Comparison of Proportional Hazards and Accelerated Failure Time Models

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the Degree of
Master of Science
in the
Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon, Saskatchewan

By
Jiezhi Qi
Mar. 2009

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5E6

Abstract

The field of survival analysis has experienced tremendous growth during the latter half of the 20th century. The methodological developments of survival analysis that have had the most profound impact are the Kaplan-Meier method for estimating the survival function, the log-rank test for comparing the equality of two or more survival distributions, and the Cox proportional hazards (PH) model for examining the covariate effects on the hazard function. The accelerated failure time (AFT) model was proposed but seldom used. In this thesis, we present the basic concepts, nonparametric methods (the Kaplan-Meier method and the log-rank test), semiparametric methods (the Cox PH model, and Cox model with time-dependent covariates) and parametric methods (Parametric PH model and the AFT model) for analyzing survival data.

We apply these methods to a randomized placebo-controlled trial to prevent Tuberculosis (TB) in Ugandan adults infected with Human Immunodificiency Virus (HIV). The objective of the analysis is to determine whether TB preventive therapies affect the rate of AIDS progression and survival in HIV-infected adults. Our conclusion is that TB preventive therapies appear to have no effect on AIDS progression, death and combined event of AIDS progression and death. The major goal of this paper is to support an argument for the consideration of the AFT model as an alternative to the PH model in the analysis of some survival data by means of this real dataset. We critique the PH model and assess the lack of fit. To overcome the violation of proportional hazards, we use the Cox model with time-dependent covariates, the piecewise exponential model and the accelerated failure time model. After comparison of all the models and the assessment of goodness-of-fit, we find that the log-logistic AFT model fits better for this data set. We have seen that the AFT model is a more valuable and realistic alternative to the PH model in some situations. It can provide the predicted hazard functions, predicted survival functions, median survival times and time ratios. The AFT model can easily interpret the results into the

effect upon the expected median duration of illness for a patient in a clinical setting. We suggest that the PH model may not be appropriate in some situations and that the AFT model could provide a more appropriate description of the data.

Acknowledgements

This thesis grew out of a research project provided by my co-supervisor Dr. Hyun Ja Lim. I'm deeply indebted to Dr. Lim, who opened my eyes for survival analysis and guided me through. I am sincerely grateful to my co-supervisor, Dr. Mikelis G. Bickis, for his invaluable advice and patient guidance. This thesis could not have been written without their constant help and support. I would like to thank the members of my committee, Prof. Raj Srinivasan and Prof. Chris Soteros and my external examiner, Prof. Xulin Guo for reading my thesis and valuable suggestions. Last but not least, I want to thank my family and friends, for their support and encouragement.

To

My parents

Yuhua Shang and Tonglian Qi

My husband

Zhidong Zhang

My daughter

Erin Jiaqi Zhang

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Survival analysis is a statistical method for data analysis where the outcome variable of interest is the time to the occurrence of an event [35]. Hence, survival analysis is also referred to as "time to event analysis", which is applied in a number of applied fields, such as medicine, public health, social science, and engineering. In medical science, time to event can be time until recurrence in a cancer study, time to death, or time until infection. In the social sciences, interest can lie in analyzing time to events such as job changes, marriage, birth of children and so forth. The engineering sciences have also contributed to the development of survival analysis which is called failure time analysis since the main focus is in modelling the lifetimes of machines or electronic components [37]. The developments from these diverse fields have for the most part been consolidated into the field of survival analysis. Because these methods have been adapted by researchers in different fields, they also have several different names: event history analysis (sociology), failure time analysis (engineering), duration analysis or transition analysis (economics). These different names do not imply any real difference in techniques, although different disciplines may emphasize slightly different approaches. Survival analysis is the name that is most widely used and recognized [38].

The complexities provided by the presence of censored observations led to the development of a new field of statistical methodology. The methodological developments in survival analysis were largely achieved in the latter half of the 20th century. Although Bayesian methods in survival analysis [26] are well developed and are becoming quite common for survival data, our application will focus on frequentist methods. There have been several textbooks written that address survival analysis from a frequentist perspective. These include Lawless [37], Cox and Oakes [14], Fleming and Harrington [18], and Klein and Moeschberger [34].

One of the oldest and most straightforward non-parametric methods for analyzing

survival data is to compute the life table, which was proposed by Berkson and Gage [6] for studying cancer survival. One important development in non-parametric analysis methods was obtained by Kaplan and Meier [33]. While non-parametric methods work well for homogeneous samples, they do not determine whether or not certain variables are related to the survival times. This need leads to the application of regression methods for analyzing survival data. The standard multiple linear regression model is not well suited to survival data for several reasons. Firstly, survival times are rarely normally distributed. Secondly, censored data result in missing values for the dependent variable (survival time) [35]. The Cox proportional hazards (PH) model is now the most widely used for the analysis of survival data in the presence of covariates or prognostic factors. This is the most popular model for survival analysis because of its simplicity, and not being based on any assumptions about the survival distribution. The model assumes that the underlying hazard rate is a function of the independent covariates, but no assumptions are made about the nature or shape of the hazard function. In the last several years, the theoretical basis for the model has been solidified by connecting it to the study of counting processes and martingale theory, which was discussed in the books of Fleming and Harrington [18] and of Andersen et al [2]. These developments have led to the introduction of several new extensions to the original model. However the Cox PH model may not be appropriate in many situations and other modifications such as stratified Cox model [35] or Cox model with time-dependent variables [10] can be used for the analysis of survival data. The accelerated failure time (AFT) [10] model is another alternative method for the analysis of survival data.

The purpose of this thesis is to compare the performance of the Cox models and the AFT models. This will be studied by means of real dataset which is from a randomized placebo-controlled trial to prevent tuberculosis (TB) in Ugandan adults infected with human immunodeficiency virus (HIV).

The rest of thesis is organized as follows. In the rest of this chapter, we introduce the main concepts and survival distributions in survival analysis. In Chapter 2, we discuss Kaplan-Meier survival curves and non-parametric test such as the log-rank test [40]. In Chapter 3, we start with an introduction of the Cox PH model which is the most popular regression model in survival analysis. Then we will discuss the estimation and assumptions in the Cox PH model. Model checking using residuals is also described. At last we describe

the methodology when the PH assumption is violated. In Chapter 4, we will describe the parametric PH model and the AFT model. The main objective of the first four chapters is to develop the background of survival analysis that we will apply to our TB/HIV dataset. In Chapter 5, we first describe some background knowledge of TB/HIV and the dataset we will use. Then we fit all methods described in the first four chapters to the dataset and give the results. At last, we summarize our experience of using the Cox models versus the AFT models. Chapter 6 provides a summary of the discussion on this study and further research on the subject is discussed.

1.1 Basic concepts

Before going into details about survival analysis, we discuss the following basic definitions. The primary concept in survival analysis is survival time which is also called failure time.

Definition 1.1.1 *survival time is a length of time that is measured from time origin to the time the event of interest occurred.*

To determine survival time precisely, there are three requirements: A time origin must be unambiguously defined, a scale for measuring the passage of time must be agreed upon and finally the definition of event (often called failure) must be entirely clear.

The specific difficulties in survival analysis arise largely from the fact that only some individuals have experienced the event and other individuals have not had the event in the end of study and thus their actual survival times are unknown. This leads to the concept of censoring.

Definition 1.1.2 *Censoring occurred when we have some information about individual survival time, but we do not know the survival time exactly.*

There are three types of censoring: 1) right censoring, 2) left censoring, and 3) interval censoring.

Right censoring is said to occur if the event occurs after the observed survival time. Let C denote the censoring time, that is, the time beyond which the study subject cannot be observed. The observed survival time is also referred to as follow up time. It starts at time 0 and continues until the event X or a censoring time C, whichever comes first.

The observed data are denoted by (T, δ), where $T = \min(X, C)$ is the follow-up time, and $\delta = I_{(X \leq C)}$ is an indicator for status at the end of follow-up,

$$\delta = I_{(X \leq C)} := \begin{cases} 0 & \text{if } X > C \text{ (observed censoring)} \\ 1 & \text{if } X \leq C \text{ (observed failure)} \end{cases} .$$

There are some reasons why right censoring may occur, for example, no event before the study ends, loss to follow-up during study period, or withdrawal from the study because of some reasons. The last reason may be caused by competing risks. The right censored survival time is then less than the actual survival time.

Censoring can also occur if we observe the presence of a condition but do not know where it began. In this case we call it left censoring, and the actual survival time is less than the observed censoring time.

If an individual is known to have experienced an event within an interval of time but the actual survival time is not known, we say we have interval censoring. The actual occurrence time of event is known within an interval of time.

Right censoring is very common in survival time data, but left censoring is fairly rare. The term "censoring" will be used in this thesis to mean in all instances "right censoring". An important assumption for methods presented in this thesis for the analysis of censored survival data is that the individuals who are censored are at the same risk of subsequent failure as those who are still alive and uncensored. i.e., a subject whose survival time is censored at time C must be representative of all other individuals who have survived to that time. If this is the case, the censoring process is called non-informative. Statistically, if the censoring process is independent of the survival time, i.e.,

$$P(X \geq x, C \geq x) = P(X \geq x)P(C \geq x),$$

then we will have non-informative censoring. Independence censoring is a special case of non-informative censoring. In this thesis, we assume that the censoring is non-informative right censoring.

1.2 Survival time distribution

Let T be a random variable denoting the survival time. The distribution of survival times is characterized by any of three functions: the survival function, the probability density function or the hazard function. The following definitions are based on textbook [32].

Note the survival function is defined for both discrete and continuous T, and the probability density and hazard functions are easily specified for discrete and continuous T.

Definition 1.2.1 *The survival function is defined as the probability that the survival time is greater or equal to t.*

$$S(t) = P(T \geq t),\ t \geq 0.$$

1.2.1 T discrete

For a discrete random variable T taking well-ordered values $0 \leq t_1 < t_2 < \cdots$, let the probability mass function be given by $P(T = t_i) = f(t_i)$, $i = 1, 2, ...$, then the survival function is

$$\begin{aligned} S(t) &= \sum_{j|t_j \geq t} f(t_j) \\ &= \sum f(t_j) I_{(t_j \geq t)}, \end{aligned}$$

where the indicator function $I_{(t_j \geq t)} := \begin{cases} 0 & \text{if } t_j < t \\ 1 & \text{if } t_j \geq t \end{cases}$.

In this case, the hazard function $h(t)$ is defined as the conditional probability of failure at time t_j given that the individual has survived up to time t_j,

$$h_j = h(t_j) = P(T = t_j | T \geq t_j) = \frac{f(t_j)}{S(t_j)} = \frac{S(t_j) - S(t_{j+1})}{S(t_j)} = 1 - \frac{S(t_{j+1})}{S(t_j)}.$$

Thus,

$$1 - h(t_j) = \frac{S(t_{j+1})}{S(t_j)},$$

and

$$\prod_{j|t_j < t} (1 - h(t_j)) = \frac{S(t_2)}{S(t_1)} \times \frac{S(t_3)}{S(t_2)} \times ... \times \frac{S(t_{j+1})}{S(t_j)} = S(t), \tag{1.1}$$

because $S(t_1) = 1$ and $S(t) = S(t_{j+1})$.

Moreover,

$$\begin{aligned} f(t_j) &= h(t_j) \times S(t_j) \\ &= h(t_j) \prod_{i=1}^{j-1} (1 - h(t_i)). \end{aligned} \tag{1.2}$$

1.2.2 T absolutely continuous

For an absolutely continuous variable T, The probability density function of T is

$$f(t) = F^{'}(t) = -S'(t),\ t \geq 0.$$

Definition 1.2.2 *The hazard function gives the instantaneous failure rate at t given that the individual has survived up to time t, i.e.,*

$$h(t) = \lim_{\Delta t \downarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t},\ t \geq 0.$$

There is a clearly defined relationship between $S(t)$ and $h(t)$, which is given by the formula

$$h(t) = f(t)/S(t) = \frac{-d \log S(t)}{dt}, \tag{1.3}$$

$$S(t) = \exp\left[-\int_0^t h(u)du\right] = \exp(-H(t)),\ t \geq 0, \tag{1.4}$$

where $H(t) = \int_0^t h(u)du$ is called the cumulative hazard function, which can be obtained from the survival function since $H\left(t\right) = -\log S\left(t\right)$.

The probability density function of T can be written

$$f(t) = h(t) \exp\left[-\int_0^t h(u)du\right],\ t \geq 0.$$

These three functions give mathematically equivalent specification of the distributions of the survival time T. If one of them is known, the other two are determined. One of these functions can be chosen as the basis of statistical analysis according to the particular situations. The survival function is most useful for comparing the survival progress of two or more groups. The hazard function gives a more useful description of the risk of failure at any time point.

Chapter 2

Non-parametric methods

In survival analysis, it is always a good idea to present numerical or graphical summaries of the survival times for the individuals. In general, survival data are conveniently summarized through estimates of the survival function and hazard function. The estimation of the survival distribution provides estimates of descriptive statistics such as the median survival time [10]. These methods are said to be non-parametric methods since they require no assumptions about the distribution of survival time. In order to compare the survival distribution of two or more groups, log-rank tests [40] can be used.

2.1 The Kaplan-Meier estimate of the survival function

The life table [6] is the earliest statistical method to study human mortality rigorously, but its importance has been reduced by the modern methods, like the Kaplan-Meier (K-M) method [33]. In clinical studies, individual data is usually available on time to death or time to last seen alive. The K-M estimator for the survival curves is usually used to analyze individual data, whereas the life table method applies to grouped data. Since the life table method is a grouped data statistic, it is not as precise as the K-M estimate, which uses the individual values. We only describe the K-M estimate here.

Suppose that r individuals have failures in a group of individuals. Let $0 \leq t_{(1)} < ... < t_{(r)} < \infty$ be the observed ordered death times. Let r_j be the size of the risk set at $t_{(j)}$, where risk set denotes the collection of individuals alive and uncensored just before $t_{(j)}$. Let d_j be the number of observed deaths at $t_{(j)}$, $j = 1, ..., r$. Then the K-M estimator of $S(t)$ is defined by

$$\widehat{S}(t) = \prod_{j:t_{(j)}<t} (1 - \frac{d_j}{r_j}).$$

This estimator is a step function that changes values only at the time of each death. In fact, K-M estimator will be shown next to maximize the likelihood in the discrete case

[14].

