 1.8 years. With respect to grade, patients with Grade 3 have experienced the worst survival among the three grades of ovarian cancer for patients included in this study (Figure 6.5(b)).

It can be seen from Figure 6.6(a) that patients diagnosed relatively younger (below 50 years) had better survival than older patients. Patients who were diagnosed at 70 years or older had experienced the worst survival as expected. With respect to diagnosis year, patients who were diagnosed after 1997 had experienced better survival than patients who were diagnosed earlier as seen in Figure 6.6(b). This may be, at least to some extent, due to improved treatment (surgery) scheme, and care and quality of treatment following the surgery.

6.3.4 Inter-relation between prognostic factors

To measure of the possible association between prognostic factors in RPA ovarian cancer data, chi-square test is used. P-values based on chi-square tests for each pair

of categorised prognostic variables are given in Table 6.2 only as an indication for association. Note that when expected cell frequencies in some cases were less than 5, Monte Carlo simulations are used to determine the p-values, marked with asterisk ($*$) in Table 6.2.

Table 6.2: Results from assessing associations between prognostic factors.

	FIGO	Histology	Grade	CA125	Age group	Diagnosis year
RD	$< 0.001^*$	$< 0.001^*$	$< 0.001^*$	< 0.001	< 0.001	0.018
FIGO		$< 0.001^*$	$< 0.001^*$	< 0.001	< 0.001	0.775
Histology			$< 0.001^*$	$< 0.001^*$	$< 0.001^*$	0.886
Grade				$< 0.001^*$	0.004	0.751
CA125					0.176	0.429
Age group						0.369

As shown in the table, RD, FIGO stage, histology, grade and CA125 are highly associated with each other (p-values $\ll 0.05$, the usual level of significance). Diagnosis year appears only associated with RD (p-value < 0.05). This may reflect that some changes might have occurred in surgical procedure in debulking ovarian cancer tumour from the earlier to the later period. Moreover, RD is the only prognostic factor that is highly associated with all other factors in RPA data. Therefore, it may have a dominant place in the model for the RPA data.

The fact that most of the factors in the RPA data are associated implies that a few of them may be sufficient in representing the others in explaining the survival time in a model. We may only consider a subset of them in our analysis to avoid possible multicollinearity problem. This is one of common approaches to deal with multicollinearity in practice. Also note that, according to Allison (2010), multicollinearity is more about linear relations among the covariates, but not necessary to be evaluated within the context of a survival analysis. Moreover, our focus here is to illustrate our established AFT model selection framework.

6.4 Application of the AFT model selection framework to the RPA data

For the RPA ovarian cancer data, survival time T, with censoring status, is used as the response variable, and the covariates (i.e., prognostic factors) considered are RD,

FIGO stage, histology, grade, CA125, age, and diagnosis year (see definitions in Section 6.2). Note that this survival data contains continuous (e.g., age), binary (e.g., CA125) and categorical (ordinal or nominal) covariates with more than two categories (e.g., FIGO). We denote the model with all the seven covariates by {1, 2, 3, 4, 5, 6, 7, 8}, where "1" indicates the intercept term of the model and other numbers $(2-8)$ indicate the covariates mentioned in the respective order. For example, {1, 2} represents the model with intercept term and RD, and {1, 2, 3} for the model with intercept term, RD and FIGO stage. Note that all models considered here include an intercept term.

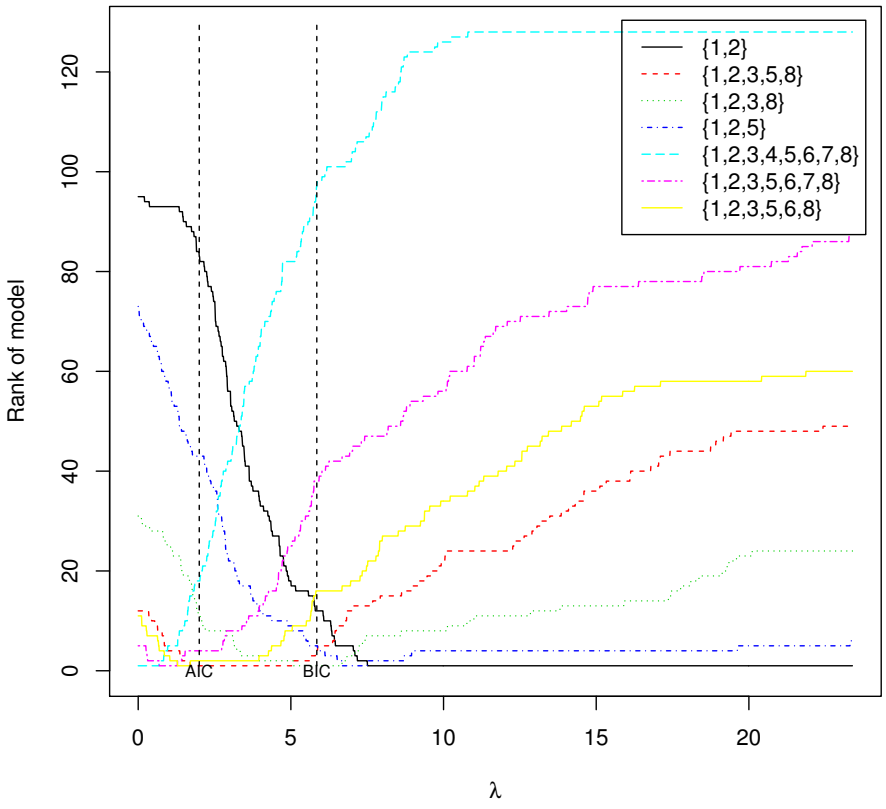


Figure 6.7: Rank 1 plots showing cathetus for the RPA ovarian cancer data.

With seven covariates considered, there are 128 possible AFT models to be fitted (see Section 4.3 for how a categorical covariate with more than two categories is handled), containing none (0) to all the seven (7) covariates. Our model selection framework for AFT models is applied to compare these 128 models and select the best model, i.e., the most stable model (see Section 1.1). Among all 128 possible models, models that have achieved rank 1 within the range of λ (0 to $4\log(n)$) with respect to

their GIC values are shown in Figure 6.7. It can be seen that 7 models from the 128 possible models have reached rank 1 at some λ values or intervals such as model {1, 2} shown on the top of the list. The model with a single covariate RD has the longest cathetus, so it is picked if using LCC. However, model with three covariates RD, FIGO stage and diagnosis year is selected if using BIC.

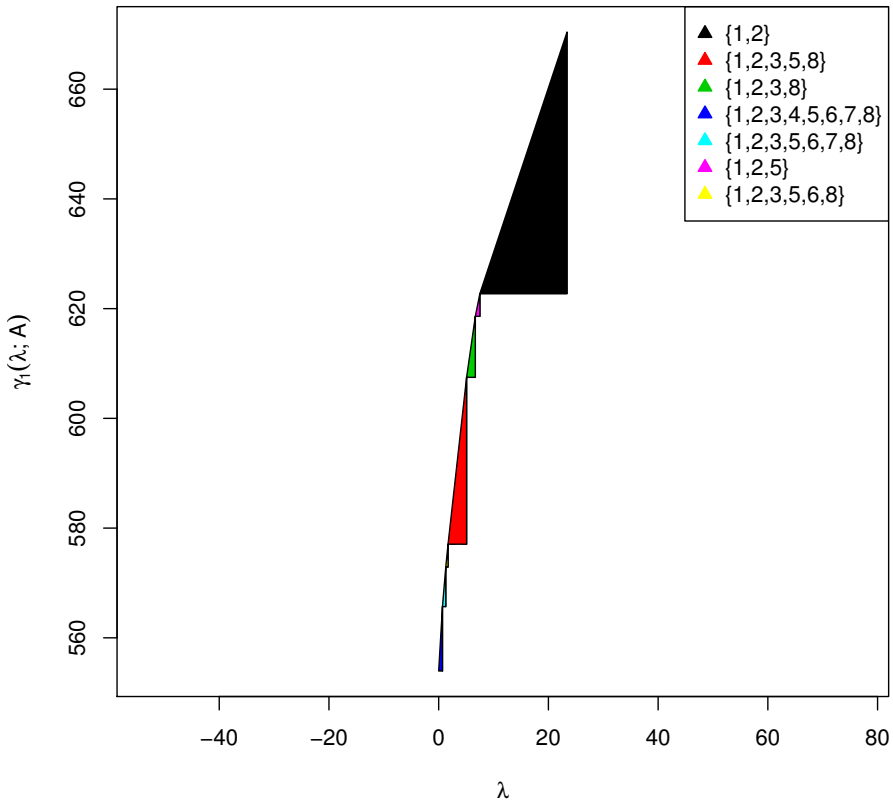


Figure 6.8: Model selected by TAC and other criteria under the model selection framework for the RPA ovarian cancer data.

Beside LCC, let us examine the other two model selection criteria LHC and TAC under the model selection framework. As shown in Figure 6.8, seven (7) of the 128 possible models have achieved rank 1, and thus appear on the model selection curve (the locus of hypotenuses of all triangles in the Figure 6.8). The model with single covariate RD, i.e., model {1, 2}, is selected not only by TAC but also LHC and LCC as the area of triangle, hypotenuse length and cathetus length corresponding to this model is the largest across all candidate models. However, AIC and BIC select different models, {1, 2, 3, 5, 8} and {1, 2, 3, 8} respectively (see Figure 6.7).

Alternatively, the same conclusion can be drawn from the summary statistics of these criteria. Summary statistics, including model dimension, cathetus length, hypotenuse length and $\sqrt{TAC}$, for the seven models that have achieved rank 1 are presented in Table 6.3. Clearly, the maximum values of cathetus length, hypotenuse length and $\sqrt{TAC}$ all correspond to model $\{1, 2\}$, as illustrated in Figure 6.8. As expected, the cathetus length C_{L_α} and $\sqrt{TAC}$ for the chosen model $\{1, 2\}$ coincide in this case where $p_\alpha = 2$. It was shown earlier that LCC and $\sqrt{TAC}$ would be identical if both of them select a model with dimension $p_\alpha = 2$. Another interesting point to note from Table 6.3 is that the $\sqrt{TAC}$ value always lies in between the corresponding LCC and LHC values as expected $\left(\sqrt{TAC} \in (C_{L_\alpha}, H_{L_\alpha})\right)$. This supports the conservative nature of TAC as a model selection criterion, which was also reported in our simulation study in Chapter 4, i.e., TAC is affected less by large variation in the data compared to LCC and LHC.