Suppose that the distribution is discrete, with atoms h_j at finitely many specified points $0 \leq \tau_1 < \tau_2 < \cdots < \tau_j$. As described in Section 1.2, the survival function $S(t)$ may be expressed in terms of the discrete hazard function h_j as

$$S(t) = \prod_{j|\tau_j<t} (1 - h_j).$$

To derive the full likelihood from a sample of n observations, we first collect all the terms corresponding to the atom τ_j. Let $b_i = j$ if the ith individual dies at τ_j. Using (1.2), the contribution to the total log likelihood is

$$\log h_{b_i} + \sum_{k<b_i} \log(1 - h_k).$$

Let $e_i = j$ if the ith individual is censored at τ_j, using the equation (1.1), the log likelihood contribution to the total likelihood is

$$\sum_{k \leq e_i} \log(1 - h_k).$$

Then the total log likelihood is given by

$$\begin{aligned} l &= \sum_{\text{death } i} \log h_{b_i} + \sum_{\text{death } i} \left[\sum_{k<b_i} \log(1 - h_k) \right] + \sum_{\text{censor } i} \left[\sum_{k \leq e_i} \log(1 - h_k) \right] \\ &= \sum_j d_j \log h_j + \sum_k \left[\sum_{j>k} d_j \right] \log(1 - h_k) + \sum_k \left[\sum_{j \geq k} c_j \right] \log(1 - h_k) \\ &= \sum_j [d_j \log h_j + (r_j - d_j) \log(1 - h_j)], \end{aligned}$$

where d_j is the number of observed death at τ_j, c_j is the number censored at $[\tau_j, \tau_{j+1})$, and r_j is the number of living and uncensored at τ_j.

If h_j is the solution of

$$\frac{\partial l}{\partial h_j} = \frac{d_j}{h_j} - \frac{r_j - d_j}{1 - h_j} = 0,$$

then

$$\widehat{h}_j = d_j / r_j.$$

This maximizes the likelihood since the total log likelihood function is concave down. So that the K-M estimator of the survival function is

$$\begin{aligned}\widehat{S}(t) &= \prod_{j|\tau_j<t}(1-\widehat{h}_j)\\ &= \prod_{j|\tau_j<t}(1-\frac{d_j}{r_j}).\end{aligned}$$

Therefore, the K-M estimator is the maximum likelihood estimator.

The K-M estimator gives a discrete distribution. If the observations are modelled to come from unknown continuous distribution, the maximum likelihood estimator does not exist [27].

2.1.1 Greenwood's formula

Confidence intervals for the survival probability can also be calculated by the well known Greenwood's formula [23].

First, we need the variances of the $\widehat{h}_j s$. Let the number of individual at risk at $t_{(j)}$ be r_j and the number of deaths at $t_{(j)}$ be d_j. Given r_j, the number of individuals surviving through the interval $[t_{(j)}, t_{(j+1)})$, $r_j - d_j$, can be assumed to have binomial distribution with parameters r_j and $1-h_j$. The conditional variance of $r_j - d_j$ is given by

$$V(r_j - d_j|r_j) = r_j h_j (1-h_j).$$

The variance of $\widehat{h}_j$ is

$$V\left(\widehat{h}_j|r_j\right) = V\left(1-\widehat{h}_j\right) = V\left(1-\frac{d_j}{r_j}\right) = \frac{h_j\,(1-h_j)}{r_j}.$$

Since $\widehat{h}_j$ is conditional independent of $\widehat{h}_1$,...,$\widehat{h}_{j-1}$ given $r_1, ..., r_{j-1}$, the delta method [11] can be used to obtain

$$\begin{aligned}V(\ln \widehat{S}(t)|r_j : t_{(j)} < t) &= V\left[\sum_{j:t_{(j)}<t}(\ln(1-\widehat{h}_j))|r_j\right]\\ &= \sum_{j:t_{(j)}<t} V\left[\ln(1-\widehat{h}_j)|r_j\right]\\ &\approx \sum_{j:t_{(j)}<t}(\frac{d}{dx}\ln(1-x))^2_{x=\widehat{h}_j} V\left(\widehat{h}_j|r_j\right)\\ &= \sum_{j:t_{(j)}<t}\left\{-\frac{1}{1-\widehat{h}_j}\right\}^2 \frac{h_j\,(1-h_j)}{r_j},\ j=1,...,r.\end{aligned}$$

We can estimate this by simply replacing h_j with $\widehat{h}_j = d_j/r_j$, which gives

$$\widehat{V}\left(\ln \widehat{S}(t)\right) = \sum_{j:t_{(j)}<t} \frac{d_j}{r_j\left(r_j - d_j\right)},\ j = 1, ..., r.$$

Let $Y = \ln \widehat{S}(t)$, again using the delta method, we get

$$\widehat{V}\left(\widehat{S}(t)\right) \approx \left[\widehat{S}(t)\right]^2 \sum_{j:t_{(j)}<t} \frac{d_j}{r_j\left(r_j - d_j\right)}. \tag{2.1}$$

This is known as *Greenwood's formula*. The K-M estimator and functions of it have been proved to be asymptotically normal distributed [2], [18]. Thus the confidence intervals can be constructed by the normal approximation based on $S\left(t\right)$.

2.1.2 Estimating the median and percentile of survival time

Since the distribution of survival time tends to be positively skewed, the median is preferred for a summary measure. The median survival time is the time beyond which 50% of the individuals under study are expected to survive, i.e., the value of $t\left(50\right)$ at $\widehat{S}\left(t\left(50\right)\right) = 0.5$. The estimated median survival time is given by

$$\widehat{t}\left(50\right) = \min\left\{t_i | \widehat{S}\left(t_i\right) < 0.5\right\},$$

where t_i is the observed survival time for the ith individual, $i = 1, 2, ..., n$. In general, the estimate of the pth percentile is

$$\widehat{t}\left(p\right) = \min\left\{t_i | \widehat{S}\left(t_i\right) < 1 - \frac{p}{100}\right\}.$$

A confidence interval for the percentiles using delta method was discussed in the text-books [2], [10], The variance of the estimator of the pth percentile is

$$\begin{aligned} V[\widehat{S}\left(t\left(p\right)\right)] &= \left(\frac{d\widehat{S}\left(t\left(p\right)\right)}{dt\left(p\right)}\right)^2 V\left\{t\left(p\right)\right\} \\ &= \left(-\widehat{f}\left(t\left(p\right)\right)\right)^2 V\left\{t\left(p\right)\right\}. \end{aligned}$$

The standard error of $\widehat{t}\left(p\right)$ is therefore given by

$$\mathrm{SE}\left[\widehat{t}\left(p\right)\right] = \frac{1}{\widehat{f}\left(t\left(p\right)\right)} \mathrm{SE}\left[\widehat{S}\left(\widehat{t}\left(p\right)\right)\right].$$

The standard error of $\widehat{S}\left(\widehat{t}\,(p)\right)$ can be obtained using Greenwood's formula, given in equation (2.1). An estimate of the probability density function at the pth percentile $\widehat{t}\,(p)$ is used by many software packages

$$\widehat{f}\left[\widehat{t}\,(p)\right] = \frac{\widehat{S}\left[\widehat{u}\,(p)\right] - \widehat{S}\left[\widehat{l}\,(p)\right]}{\widehat{l}\,(p) - \widehat{u}\,(p)},$$

where

$$\widehat{u}\,(p) = \max\left\{t_{(j)}|\widehat{S}\left(t_{(j)}\right) \geq 1 - \frac{p}{100} + \varepsilon\right\},$$
$$\widehat{l}\,(p) = \min\left\{t_{(j)}|\widehat{S}\left(t_{(j)}\right) \leq 1 - \frac{p}{100} - \varepsilon\right\},$$

$t_{(j)}$ is jth ordered death time, $j = 1, 2, ...r$. $\varepsilon = 0.05$ is typically used by a number of statistical packages. Therefore, for median survival time, $\widehat{u}\,(50)$ is the largest observed survival time from the K-M curve for which $\widehat{S}\,(t) \geq 0.55$, and $\widehat{l}\,(50)$ is the smallest observed survival time from the K-M curve for which $\widehat{S}\,(t) \leq 0.45$.

The 95% confidence interval for the pth percentile $\widehat{t}\,(p)$ has limits of

$$\widehat{t}\,(p) \pm 1.96\text{SE}\{\widehat{t}\,(p)\}.$$

2.2 Nonparametric comparison of survival distributions

The K-M survival curves can give us an insight about the difference of survival functions in two or more groups, but whether this observed difference is statistically significant requires a formal statistical test. There are a number of methods that can be used to test equality of the survival functions in different groups. One commonly used non-parametric tests for comparison of two or more survival distributions is the log-rank test [40].

Let's take two groups as an example. Let $t_{(1)} < t_{(2)} < ... < t_{(k)}$ be the ordered death times across two groups. Suppose that d_j failures occur at $t_{(j)}$ and that r_j subjects are at risk just prior to $t_{(j)}$ $(j = 1, 2, ..., k)$. Let d_{ij} and r_{ij} be the corresponding numbers in group i $(i = 1, 2)$.

The log-rank test compares the observed number of deaths with the expected number of deaths for group i. Consider the null hypothesis: $S_1(t) = S_2(t)$, i.e. there is no difference between survival curves in two groups.

Given r_j and d_j, the random variable d_{1j} has the hypergeometric distribution

$$\frac{\binom{d_j}{d_{1j}}\binom{r_j - d_j}{r_{1j} - d_{1j}}}{\binom{r_j}{r_{1j}}}.$$

Under the null hypothesis, the probability of death at $t_{(j)}$ does not depend on the group, i.e., the probability of death at $t_{(j)}$ is $\frac{d_j}{r_j}$. So that the expected number of deaths in group one is

$$E(d_{1j}) = e_{1j} = r_{1j} d_j r_j^{-1}.$$

The test statistic is given by the difference between the total observed and expected number of deaths in group one

$$U_L = \sum_{j=1}^{r} (d_{1j} - e_{1j}). \tag{2.2}$$

Since d_{1j} has the hypergeometric distribution, the variance of d_{1j} is given by

$$v_{1j} = V(d_{1j}) = \frac{r_{1j} r_{2j} d_j \left(r_j - d_j\right)}{r_j^2 \left(r_j - 1\right)}. \tag{2.3}$$

So that the variance of U_L is

$$V(U_L) = \sum_{j=1}^{r} v_{1j} = V_L.$$

Under the null hypothesis, statistic (2.2) has an approximate normal distribution with zero mean and variance V_L. This then follows

$$\frac{U_L^2}{V_L} \sim \varkappa_1^2.$$

There are several alternatives to the log-rank test to test the equality of survival curves, for example, the Wilcoxon test [20]. These tests may be defined in general as follows:

$$\frac{\sum_{j=1}^{r} w_j (d_{1j} - e_{1j})}{\sum_{j=1}^{r} w_j^2 v_{1j}},$$

where w_j are weights whose values depend on the specific test.

The Wilcoxon test uses weights equal to risk size at $t_{(j)}$, $w_j = r_j$. This gives less weight to longest survival times. Early failures receive more weight than later failures. The Wilcoxon test places more emphasis on the information at the beginning of the survival curve where the number at risk is large. This type of weighting may be used to assess whether the effect of treatment on survival is strongest in the earlier phases of administration and tends to be less effective over time. Whereas the log-rank test uses weights equal to one at $t_{(j)}$, $w_j = 1$. This gives the same weight to each survival time. Therefore, Wilcoxon statistic is less sensitive than the log-rank statistic to difference of d_{1j} from e_{1j} in the tail of the distribution of survival times.

The log-rank test is appropriate when hazard functions for two groups are proportional over time, i.e., $h_1(t) = \phi h_2(t)$. So it is the most likely to detect a difference between groups when the risk of a failure is consistently greater for one group than another.

Chapter 3

Cox regression model

3.1 Introduction

The non-parametric method does not control for covariates and it requires categorical predictors. When we have several prognostic variables, we must use multivariate approaches. But we cannot use multiple linear regression or logistic regression because they cannot deal with censored observations. We need another method to model survival data with the presence of censoring. One very popular model in survival data is the Cox proportional hazards model, which is proposed by Cox [12].

Definition 3.1.1 *The Cox Proportional Hazards model is given by*

$$h(t|\mathbf{x}) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p) = h_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{x}),$$

where $h_0(t)$ is called the baseline hazard function, which is the hazard function for an individual for whom all the variables included in the model are zero., $\mathbf{x} = (x_1, x_2, ..., x_p)'$ is the values of the vector of explanatory variables for a particular individual, and $\boldsymbol{\beta}' = (\beta_1, \beta_2, ..., \beta_p)$ is a vector of regression coefficients.

The corresponding survival functions are related as follows:

$$S(t|\mathbf{x}) = S_0\left(t\right)^{\exp(\sum_{i=1}^{p}\beta_i \mathbf{x}_i)}.$$

This model, also known as the Cox regression model, makes no assumptions about the form of $h_0(t)$ (non-parametric part of model) but assumes parametric form for the effect of the predictors on the hazard (parametric part of model). The model is therefore referred to as a semi-parametric model.

The beauty of the Cox approach is that this vagueness creates no problems for estimation. Even though the baseline hazard is not specified, we can still get a good estimate for regression coefficients $\boldsymbol{\beta}$, hazard ratio, and adjusted hazard curves.

The measure of effect is called hazard ratio. The hazard ratio of two individuals with different covariates $\mathbf{x}$ and $\mathbf{x}^*$ is

$$\widehat{HR} = \frac{h_0(t)\exp(\widehat{\boldsymbol{\beta}}'\mathbf{x})}{h_0(t)\exp(\widehat{\boldsymbol{\beta}}'\mathbf{x}^*)} = \exp\left(\sum \widehat{\boldsymbol{\beta}}'(\mathbf{x}-\mathbf{x}^*)\right).$$

This hazard ratio is time-independent, which is why this is called the proportional hazards model.

3.2 Partial likelihood estimate for Cox proportional hazards model

Fitting the Cox proportional hazards model, we wish to estimate $h_0(t)$ and $\boldsymbol{\beta}$. One approach is to attempt to maximize the likelihood function for the observed data simultaneously with respect to $h_0(t)$ and $\boldsymbol{\beta}$. A more popular approach is proposed by Cox [13] in which a partial likelihood function that does not depend on $h_0(t)$ is obtained for $\boldsymbol{\beta}$. Partial likelihood is a technique developed to make inference about the regression parameters in the presence of nuisance parameters ($h_0(t)$ in the Cox PH model). In this section, we will construct the partial likelihood function based on the proportional hazards model.

Let $t_1, t_2, ..., t_n$ be the observed survival time for n individuals. Let the ordered death time of r individuals be $t_{(1)} < t_{(2)} < ... < t_{(r)}$ and let $R(t_{(j)})$ be the risk set just before $t_{(j)}$ and r_j for its size. So that $R(t_{(j)})$ is the group of individuals who are alive and uncensored at a time just prior to $t_{(j)}$. The conditional probability that the ith individual dies at $t_{(j)}$

given that one individual from the risk set on $R(t_{(j)})$ dies at $t_{(j)}$ is

$$
\begin{aligned}
&P(\text{individual } i \text{ dies at } t_{(j)} | \text{ one death from the risk set } R(t_{(j)}) \text{ at } t_{(j)}) \\
&= \frac{P\left(\text{individual } i \text{ dies at } t_{(j)}\right)}{P\left(\text{one death at } t_{(j)}\right)} \\
&= \frac{P\left(\text{individual } i \text{ dies at } t_{(j)}\right)}{\sum\limits_{k \in R(t_{(j)})} P\left(\text{individual } k \text{ dies at } t_{(j)}\right)} \\
&\simeq \frac{P\{\text{individual } i \text{ dies at } (t_{(j)}, t_{(j)} + \Delta t)\}/\Delta t}{\sum\limits_{k \in R(t_{(j)})} P\{\text{individual } k \text{ dies at } (t_{(j)}, t_{(j)} + \Delta t)\}/\Delta t} \\
&= \frac{\lim_{\Delta t \downarrow 0} P\{\text{individual } i \text{ dies at } (t_{(j)}, t_{(j)} + \Delta t)\}/\Delta t}{\lim_{\Delta t \downarrow 0} \sum\limits_{k \in R(t_{(j)})} P\{\text{individual } k \text{ dies at } (t_{(j)}, t_{(j)} + \Delta t)\}/\Delta t} \\
&= \frac{h_i(t_{(j)})}{\sum\limits_{k \in R(t_{(j)})} h_k(t_{(j)})} \\
&= \frac{h_0(t_{(j)}) \exp(\boldsymbol{\beta}' \mathbf{x}_i(t_{(j)})}{\sum\limits_{k \in R(t_{(j)})} h_0(t_{(j)}) \exp(\boldsymbol{\beta}' \mathbf{x}_k(t_{(j)})} \\
&= \frac{\exp(\boldsymbol{\beta}' \mathbf{x}_i(t_{(j)})}{\sum\limits_{k \in R(t_{(j)})} \exp(\boldsymbol{\beta}' \mathbf{x}_k(t_{(j)})}.
\end{aligned}
$$

Then the partial likelihood function for the Cox PH model is given by

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}' \mathbf{x}_i(t_{(j)})}{\sum\limits_{k \in R(t_{(j)})} \exp(\boldsymbol{\beta}' \mathbf{x}_k(t_{(j)})}, \tag{3.1}$$

in which $\mathbf{x}_i(t_{(j)})$ is the vector of covariate values for individual i who dies at $t_{(j)}$. The general method of partial likelihood was discussed by Cox [13].

Note that this likelihood function is only for the uncensored individuals. Let $t_1, t_2, ..., t_n$ be the observed survival time for n individuals and δ_i be the event indicator, which is zero if the ith survival time is censored, and unity otherwise. The likelihood function in equation (3.1) can be expressed by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{\beta}' \mathbf{x}_i(t_i)}{\sum\limits_{k \in R(t_i)} \exp(\boldsymbol{\beta}' \mathbf{x}_k(t_i)} \right]^{\delta_i}, \tag{3.2}$$

where $R(t_i)$ is the risk set at time t_i.

The partial likelihood is valid when there are no ties in the dataset. That means there is no two subjects who have the same event time.

3.3 Proportional hazard assumption checking

The main assumption of the Cox proportional hazards model is proportional hazards. Proportional hazards means that the hazard function of one individual is proportional to the hazard function of the other individual, i.e., the hazard ratio is constant over time. There are several methods for verifying that a model satisfies the assumption of proportionality.

3.3.1 Graphical method

We can obtain Cox PH survival function by the relationship between hazard function and survival function

$$S(t, \mathbf{x}) = S_0(t)^{\exp(\sum_{i=1}^{p} \beta_i x_i)},$$

where $\mathbf{x} = (x_1, x_2, ..., x_p)'$ is the values of the vector of explanatory variables for a particular individual. When taking the logarithm twice, we can easily get

$$\ln[-\ln S(t, \mathbf{x})] = \sum_{i=1}^{p} \beta_i x_i + \ln[-\ln S_0(t)].$$

Then the difference in log-log curves corresponding to two different individuals with variables $\mathbf{x}_1 = (x_{11}, x_{12}, ..., x_{1p})$ and $\mathbf{x}_2 = (x_{21}, x_{22}, ..., x_{2p})$ is given by

$$\ln[-\ln S(t, \mathbf{x}_1)] - \ln[-\ln S(t, \mathbf{x}_2)] = \sum_{i=1}^{p} \beta_i (x_{1i} - x_{2i}),$$

which does not depend on t. This relationship is very helpful to help us identify situations where we may have proportional hazards. By plotting estimated $\log(-\log(\text{survival}))$ versus survival time for two groups we would see parallel curves if the hazards are proportional. This method does not work well for continuous predictors or categorical predictors that have many levels because the graph becomes "cluttered". Furthermore, the curves are sparse when there are few time points and it may be difficult to tell how close to parallel is close enough.

However, looking at the K-M curves and $\log(-\log(\text{survival}))$ is not enough to be certain of proportionality since they are univariate analysis and do not show whether hazards will still be proportional when a model includes many other predictors. But they support our argument for proportionality. We will show some other statistical methods for checking the proportionality.

3.3.2 Adding time-dependent covariates in the Cox model

We create time-dependent covariates by creating interactions of the predictors and a function of survival time and including them in the model. For example, if the predictor of interest is X_j, then we create a time-dependent covariate $X_j(t)$, $X_j(t) = X_j \times g\,(t)$, where $g\,(t)$ is a function of time, e.g., t, $\log t$ or Heaviside function of t. The model assessing PH assumption for X_j adjusted for other covariates is

$$h(t, \mathbf{x}\,(t)) = h_0\,(t) \exp[\beta_1 x_1 + \beta_2 x_2 + ... + \beta_j x_j + ... + \beta_p x_p + \delta x_j \times g\,(t)],$$

where $\mathbf{x}\,(t) = (x_1, x_2, ..., x_p, x_j\,(t))'$ is the values of the vector of explanatory variables for a particular individual. The null hypothesis to check proportionality is that $\delta = 0$. The test statistic can be carried out using either a Wald test or a likelihood ratio test. In the Wald test, the test statistic is $W = \left(\widehat{\delta}/se\left(\widehat{\delta}\right)\right)^2$. The likelihood ratio test calculates the likelihood under null hypothesis, L_0 and the likelihood under the alternative hypothesis, L_a. The LR statistic is then $LR = -2\ln\left(L_0/L_a\right) = -2\left(l_0 - l_a\right)$, where l_0, l_a are log likelihood under two hypothesis respectively. Both statistics have a chi-square distribution with one degree of freedom under the null hypothesis. If the time-dependent covariate is significant, i.e., the null hypothesis is rejected, then the predictor is not proportional. In the same way, we can also assess the PH assumption for several predictors simultaneously.

3.3.3 Tests based on the Schoenfeld residuals

The other statistical test of the proportional hazards assumption is based on the Schoenfeld residual [48]. The Schoenfeld residuals are defined for each subject who is observed to fail. We will talk about it in detail in Section 3.4.2. If the PH assumption holds for a particular covariate then the Schoenfeld residual for that covariate will not be related to survival time. So this test is accomplished by finding the correlation between the Schoenfeld residuals for a particular covariate and the ranking of individual survival times. The null hypothesis is that the correlation between the Schoenfeld residuals and the ranked survival time is zero. Rejection of null hypothesis concludes that PH assumption is violated.

3.4 Cox proportional hazards model diagnostics

After a model has been fitted, the adequacy of the fitted model needs to be assessed. The model checking procedures below are based on residuals. In linear regression methods, residuals are defined as the difference between the observed and predicted values of the dependent variable. However, when censored observations are present and partial likelihood function is used in the Cox PH model, the usual concept of residual is not applicable. A number of residuals have been proposed for use in connection with the Cox PH model. We will describe three major residuals in the Cox model: the Cox-Snell residual, the deviance residual, and the Schoenfeld residual. Then we will talk about influence assessment.

3.4.1 Cox-Snell residuals and deviance residuals

The Cox-Snell residual is given by Cox and Snell [15]. The Cox-Snell residual for the ith individual with observed survival time t_i is defined as

$$r_{c_i} = \exp\left(\widehat{\boldsymbol{\beta}}' \mathbf{x}_i\right) \widehat{H}_0\left(t_i\right) = \widehat{H}_i\left(t_i\right) = -\log \widehat{S}_i\left(t_i\right),$$

where $\widehat{H}_0\left(t_i\right)$ is an estimate of the baseline cumulative hazard function at time t_i, which was derived by Kalbfleisch and Prentice [31].

This residual is motivated by the following result: Let T have continuous survival distribution $S(t)$ with the cumulative hazard $H(t) = -\log(S(t))$. Thus, $S_T(t) = \exp(-H(t))$. Let $Y = H(T)$ be the transformation of T based on the cumulative hazard function. Then the survival function for Y is

$$\begin{aligned} S_Y(y) &= P(Y > y) = P(H(t) > y) \\ &= P(T > H_T^{-1}(y)) = S_T(H_T^{-1}(y)) \\ &= \exp(-H_T(H_T^{-1}(y))) = \exp(-y). \end{aligned}$$

Thus, regardless of the distribution of T, the new variable $Y = H(T)$ has an exponential distribution with unit mean. If the model was well fitted, the value $\widehat{S}_i\left(t_i\right)$ would have similar properties to those of $S_i\left(t_i\right)$. So $r_{c_i} = -\log \widehat{S}_i\left(t_i\right)$ will have a unit exponential distribution with $f_R\left(r\right) = \exp\left(-r\right)$. Let $S_R\left(r\right)$ denote the survival function of Cox-Snell residual r_{c_i}. Then

$$S_R\left(r\right) = \int_r^\infty f_R\left(x\right) dx = \int_r^\infty \exp\left(-x\right) dx = \exp\left(-r\right),$$

and

$$H_R(r) = -\log S_R(r) = -\log(\exp(-r)) = r.$$

Therefore, we use a plot of $H(r_{c_i})$ versus r_{c_i} to check the fit of the model. This gives a straight line with unit slope and zero intercept if the fitted model is correct. Note the Cox-Snell residuals will not be symmetrically distributed about zero and cannot be negative.

The deviance residual [53] is defined by

$$r_{D_i} = sign(r_{m_i})[-2\{r_{m_i} + \delta_i \log(\delta_i - r_{m_i}\}]^{1/2},$$

where the function $sign(.)$ is the sign function which takes the value 1 if r_{m_i} is positive and -1 if r_{m_i} is negative; $r_{m_i} = \delta_i - r_{c_i}$ is the martingale residuals [5] for the i_{th} individual; and $\delta_i = 1$ for uncensored observation, $\delta_i = 0$ for censored observation.

The martingale residuals take values between negative infinity and unity. They have a skewed distribution with mean zero [3]. The deviance residuals are a normalized transform of the martingale residuals [53]. They also have a mean of zero but are approximately symmetrically distributed about zero when the fitted model is appropriate. Deviance residual can also be used like residuals from linear regression. The plot of the deviance residuals against the covariates can be obtained. Any unusual patterns may suggest features of the data that have not been adequately fitted for the model. Very large or very small values suggest that the observation may be an outlier in need of special attention. In a fitted Cox PH model, the hazard of death for the ith individual at any time depends on the value of $\exp(\boldsymbol{\beta}'\mathbf{x}_i)$ which is called the risk score. A plot of the deviance residuals versus the risk score is a helpful diagnostic to assess a given individual on the model. Potential outliers will have deviance residuals whose absolute values are very large. This plot will give the information about the characteristic of observations that are not well fitted by the model.

3.4.2 Schoenfeld residuals

All the above three residuals are residuals for each individual. We will describe covariate-wise residuals: Schoenfeld residuals [48]. The Schoenfeld residuals were originally called partial residuals because the Schoenfeld residuals for ith individual on the jth explanatory variable X_j is an estimate of the ith component of the first derivative of the logarithm of the partial likelihood function with respect to β_j. From equation (3.2), this logarithm of

the partial likelihood function is given by

$$\frac{\partial \log L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{p} \delta_i \left\{ x_{ij} - a_{ij} \right\},$$

where x_{ij} is the value of the jth explanatory variable $j = 1, 2, ..., p$ for the ith individual and

$$a_{ji} = \frac{\sum_{l \in R(t_i)} x_{jl} \exp(\boldsymbol{\beta}' \mathbf{x}_l)}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}' \mathbf{x}_l)}.$$

The Schoenfeld residual for ith individual on X_j is given by $r_{p_{ji}} = \delta_i \left\{ x_{ji} - a_{ji} \right\}$. The Schoenfeld residuals sum to zero.

3.4.3 Diagnostics for influential observations

Observations that have an undue effect on model-based inference are said to be influential. In the assessment of model adequacy, it is important to determine whether there are any influential observations. The most direct measure of influence is $\widehat{\beta}_j - \widehat{\beta}_{j(i)}$, where $\widehat{\beta}_j$ is the jth parameter, $j = 1, 2, ..., p$ in a fitted Cox PH model and $\widehat{\beta}_{j(i)}$ is obtained by fitting the model after omitting observation i. In this way, we have to fit the $n + 1$ Cox models, one with the complete data and n with each observation eliminated. This procedure involves a significant amount of computation if the sample size is large. We would like to use an alternative approximate value that does not involve an iterative refitting of the model. To check the influence of observations on a parameter estimate, Cain and Lange [9] showed that an approximation to $\widehat{\beta}_j - \widehat{\beta}_{j(i)}$ is the jth component of the vector

$$r'_{S_i} V(\widehat{\boldsymbol{\beta}}),$$

where r_{S_i} is the $p \times 1$ vector of score residuals for the ith observation [10], which are modifications of Schoenfeld residuals and are defined for all the observations, and $V(\widehat{\boldsymbol{\beta}})$ is the variance-covariance matrix of the vector of parameter estimates in the fitted Cox PH model. The jth element of this vector is called delta-beta statistic for the jth explanatory variable, i.e., $\Delta_i \widehat{\beta}_j \approx \widehat{\beta}_j - \widehat{\beta}_{j(i)}$, which tells us how much each coefficient will change by removal of a single observation. Therefore, we can check whether there are influential observations for any particular explanatory variable.

3.5 Strategies for analysis of nonproportional data

Suppose that statistic tests or other diagnostic techniques give strong evidence of nonproportionality for one or more covariates. To deal with this we will describe two popular methods: stratified Cox model and Cox regression model with time-dependent variables which are particularly simple and can be done using available software. Another way to consider is to use a different model. A parametric model such as an AFT model, which we will describe in Chapter 4, might be more appropriate for the data.