Table 6.3: Summary statistics extracted from the model selection framework for RPA ovarian cancer data.

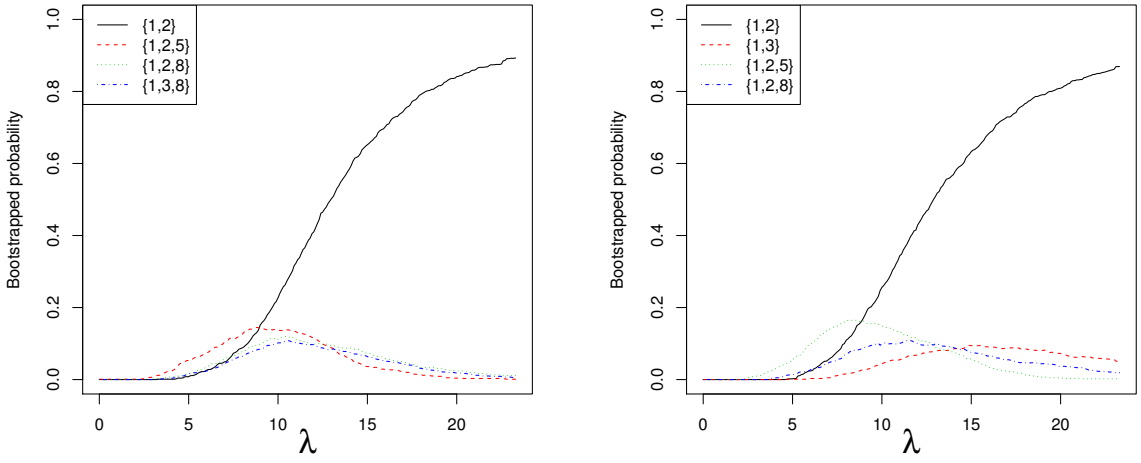
Model α	p_α	C_{L_α}	H_{L_α}	$\sqrt{TAC}$
{1,2}	2	15.89	35.53	15.89
{1,2,5}	3	0.83	2.62	1.02
{1,2,3,8}	4	1.59	6.56	2.25
{1,2,3,5,8}	5	3.38	17.23	5.34
{1,2,3,5,6,8}	6	0.42	2.55	0.73
{1,2,3,5,6,7,8}	7	0.60	4.24	1.12
{1,2,3,4,5,6,7,8}	8	0.68	5.48	1.36

Table 6.4: Models with $\pi^*(\alpha) > 4\%$ based on 1,000 bootstrap replications of RPA ovarian cancer data.

Model α	p_α	$\pi^*(\alpha)$	
		Ordinary bootstrap	Stratified bootstrap
{1,2}	2	0.402	0.395
{1,3}	2	–	0.045
{1,2,5}	3	0.051	0.058
{1,2,8}	3	0.048	0.048
{1,3,8}	3	0.041	–

We have also repeated the model selection framework for AFT model on 1,000 bootstrap replications of the RPA ovarian cancer data. Both ordinary and stratified

bootstrapping schemes are considered. In ordinary bootstrapping, censoring proportions may not be similar across all bootstrap samples while the stratified bootstrapping samples are also obtained to ensure censoring proportions matching with the original RPA ovarian data. The marginal probability of selecting a model across the 1,000 replications via LCC is presented in Table 6.4, only for models that have $\pi^*(\alpha) > 4\%$ under either the ordinary or stratified bootstrapping scheme. Many other models, not appeared when using original RPA data, have also shown up in the rank 1 model selection curves in one or more of those 1,000 bootstrap replications, but are not reported here because of very small $\pi^*(\alpha)$ values. Among all 128 possible models, model {1, 2} has the highest chance (about 40%) of being selected across 1,000 replications using the LCC. This can be visualised using the model detection plot as shown in Figure 6.9. Clearly, the area under the curve for model {1, 2} is the largest, i.e., the most frequently selected model. This is also evident in Table 6.4. Note that this model has also been selected by TAC and LHC, as shown in Figure 6.8.



(a) (b)

Figure 6.9: The model detection plot contains only models with $\pi^*(\alpha) > 4\%$: (a) based on ordinary bootstrap replications; (b) based on stratified bootstrap replications.

The proportion for each studied covariate that has been included in rank 1 models across 1,000 bootstrap replications is computed. The proportions of all covariates considered are then plotted against the penalty multiplier as shown in Figure 6.10. This plot, known as variable inclusion plot, provides information about the importance

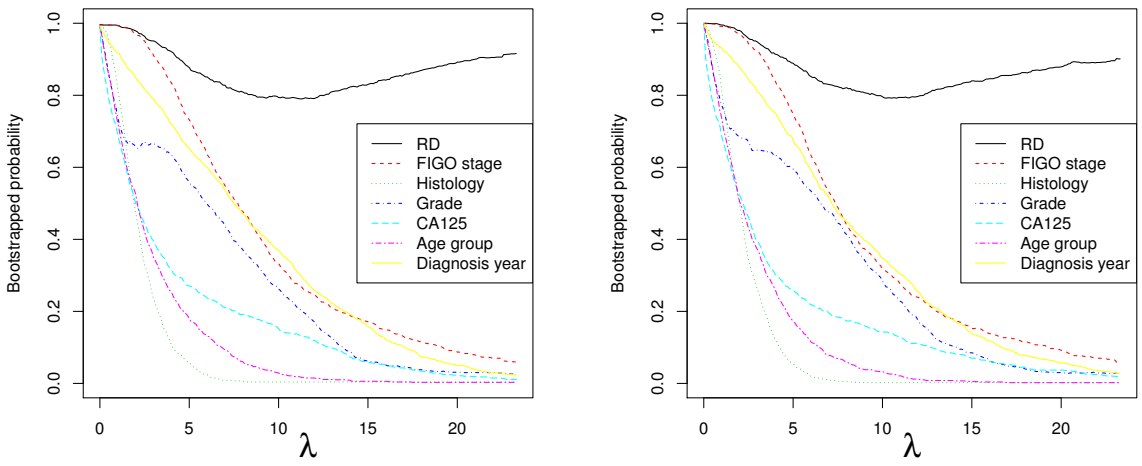


Figure 6.10: Variable inclusion plot: (a) based on ordinary bootstrap replications; (b) based on stratified bootstrap replications.

of each variable considered. The residual disease (RD) is the covariate (i.e., prognostic factor) that appeared most frequently in rank 1 models across all bootstrap replications, and other covariates have relatively much lower chances of being included in rank 1 models. As mentioned earlier, RD is also the only covariate selected by LCC, LHC and TAC (see Figure 6.8) and one of the covariates selected by AIC and BIC (see Figure 6.7). The next probable covariates being included in rank 1 models are FIGO stage and diagnosis year, which were included in the model selected by AIC or BIC. Note that RD is highly associated with all other prognostic factors, as discussed earlier. Therefore, it may sufficiently represent its related prognostic factors in the model, similar to the example where the data contains highly correlated covariates as shown in the simulation study.

7

Conclusion

A model selection framework for AFT models of survival data with right censoring has been established in this thesis. It consists of two recently suggested model selection criteria by Müller and Welsh (2010), LCC and LHC, and one newly developed criterion in our study, TAC. It also includes two commonly known criteria, AIC and BIC. Under this framework, the general information criterion is studied as a function of penalty multiplier over a range of values rather than only at a single value. This framework enables, to some extent, an evaluation on the stability of a model selection criterion used in a model selection process.

Similar to the findings reported on linear regression models in Müller and Welsh's paper, our study shows that a combination of the three MSC based criteria, LCC, LHC and TAC, has the potential to outperform other model selection criteria, such as AIC and BIC, in selecting the specificied or true (i.e., correct) AFT model. Our proposed model selection criterion, TAC, as well as LHC by Müller and Welsh, tends to do better than LCC or other model selection criteria in identifying the correct model, especially

if the survival data studied is heavily right-skewed. Furthermore, TAC has displayed a more consistent performance and appeared less affected by the distribution shape of the data.

The three MSC based model selection criteria under our framework often ($> 50\%$) select same model although not always. When they do not select same model, we suggest that a model may be chosen on the basis of voting, i.e., the model selected by more criteria may be chosen. This may lead to the study of model averaging approach studied by Claeskens and Hjort (2008).

A user-friendly tool for our model selection framework established has been developed in the statistical computing project `R`. The tool incorporates all model selection criteria under the framework, and results in a few graphs and tables. This makes model selection using our framework relatively easy. Users only need to apply the tool in `R` appropriately, and then examine the graphs and tables generated from it to identify the model to be selected. Having the tool ready for use would encourage the application of our model selection framework in practice.

Two bootstrapping schemes suggested in this thesis can be utilised to provide additional information for model or variable selection. Using one of the two bootstrapping schemes, a model detection plot can be constructed to help selecting the best model. In addition, the order of importance among covariates considered can be visualised via a variable inclusion plot that can be produced. These are particularly useful when there are more than one best models identified by the framework. The programming codes written and used for those bootstrapping schemes have not yet been included into the `R` tool developed for our AFT model selection framework, since running these programming codes for bootstrapping requires impractically long time. Thus including those programming codes into the tool would limit its use in practice, unless some parallel computing techniques can be utilised in the future to make it more efficient (Matloff, 2016).

The model selection framework for AFT models has opened up new insight into model selection in survival analysis. This framework can handle models that have continuous, binary and general categorical (ordinal or nominal with more than two categories) covariates. Using this framework, a large number of covariates can be handled, although the number of possible models increases exponentially with the number

of covariates considered. One way to overcome this problem is to reduce the number of models considered via some initial screening of all possible models, such as incorporating the lasso approach (Tibshirani, 1997) inside the model selection framework. This can be investigated in a future study. Note that the model selection framework for AFT models established in this thesis has not been applied to models with interaction terms between covariates due to time limit, but will be investigated in the future. The study of group lasso by Meier et al. (2008) may give some ideas for investigation of such interactions. It is also notable that the bootstrapping for model selection used in this thesis is based on LCC. However, the potential of using TAC in the bootstrapping for model selection is worth to be investigated in the future. This may be done by computing, for each model considered, the proportion that the model appears on the 1 rank model selection curve with the largest triangle area across all bootstrap replications. It needs to mention again that the three MSC based model selection criteria may not be suitable for data with relatively small sample size (n) and very large number of covariates (p), i.e., $p \gg n$. Model selection for such high dimensional data in survival analysis has been considered in the literature (e.g., Gui and Li (2005), Ma and Huang (2007), Wang et al. (2008), Huang and Ma (2010) etc.) and combining them with the three MSC based model selection criteria will again be a future research topic.