3.5.1 Stratified Cox model

One method that we can use is the stratified Cox model, which stratifies on the predictors not satisfying the PH assumption. The data are stratified into subgroups and the model is applied for each stratum. The model is given by

$$h_{ig}(t) = h_{og}(t) \exp\left(\boldsymbol{\beta}' \mathbf{x}_{ig}\right),$$

where g represents the stratum.

Note that the hazards are non-proportional because the baseline hazards may be different between strata. The coefficients $\boldsymbol{\beta}$ are assumed to be the same for each stratum g. The partial likelihood function is simply the product of the partial likelihoods in each stratum. A drawback of this approach is that we cannot identify the effect of this stratified predictor. This technique is most useful when the covariate with non-proportionality is categorical and not of direct interest.

3.5.2 Cox regression model with time-dependent variables

Until now we have assumed that the values of all covariates did not change over the period of observation. However, the values of covariates may change over time t. Such a covariate is called a time-dependent covariate. The second method to consider is to model nonproportionality by time-dependent covariates. The violation of PH assumptions are equivalent to interactions between covariates and time. That is, the PH model assumes that the effect of each covariate is the same at all points in time. If the effect of a variable varies with time, the PH assumption is violated for that variable. To model a time-dependent effect, one can create a time-dependent covariate $X(t)$, then $\beta X(t) = \beta X \times g(t)$. $g(t)$ is

a function of t such as t, $\log t$ or Heaviside functions, etc. The choice of time-dependent covariates may be based on theoretical considerations and strong clinical evidence.

The Cox regression with both time independent predictors X_i and time-dependent covariates $X_j(t)$ can be written

$$h(t|\mathbf{x}(t)) = h_0(t) \exp\left[\sum_{i=1}^{p_1} \beta_i x_i + \sum_{j=1}^{p_2} \alpha_j x_j(t)\right].$$

The hazard ratio at time t for the two individuals with different covariates $\mathbf{x}$ and $\mathbf{x}^*$ is given by

$$\widehat{HR}\,(t) = \exp\left[\sum_{i=1}^{p_1} \widehat{\beta}_i(x_i^* - x_i) + \sum_{j=1}^{p_2} \widehat{\alpha}_j \left(x_j^*(t) - x_j(t)\right)\right].$$

Note that, in this hazard ratio formula, the coefficient $\widehat{\alpha}_j$ is not time-dependent. $\widehat{\alpha}_j$ represents overall effect of $X_j(t)$ considering all times at which this variable has been measured in this study. But the hazard ratio depends on time t. This means that the hazards of event at time t is no longer proportional, and the model is no longer a PH model.

In addition to considering time-dependent variable for analyzing a time-independent variable not satisfying the PH assumption, there are variables that are inherently defined as time-dependent variables. One of the earliest applications of the use of time-dependent covariates is in the report by Crowley and Hu [16] on the Stanford Heart Transplant study. Time-dependent variables are usually classified to be internal or external. An internal time-dependent variable is one that the change of covariate over time is related to the characteristics or behavior of the individual. For example, blood pressure, disease complications, etc. The external time-dependent variable is one whose value at a particular time does not require subjects to be under direct observations, i.e., values changes because of external characteristics to the individuals. For example, level of air pollution.

Chapter 4

Parametric model

The Cox PH model described in Chapter 3 is the most common way of analyzing prognostic factors in clinical data. This is probably due to the fact that this model allows us to estimate and make inference about the parameters without assuming any distribution for the survival time. However, when the proportional hazards assumption is not tenable, these models will not be suitable. In this section, we will introduce parametric model, in which specific probability distribution is assumed for the survival times. In Section 4.1, we will introduce the parametric proportional hazards (PH) model. In Section 4.2, we will present the accelerated failure time (AFT) model and more detailed discussions of exponential, Weibull, log-logistic, log-Normal and gamma AFT models.

4.1 Parametric proportional hazards model

The parametric proportional hazards model is the parametric versions of the Cox proportional hazards model. It is given with the similar form to the Cox PH models. The hazard function at time t for the particular patient with a set of p covariates $(x_1, x_2...x_p)$ is given as follows:

$$h(t|\mathbf{x}) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p) = h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}).$$

The key difference between the two kinds of models is that the baseline hazard function is assumed to follow a specific distribution when a fully parametric PH model is fitted to the data, whereas the Cox model has no such constraint. The coefficients are estimated by partial likelihood in Cox model but maximum likelihood in parametric PH model. Other than this, the two types of models are equivalent. Hazard ratios have the same interpretation and proportionality of hazards is still assumed. A number of different parametric PH models may be derived by choosing different hazard functions. The commonly applied models are exponential, Weibull, or Gompertz models.

4.1.1 Weibull PH model

Suppose that survival times are assumed to have a Weibull distribution with scale parameter λ and shape parameter γ, so the survival and hazard function of a $W(\lambda, \gamma)$ distribution are given by

$$S(t) = \exp(-\lambda t^{\gamma}),\ h(t) = \lambda\gamma(t)^{\gamma-1},$$

with λ, $\gamma > 0$. The hazard rate increases when $\gamma > 1$ and decreases when $\gamma < 1$ as time goes on. When $\gamma = 1$, the hazard rate remains constant, which is the special exponential case.

Under the Weibull PH model, the hazard function of a particular patient with covariates $(x_1, x_2, ..., x_p)$ is given by

$$h(t|\mathbf{x}) = \lambda\gamma(t)^{\gamma-1}\exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p) = \lambda\gamma(t)^{\gamma-1}\exp(\boldsymbol{\beta}'\mathbf{x}).$$

We can see that the survival time of this patient has the Weibull distribution with scale parameter $\lambda\exp(\boldsymbol{\beta}'\mathbf{x})$ and shape parameter γ. Therefore the Weibull family with fixed γ possesses PH property. This shows that the effects of the explanatory variables in the model alter the scale parameter of the distribution, while the shape parameter remains constant.

From equation (1.4), the corresponding survival function is given by

$$S\,(t|\mathbf{x}) = \exp\{-\exp(\boldsymbol{\beta}'\mathbf{x})\lambda t^{\gamma}\}. \tag{4.1}$$

After a transformation of the survival function for a Weibull distribution, we can obtain

$$\log\{-\log S(t)\} = \log\lambda + \gamma\log t.$$

The $\log\{-\log S(t)\}$ versus $\log(t)$ should give approximately a straight line if the Weibull distribution assumption is reasonable. The intercept and slope of the line will be rough estimate of $\log\lambda$ and γ respectively. If the two lines for two groups in this plot are essentially parallel, this means that the proportional hazards model is valid. Furthermore, if the straight line has a slope nearly one, the simpler exponential distribution is reasonable. In the other way, for a exponential distribution, there is $\log S(t) = -\lambda t$. Thus we can consider the graph of $\log S(t)$ versus t. This should be a line that goes through the origin if exponential distribution is appropriate.

Another approach to assess the suitability of a parametric model is to estimate the hazard function using the non-parametric method. If the hazard function were reasonably constant over time, this would indicate that the exponential distribution might be appropriate. If the hazard function increased or decreased monotonically with increasing survival time, a Weibull distribution or Gompertz distribution might be considered.

4.1.2 Exponential PH model

The exponential PH model is a special case of the Weibull model when $\gamma = 1$. The hazard function under this model is to assume that it is constant over time. The survival and hazard function are written as

$$S(t) = \exp(-\lambda t),\ h(t) = \lambda.$$

Under the exponential PH model, the hazard function of a particular patient is given by

$$h(t|\mathbf{x}) = \lambda \exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p) = \lambda \exp(\boldsymbol{\beta}^{'}\mathbf{x}).$$

The piecewise exponential model [7] is an extension of the exponential PH model. For the piecewise exponential model, the period of follow-up is divided into k intervals $(t_j, t_{j+1}]$, $j = 1, 2...k$, $t_1 = 0$. Assume that the baseline hazard is constant within each interval but can vary across intervals, so that $h_0\ (t) = \exp(\alpha_j) = \lambda_j$ for $t_j < t \leq t_{j+1}$, i.e., the baseline hazard function is approximated by a step function.

The piecewise exponential model is given by

$$\lambda_{ij} = \lambda_j \exp(\boldsymbol{\beta}^{'}\mathbf{x}_i),$$

where λ_{ij} is the hazard corresponding to individual i in interval j and $\exp(\boldsymbol{\beta}^{'}\mathbf{x}_i)$ is the relative risk for an individual with covariate value $\mathbf{x}_i$ compared to the baseline at any given time.

In the piecewise exponential approach, a log-linear model is used to model both the effects of the covariates and the underlying hazard function. Estimates of the underlying hazard function and the regression parameters can be obtained using maximum likelihood. The maximum likelihood estimate of the baseline hazard function in interval i for given regression coefficients $\boldsymbol{\beta}$ is given by

$$\widehat{\lambda}_j = \frac{d_j}{\sum\limits_{i \in R_j} \exp(\boldsymbol{\beta}^{'}\mathbf{x}_i) t_{ij}},$$

where d_j is the number of events in interval j, R_j is the risk set entering interval j, and t_{ij} is the observed survival time for individual i in interval j. This approach was first studied by Holford [24], and is also the subject of work by Holford [25] and Laird and Olivier [36].

One of the greatest challenge related to the use of the piecewise exponential model is to find an adequate grid of time-points needed in its construction. One of the advantage of this method is the ability to incorporate time-dependent covariates. If there were any time-dependent covariates, their values at the beginning of each interval could be assigned to the records for that time interval.

4.1.3 Gompertz PH model

The survival and hazard function of the Gompertz distribution are given by

$$S(t) = \exp\left(\frac{\lambda}{\theta}(1 - e^{\theta t}\right), \; h(t) = \lambda \exp(\theta t),$$

for $0 \leq t < \infty$ and $\lambda > 0$. The parameter θ determines the shape of the hazard function. When $\theta = 0$, the survival time then have an exponential distribution, i.e., the exponential distribution is also a special case of the Gompertz distribution. Like the Weibull hazard function, the Gompertz hazard increases or decreases monotonically. For the Gompertz distribution, $\log(h(t))$ is linear with t.

Under the Gompertz PH model, the hazard function of a particular patient is given by

$$h(t|\mathbf{x}) = \lambda \exp(\theta t) \exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p) = \lambda \exp(\boldsymbol{\beta}' \mathbf{x}) \exp(\theta t).$$

It is straightforward to see that the Gompertz distribution has the PH property. But the Gompertz PH model is rarely used in practice.

Most computer software for fitting the exponential and Weibull models uses a different form of the model, AFT model, which we will describe it in the next section.

4.2 Accelerated failure time model

4.2.1 Introduction

Although parametric PH models are very applicable to analyze survival data, there are relatively few probability distribution for the survival time that can be used with these models. In these situations, the accelerated failure time model (AFT) is an alternative

to the PH model for the analysis of survival time data. Under AFT models we measure the direct effect of the explanatory variables on the survival time instead of hazard, as we do in the PH model. This characteristic allows for an easier interpretation of the results because the parameters measure the effect of the correspondent covariate on the mean survival time. Currently, the AFT model is not commonly used for the analysis of clinical trial data, although it is fairly common in the field of manufacturing. Similar to the PH model, the AFT model describes the relationship between survival probabilities and a set of covariates.

Definition 4.2.1 *For a group of patients with covariate* $(X_1, X_2, ..., X_p)$ *, the model is written mathematically as* $S(t|\mathbf{x}) = S_0(t/\eta(\mathbf{x}))$*, where* $S_0(t)$ *is the baseline survival function and* η *is an 'acceleration factor' that is a ratio of survival times corresponding to any fixed value of* $S(t)$*. The acceleration factor is given according to the formula* $\eta(\mathbf{x}) = \exp(\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + ... + \alpha_p\mathbf{x}_p)$.

Under an accelerated failure time model, the covariate effects are assumed to be constant and multiplicative on the time scale, that is, the covariate impacts on survival by a constant factor (acceleration factor).

According to the relationship of survival function and hazard function, the hazard function for an individual with covariate $X_1, X_2, ..., X_p$ is given by

$$h(t|\mathbf{x}) = [1/\eta(\mathbf{x})]h_0[t/\eta(\mathbf{x})]. \tag{4.2}$$

The corresponding log-linear form of the AFT model with respect to time is given by

$$\log T_i = \mu + \alpha_1 X_{1i} + \alpha_2 X_{2i} + ... + \alpha_p X_{pi} + \sigma\varepsilon_i,$$

where μ is intercept, σ is scale parameter and ε_i is a random variable, assumed to have a particular distribution. This form of the model is adopted by most software package for the AFT model.

For each distribution of ε_i, there is a corresponding distribution for T. The members of the AFT model class include the exponential AFT model, Weibull AFT model, log-logistic AFT model, log-normal AFT model, and gamma AFT model. The AFT models are discussed in details in textbooks [10], [14], [37]. The AFT models are named for the distribution of T rather than the distribution of ε_i or $\log T$.

Distribution of ε	Distribution of T
Extreme value(1 parameters)	Exponential
Extreme value(2 parameters)	Weibull
Logistic	Log-logistic
Normal	Log-normal
Log-Gamma	Gamma

Table 4.1: Summary of parametric AFT models

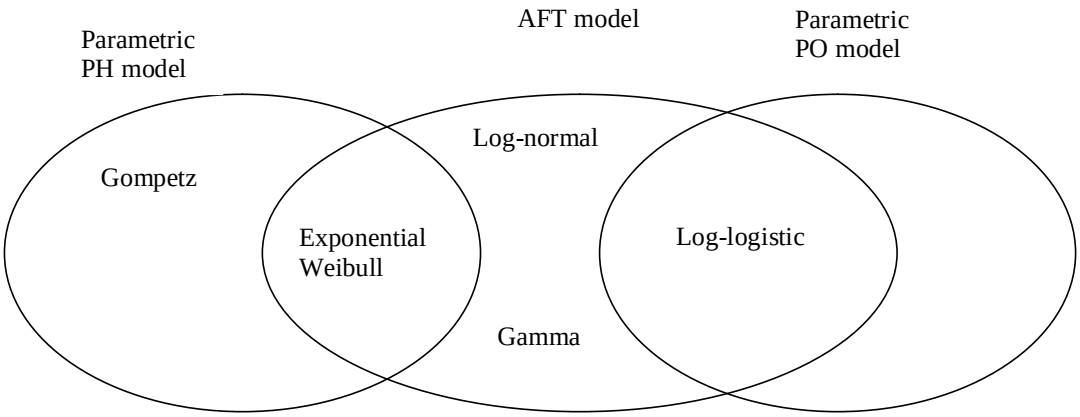

Figure 4.1: Summary of parametric models

The survival function of T_i can be expressed by the survival function of ε_i :

$$\begin{aligned} S_i(t) &= P\left(T_i \geq t\right) \\ &= P\left(\log T_i \geq \log t\right) \\ &= P\left(\mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + ... + \alpha_p x_{pi} + \sigma\varepsilon_i \geq \log t\right) \\ &= P\left(\varepsilon_i \geq \frac{\log t - \mu - \alpha\mathbf{x}}{\sigma}\right) \\ &= S_{\varepsilon_i}\left(\frac{\log t - \mu - \alpha\mathbf{x}}{\sigma}\right). \end{aligned} \tag{4.3}$$

The distributions of ε_i and the corresponding distributions of T_i are summarized in Table (4.1). And the summary of the commonly used parametric models are described in Figure (4.1).