To conclude, our established model selection framework for AFT models is considered a good addition for handling model selection problems in survival analysis. The user-friendly `R` tool developed for this framework improves the model selection process for users and makes the framework reasonably easy to apply in practice.

A

Main Functions for the R Tool

aftmsc is the main function of the R tool developed for the proposed model selection framework for AFT model selection. R code for this function is given below.

```
require(survival) # Need survreg and Surv
aftmsc = function(formula, data, distn = 'weibull'
                  ,lambda, intercept = TRUE){
  Call <- match.call()
  indx <- match(c("formula", "data"), names(Call), nomatch = 0)
  if (indx[1]==0)
    stop("A formula argument is required")
  temp <- Call[c(1L, indx)]
  temp[[1L]] = quote(stats::model.frame)
  if (missing(data))
  { temp$formula <- terms(formula)}
  else{ temp$formula = terms(formula,data=data)}
```

```
m <- eval(temp,parent.frame())
Terms <- attr(m, "terms")
Xlabels = attr(Terms,"term.labels")
Y <- model.extract(m, "response")
if (!inherits(Y, "Surv"))
{stop("Response must be a survival object")}
X <- model.matrix(Terms, m)
mu = length(unique(attr(X,"assign")))-1
n = nrow(X)
if (intercept == TRUE) {D=DSM(mu)} else
{D = DSM(mu+1,FALSE)}        #No. of covariates without the intercept
if (missing(lambda)){lambda=seq(0,4*log(n),.01)}
M=surv.gic(Y,X,distn,lambda,D)
r=apply(M,2,rank)
# models that have rank 1 at each lambda
rank1model = which(r == 1, arr.ind=TRUE)[,1]
# finding out where the changes of model occur
lambdaindex.changeinmodel = (1:length(rank1model))
                            [c(FALSE, diff(rank1model)!=0)]
if (length(lambdaindex.changeinmodel)>1){
  if (lambdaindex.changeinmodel[length(lambdaindex.changeinmodel)
     -1] == lambdaindex.changeinmodel[length(lambdaindex.changeinmodel)
       ]){
    lambdaindex.changeinmodel = lambdaindex.changeinmodel
                                [-length(lambdaindex.changeinmodel)]
  }
}
q=sort(table(which(r == 1, arr.ind=TRUE)[,1]),decreasing = TRUE)
if (sum(duplicated(q))!=0){
  duplicatedmodel = names(q[q[duplicated(q)] == q])
  orderedduplicatedmodel = duplicatedmodel[order(apply
                            (D[as.numeric(duplicatedmodel),],1,sum))]
```

```
    q[duplicatedmodel] = q[ordereddduplicatedmodel]
  }
  rank1model.sub = rank1model[c(1,lambdaindex.changeinmodel)]
  #Finding lambda values where models have rank 1
  new.l.v=sort(lambda[c(1,lambdaindex.changeinmodel
                        , length(lambda))],decreasing=TRUE)
  #Determining Cathetus lengths, hypotenuse lengths, and TAC values
  cat.length=sort(-diff(new.l.v),decreasing=TRUE)
  # palpha in hypotenuse formula
  if (is.null(dim(D[c(as.numeric(names(q))),])))
                    {model.dim = sum(D)}   else
     {model.dim=apply(D[c(as.numeric(names(q))),],1,sum)}
  if (length(cat.length) != length(model.dim))
     {cat.length = cat.length[1:length(model.dim)]}
  hyp.length=cat.length*sqrt(1+model.dim^2)
  TAC=(1/2)*cat.length^2*model.dim
  TAC.sqrt=sqrt(TAC)
  Model.num=as.numeric(names(q))
  lengths=cbind(Model.num,cat.length,hyp.length,TAC.sqrt)
  #Used in TAC and hyp plot to get legends_row according to ascending
  q.TAC=q[order(-TAC.sqrt)]
  q.hyp=q[order(-hyp.length)]
  out=list(M = M, r = r, D = D, rank1model.sub = rank1model.sub
           , lambda = lambda, mu=mu , rank1model=rank1model
           , lambdaindex.changeinmode l= lambdaindex.changeinmodel
           , q = q, q.hyp = q.hyp, q.TAC = q.TAC
           , new.l.v=new.l.v,cathetus.hypotenuse.TAC = lengths,
           Y = Y, X = X, Xlab = c("Intercept",Xlabels))
  class(out) = "msc"
  return(out)
}
```

B

Sub-functions in the R Tool

Two important sub-functions, `DSM` and `surv.gic`, are described below.

DSM: `DSM` is an internal sub-function to generate a list of all submodels. This sub-function is meant to be used within some other function and sub-function (e.g., `aftmsc` and `surv.gic`). A call to this function is

```
DSM(n.s, intcpt = TRUE).
```

Here, `n.s` is the number of variables to be considered in the model selection framework. If `intcpt = TRUE`, then an intercept term will be added to all submodels. For example, if we have four covariates and all models include the intercept term, then we set `n.s = 4` in the `DSM` sub-function to get list of models. It returns an indicator matrix of 0's and 1's with 0's indicating the variable is not included and 1 the variable is included.

surv.gic: `surv.gic` is also a sub-function, which is used within a main function (e.g., `aftmsc`). This function is called by

```
surv.gic(Y, X, distn, lambda, D).
```

Here Y is a response variable, and it must be a survival object, which may be obtained by using the Surv function from the R package "survival", X is the covariate matrix, any distribution of response variable is specified at distn. The argument lambda (λ) is the penalty multiplier, which can take a range of values (e.g., $\lambda \in [0,\ 4\log(n)]$), and D is the matrix obtained from the DSM sub-function. The sub-function surv.gic computes the GIC under the AFT model.

The R code for the two sub-functions are given below.

```
DSM = function(n.s,intcpt = TRUE){
    D = NULL
    for(i in 1:n.s){
    D0 = cbind(0L,D)
    D1 = cbind(1L,D)
    D  = rbind(D0,D1)
     }
  if (intcpt == TRUE){D = cbind(1L,D)} else {D = D[-1,]}
  return(D[order(apply(D,1,sum)),])
 }  # end DSM function

surv.gic=function(Y,X,distn,lambda,D){
  RES = NULL;
  nr.D = dim(D)[1];
  n = nrow(X);
  D.temp = NULL
  for (i in 1:nr.D){
  D.temp = rbind(D.temp,rep(D[i,],table(attr(X,"assign"))))}
  nc.D = dim(D.temp)[2];
  D.c = D.temp %*% diag(1:nc.D)
  for(d in 1:nr.D){
    d.c = D.c[d,]
    d.c = d.c[d.c>0]
    X.dat = X[,d.c]
    out=survreg(Y~-1+X.dat,dist=distn)
```

```
    p=ncol(model.matrix(out))
    gic=-2*out$loglik[2]+lambda*p
    RES = rbind(RES,gic)
  }
  return(RES)
}  # end surv.gic function
```

C

Additional R Programs

In order to run the following codes we require several R packages, including survival, boot (Canty and Ripley, 2015; Davison and Hinkley, 1997), plyr (Wickham, 2011) and data.table (Dowle et al., 2015), MASS (Venables and Ripley, 2002) and matrixcalc (Novomestky, 2012).

C.1 Sample R codes for bootstrapping

```
##some R Packages required
require(survival);require(boot); require(plyr); require(data.table)
surv.gic.simul=function(D,dat,lambda){
RES = NULL;
   nr.D = dim(D)[1];
   nc.D = dim(D)[2];
   n = nrow(dat);
```

```
   D.c = D %*% diag(1:nc.D)
   for(d in 1:nr.D){
     d.c = D.c[d,]
     d.c = d.c[d.c>0]
X=dat[,-c(1,2)]
Xdata = cbind(1L,X)
X.dat = Xdata[,d.c]
if(dim(as.matrix(X.dat))[2] == 1)
out = survreg(Surv(time  =T, event = event) ~ -1+., data = data.frame(T
                                    = dat$T, event=dat$event,X.dat)) else
out=survreg(Surv(time=T,event=event)~., data = data.frame(T
                              =dat$T, event = dat$event, X.dat[,-1]))
p=ncol(model.matrix(out))
gic=-2*out$loglik[2]+lambda*p
     RES = rbind(RES,gic)
     }
   return(RES)
   }
T = ovarian$futime; event = ovarian$fustat
work.dat = data.frame(T, event, ovarian$age, ovarian$resid.ds
                       , ovarian$rx, ovarian$ecog.ps)
n = nrow(work.dat)
lambda = seq(0, 4*log(n), .01)
mu = ncol(work.dat)-2
D = DSM(mu)
##Bootstraping
R=1000
b.fun <- function(data) {
mean.dat=mean(data$T)
mean.dat
}
set.seed(100)
```

```
#Ordinary bootstrapping
b.work.dat <- censboot(work.dat, b.fun, R)
#stratified bootstrapping
#b.work.dat <- censboot(work.dat, b.fun, R, strata = work.dat$event)
bt = boot.array(b.work.dat, indices = TRUE)
#prob function calculates probability of selecting a model
prob = function(V){
out = matrix(data = NA, nrow = nrow(D), ncol = 1)
for (j in 1:1){
for (i in 1:nrow(D)){
r = apply(V, 2, rank)
out[i, j] = ifelse(length(V[i, ][r[i, ] == 1]) > 0
           , length(V[i, ][r[i, ] == 1])/length(lambda), 0)
}}
return(out)
}
Pmat = matrix(0, nrow = nrow(D), ncol = R)
for (h in 1:R){
boot.dat = work.dat[bt[h, ], ]
M2=surv.gic.simul(D, boot.dat, lambda)
Pmat[,h]=prob(M2)
}
Pmat
P.alpha = matrix(0, nrow = nrow(D), ncol = 1)
for (q in 1:nrow(D)){
pbar=mean(Pmat[q, ])
P.alpha[q, ]=pbar
}
P.alpha
model.show = function(W){
test = matrix(0, nrow = length(lambda), ncol = 1)
for (t in 1:1){
```

```
for (s in 1:length(lambda)){
test[s, t]=order(W[, s])[1]
}}
return(test)
}
Mmat = matrix(0, nrow = length(lambda), ncol = R)
for (h in 1:R){
boot.dat=work.dat[bt[h, ], ]
M3=surv.gic.simul(D, boot.dat, lambda)
Mmat[, h]=model.show(M3)
}
Mmat
Mcount.lambda = function(Mmat){
Model.freq=matrix(0, nrow = nrow(Mmat), ncol = max(Mmat))
Model.freq
for (k in 1:nrow(Mmat)){
for (i in 1:max(Mmat)){
for (j in 1:ncol(Mmat)){
if (Mmat[k, j] == i){Model.freq[k, i] = count(Mmat[k, ] == i)[2, 2]}
}
}
Model.freq
for (l in 1:max(Mmat)){
index = which(is.na(Model.freq[, l]) == TRUE)
Model.freq[index, l] = length(Mmat[index, ])
}
}
Model.freq
}
Modelcount = Mcount.lambda(Mmat)
Model.p = Modelcount/R
#Find model which have selection probaility >4%
```

```
mdl.no = c()
for (e in 1:nrow(D)){
if (P.alpha[e, ] > .04)  mdl = e else mdl = 0
mdl.no = c(mdl.no, mdl)
mdl.no
}
mdl.no[mdl.no > 0]
M1.p = Model.p[, 1]; M5.p = Model.p[, 5]
M10.p = Model.p[, 10]; M11.p = Model.p[, 11]
M15.p = Model.p[, 15]; M16.p = Model.p[,16]
plot(lambda, M5.p, type = 'l', col =1, ylim = c(0,1),
xlab = expression(paste(lambda)), ylab = "Bootstrapped  probability")
lines(lambda,M1.p, lty = 2, col = 2)
lines(lambda, M10.p, lty = 3, col = 3)
lines(lambda, M11.p, lty = 4, col = 4)
lines(lambda, M15.p, lty = 5, col = 5)
lines(lambda,M16.p, lty = 6, col = 6)
legend("topright", legend = c("{1,2}", "{1}", "{1,2,4}", "{1,2,3}"
                , "{1,2,3,4}", "{1,2,3,4,5}"), lty = 1:6, col = 1:6)
Vcount = matrix(0, nrow = nrow(Modelcount), ncol = ncol(D))
for (i in 1:nrow(Modelcount)){
L12 = matrix(0, nrow = ncol(Modelcount), ncol = ncol(D))
for (j in 1:ncol(Modelcount)){
L12[j, ] = Modelcount[i, j]*D[j, ]
}
L12
L1 = colSums(L12)
L1
Vcount[i, ] = L1
}
Vcount
V.p = Vcount/R
```

```
x1.p = V.p[, 2]
x2.p = V.p[, 3]
x3.p = V.p[, 4]
x4.p = V.p[, 5]
plot(lambda, x1.p, type = 'l', xlab = expression(paste(lambda))
     ,ylab = "Bootstrapped  probability", ylim = c(0, 1), col=1)
lines(lambda, x2.p, lty = 2, col = 2)
lines(lambda, x3.p, lty = 3, col = 3)
lines(lambda, x4.p, lty = 4, col = 4)
legend("topright", legend = c("x1","x2","x3","x4")
         , lty = 1:4, col = c(1,2,3,4))
```