The effect size for the AFT model is the time ratio. The time ratio comparing two levels of covariate x_i ($x_i = 1$ vs. $x_i = 0$), after controlling all the other covariates is $\exp(\alpha_i)$, which is interpreted as the estimated ratio of the expected survival times for two groups. A time ratio above 1 for the covariate implies that this covariate prolongs the time to event, while a time ratio below 1 indicates that an earlier event is more likely. Therefore, the AFT models can be interpreted in terms of the speed of progression of a disease. The effect of the covariates in an accelerated failure time model is to change the scale, and not the location of a baseline distribution of survival times.

4.2.2 Estimation of AFT model

AFT models are fitted using the maximum likelihood method. The likelihood of the n observed survival times, $t_1, t_{2,...} t_n$ is given by

$$L(\alpha, \mu, \sigma) = \prod_{i=1}^{n} \{f_i(t_i)\}^{\delta_i} \{S_i(t_i)\}^{1-\delta_i},$$

where $f_i(t_i)$ and $S_i(t_i)$ are the density and survival functions for the ith individual at t_i and δ_i is the event indicator for the ith observation. Using equation (4.3), the log-likelihood function is then given by

$$\log L(\alpha, \mu, \sigma) = \sum_{i=1}^{n} \{-\delta_i \log(\sigma t_i + \delta_i \log f_{\varepsilon_i}(z_i) + (1 - \delta_i) \log S_{\varepsilon_i}(z_i))\},$$

where $z_i = (\log t_i - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - ... - \alpha_p x_{pi})/\sigma$. The maximum likelihood estimates of the $p+2$ unknown parameters, $\mu, \sigma, \alpha_1, \alpha_2, ..., \alpha_p$, are found by maximizing this function using the Newton-Raphson procedure in SAS, which is the same method used to maximize the partial likelihood in the Cox regression model.

Several other approaches have been proposed for the estimation and inference on the AFT model in the literature. Classical semi-parametric approaches to the AFT model that emphasize estimation of the regression parameters include the method of Buckley and James [8] and linear-rank-test-based estimators [32]. Despite theoretical advances, all these approaches are numerically complicated and difficult to implement, especially when the number of covariates is large.

4.2.3 Weibull AFT model

Suppose the survival time T has $W(\lambda, \gamma)$ distribution with scale parameter λ and shape parameter γ. From equation (4.2), under AFT model, the hazard function for the ith individual is

$$\begin{aligned} h_i(t) &= [1/\eta_i(\mathbf{x})]h_0[t/\eta_i(\mathbf{x})] \\ &= [1/\eta_i(\mathbf{x})]\lambda\gamma(t/\eta_i(\mathbf{x}))^{\gamma-1} \\ &= 1/[\eta_i(\mathbf{x})]^{\gamma}\lambda\gamma(t)^{\gamma-1}, \end{aligned}$$

where $\eta_i = \exp(\alpha_1 x_{1i} + \alpha_2 x_{2i} + ... + \alpha_p x_{pi})$ for individual i with p explanatory variables. So the survival time for the ith patient is $W(1/[\eta_i(\mathbf{x})]^{\gamma}\lambda, \gamma)$. The Weibull distribution has the AFT property.

If T_i has a Weibull distribution, then ε_i has an extreme value distribution (Gumbel distribution). The survival function of Gumbel distribution is given by

$$S_{\varepsilon_i}(\varepsilon) = \exp(-\exp(\varepsilon)).$$

From equation (4.3), the AFT representation of the survival function of the Weibull model is given by

$$\begin{aligned} S_i(t) &= \exp\left[-\exp(\frac{\log t - \mu - \alpha_1 x_{1i} - ... - \alpha_p x_{pi})}{\sigma})\right] \\ &= \exp\left[-\exp\left(\frac{-\mu - \alpha_1 x_{1i} - ... - \alpha_p x_{pi})}{\sigma}\right) t^{1/\sigma}\right]. \end{aligned} \tag{4.4}$$

From equation (4.1), the PH representation of the survival function of the Weibull model is given by

$$S_i(t) = \exp\{-\exp(\beta_1 x_{1i} + ... + \beta_p x_{pi})\lambda t^{\gamma}\}. \tag{4.5}$$

Comparing the above two formulas (4.4) and (4.5), we can easily see that the parameter λ, γ, β_j in the PH model can be expressed by the parameters μ, σ, α_j in the AFT model:

$$\lambda = \exp(-\mu/\sigma),\ \gamma = 1/\sigma,\ \beta_j = -\alpha_j/\sigma. \tag{4.6}$$

Using equation (1.3), the AFT representation of hazard function of the Weibull model is given by

$$h_i(t) = \frac{1}{\sigma} t^{\frac{1}{\sigma}-1} \exp\left(\frac{-\mu - \alpha_1 x_{1i} - ... - \alpha_p x_{pi})}{\sigma}\right). \tag{4.7}$$

Suppose the pth percentile of the survival distribution for the ith individual is $t_i(p)$, which is the value such that $S_i(t_i(p)) = \frac{100-p}{100}$. From equation (4.4), we can easily get

$$t_i(p) = \exp\left[\sigma \log\left\{-\log\left(\frac{100-p}{100}\right)\right\} + \mu + \alpha' \mathbf{x}_i\right].$$

The median survival time is

$$t_i(50) = \exp\left[\sigma \log(\log 2) + \mu + \alpha' \mathbf{x}_i\right]. \tag{4.8}$$

To calculate the standard error of $\widehat{\beta}_j$, we can use the approximate variance of a function of two parameter estimate θ_1, θ_2, which is given by

$$\left(\frac{\partial g}{\partial \widehat{\theta}_1}\right)^2 V\left(\widehat{\theta}_1\right) + \left(\frac{\partial g}{\partial \widehat{\theta}_2}\right)^2 V\left(\widehat{\theta}_2\right) + 2\left(\frac{\partial g}{\partial \widehat{\theta}_1}\frac{\partial g}{\partial \widehat{\theta}_2}\right)\mathrm{Cov}(\theta_1, \theta_2).$$

The approximate variance of $\widehat{\beta}_j$ is expressed as

$$V(\beta_j) = \left(\frac{-1}{\widehat{\sigma}}\right)^2 V(\widehat{\alpha}_j) + \left(\frac{\widehat{\alpha}_j}{\widehat{\sigma}^2}\right)^2 V(\widehat{\sigma}) + 2\left(\frac{-1}{\widehat{\sigma}}\right)\left(\frac{\widehat{\alpha}_j}{\widehat{\sigma}^2}\right)\mathrm{Cov}(\widehat{\alpha}_j, \widehat{\sigma}).$$

The square root of this is the standard error of $\widehat{\beta}_j$. Then the 95% confidence interval can be calculated.

4.2.4 Log-logistic AFT model

One limitation of the Weibull hazard function is that it is a monotonic function of time. However, the hazard function can change direction in some situations. We will describe the log-logistic model in this section. The log-logistic survival and hazard function are given by

$$S(t) = \frac{1}{1 + e^{\theta} t^k}, h(t) = \frac{e^{\theta} \kappa t^{k-1}}{1 + e^{\theta} t^k},$$

where θ and k are unknown parameters and $k > 0$. When $k \leq 1$, the hazard rate decreases monotonically and when $k > 1$, it increases from zero to a maximum and then decreases to zero.

Suppose that the survival times have a log-logistic distribution with parameter θ and k, then from equation (4.2), under the AFT model, the hazard function for the ith individual is

$$\begin{aligned} h_i(t) &= (1/\eta_i)\, h_0(t/\eta_i) \\ &= \frac{e^{\theta} \kappa (t/\eta_i)^{k-1}}{\eta_i (1 + e^{\theta} (t/\eta_i)^{\kappa})} \\ &= \frac{e^{\theta - \kappa \log \eta_i} \kappa t^{k-1}}{1 + e^{\theta - \kappa \log \eta_i} t^k}, \end{aligned}$$

where $\eta_i = \exp(\alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_p x_p)$ for individual i with p explanatory variables. Therefore, the survival time for the ith individual has a log-logistic distribution with parameter $\theta - k \log \eta_i$ and k, log-logistic distribution has AFT property.

If the baseline survival function is $S_0(t) = \{1 + e^{\theta} t^k\}^{-1}$, where θ and k are unknown parameters, then the baseline odds of surviving beyond time t are given by

$$\frac{S_0(t)}{1 - S_0(t)} = e^{-\theta} t^{-k}.$$

The survival time for the ith individual also has a log-logistic distribution, which is

$$S_i(t) = \frac{1}{1 + e^{\theta - k \log \eta_i} t^k}. \tag{4.9}$$

Therefore, the odds of the ith individual surviving beyond time t is given by

$$\frac{S_i(t)}{1 - S_i(t)} = e^{\log \eta_i - \theta} t^{-k}. \tag{4.10}$$

We can see that the log-logistic distribution has the proportional odds (PO) property. So this model is also a proportional odds model, in which the odds of an individual surviving beyond time t are expressed as

$$\frac{S_i(t)}{1 - S_i(t)} = \exp(\beta_1 x_{1i} + ... + \beta_p x_{pi}) \frac{S_0(t)}{1 - S_0(t)}.$$

In a two group study, using (4.10), the log (odds) of the ith individual surviving beyond time t are

$$\log \left[\frac{S_i(t)}{1 - S_i(t)} \right] = \beta x_i - \theta - k \log t,$$

where x_i is the value of a categorical variable which takes the value one in one group and zero in the other group. A plot of $\log[(1 - S(t))/S(t)]$ versus $\log t$ should be linear if log-logistic distribution is appropriate. Therefore we can check the suitability of log-logistic distribution using the PO property.

If T_i has a log-logistic distribution, then ε_i has a logistic distribution. The survival function of logistic distribution is given by

$$S_{\varepsilon_i}(\varepsilon) = \frac{1}{1 + \exp(\varepsilon)}.$$

Using equation (4.3), the AFT representation of survival function of the log-logistic model is given by

$$S_i(t) = \left[1 + t^{1/\sigma} \exp\left(\frac{-\mu - \alpha_1 x_{1i} - ... - \alpha_p x_{pi}}{\sigma} \right) \right]^{-1}. \tag{4.11}$$

Comparing the formula (4.9) and (4.11), we can easily find a $\theta = -\mu/\sigma, k = \sigma^{-1}$.

According to the relationship of survival and hazard function, the hazard function for the ith individual is given by

$$h_i(t) = \frac{1}{\sigma t}\left\{1 + t^{-1/\sigma}\exp\left(\frac{\mu + \alpha_1 x_{1i} + ... + \alpha_p x_{pi}}{\sigma}\right)\right\}^{-1}. \tag{4.12}$$

The pth percentile of the survival distribution for the ith individual is $t_i(p)$, from equation (4.11), is

$$t_i(p) = \exp\left[\sigma \log\left(\frac{100-p}{100}\right) + \mu + \boldsymbol{\alpha}^{'}\mathbf{x}_i\right].$$

The median survival time is

$$t_i(50) = \exp(\mu + \boldsymbol{\alpha}^{'}\mathbf{x}_i). \tag{4.13}$$

4.2.5 Log-normal AFT model

If the survival times are assumed to have a log-normal distribution, the baseline survival function and hazard function are given by

$$S_0(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right),\ h_0(t) = \frac{\phi\left(\frac{\log t}{\sigma}\right)}{\left[1 - \Phi\left(\frac{\log t}{\sigma}\right)\right]\sigma t},$$

where μ and σ are parameters, $\phi(x)$ is the probability density function and $\Phi(x)$ is the cumulative density function of the standard normal distribution. The survival function for the ith individual is

$$\begin{aligned} S_i(t) &= S_0(t/\eta_i) \\ &= 1 - \Phi\left(\frac{\log t - \boldsymbol{\alpha}^{'}\mathbf{x}_i - \mu}{\sigma}\right), \end{aligned}$$

where $\eta_i = \exp(\alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_p x_p)$. Therefor the log survival time for the ith individual has normal $(\mu + \boldsymbol{\alpha}^{'}\mathbf{x}_i, \sigma)$. The log-normal distribution has the AFT property.

In a two group study, we can easily get

$$\Phi^{-1}[1 - S(t)] = 1/\sigma\left(\log t - \boldsymbol{\alpha}^{'}\mathbf{x}_i - \mu\right),$$

where x_i is the value of a categorical variable which takes the value one in one group and zero in the other group. This implies that a plot of $\Phi^{-1}[1 - S(t)]$ versus $\log t$ will be linear if the log-normal distribution is appropriate.

4.2.6 Gamma AFT model

There are two different gamma models discussed in survival analysis literature. The standard (2-parameter) and the generalized (3-parameter) gamma model. The gamma model means the generalized gamma model in this thesis. The probability density function of the generalized gamma distribution with three parameters, λ, α and γ is defined by

$$f(t) = \frac{\alpha\lambda^{\alpha\gamma}}{\Gamma(\gamma)} t^{\alpha\gamma-1} \exp\left[-(\lambda t)^{a}\right] \qquad t > 0, \gamma > 0, \lambda > 0, \alpha > 0,$$

where γ is the shape parameter of the distribution. The survival function and the hazard function do not have a closed form for the generalized gamma distribution. The exponential, Weibull and log-normal models are all special cases of the generalized gamma model. It is easily to seen that this generalized gamma distribution becomes the exponential distribution if $\alpha = \gamma = 1$, the Weibull distribution if $\gamma = 1$, and the log-normal distribution if $\gamma \to \infty$. The generalized gamma model can take on a wide variety of shapes except for any of the special cases. For example, it can have a hazard function with U or bathtub shapes in which the hazard declines reaches a minimum and then increases.