C.2 Sample R codes for Monte Carlo simulation

```
library(boot)
library(plyr)  #for count function
library("survival")
library("MASS")
library("data.table")
library("matrixcalc")
DSM = function(n.s,intcpt = TRUE){
  if (intcpt == FALSE){n.s = n.s}
  D = NULL
  for(i in 1:n.s){
    D0 = cbind(0L,D)
    D1 = cbind(1L,D)
    D  = rbind(D0,D1)
    }
  if (intcpt == TRUE){D = cbind(1L,D)}
  return(D[order(apply(D,1,sum)),])
}
surv.gic.simul=function(D,dat,lambda){
```

```
RES = NULL;
    nr.D = dim(D)[1];
   nc.D = dim(D)[2];
   n = nrow(dat);
   D.c = D %*% diag(1:nc.D)
   for(d in 1:nr.D){
     d.c = D.c[d,]
     d.c = d.c[d.c>0]
X=dat[,-c(1,2)]
Xdata=cbind(1L,X)
X.dat = Xdata[,d.c]
if(dim(as.matrix(X.dat))[2]==1)
out=survreg(Surv(time=T,event=event)~-1+.,data=data.frame
                   (T=dat$T,event=dat$event,X.dat)) else
out=survreg(Surv(time=T,event=event)~.,data=data.frame
                  (T=dat$T,event=dat$event,X.dat[,-1]))
p=ncol(model.matrix(out))
gic=-2*out$loglik[2]+lambda*p
     RES = rbind(RES,gic)
     }
   return(RES)
   }
#generating observations for model {1, 4, 5}
n=500000
mu=c(0,0,0,0)
cor.mat=as.matrix(rbind(c(1,.001,.001,.001),
                        c(.001,1,.001,.001),
                        c(.001,.001,1,.001),
                        c(.001,.001,.001,1)))
SD<-rep(1,length(mu))
S<-cor.mat*(SD%*%t(SD))
if (is.positive.definite(S)==!TRUE)
```

```
  stop("A positive definite S is required")
set.seed(600)
X=mvrnorm(n,mu,S)
colnames(X)=c("X1","X2","X3","X4")
y=1:nrow(X)
d=data.frame(y,X)
out5=lm(y~.,data=d)
designX=model.matrix(out5)
beta=c(0.1, 0, 0, 0.9, 0.8)
inter=log(3)
scale.event=exp((designX%*%beta))
scale.cens=scale.event*exp(inter*rep(1,n))
T = rweibull(n, shape=2, scale=scale.event)
C = rweibull(n, shape=2, scale=scale.cens)
time = pmin(T,C)
event = time==T
cens.prop=1-mean(event)
newd=d[,-1]
gen.dat_145<-data.frame(T=time,event,newd)
set.seed(600)
r1.model=NULL;AIC.m=NULL;BIC.m=NULL
LHC.m_300s1=NULL;TAC.m_300s1=NULL
for (s in 1:100){
n1=300
s1_270=gen.dat_145[sample(which(gen.dat_145$event=="TRUE")
                                    ,n1*.90,replace=TRUE),]
s2_30=gen.dat_145[sample(which(gen.dat_145$event=="FALSE")
                                    ,n1*.10,replace=TRUE),]
work.dat1=rbind(s1_270,s2_30)
D=DSM(length(mu))
lambda=seq(0,4*log(n1),.01)
M1_300s1=surv.gic.simul(D,work.dat1,lambda)
```

```
r_300s1=apply(M1_300s1,2,rank)
q_300s1=table(which(r_300s1 == 1, arr.ind=TRUE)[,1])
q_300s1
lam.r1.300s1=c()
for (i in 1:nrow(D)){
lam.r1.300s1[i]=max(lambda[r_300s1[i,]<=1])
}
lam.r1.300s1
lam.r1.300s1[which(lam.r1.300s1!=-Inf)]
hyp.length.300s1=c()
cat.length.300s1=c()
for (i in 1:length(lam.r1.300s1[which(lam.r1.300s1!=-Inf)])){
if (i+1>length(lam.r1.300s1[which(lam.r1.300s1!=-Inf)]))
c1=min(lam.r1.300s1[which(lam.r1.300s1!=-Inf)])*sqrt(1+(sum(D
                          [which(lam.r1.300s1!=-Inf)[i],]))^2) else
c1=(lam.r1.300s1[which(lam.r1.300s1!=-Inf)][i]-lam.r1.300s1[which
(lam.r1.300s1!=-Inf)][i+1])*sqrt(1+(sum(D
           [which(lam.r1.300s1!=-Inf)[i],]))^2)
hyp.length.300s1[i]=c1
if (i+1>length(lam.r1.300s1[which(lam.r1.300s1!=-Inf)]))
cat.length=min(lam.r1.300s1[which(lam.r1.300s1!=-Inf)]) else
cat.length=(lam.r1.300s1[which(lam.r1.300s1!=-Inf)][i]
    -lam.r1.300s1[which(lam.r1.300s1!=-Inf)][i+1])
cat.length.300s1[i]=cat.length
}
hyp.length.300s1
cat.length.300s1
hyp.longest.m=which(lam.r1.300s1!=-Inf)[which
       (hyp.length.300s1==max(hyp.length.300s1))]
LHC.m_300s1=c(hyp.longest.m,LHC.m_300s1)
LHC.m_300s1
if (is.matrix(D[which(lam.r1.300s1!=-Inf),])==TRUE)
```

```
area=(1/2)*(cat.length.300s1^2)*apply(D
   [which(lam.r1.300s1!=-Inf),],1,sum) else
area=(1/2)*(cat.length.300s1^2)*sum(D[which(lam.r1.300s1!=-Inf),])
area.largest=which(lam.r1.300s1!=-Inf)[which(area==max(area))]
TAC.m_300s1=c(area.largest,TAC.m_300s1)
TAC.m_300s1
BICm=which(r_300s1==1,arr.ind=TRUE)[,1][lambda==round(log(n1),1)]
BIC.m=c(BIC.m,BICm)
BIC.m
AICm=which(r_300s1==1,arr.ind=TRUE)[,1][lambda==2]
AIC.m=c(AIC.m,AICm)
AIC.m
o=order(q_300s1);bc=q_300s1[o[length(q_300s1)]];
r1.model=c(r1.model,bc)
}
r1.model;f=count(labels(r1.model))
LCC.ppn_300s1=count(labels(r1.model))$freq/sum(count
                          (labels(r1.model))$freq)
LCC.tab_300s1=cbind(f,LCC.ppn_300s1)
a1=data.table(LCC.tab_300s1)
LCC.tab_300s1=a1[order(-a1$LCC.ppn_300s1)]
LCC.tab_300s1
LHC.ppn_300s1=table(LHC.m_300s1)/sum(table(LHC.m_300s1))
LHC.tab_300s1=sort(LHC.ppn_300s1,decreasing=TRUE)
LHC.tab_300s1
TAC.ppn_300s1=table(TAC.m_300s1)/sum(table(TAC.m_300s1))
TAC.tab_300s1=sort(TAC.ppn_300s1,decreasing=TRUE)
TAC.tab_300s1
freq.AIC=count(AIC.m)
AIC.ppn_300s1=freq.AIC$freq/sum(freq.AIC$freq)
AIC.tab_300s1=cbind(freq.AIC,AIC.ppn_300s1)
a2=data.table(AIC.tab_300s1)
```

```
AIC.tab_300s1=a2[order(-a2$AIC.ppn_300s1)]
AIC.tab_300s1
freq.BIC=count(BIC.m)
BIC.ppn_300s1=freq.BIC$freq/sum(freq.BIC$freq)
BIC.tab_300s1=cbind(freq.BIC,BIC.ppn_300s1)
a3=data.table(BIC.tab_300s1)
BIC.tab_300s1=a3[order(-a3$BIC.ppn_300s1)]
BIC.tab_300s1
#generating observations for model {1, 4}
n=500000
mu=c(0,0,0,0)
cor.mat=as.matrix(rbind(c(1,.001,.001,.001),
                        c(.001,1,.001,.001),
                        c(.001,.001,1,.001),
                        c(.001,.001,.001,1)))
SD<-rep(1,length(mu))
S<-cor.mat*(SD%*%t(SD))
if (is.positive.definite(S)==!TRUE)
  stop("A positive definite S is required")
set.seed(600)
X=mvrnorm(n,mu,S)
colnames(X)=c("X1","X2","X3","X4")
y=1:nrow(X)
d=data.frame(y,X)
out5=lm(y~.,data=d)
designX=model.matrix(out5)
beta=c(0.1, 0, 0, 0.9, 0)
inter=log(3)
scale.event=exp((designX%*%beta))
scale.cens=scale.event*exp(inter*rep(1,n))
T = rweibull(n, shape=2, scale=scale.event)
C = rweibull(n, shape=2, scale=scale.cens)
```

```
time = pmin(T,C)
event = time==T
cens.prop=1-mean(event)
newd=d[,-1]
gen.dat_14<-data.frame(T=time,event,newd)
set.seed(600)
r1.model=NULL; AIC.m=NULL; BIC.m=NULL
LHC.m_300s2=NULL;TAC.m_300s2=NULL;
for (s in 1:100){
n2=300
s1_270=gen.dat_14[sample(which(gen.dat_14$event=="TRUE")
                                ,n2*.90,replace=TRUE),]
s2_30=gen.dat_14[sample(which(gen.dat_14$event=="FALSE")
                               ,n2*.10,replace=TRUE),]
work.dat2=rbind(s1_270,s2_30)
D=DSM(length(mu))
lambda=seq(0,4*log(n2),.01)
M1_300s2=surv.gic.simul(D,work.dat2,lambda)
r_300s2=apply(M1_300s2,2,rank)
q_300s2=table(which(r_300s2 == 1, arr.ind=TRUE)[,1])
q_300s2
lam.r1.300s2=c()
for (i in 1:nrow(D)){
lam.r1.300s2[i]=max(lambda[r_300s2[i,]<=1])
}
lam.r1.300s2
hyp.length.300s2=c()
cat.length.300s2=c()
for (i in 1:length(lam.r1.300s2[which(lam.r1.300s2!=-Inf)])){
if (i+1>length(lam.r1.300s2[which(lam.r1.300s2!=-Inf)]))
c1=min(lam.r1.300s2[which(lam.r1.300s2!=-Inf)])*sqrt(1+(sum(D
                                [which(lam.r1.300s2!=-Inf)[i],]))^2) else
```

```
c1=(lam.r1.300s2[which(lam.r1.300s2!=-Inf)][i]
-lam.r1.300s2[which(lam.r1.300s2!=-Inf)][i+1])*sqrt(1+(sum(D
                        [which(lam.r1.300s2!=-Inf)[i],]))^2)
hyp.length.300s2[i]=c1
if (i+1>length(lam.r1.300s2[which(lam.r1.300s2!=-Inf)]))
cat.length=min(lam.r1.300s2[which(lam.r1.300s2!=-Inf)]) else
cat.length=(lam.r1.300s2[which(lam.r1.300s2!=-Inf)][i]
       -lam.r1.300s2[which(lam.r1.300s2!=-Inf)][i+1])
cat.length.300s2[i]=cat.length
}
hyp.length.300s2
cat.length.300s2
hyp.longest.m=which(lam.r1.300s2!=-Inf)[which
  (hyp.length.300s2==max(hyp.length.300s2))]
LHC.m_300s2=c(hyp.longest.m,LHC.m_300s2)
LHC.m_300s2
if (is.matrix(D[which(lam.r1.300s2!=-Inf),])==TRUE)
area=(1/2)*(cat.length.300s2^2)*apply(D
  [which(lam.r1.300s2!=-Inf),],1,sum) else
area=(1/2)*(cat.length.300s2^2)*sum(D[which(lam.r1.300s2!=-Inf),])
area.largest=which(lam.r1.300s2!=-Inf)[which(area==max(area))]
TAC.m_300s2=c(area.largest,TAC.m_300s2)
TAC.m_300s2
BICm=which(r_300s2==1,arr.ind=TRUE)[,1][lambda==round(log(n2),1)]
BIC.m=c(BIC.m,BICm)
BIC.m
AICm=which(r_300s2==1,arr.ind=TRUE)[,1][lambda==2]
AIC.m=c(AIC.m,AICm)
AIC.m
o=order(q_300s2);bc=q_300s2[o[length(q_300s2)]];
r1.model=c(r1.model,bc)
}
```

```
r1.model;f=count(labels(r1.model))
LCC.ppn_300s2=count(labels(r1.model))$freq/sum(count
                              (labels(r1.model))$freq);
LCC.tab_300s2=cbind(f,LCC.ppn_300s2)
a4=data.table(LCC.tab_300s2)
LCC.tab_300s2=a4[order(-a4$LCC.ppn_300s2)]
LCC.tab_300s2
LHC.ppn_300s2=table(LHC.m_300s2)/sum(table(LHC.m_300s2))
LHC.tab_300s2=sort(LHC.ppn_300s2,decreasing=TRUE)
LHC.tab_300s2
TAC.ppn_300s2=table(TAC.m_300s2)/sum(table(TAC.m_300s2))
TAC.tab_300s2=sort(TAC.ppn_300s2,decreasing=TRUE)
TAC.tab_300s2
freq.AIC=count(AIC.m)
AIC.ppn_300s2=freq.AIC$freq/sum(freq.AIC$freq)
AIC.tab_300s2=cbind(freq.AIC,AIC.ppn_300s2)
a5=data.table(AIC.tab_300s2)
AIC.tab_300s2=a5[order(-a5$AIC.ppn_300s2)]
AIC.tab_300s2
freq.BIC=count(BIC.m)
BIC.ppn_300s2=freq.BIC$freq/sum(freq.BIC$freq)
BIC.tab_300s2=cbind(freq.BIC,BIC.ppn_300s2)
a6=data.table(BIC.tab_300s2)
BIC.tab_300s2=a6[order(-a6$BIC.ppn_300s2)]
BIC.tab_300s2
set.seed(600)
r1.model=NULL; AIC.m=NULL; BIC.m=NULL
LHC.m_150s1=NULL;TAC.m_150s1=NULL
for (s in 1:100){
n3=150
s1_135=gen.dat_145[sample(which(gen.dat_145$event=="TRUE")
                                    ,n3*.90,replace=TRUE),]
```

```
s2_15=gen.dat_145[sample(which(gen.dat_145$event=="FALSE")
                                    ,n3*.10,replace=TRUE),]
work.dat3=rbind(s1_135,s2_15)
D=DSM(length(mu))
lambda=seq(0,4*log(n3),.01)
M1_150s1=surv.gic.simul(D,work.dat3,lambda)
r_150s1=apply(M1_150s1,2,rank)
q_150s1=table(which(r_150s1 == 1, arr.ind=TRUE)[,1])
q_150s1
lam.r1.150s1=c()
for (i in 1:nrow(D)){
lam.r1.150s1[i]=max(lambda[r_150s1[i,]<=1])
}
lam.r1.150s1
hyp.length.150s1=c()
cat.length.150s1=c()
for (i in 1:length(lam.r1.150s1[which(lam.r1.150s1!=-Inf)])){
if (i+1>length(lam.r1.150s1[which(lam.r1.150s1!=-Inf)]))
c1=min(lam.r1.150s1[which(lam.r1.150s1!=-Inf)])*sqrt(1+(sum(D
             [which(lam.r1.150s1!=-Inf)[i],]))^2) else
c1=(lam.r1.150s1[which(lam.r1.150s1!=-Inf)][i]
-lam.r1.150s1[which(lam.r1.150s1!=-Inf)][i+1])*sqrt(1+(sum(D
                     [which(lam.r1.150s1!=-Inf)[i],]))^2)
hyp.length.150s1[i]=c1
if (i+1>length(lam.r1.150s1[which(lam.r1.150s1!=-Inf)]))
cat.length=min(lam.r1.150s1[which(lam.r1.150s1!=-Inf)]) else
cat.length=(lam.r1.150s1[which(lam.r1.150s1!=-Inf)][i]
        -lam.r1.150s1[which(lam.r1.150s1!=-Inf)][i+1])
cat.length.150s1[i]=cat.length
}
hyp.length.150s1
cat.length.150s1
```

```
hyp.longest.m=which(lam.r1.150s1!=-Inf)[which
  (hyp.length.150s1==max(hyp.length.150s1))]
LHC.m_150s1=c(hyp.longest.m,LHC.m_150s1)
LHC.m_150s1
if (is.matrix(D[which(lam.r1.150s1!=-Inf),])==TRUE)
area=(1/2)*(cat.length.150s1^2)*apply(D
       [which(lam.r1.150s1!=-Inf),],1,sum) else
area=(1/2)*(cat.length.150s1^2)*sum(D[which(lam.r1.150s1!=-Inf),])
area.largest=which(lam.r1.150s1!=-Inf)[which(area==max(area))]
TAC.m_150s1=c(area.largest,TAC.m_150s1)
TAC.m_150s1
BICm=which(r_150s1==1,arr.ind=TRUE)[,1][lambda==round(log(n3),1)]
BIC.m=c(BIC.m,BICm)
BIC.m
AICm=which(r_150s1==1,arr.ind=TRUE)[,1][lambda==2]
AIC.m=c(AIC.m,AICm)
AIC.m
o=order(q_150s1);bc=q_150s1[o[length(q_150s1)]];
r1.model=c(r1.model,bc)
}
r1.model;f=count(labels(r1.model))
LCC.ppn_150s1=count(labels(r1.model))$freq/sum(count
                                    (labels(r1.model))$freq);
LCC.tab_150s1=cbind(f,LCC.ppn_150s1)
a7=data.table(LCC.tab_150s1)
LCC.tab_150s1=a7[order(-a7$LCC.ppn_150s1)]
LCC.tab_150s1
LHC.ppn_150s1=table(LHC.m_150s1)/sum(table(LHC.m_150s1))
LHC.tab_150s1=sort(LHC.ppn_150s1,decreasing=TRUE)
LHC.tab_150s1
TAC.ppn_150s1=table(TAC.m_150s1)/sum(table(TAC.m_150s1))
TAC.tab_150s1=sort(TAC.ppn_150s1,decreasing=TRUE)
```

```
TAC.tab_150s1
freq.AIC=count(AIC.m)
AIC.ppn_150s1=freq.AIC$freq/sum(freq.AIC$freq)
AIC.tab_150s1=cbind(freq.AIC,AIC.ppn_150s1)
a8=data.table(AIC.tab_150s1)
AIC.tab_150s1=a8[order(-a8$AIC.ppn_150s1)]
AIC.tab_150s1
freq.BIC=count(BIC.m)
BIC.ppn_150s1=freq.BIC$freq/sum(freq.BIC$freq)
BIC.tab_150s1=cbind(freq.BIC,BIC.ppn_150s1)
a9=data.table(BIC.tab_150s1)
BIC.tab_150s1=a9[order(-a9$BIC.ppn_150s1)]
BIC.tab_150s1
set.seed(600)
r1.model=NULL; AIC.m=NULL; BIC.m=NULL
LHC.m_150s2=NULL;TAC.m_150s2=NULL;
for (s in 1:100){
n4=150
s1_135=gen.dat_14[sample(which(gen.dat_14$event=="TRUE")
                                  ,n4*.90,replace=TRUE),]
s2_15=gen.dat_14[sample(which(gen.dat_14$event=="FALSE")
                                 ,n4*.10,replace=TRUE),]
work.dat4=rbind(s1_135,s2_15)
D=DSM(length(mu))
lambda=seq(0,4*log(n4),.01)
M1_150s2=surv.gic.simul(D,work.dat4,lambda)
r_150s2=apply(M1_150s2,2,rank)
q_150s2=table(which(r_150s2 == 1, arr.ind=TRUE)[,1])
q_150s2
lam.r1.150s2=c()
for (i in 1:nrow(D)){
lam.r1.150s2[i]=max(lambda[r_150s2[i,]<=1])
```

```
}
lam.r1.150s2
hyp.length.150s2=c()
cat.length.150s2=c()
for (i in 1:length(lam.r1.150s2[which(lam.r1.150s2!=-Inf)])){
if (i+1>length(lam.r1.150s2[which(lam.r1.150s2!=-Inf)]))
c1=min(lam.r1.150s2[which(lam.r1.150s2!=-Inf)])*sqrt(1+(sum(D
               [which(lam.r1.150s2!=-Inf)[i],]))^2) else
c1=(lam.r1.150s2[which(lam.r1.150s2!=-Inf)][i]
-lam.r1.150s2[which(lam.r1.150s2!=-Inf)][i+1])*sqrt(1+(sum(D
                     [which(lam.r1.150s2!=-Inf)[i],]))^2)
hyp.length.150s2[i]=c1
if (i+1>length(lam.r1.150s2[which(lam.r1.150s2!=-Inf)]))
cat.length=min(lam.r1.150s2[which(lam.r1.150s2!=-Inf)]) else
cat.length=(lam.r1.150s2[which(lam.r1.150s2!=-Inf)][i]
        -lam.r1.150s2[which(lam.r1.150s2!=-Inf)][i+1])
cat.length.150s2[i]=cat.length
}
hyp.length.150s2
cat.length.150s2
hyp.longest.m=which(lam.r1.150s2!=-Inf)[which
  (hyp.length.150s2==max(hyp.length.150s2))]
LHC.m_150s2=c(hyp.longest.m,LHC.m_150s2)
LHC.m_150s2
if (is.matrix(D[which(lam.r1.150s2!=-Inf),])==TRUE)
area=(1/2)*(cat.length.150s2^2)*apply(D
  [which(lam.r1.150s2!=-Inf),],1,sum) else
area=(1/2)*(cat.length.150s2^2)*sum(D[which(lam.r1.150s2!=-Inf),])
area.largest=which(lam.r1.150s2!=-Inf)[which(area==max(area))]
TAC.m_150s2=c(area.largest,TAC.m_150s2)
TAC.m_150s2
BICm=which(r_150s2==1,arr.ind=TRUE)[,1][lambda==round(log(n4),1)]
```

```
BIC.m=c(BIC.m,BICm)
BIC.m
AICm=which(r_150s2==1,arr.ind=TRUE)[,1][lambda==2]
AIC.m=c(AIC.m,AICm)
AIC.m
o=order(q_150s2);bc=q_150s2[o[length(q_150s2)]];
r1.model=c(r1.model,bc)
}
r1.model;f=count(labels(r1.model))
LCC.ppn_150s2=count(labels(r1.model))$freq/sum(count
                            (labels(r1.model))$freq);
LCC.tab_150s2=cbind(f,LCC.ppn_150s2)
LCC.tab_150s2
a10=data.table(LCC.tab_150s2)
LCC.tab_150s2=a10[order(-a10$LCC.ppn_150s2)]
LCC.tab_150s2
LHC.ppn_150s2=table(LHC.m_150s2)/sum(table(LHC.m_150s2))
LHC.tab_150s2=sort(LHC.ppn_150s2,decreasing=TRUE)
LHC.tab_150s2
TAC.ppn_150s2=table(TAC.m_150s2)/sum(table(TAC.m_150s2))
TAC.tab_150s2=sort(TAC.ppn_150s2,decreasing=TRUE)
TAC.tab_150s2
freq.AIC=count(AIC.m)
AIC.ppn_150s2=freq.AIC$freq/sum(freq.AIC$freq)
AIC.tab_150s2=cbind(freq.AIC,AIC.ppn_150s2)
a11=data.table(AIC.tab_150s2)
AIC.tab_150s2=a11[order(-a11$AIC.ppn_150s2)]
AIC.tab_150s2
freq.BIC=count(BIC.m)
BIC.ppn_150s2=freq.BIC$freq/sum(freq.BIC$freq)
BIC.tab_150s2=cbind(freq.BIC,BIC.ppn_150s2)
a12=data.table(BIC.tab_150s2)
```

```
BIC.tab_150s2=a12[order(-a12$BIC.ppn_150s2)]
BIC.tab_150s2
LiM=list(LCC.tab_300s1,LHC.tab_300s1,TAC.tab_300s1
,AIC.tab_300s1,BIC.tab_300s1,
LCC.tab_300s2,LHC.tab_300s2,TAC.tab_300s2
,AIC.tab_300s2,BIC.tab_300s2,
LCC.tab_150s1,LHC.tab_150s1,TAC.tab_150s1
,AIC.tab_150s1,BIC.tab_150s1,
LCC.tab_150s2,LHC.tab_150s2,TAC.tab_150s2
,AIC.tab_150s2,BIC.tab_150s2)
LiM
my_names=c("LCC","LHC","TAC","AIC","BIC"
           ,"LCC","LHC","TAC","AIC","BIC"
           ,"LCC","LHC","TAC","AIC","BIC"
           ,"LCC","LHC","TAC","AIC","BIC")
setNames(LiM, my_names)
```

D

Data and Acronyms

D.1 Ovarian cancer data from R package survival

```
> library(survival)
> ovarian
   futime fustat     age resid.ds rx ecog.ps
1      59      1 72.3315        2  1       1
2     115      1 74.4932        2  1       1
3     156      1 66.4658        2  1       2
4     421      0 53.3644        2  2       1
5     431      1 50.3397        2  1       1
6     448      0 56.4301        1  1       2
7     464      1 56.9370        2  2       2
8     475      1 59.8548        2  2       2
9     477      0 64.1753        2  1       1
```

10	563	1	55.1781	1	2	2
11	638	1	56.7562	1	1	2
12	744	0	50.1096	1	2	1
13	769	0	59.6301	2	2	2
14	770	0	57.0521	2	2	1
15	803	0	39.2712	1	1	1
16	855	0	43.1233	1	1	2
17	1040	0	38.8932	2	1	2
18	1106	0	44.6000	1	1	1
19	1129	0	53.9068	1	2	1
20	1206	0	44.2055	2	2	1
21	1227	0	59.5890	1	2	2
22	268	1	74.5041	2	1	2
23	329	1	43.1370	2	1	1
24	353	1	63.2192	1	2	2
25	365	1	64.4247	2	2	1
26	377	0	58.3096	1	2	1

D.2 List of abbreviations and acronyms

Table D.1: Abbreviations and acronyms

AFT	Accelerated failure time
AIC	Akaike information criterion
ALP	Alkaline phosphatise
BIC	Bayesian information criterion
CA125	Cancer Antigen 125
CI	Confidence interval
EGFR	Epidermal growth factor receptor
GIC	Generalised information criterion
IQR	Inter-quartile range
K-M	Kaplan-Meier
lasso	least absolute shrinkage and selection operator
LCC	Longest cathetus criterion
LHC	Longest hypotenuse criterion
MAD	Median absolute deviation
MLE	Maximum likelihood estimate
MSC	Model selection curves
pdf	probability density function
RD	Residual disease
RPA	Royal Prince Alfred
SCAD	Smoothly clipped absolute deviation
SD	Standard deviation
SE	Standard error
TAC	Triangle area criterion