4.2.7 Model checking

The graphical methods can be used to check if a parametric distribution fits the observed data. Specifically, if the survival time follows an exponential distribution, a plot of $\log[-\log S(t)]$ versus $\log t$ should yield a straight line with slope of 1. If the plots are parallel but not straight then PH assumption holds but not the Weibull. If the lines for two groups are straight but not parallel, the Weibull assumption is supported but the PH and AFT assumptions are violated. The log-logistic assumption can be graphically evaluated by plotting $\log[(1 - S(t))/S(t)]$ versus $\log t$. If the distribution of survival function is log-logistic, then the resulting plot should be a straight line. For the log-normal distribution, a plot of $\Phi^{-1}[1 - S(t)]$ versus $\log t$ should be linear. All these plots are based on the assumption that the sample is drawn from a homogeneous population, implying that no covariates are taken into account. So this graphical method is not very reliable in practice. There are other methods to check the fitness of the model.

Using quantile-quantile plot

An initial method for assessing the potential for an AFT model is to produce a quantile-quantile plot. For any value of p in the interval $(0, 100)$, the pth percentile is

$$t(p) = S^{-1}\left(\frac{100-p}{100}\right).$$

Let $t_0(p)$ and $t_1(p)$ be the pth percentiles estimated from the survival functions of the two groups of survival data. The percentiles for the two groups may be expressed as

$$t_0(p) = S_0^{-1}\left(\frac{100-p}{100}\right),\ t_1(p) = S_1^{-1}\left(\frac{100-p}{100}\right),$$

where $S_0(t)$ and $S_1(t)$ are the survival functions for the two groups. So we can get

$$S_1[t_1(p)] = S_0[t_0(p)].$$

Under the AFT model, $S_1(t) = S_0(t/\eta)$, and so

$$S_1[t_1(p)] = S_0[t_1(p)/\eta].$$

Therefore, we get

$$t_0(p) = \eta^{-1} t_1(p).$$

The percentiles of the survival distributions for two groups can be estimated by the K-M estimates of the respective survival functions. A plot of percentiles of the K-M estimated survival function from one group against another should give an approximate straight line through the origin if the accelerated failure time model is appropriate. The slope of this line will be an estimate of the acceleration factor η^{-1}.

Using statistical criteria

We can use statistical tests or statistical criteria to compare all these AFT models. Nested models can be compared using the likelihood ratio test. The exponential model, the Weibull model and log-normal model are nested within gamma model. For comparing models that are not nested, the Akaike information criterion (AIC) can be used instead, which is defined as

$$AIC = -2l + 2(k + c),$$

where l is the log-likelihood, k is the number of covariates in the model and c is the number of model-specific ancillary parameters. The addition of $2(k + c)$ can be thought of

as a penalty if nonpredictive parameters are added to the model. Lower values of the AIC suggest a better model. But there is a difficulty in using the AIC in that there are no formal statistical tests to compare different AIC values. When two models have very similar AIC values, the choice of model may be hard and external model checking or previous results may be required to judge the relative plausibility of the models rather than relying on AIC values alone.

Using residual plots

Residual plots can be used to check the goodness of fit of the model. Procedures based on residuals in the AFT model are particularly relevant with the Cox PH model. One of the most useful plots is based on comparing the distribution of the Cox-Snell residuals with the unit exponential distribution. The Cox-Snell residual for the ith individual with observed time t_i is defined as

$$r_{c_i} = \widehat{H}\left(t_i|\mathbf{x}_i\right) = -\log\left[\widehat{S}\left(t_i|\mathbf{x}_i\right)\right],$$

where t_i is the observed survival time for individual i, $\mathbf{x}_i$ is the vector of covariate values for individual i, and $\widehat{S}(t_i)$ is the estimated survival function on the fitted model. From equation (4.3), the estimated survival function for the ith individual is given by

$$\widehat{S}_i\left(t\right) = S_{\varepsilon_i}\left(\frac{\log t - \widehat{\mu} - \widehat{\alpha}\mathbf{x}_i}{\widehat{\sigma}}\right),$$

where $\widehat{\mu}$, $\widehat{\alpha}$ and $\widehat{\sigma}$ are the maximum likelihood estimator of μ, α and σ respectively, $S_{\varepsilon_i}\left(\varepsilon\right)$ is the survival function of ε_i in the AFT model, and $\frac{\log t - \widehat{\mu} - \widehat{\alpha}\mathbf{x}_i}{\widehat{\sigma}} = r_{s_i}$ is referred to as standardized residual.

The Cox-Snell residual can be applied to any parametric model. The corresponding form of residual based particular AFT model can be obtained. For example, under the Weibull AFT model, since $S_{\varepsilon_i}\left(\varepsilon\right) = \exp\left(-e^{\varepsilon}\right)$, the Cox-Snell residual is then

$$r_{c_i} = -\log\{\widehat{S}\left(t_i\right)\} = -\log S_{\varepsilon_i}\left(r_{s_i}\right) = \exp\left(r_{s_i}\right).$$

Under the log-logistic AFT model, since $S_{\varepsilon_i}\left(\varepsilon\right) = \left(1+e^{\varepsilon}\right)^{-1}$, the Cox-Snell residual is then

$$r_{c_i} = \log[1 + \exp\left(r_{s_i}\right)].$$

If the fitted model is appropriate, the plot of $\log(-\log S(r_{c_i}))$ versus $\log r_{c_i}$ is a straight line with unit slope through the origin.

These residuals lead to the deviance residuals for the particular AFT model. A plot of deviance residuals against the survival time or explanatory variables can be used to check whether there are particular times, or particular values of explanatory variables, for which the model is not a good fit.

Chapter 5

Application to TB/HIV data

5.1 Introduction

In this chapter, after providing the details of a randomized placebo-controlled trial to prevent TB in Ugandan adults infected with HIV, we will apply non-parametric methods, the Cox PH model, Cox model with time-dependent variables, piecewise exponential model and the AFT model to this dataset. We also give all the corresponding results and compare the main methods: Cox model and AFT model.

Tuberculosis (TB) and acquired immunodeficiency syndrome (AIDS) are two completely different diseases. TB is caused by the tubercle bacillus, *Mycobacterium tuberculosis* and spread through the air. TB attacks the lungs, but sometimes other parts of the body too. AIDS is caused by human immunodeficiency virus (HIV). HIV can be spread by unprotected sexual relations, contaminated needles, breast milk, and transmission from an infected mother to her baby at birth. HIV attacks the immune system. As the two leading causes of infectious disease-associated mortality worldwide, TB and HIV diseases have been bound together from the early years of the HIV/AIDS epidemic [19]. It has been known that there is an interaction of TB and HIV infection. HIV infection is the greatest known risk factor for progression of active TB disease [17] and reactivation of latent TB infection [49], [44]. People infected with TB have only a 10% chance of ever getting active TB in their lifetime. However, people infected with both HIV and TB have a 50% chance of getting active TB in their lifetime [51]. HIV infection also increases TB case fatality [1]. On the other hand, TB is a leading cause of morbidity and mortality in population with high HIV prevalence. It has been shown that active TB promotes HIV replication, increases virus load, and so accelerates HIV disease progression and mortality [21], [54]. For these reasons, the prevention and treatment of TB in HIV infected people is an important concern.

Both the TB and HIV/AIDS global pandemics are staggering, particularly at the points of HIV and TB coinfection [19]. The Joint United Nations Programme on HIV/AIDS (UNAIDS) estimated that there were 33.2 million people living with HIV infection at the end of 2007, increasing from 29 million in 2001, 2.5 million newly infected with HIV in 2007, and 2.1 million deaths caused by HIV [29]. Sub-Saharan Africa remains the most seriously affected region by the AIDS pandemic, more than two-thirds of HIV-infected adults live there and more than three fourth AIDS death in 2007 occurred there. With regard to TB, the World Health Organization (WHO) estimated that there were 9.2 million new cases, increasing from 9.1 million in 2005 and 1.7 millions deaths from TB worldwide in 2006, of which around 0.7 millions (7.7%) cases and 0.2 millions deaths were HIV infected people [61]. Twelve of 15 countries with the highest estimated TB incidence rates are in Africa, which are partly explained by the relatively high rates of HIV coinfection [61].

The treatment of both TB and HIV infection have been enormously successful, but there are still major diagnostic and therapeutic deficiencies. After the introduction of highly active antiretroviral therapy (HAART), the morbidity and mortality rates of HIV infection are decreasing enormously in Europe and the USA [46]. However, HIV morbidity and mortality rates are still increasing in some African and Asian countries, because of the ineffective implementation of prevention and intervention policies [52]. The access to antiretroviral drugs is limited in developing countries because of the high cost [30].

TB treatment comprises case treatment and preventive treatment. TB preventive therapies has been recommended as a means of preventing TB in HIV-infected subjects [60]. It aims to treat latent infection with *Mycobacterium tuberculosis* before active disease develops. Latent TB is diagnosed by a positive reaction to intradermal injection with purified protein derivative (PPD, tuberculosis skin test). Because of the association of HIV with TB, TB preventive therapy may be an important intervention to reduce the rising incidence of TB and HIV in developing countries. Some studies [47], [56], [22], [58] have shown that TB preventive therapy reduces the incidence of TB in HIV-infected adults. However, few studies [41], [39] have assessed the effect of TB preventive therapy on HIV progression and mortality. No studies have shown a significant effect of TB preventive therapy on mortality.

5.2 Description of the dataset

5.2.1 Study population and objective

In March 1993 the Uganda-Case Western Reserve Research Collaboration began a randomized, placebo-controlled clinical trial to assess the efficacy of three regimens for the preventions of TB in PPD-positive (at least 5 mm induration in PPD skin test) HIV-infected Uganda adults. The three regimens were isoniazid for 6 months, isoniazid plus rifampicin for 3 months, and isoniazid, rifampicin plus pyrazinamide for 3 months. In October 1993 the study was expanded to include persons anergic (0 induration in reaction to PPD and candida antigens) to both PPD and candida antigens based on studies suggesting that there was an increased risk of tuberculosis in anergic HIV-infected adults compared to PPD-positive groups [43], [50]. Between March 1993 and April 1995, 9095 subjects were screened and 2736 HIV-infected subjects were enrolled into the study. These eligible subjects included 18-50 year old HIV infected males and non-pregnant females with a positive PPD skin test or anergy, and a Karnofsky performance scale score greater than 50. Of the 6309 subjects screened but excluded for the study, 4306 subjects did not complete the baseline evaluation and 2053 subjects were ineligible from the study for various detailed in Figure 5.1. More detailed information regarding the study design, intervention, measurement and the study profiles has been published [58], [28].

PPD-positive subjects were randomly assigned to receive either placebo (250 mg of ascorbic acid) daily for 6 months; isoniazid 300 mg daily for 6 months (6H); isoniazid 300 mg and rifampicin 600 mg daily for 3 months (3HR); or isoniazid 300 mg, rifampicin 600 mg and pyrazinamide 200 mg daily for 3 months (3HRZ). Anergic subjects were randomly assigned to receive either placebo (250 mg of ascorbic acid), or isoniazid 300 mg daily for 6 months (6H) by a separate but identical process. The study medications were dispensed monthly. Subjects were followed monthly during study therapy and every 3 months thereafter. The last date of follow-up was August 8, 1998.

In original studies [58], [28], the study objectives were to determine the efficacy of the three regimens of TB preventive therapy for TB in HIV-infected adults. In another study [39], the effect of the TB preventive therapy on HIV disease progression to AIDS and survival was studied, and the significant prognostic factors for HIV disease progression to AIDS were identified. Only non-parametric method (K-M estimate and the log-rank test)

and the Cox PH model were applied in this study. The association between the number of initial HIV-related signs at baseline and the AIDS progression or mortality was shown to be significant [39]. Based on this result, we use a sub-dataset excluding subjects with any baseline signs related to HIV. Of 2736 original subjects, 2158 participants have no HIV-related signs and 491 participants have one and 81 participants have two HIV-related signs. We only include 2158 subjects without any baseline signs or symptoms in current analysis. Figure 5.1 gives the details of study profile in the current analysis. We apply non-parametric methods, the Cox PH model, Cox model with time-dependent variables, piecewise exponential model and the AFT model to this sub-dataset.

The objective of this study is to see the effect of the TB preventive therapy on HIV disease progression to AIDS and survival in HIV-infected adults without baseline HIV-related signs. Using this analysis, we will compare the Cox models and the AFT models.

5.2.2 Study outcomes

In this analysis, the primary outcome is (a) HIV disease progression to AIDS, (b) death, and (c) the combined event of AIDS progression and death. The outcome measures are stratified by anergy status (PPD positive and anergic).

In our study, the criteria for AIDS progression are based mainly on the 1990 WHO classification system for HIV/AIDS [59] and AIDS progression event is defined when a patient develops any one major sign or three or more minor signs during follow up. Major signs are defined as Karnofsky performance score less than or equal to 60, Kaposi's sarcoma, and esophageal candidiasis. The minor signs are defined as pruritic papules, oral candidiasis, *Varicella zoster* virus, genital *Herpes simplex*, oral *Herpes simplex*, and TB. This classification is obtained by a severity order of HIV-related signs based on CD4 lymphocyte counts. The classification is supported by the studies suggesting that HIV-infected persons show many clinical signs as the disease progresses and there is a strong association between HIV-related signs and CD4 lymphocyte counts [59], [42], [55].

The individual survival times for AIDS progression outcome are defined as the period between enrollment in the study and the incidence date of the AIDS progression the last clinic visit before the end of the study. The individual survival times for death outcome are calculated from the start of the study to the date of death or to the date of the last clinic visit before the end of study. The survival time for the combined event of AIDS

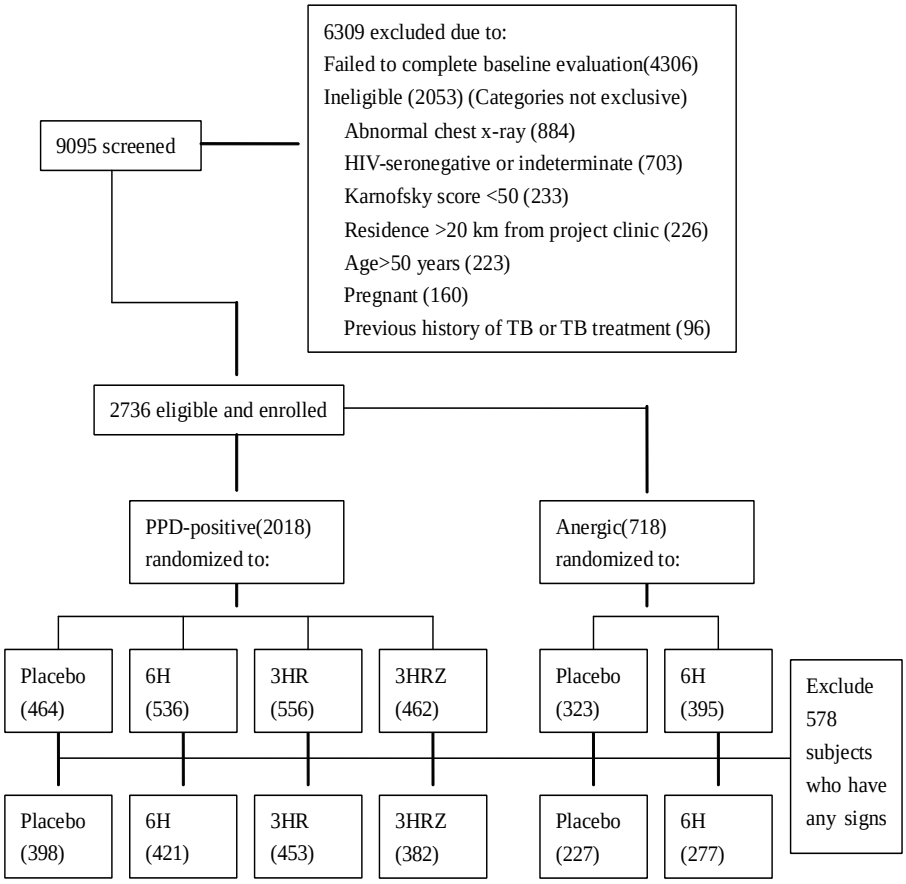

Figure 5.1: Subjects enrolled in the study

progression and death are obtained from the enrollment in the study up to the date of death or incidence date of the AIDS progression. Patients are censored at the last clinic visit.