D.3 Number of patients at risk (n.risk) and dying (n.event) at each time point, $\hat{S}(t)$ (survival function) with SE and 95% CI for RPA data

time	n.risk	n.event	survival	SE	lower 95% CI	upper 95% CI
0.0438	333	1	0.997	0.00300	0.979	1.000
0.0575	331	1	0.994	0.00424	0.976	0.998
0.0767	329	1	0.991	0.00519	0.972	0.997
0.0794	328	1	0.988	0.00599	0.968	0.995
0.1314	324	1	0.985	0.00670	0.964	0.994
0.1834	322	1	0.982	0.00735	0.960	0.992
0.1971	321	1	0.979	0.00794	0.956	0.990
0.2218	319	1	0.976	0.00848	0.952	0.988
0.2245	318	1	0.973	0.00900	0.948	0.986
0.2656	314	1	0.970	0.00949	0.944	0.983
0.2683	312	1	0.966	0.00995	0.940	0.981
0.3066	309	1	0.963	0.01040	0.936	0.979
0.3368	307	1	0.960	0.01083	0.932	0.977
0.3641	305	1	0.957	0.01124	0.928	0.974
0.3723	304	1	0.954	0.01164	0.925	0.972
0.4873	291	1	0.951	0.01205	0.921	0.969
0.5339	287	1	0.947	0.01245	0.917	0.967
0.5941	284	1	0.944	0.01285	0.912	0.964
0.6379	281	1	0.941	0.01323	0.908	0.962
0.6845	273	1	0.937	0.01363	0.904	0.959
0.7365	269	1	0.934	0.01402	0.900	0.956
0.7474	268	1	0.930	0.01439	0.896	0.954
0.7584	266	1	0.927	0.01475	0.892	0.951
0.7693	264	2	0.920	0.01545	0.883	0.945
0.7776	262	1	0.916	0.01579	0.879	0.942
0.8433	259	1	0.913	0.01612	0.875	0.939

D.3 Number of patients at risk (n.risk) and dying (n.event) at each time point, $\hat{S}(t)$ (survival function) with SE and 95% CI for RPA data

time	n.risk	n.event	survival	SE	lower 95% CI	upper 95% CI
0.9172	258	1	0.909	0.01644	0.871	0.936
0.9254	257	1	0.906	0.01675	0.867	0.933
0.9473	255	1	0.902	0.01706	0.863	0.931
0.9829	254	1	0.898	0.01736	0.859	0.928
1.0705	246	1	0.895	0.01767	0.854	0.925
1.1006	245	1	0.891	0.01797	0.850	0.921
1.1362	243	1	0.887	0.01827	0.846	0.918
1.2183	239	1	0.884	0.01856	0.842	0.915
1.2238	238	1	0.880	0.01885	0.837	0.912
1.2704	236	1	0.876	0.01914	0.833	0.909
1.2923	235	1	0.873	0.01942	0.829	0.906
1.3114	233	1	0.869	0.01969	0.825	0.903
1.3142	232	1	0.865	0.01996	0.820	0.899
1.3443	227	1	0.861	0.02023	0.816	0.896
1.3635	226	1	0.857	0.02050	0.812	0.893
1.3881	223	1	0.854	0.02076	0.807	0.889
1.4073	220	1	0.850	0.02103	0.803	0.886
1.4620	218	1	0.846	0.02129	0.799	0.883
1.4702	217	1	0.842	0.02155	0.794	0.879
1.4894	216	1	0.838	0.02180	0.790	0.876
1.5031	212	1	0.834	0.02205	0.786	0.873
1.5524	210	1	0.830	0.02230	0.781	0.869
1.5770	208	1	0.826	0.02255	0.777	0.866
1.6400	207	1	0.822	0.02279	0.772	0.862
1.6756	204	1	0.818	0.02303	0.768	0.859
1.7084	202	1	0.814	0.02327	0.763	0.855
1.7112	201	1	0.810	0.02350	0.759	0.851
1.7468	198	2	0.802	0.02397	0.750	0.844
1.7878	195	1	0.798	0.02419	0.745	0.841
1.8398	192	1	0.794	0.02442	0.741	0.837
1.9302	185	1	0.789	0.02466	0.736	0.833

time	n.risk	n.event	survival	SE	lower 95% CI	upper 95% CI
1.9795	183	1	0.785	0.02490	0.731	0.829
1.9932	182	2	0.776	0.02537	0.722	0.822
2.0479	178	1	0.772	0.02560	0.717	0.818
2.0589	176	1	0.768	0.02582	0.712	0.814
2.0862	175	1	0.763	0.02605	0.707	0.810
2.0999	173	1	0.759	0.02627	0.703	0.806
2.1766	170	1	0.754	0.02649	0.698	0.802
2.2971	167	1	0.750	0.02671	0.693	0.798
2.3381	165	1	0.745	0.02693	0.688	0.794
2.3847	164	1	0.741	0.02715	0.683	0.790
2.5051	157	1	0.736	0.02738	0.678	0.785
2.5489	155	1	0.731	0.02762	0.673	0.781
2.5599	154	1	0.727	0.02784	0.668	0.777
2.5763	153	1	0.722	0.02806	0.662	0.773
2.6639	151	1	0.717	0.02828	0.657	0.768
2.7351	147	1	0.712	0.02851	0.652	0.764
2.7844	145	1	0.707	0.02873	0.647	0.759
2.8939	142	1	0.702	0.02895	0.641	0.755
2.9487	140	1	0.697	0.02918	0.636	0.750
2.9952	137	1	0.692	0.02941	0.630	0.746
3.0527	133	1	0.687	0.02964	0.625	0.741
3.0554	132	1	0.682	0.02987	0.619	0.736
3.0992	130	1	0.676	0.03010	0.613	0.731
3.1266	126	1	0.671	0.03033	0.608	0.727
3.1321	124	1	0.666	0.03057	0.602	0.722
3.1485	123	1	0.660	0.03080	0.596	0.717
3.1759	122	1	0.655	0.03101	0.590	0.712
3.3265	119	1	0.649	0.03124	0.584	0.707
3.6112	116	1	0.644	0.03147	0.578	0.702
3.6550	114	1	0.638	0.03169	0.572	0.696
3.8877	110	1	0.632	0.03193	0.566	0.691

time	n.risk	n.event	survival	SE	lower 95% CI	upper 95% CI
4.0876	104	1	0.626	0.03220	0.560	0.686
4.6133	96	1	0.620	0.03252	0.553	0.680
4.8789	90	1	0.613	0.03288	0.545	0.674
5.0541	83	1	0.605	0.03330	0.537	0.667
5.4894	71	1	0.597	0.03390	0.527	0.660
5.8809	65	1	0.588	0.03460	0.517	0.652
7.6003	41	1	0.573	0.03661	0.498	0.641
8.7912	25	1	0.550	0.04171	0.465	0.628
9.9302	13	1	0.508	0.05601	0.394	0.611

Bibliography

Aitkin, M., Francis, B., Hinde, J., and Darnell, R. (2009). Statistical Modelling in R. Oxford University Press, New York.

Aitkin, M., Francis, B., Hinde, J., and Darnell, R. (2012). SMIR: Companion to Statistical Modelling in R. R package version 0.02. http://CRAN.R-project.org/package=SMIR.

Aitkin, M., Laird, N., and Francis, B. (1983). A reanalysis of the Stanford heart transplant data. Journal of the American Statistical Association, 78, 264–274.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Proceedings of the Second International Symposium of Information Theory, Eds. B.N. Petrov, F. Csáki, pp. 267–281. Akadémiai Kiadó: Budapest.

Allison, P. D. (2010). Survival Analysis, in The Reviewer's Guide to Quantitative Methods in the Social Sciences, edited by G. R. Hancock and R. O. Mueller. Routledge, New York.

Barnholtz-Sloan, J. S., Schwartz, A. G., Qureshi, F., Jacques, S., Malone, J., and Munkarah, A. R. (2003). Ovarian Cancer: Changes in patterns at diagnosis and relative survival over the last three decades. American Journal of Obstetrics and Gynecology, 189, 1120–1127.

Bast, R. C. Jr, Feeney, M., Lazarus, H., Nadler, L. M., Colvin, R. B., and Knapp, R. C. (1981). Reactivity of a monoclonal antibody with human ovarian carcinoma. J Clin Invest, 68, 1331–1337.

Bennett, S. (1983). Log-logistic regression models for survival data. Applied Statistics, 32, 165–171.

Berretoni, J. N. (1964). Practical applications of the Weibull distribution. Industrial Quality Control, 21, 71–79.

Boag, J. W. (1949). Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy. Journal of the Royal Statistical Society, Series B, 11, 15–53.

Braun, W. J. (2015). MPV: Data Sets from Montgomery, Peck and Vining's Book. R package version 1.38. http://CRAN.R-project.org/package=MPV.

Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006). The Design of Simulation Studies in Medical Statistics. Statistics in Medicine, 25, 4279–4292.

Canty, A., and Ripley, B. (2015). boot: Bootstrap R (S-Plus) Functions. R package version 1.3–16.

Chan, J. K., Cheung, M. K., Husain, A., Teng, N. N., West, D., Whittemore, A. S., Berek, J. S., and Osann, K. (2006). Patterns and progress in ovarian cancer over 14 years. Obstetrics & Gynecology, 108, 521–528.

Claeskens, G., and Hjort, N. L. (2008). Model Selection and Model Averaging. Cambridge University Press, Cambridge.

Clark, D. E., and El-Taha, M. (2015). Some Useful Properties of Log-Logistic Random Variables for Health Care Simulations. International Journal of Statistics in Medical Research, 4(1), 79.

Collett, D. (2003). Modelling Survival Data in Medical Research, 2nd ed. Chapman & Hall, London.

Colombo, P. E., Mourregot, A., Fabbro, M., Gutowski, M., Saint-Aubert, B., Quenet, F., Gourgou, S., and Rouanet, P. (2009). Aggressive surgical strategies in advanced ovarian cancer: A monocentric study of 203 stage IIIC and IV patients. The Journal of Cancer Surgery, 35, 135–143.

Cox, D.R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society, Series B, 34, 187–220.