5.2.3 Description of variables

The variables and codes for this data are provided in the following table:

Variables	Description	Codes/Values
AGE	Age	years
BCGscar	BCGscar indicator	0 = no BCGscar, 1 = BCGscar
BMI	Body mass index	kg/m^2
CREAT	creatinine level	mg/dl
EDUC	Education length	years
HCT	Hematocrit level	mg/dL
HGB	Hemoglobin	mg/dl
LYMPHABS	Absolute lymphocytes counts	cm^{-3}
MARITAL	Marital status	0 = never married, 1 = currently married, 2 = divorced/widowed.
PLT	Platelet counts	/L
Anergy	Indicator of TB skin test induration	0 = (induration $\geq 5mm$), 1 = (induration $< 5mm$).
SEX	Sex	0 = female, 1 = male.
STUDYARM	Six treatment arms	1 = Placebo, 2 = 6H, 3 = 3HR, 4 = 3HRZ, 5 = Anergic-Placebo, 6 = Anergic-6H.
SGOT	Serum glutamic oxaloacetic transaminase	U/L
WBC	White blood cell	$cells/cm^3$
death	censoring status	0 =censoring, 1 =death
survt	survival time for AIDS progression	year

5.3 Statistical analysis and results

5.3.1 Descriptive and non-parametric analysis

First, descriptive statistics are used to give us information about the distributions of the variables. We get the baseline characteristics in 2158 participants using the descriptive statistics (Table 5.1). Since there are 895 subjects missing for hematocrit level, we remove this variable from the group of potential risk factors. We then get the cross-tabulations for the remaining variables with six treatment arms respectively (Table 5.2). In both the PPD-positive and anergy cohorts the treatment groups are approximately balanced at baseline in terms of demographic factors and the laboratory test. Some continuous

Characteristic	n(%); mean(SD)
Male	680(32%)
Age	30(±6.5)
Marital status	
Never married	260(12%)
Current married	943(44%)
Divorce/widow	949(44%)
BCG scar	1413(66%)
Anergic	503(23%)
Education	8 (±3.3)
Body Mass Index(kg/m2)	22.3(±3.6)
Hemoglobin(mg/dl)	12.7(±2.8)
Plalet count(/L)	259.7(±86.9)
Creatinine(mg/dl)	0.88(±0.22)
AST/SGOT(U/L)	26.4(±10.4)
Absolute lymphocyte count	2.27(±0.94)
white blood cell	5.6(±1.8)

Note: Anergic is an indicator of TB skin test induration less than 5 mm.

Table 5.1: Baseline characteristics in 2158 participants

variables are grouped into categories according to clinical meaning. We categorize BMI into three categories BMI ≤ 19, $19 <$ BMI ≤ 25, and BMI > 25. SGOT is categorized into two categories SGOT ≤ 40 and SGOT > 40. When doing so, we use the established cut-points that have clinical meaning and we also ensure that each group contains an adequate number of individuals.

Survival time distributions for incident AIDS and death is estimated for each arm using the K-M method (Section 2.1) and compared using the log-rank test (Section 2.2). The K-M curves for each study arm provide a initial insight into the shape of the survival function for each treatment arm. The log-rank test is used to compare survival time distributions among treatment arms.

	PPD-positive				Anergy	
Characteristics	Placebo (n=398)	6H (n=421)	3HR (n=453)	3HRZ (n=382)	Placebo (n=227)	6H (n=277)
male,%	124(31%)	132(31%)	135(30%)	134(35%)	71(31%)	84(30%)
BCG scar, %	261(66%)	287(69%)	299(66%)	245(65%)	144(63%)	177(64%)
mean education, years	7.9(±3.39)	8.1(±3.20)	7.9(±3.42)	7.8(±3.31)	8.2(±3.23)	7.9(±3.56)
mean age, years	30(±6.62)	30(±6.30)	30(±6.40)	29(±5.89)	30(±6.77)	29(±7.04)
Body mass index	23.0(±3.53)	22.9(±3.78)	23.3(±3.81)	22.8(±3.27)	22.3(±3.57)	22.0(±3.34)
Hemoglobin, mg/dL	12.7(±1.9)	12.7(±1.94)	12.8(±1.89)	12.7(±1.96)	12.7(±1.82)	12.4(±1.91)
Platelet count, /L	262(±85.64)	257(±78.90)	263(±91.58)	261(±90.92)	259(±81.29)	255(±92.00)
Creatinine, mg/dL	0.89(±0.24)	0.89(±0.22)	0.88(±0.23)	0.89(±0.22)	0.83(±0.20)	0.84(±0.19)
AST/SGOT, U/L	26.6(±11.05)	26.8(±10.38)	25.4(±9.34)	25.9(±10.24)	27.1(±10.89)	27.2(±10.75)
Absolute lymphocyte count	2.29(±0.85)	2.35(±0.78)	2.34(±0.82)	2.25(±0.70)	2.07± (0.86)	2.22(±1.59)
white blood cell	5.64(±1.81)	5.73(±1.70)	5.77(±1.95)	5.52(±1.60)	5.31(±1.57)	5.54(±1.73)

Note: PPD=purified protein derivative; 6H=isoniazid (INH) for 6 months; 3HR=INH plus rifampicin for 3 months; 3HRZ=INH plus rifampicin plus pyrazinamide for 3 months.

The body mass index was calculated as the weight in kilograms divided by the square of the height in meters.

Table 5.2: Baseline characteristics by anergic status in 2158 participants

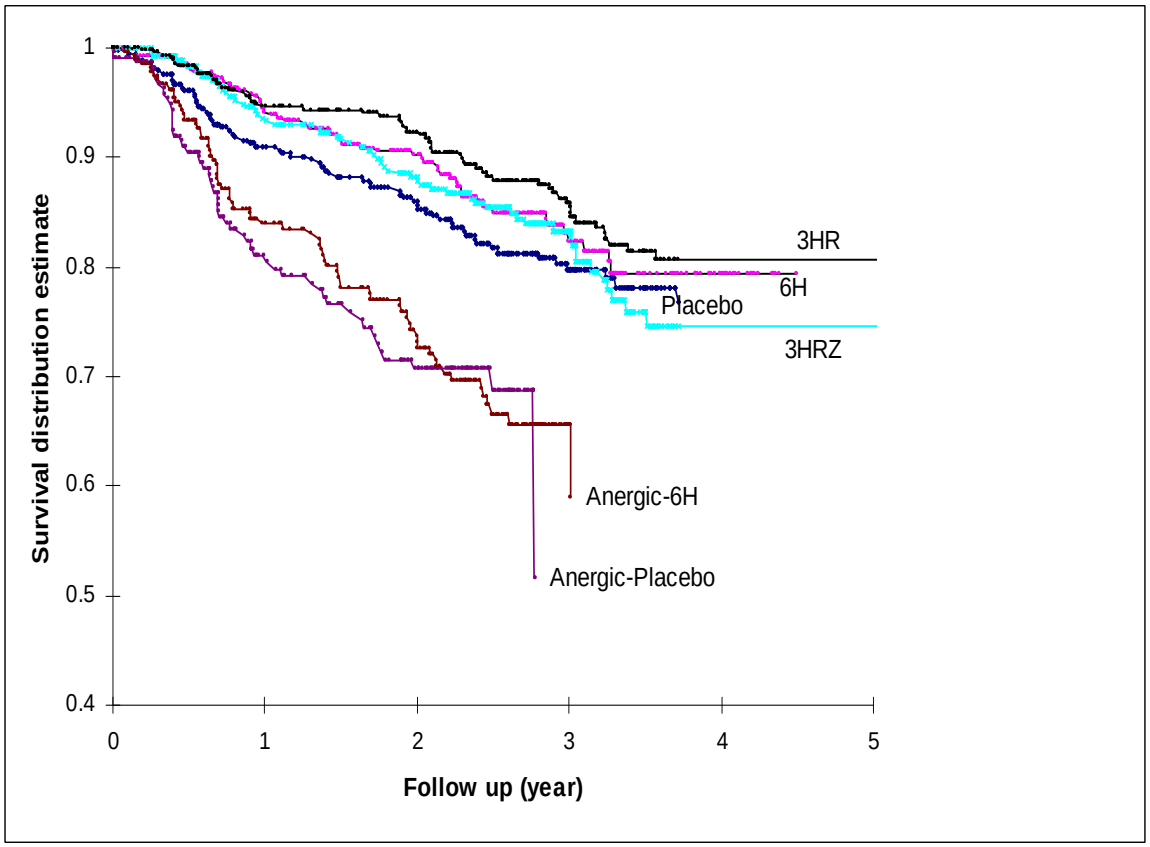

Figure 5.2: K-M curves for the time to AIDS progression among the TB preventive treatment regimens

By the log-rank test, in the PPD-positive cohorts, there is no significant difference in the cumulative incidence of AIDS progression among TB preventive treatment arms ($p = 0.2805$). The same is in the anergic cohorts ($p = 0.3922$). The K-M curves also shows the same result as the log-rank test (Figure 5.2). The K-M curves for time to death and time to combined event of AIDS progression and death are presented (Figure 5.3 and 5.4). The results of these three kinds of endpoints are very similar.

5.3.2 Cox PH model

We use univariate analysis to check all the risk factors before proceeding to more complicated models. We use a univariate Cox proportional hazards regression for every potential risk factor. The Wald test is considered in each univariate Cox PH model. Variables are identified as significant using a 0.1 significance level in the univariate model. We then fit the

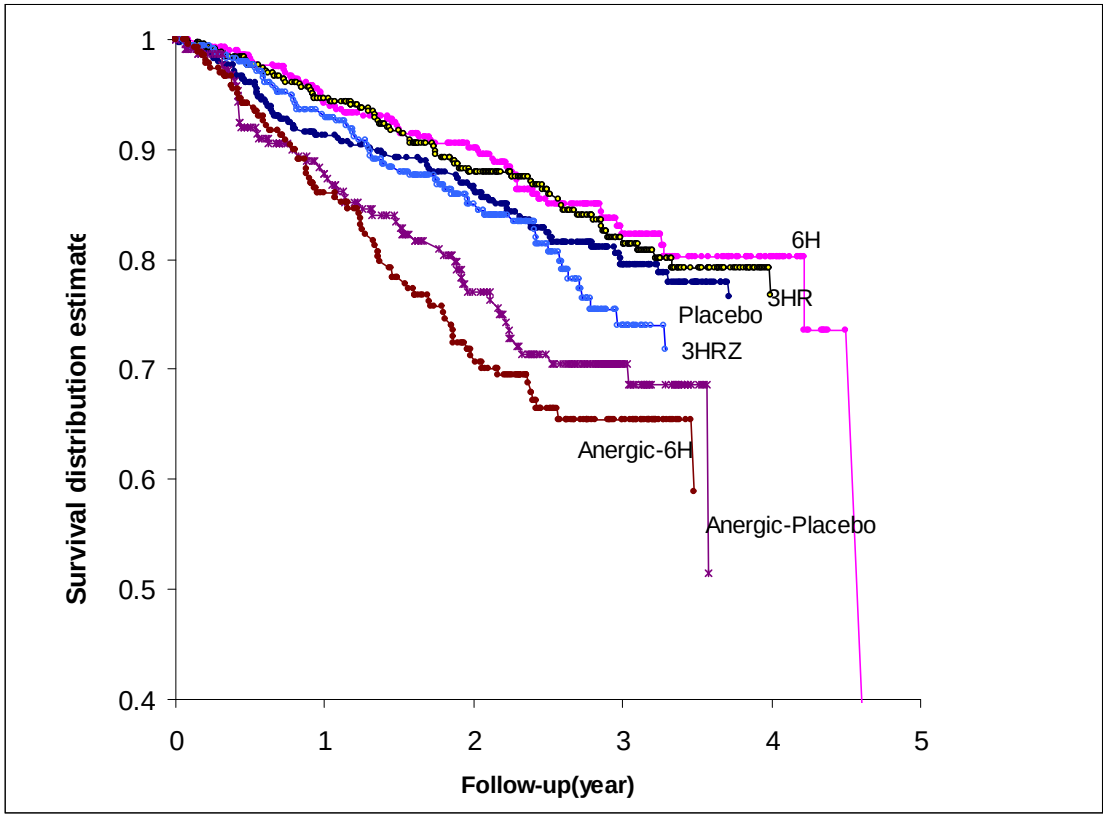

Figure 5.3: Time to death among the TB preventive treatment regimens

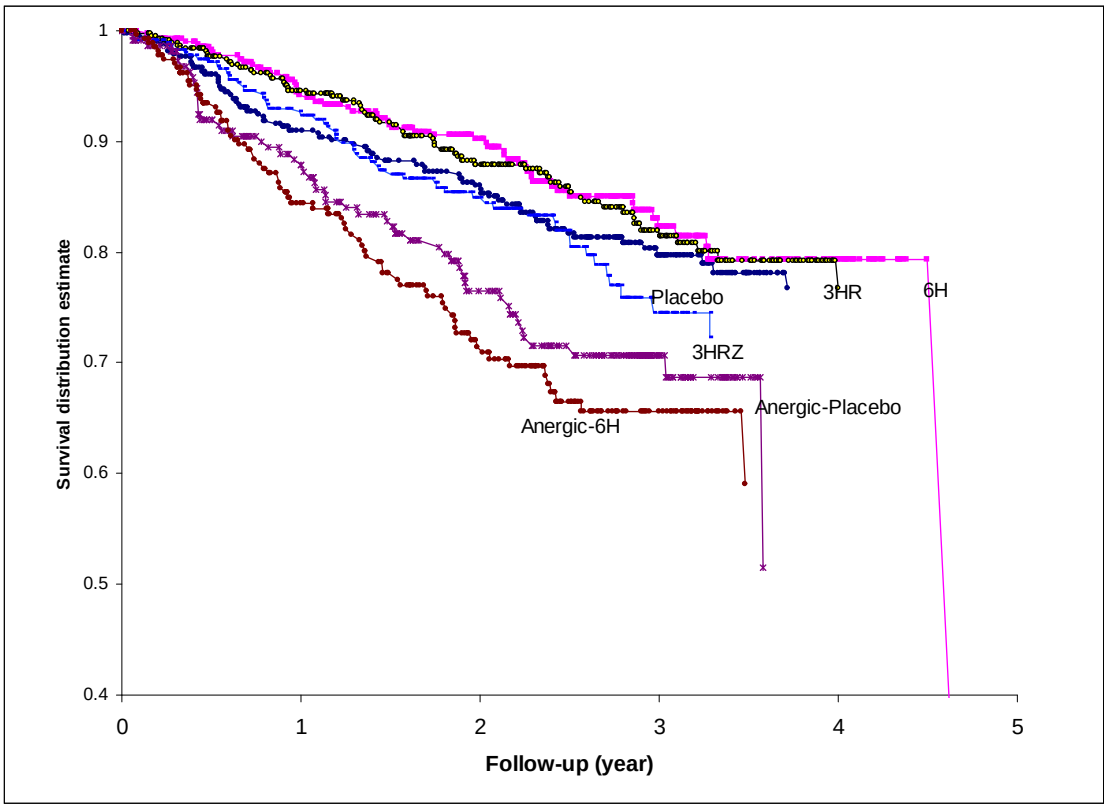

Figure 5.4: Time to AIDS progression or death among the TB preventive treatment regimens