Cox, D. R., and Oakes, D. (1984). Analysis of survival data. Chapman & Hall, New York.

Crowley, J. and Hu, M. (1977). Covariance analysis of heart transplant survival data. Journal of the American Statistical Association, 72, 27–36.

Davison, A. C., and Hinkley, D. V. (1997). Bootstrap Methods and their Applications. Cambridge University Press, Cambridge.

Dowle, M., Srinivasan, A., Short, T., Lianoglou, S., and Antonyan, E. (2015). data.table: Extension of Data.frame. R package version 1.9.6. http://CRAN.R-project.org/package=data.table.

Edmunson, J. H., Fleming, T. R., Decker, D. G., Malkasian, G. D., Jorgensen, E. O., Jefferies, J. A., Webb, M. J., and Kvols, L. K. (1979). Different chemotherapeutic sensitivities and host factors affecting prognosis in advanced ovarian carcinoma versus minimal residual disease. Cancer Treatment Reports, 63, 241–247.

Eisenhauer, E. A., Gore, M., and Neijt, J. P. (1999). Ovarian cancer: should we be managing patients with good and bad prognostic factors in the same manner? Annals of Oncology, 10, 9–15.

Fan, J., and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. The Annals of Statistics, 30, 74–99.

Faraggi, D., and Simon, R. (1998). Bayesian Variable Selection Method for Censored Survival Data. Biometrics, 54, 1475–1485.

Felder, M., Kapur, A., Gonzalez-Bosquet, J., Horibata, S., Heintz, J., Albrecht, R., Fass, L., Kaur, J., Hu, K., Shojaei, H., Whelan, R. J., and Patankar, M. S. (2014). MUC16 (CA125): tumor biomarker to cancer therapy, a work in progress. Molecular Cancer, 13, 129.

Fisk, P. R. (1961). The graduation of income distributions. Econometrica, 29, 171–185.

Greenberg, R. A., Bayard, S., and Byar, D. (1974). Selecting concomitant variables using a likelihood ratio step-down procedure and a method of testing goodness of fit in an exponential survival model. Biometrics, 30, 601–608.

Gui, J., and Li, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics, 21, 3001–3008.

Hosmer, D. W., and Lemeshow, S. (1999). Applied Survival Analysis, 2nd ed. Wiley, New York.

Hougaard, P. (1995). Frailty models for survival analysis. Lifetime Data Analysis, 1, 255–273.

Huang, J., and Ma, S. (2010). Variable selection in the accelerated failure time model via the bridge method. Lifetime Data Analysis, 16, 176–195.

Hurvich, C. M., and Tsai, C. (1989). Regression and Time Series Model Selection in Small Samples. Biometrika, 76, 297–307.

Kaplan, E. L., and Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53, 457–481.

Karami, J. H., Luo, K., and Fung, T. (2015). Model selection curves for survival analysis with accelerated failure time models. Proceedings of the 60th ISI World Statistics Congress, Rio de Janeiro, July 26–31, p.2642–2647.

Klein, J. P., and Moeschberger, M. L. (2003). Survival Analysis: techniques for censored and truncated data, 2nd ed. Springer-Verlag, New York.

Kleinbaum, D. G., and Klein, M. (2012). Survival Analysis: A Self-Learning Text, 3rd ed. Springer, New York.

Konishi, S., and Kitagawa, G. (1996). Generalised Information Criteria in Model Selection. Biometrika, 83, 875–890.

Kotsopoulos, J., Moody, J. R. K., Fan, I., Rosen, B., Risch, H. A., McLaughlin, J. R., Sun, P., and Narod, S. A. (2012). Height, weight, BMI and ovarian cancer survival. Gynecologic Oncology, 127, 83–87.

Krall, J. M., Uthoff, V. A., and Harley, J. B. (1975). A step-up procedure for selecting variables associated with survival. Biometrics, 31, 49–57.

Lawless, J. F. (1982). Statistical Models and Methods for Lifetime Data. Wiley, New York.

Lee, E. T. and Wang, J. W. (2003). Statistical Methods for Survival Data Analysis, 3rd ed. Wiley, New Jersey.

Liang, H., and Zou, G. (2008). Improved AIC Selection Strategy for Survival Analysis. Comput Stat Data Anal, 52, 2538–2548.

Lindley, D. V. (1968). The Choice of Variables in Multiple Regression (with discussion). Journal of the Royal Statistical Society, Series B, 30, 31–66.

Ma, S., and Huang, J. (2007). Additive risk survival model with microarray data. BMC Bioinformatics, 8:192.

Markmann, S., Gerber, B., and Bries, V. (2007). Prognostic Value of Ca 125 Levels during Primary Therapy. Anticancer research, 27, 1837–1840.

Matloff, N. (2016). Parallel Computing for Data Science With Examples in R, C++ and CUDA. CRC press, London.

Meier, L., van de Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. Journal of the Royal Statistical Society, Series B, 70, 53–71.

Meinshausen, N., and Bühlmann, P. (2010). Stability selection. Journal of the Royal Statistical Society, Series B, 72, 417–473.

Moriña, D., and Navarro, A. (2014). The R package survsim for the simulation of simple and complex survival data. Journal of Statistical Software, 59(2), 1–20.

Müller, S., and Welsh, A. H. (2010). On Model Selection Curves. International Statistical Review, 78, 240–256.

Murray, K., Heritier, S., and Müller, S. (2013). Graphical tools for model selection in generalized linear models. Statistics in Medicine, 32, 4438–4451.

Nelson, W., and Hahn, G. J. (1972). Linear Estimation of a Regression Relationship from Censored Data Part I–Simple Methods and Their Application. Technometrics, 14, 247–269.

Novomestky, F. (2012). matrixcalc: Collection of functions for matrix calculations. R package version 1.0-3. http://CRAN.R-project.org/package=matrixcalc.

Obermair, A., Tang, A., Kondalsamy-Chennakesavan, S., Ngan, H., Zusterzeel, P., Quinn, M., Carter, J., Leung, Y., and Janda, M. (2013). Nomogram to predict the probability of relapse in patients diagnosed with borderline ovarian tumors. International journal of gynecological cancer, 23, 264–267.

O'Quigley, J., and Struthers, L. (1982). Survival models based upon the logistic and log-logistic distributions. Computer Programs in Biomedicine, 15, 3–12.

Orbe, J., Ferreira, E., Núñez-Antón, V. (2002). Comparing proportional hazards and accelerated failure time models for survival analysis. Statistics in Medicine, 21, 3493–3510.

Peduzzi, P. N., Hardy, R. J., and Holford, T. R. (1980). A Stepwise Variable Selection Procedure for Nonlinear Regression Models. Biometrics, 36, 511–516.

Pike, M. C. (1966). A method of analysis of a certain class of experiments in carcinogenesis. Biometrics, 22, 142–161.

Polterauer, S., Vergote, I., Concin, N., Braicu, I., Chekerov, R., Mahner, S., Woelber, L., Cadron, I., Van Gorp, T., Zeillinger, R., Castillo-Tong, D. C., and Sehouli, J. (2012). Prognostic Value of Residual Tumor Size in Patients with Epithelial Ovarian Cancer FIGO Stages IIA–IV: Analysis of the OVCAD Data. International Journal of Gynecological Cancer, 22, 380–385.

Prentice, R. L. (1973). Exponential survivals with censoring and explanatory variables. Biometrika, 60, 279–288.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rinne, H. (2008). The Weibull distribution: A Handbook. CRC press, London.

Roett, M. A., and Evans, P. (2009). Ovarian Cancer: An Overview. American Family Physician, 80, 609–616.

Schmidt, C. (2011). CA-125: A Biomarker put to the test. Journal of the national cancer institute, 103, 1290–1291.

Schwarz, G. (1978). Estimating the dimension of a model. Annals of statistics, 6, 461–464.

Seiden, M. V. (2001). Ovarian Cancer. The Oncologist, 6, 327–332.

Tadikamalla, P. R., and Johnson, N. L. (1982). Systems of frequency curves generated by transformations of logistic variables. Biometrika, 69, 461–5465.

Tarr, G., Müller, S., and Welsh, A. (2015). mplot: An R package for graphical model stability and variable selection. R package version 0.7.7. http://arxiv.org/abs/1509.07583.

Therneau, T. (2015). A Package for Survival Analysis in S. Version 2.38. http://CRAN.R-project.org/package=survival.

Therneau, T. M., and Grambsch, P. M. (2000). Modeling Survival Data: Extending the Cox Model. Springer, New York.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. Statistics in medicine, 16, 385–395.

Tingulstad, S., Skjeldestad, F. E., Halvorsen, T. B., and Hagen, B. (2003). Survival and Prognostic Factors in Patients With Ovarian Cancer. Obstetrics & Gynecology, 101, 885–891.

Venables, W. N., and Ripley, B. D. (2002). Modern Applied Statistics with S, 4th ed. Springer, New York.

Volinsky, C. T., and Raftery, A. E. (2000). Bayesian information criterion for censored survival models. Biometrics, 56, 256–262.

Wang, S., Nan, B., Zhu, J., and Beer, D. G. (2008). Doubly Penalized Buckley-James Method for Survival Data with High-Dimensional Covariates. Biometrics, 64, 132–140.

Weibull, W. (1939). A statistical theory of the strength of materials. Ingeniöers vetenskaps akademiens handlingar, 151, 1–45.

Weibull, W. (1951). A statistical distribution function of wide applicability. Journal of Applied Mechanics, 18, 293–297.

Whittemore, A., and Altshuler, B. (1976). Lung Cancer Incidence in Cigarette Smokers: Further Analysis of Doll and Hill's Data for British Physicians. Biometrics, 32, 805–816.

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1–29.

Winter, W. E. 3rd., Maxwell G. L., Tian C., Sundborg M. J., Rose G. S., Rose P. G., Rubin S. C., Muggia F., and McGuire, W. P. (2008). Tumor residual after surgical cytoreduction in prediction of clinical outcome in stage IV epithelial ovarian cancer: A gynecologic oncology group study. Journal of Clinical Oncology, 26, 83–89.

Zhang, M., Xie, X., and Holman, C. J. (2005). Body weight and body mass index and ovarian cancer risk: A case-control study in China. Gynecologic Oncology, 98, 228–234.