full multivariate Cox PH model including all the potential risk factors and treatment arms. In univariate and the full multivariate proportional hazards models, age, hemoglobin, body mass index, sex, SGOT, and absolute lymphocyte count show a statistically significant association with disease progression to AIDS. But other characteristics such as education, marital status, creatinine level and platelet counts are not statistically significant, suggesting that these variable are not associated with the AIDS progression. So we will consider the model which includes all the significant predictors. The categorical predictor body mass index has three levels and therefore we will include this predictor using two dummy variables (BMIB_2, BMIB_3) with the group $19 < \text{BMI} \leq 25$ as the reference group. After we built a multivariate model of main effects, we then check all the interactions between predictors. To consider the effect of TB preventive therapy on AIDS progression, we include the study arms in the final model. The treatment regimens (STUDYARM) have six levels and therefore we will use five dummy variables (STUDYARMA_{2i}, STUDYARMA_{3i}, STUDYARMA_{4i} STUDYARMA_{5i}, STUDYARMA_{6i}) with PPD positive-Placebo as the reference group. The univariate and multivariate results of a PH model fitted to this dataset are obtained (Table 5.3). The final multivariate Cox PH model is then given by

$$\begin{aligned} h_i(t) = h_0(t) \exp(&-0.52 LYMPHABS - 0.25 STUDYARMA_{2i} - 0.15 STUDYARMA_{3i} \\ &- 0.02 STUDYARMA_{4i} + 0.26 STUDYARMA_{5i} + 0.24 STUDYARMA_{6i} + 0.02 AGE \\ &+ 0.68 BMIB_2 - 0.61 BMIB_3 - 0.32 HGB - 0.54 SEX + 0.56 SGOT). \end{aligned}$$

As expected from non-parametric test, TB preventive effect is not statistically significant in PPD-positive cohort and anergic cohort even after adjusting age, hemoglobin, body mass index, sex, SGOT and absolute lymphocyte count in a multivariate Cox PH model. Among the PPD-positive patients, the hazard ratio for AIDS progression is 0.78 in 6H group, 0.86 in 3HR group and 0.98 in 3HRZ group compared with placebo group, but they are not statistically significant. Among the anergic group, the hazard ratio for AIDS progression is $\exp(0.24 - 0.26) = 0.98$ in 6H group relative to placebo group but not statistically significant. So there is no evidence that the TB preventive therapies have the different effect on AIDS progression through time. We can observe the similar result for event of death and the combined event of AIDS progression or death after controlling age, hemoglobin, body mass index, sex, SGOT and absolute lymphocyte count(Table 5.4).

After a Cox PH model is fitted, the adequacy of this model, including the PH assump-

	Univarate analysis				Mutivariate analysis			
Covariates	β	HR	95%CI	P-value	β	HR	95%CI	P-value
PPD-positive								
Placebo		1				1		
6H	-0.23	0.80	(0.56,1.14)	0.21	-0.25	0.78	(0.55,1.12)	0.18
3HR	-0.17	0.84	(0.60,1.19)	0.34	-0.15	0.86	(0.61,1.22)	0.40
3HRZ	0.14	1.15	(0.81,1.63)	0.44	-0.02	0.98	(0.68,1.41)	0.92
Anergic								
placebo	0.52	1.68	(1.17,2.4)	0.005	0.26	1.30	(0.90,1.87)	0.16
6H	0.69	2	(1.43,2.79)	<.0001	0.24	1.28	(0.90,1.80)	0.17
AGE	0.02	1.02	(1.01,1.04)	0.01	0.02	1.02	(1.00,1.03)	0.02
BMI								
19<BMI<=25		1				1		
BMI<=19	1.05	2.87	(2.26,3.65)	<.0001	0.68	1.98	(1.54-2.56)	<.0001
BMI>25	-0.82	0.44	(0.31,0.62)	<.0001	-0.61	0.54	(0.38-0.77)	0.0006
HGB	-0.26	0.77	(0.73,0.81)	<.0001	-0.32	0.72	(0.68-0.77)	<.0001
Absolute lymphocyte count	-0.79	0.46	(0.39-0.54)	<.0001	-0.52	0.60	(0.51-0.70)	<.0001
SEX	-0.23	0.80	(0.64-0.99)	0.037	-0.54	0.58	(0.46-0.74)	<.0001
SGOT	0.84	2.31	(1.74-3.08)	<.0001	0.56	1.76	(1.30-2.38)	<.0001

Table 5.3: Univariate and multivariate Cox PH model for the relative hazard of AIDS progression

	AIDS progression			Death			AIDS progression or death		
	HR	95%CI	P-value	HR	95%CI	P-value	HR	95%CI	P-value
PPD-positive									
Placebo	1			1			1		
6H	0.78	(0.55,1.12)	0.18	0.8	(0.56,1.14)	0.21	0.77	(0.54,1.08)	0.13
3HR	0.86	(0.61,1.22)	0.40	0.88	(0.63,1.25)	0.48	0.83	(0.60,1.16)	0.27
3HRZ	0.98	(0.68,1.41)	0.92	0.99	(0.70,1.41)	0.95	0.97	(0.69,1.36)	0.84
Anergic									
Placebo	1			1			1		
6H	0.98	(0.71,1.47)	0.92	0.97	(0.67,1.39)	0.86	0.96	(0.68,1.37)	0.84

Note: The hazard ratio is adjusted for age, body mass index, hemoglobin, absolute lymphocyte count, sex, and SGOT in a Cox PH model.

Placebo is the reference group.

Table 5.4: Multivariate Cox PH model for the relative hazard of AIDS progression, death, and combination of AIDS progression or death

tion and the goodness of fit, needs to be assessed. The PH assumption checking with graphical method and two statistical test methods (adding time-dependent covariates in the Cox model and tests based on the Schoenfeld residuals have been described in Section 3.3. We use $\log(-\log(\text{survival}))$ plot (Section 3.3.1) to check the PH assumption for all the categorical variables. There is no evidence that the PH assumption is violated for any categorical variables. We also create the time-dependent covariate by creating interactions of the predictors and survival time and include them in the model (Section 3.3.2). The result indicates that the PH assumption for LYMPHABS is violated (p-value for LYMPHABS $\times$ t is less than .0001). The Schoenfeld residuals are also used to check the PH assumption (Section 3.3.3). We check the p-value for testing whether the correlation between Schoenfeld residual for this covariate and ranked survival time is zero. The p-values for LYMPHABS is less than .0001 and greater than 0.05 for all the other covariates, which suggesting that the PH assumption is violated for LYMPHABS, but reasonable for all the other covariates.

We assess goodness of fit by residual plots (Section 3.4). A plot of the Cox-Snell residuals against the cumulative hazard of Cox-Snell residuals is presented (Figure 5.5). There is some evidence of a systematic deviation from the straight line, which gives us some

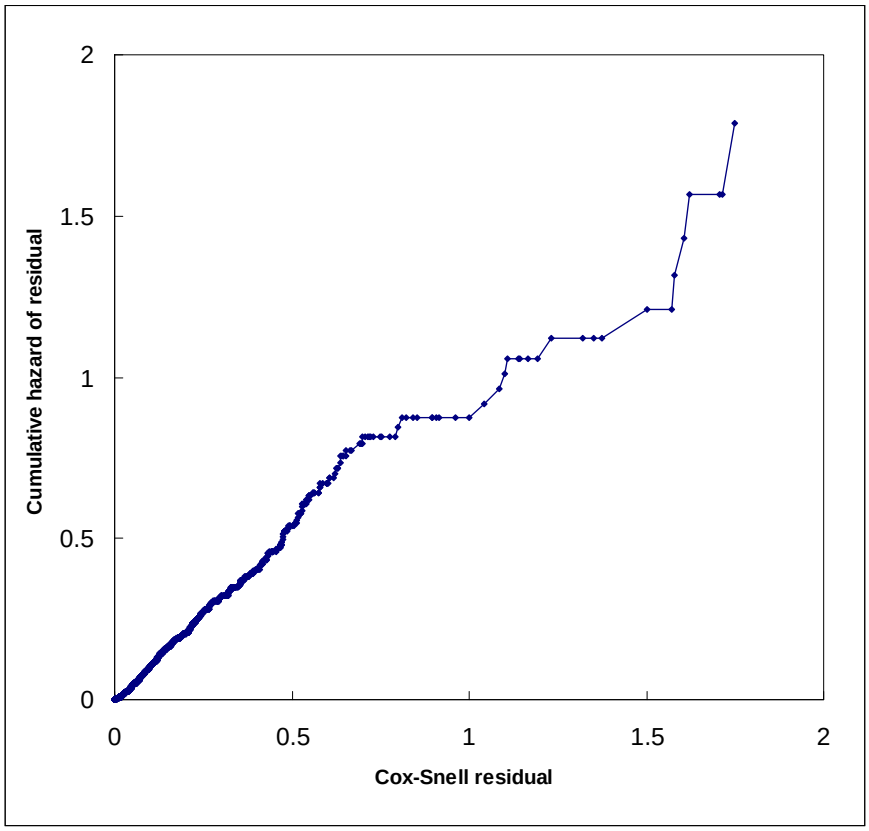

Figure 5.5: Cumulative hazard plot of the Cox-Snell residual for Cox PH model

concern about the adequacy of the fitted model. The plot of deviance residual against the risk score shows that the deviance residuals seem not to be symmetrically distributed about zero. There are very high or very low deviance residuals which suggest that these observations may be outliers. (Figure 5.6). Therefore, we have some concern about the adequacy of the fitted Cox PH model. We also use delta-beta statistic to measure the influential observations on the model as a whole. It shows that the coefficient does not change too much when the observations corresponding to the largest delta-beta statistics are removed. Therefore, we do not remove them from the dataset and conclude that there are no influential observations.

5.3.3 Cox model with time-dependent variables

We have shown that the Cox model displayed nonproportionality for variable LYMPHABS. We believe that there is an interaction between LYMPHABS and time. It is not appro-

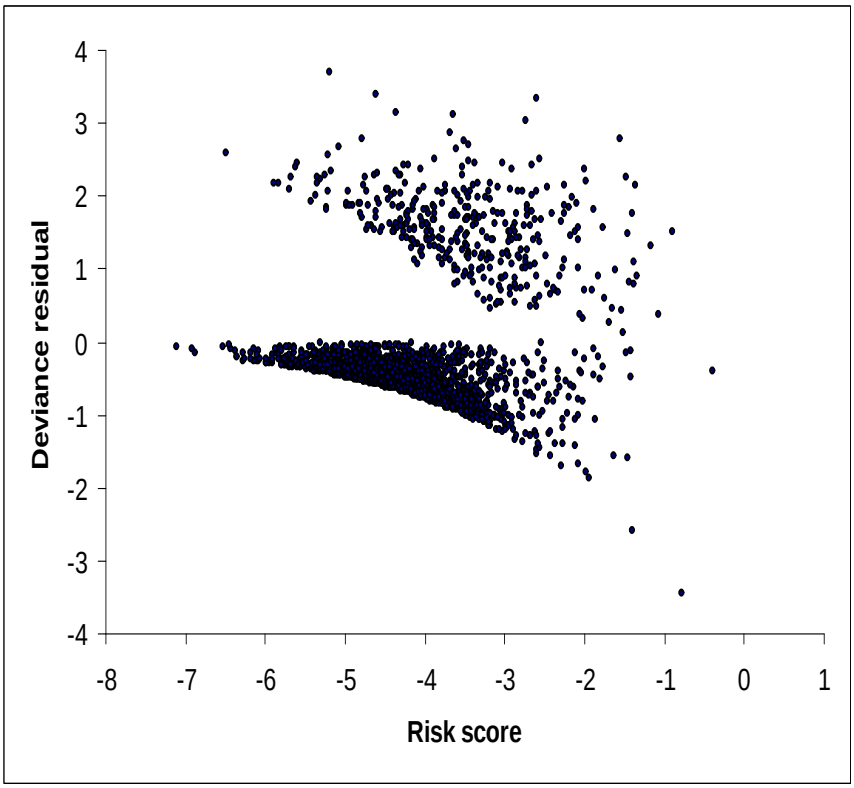

Figure 5.6: Deviance residuals plotted against the risk score for Cox PH model

	Year				
variable	0-0.5	0.5-1year	1-2 years	2-3 years	>3 years
g1(t)	0	1	0	0	0
g2(t)	0	0	1	0	0
g3(t)	0	0	0	1	0
g4(t)	0	0	0	0	1

Table 5.5: Time-dependent covariates represent different time periods

priate to use stratified Cox model because LYMPHABS is a continuous variable. We then incorporate time-dependent covariate into the model (Section 3.5.2).

We may define $X(t)$ =LYMPHABS$\times t$ and formulate a model (model A):

$$h(t, \mathbf{x}(t)) = h_0\,(t)\,\{\exp(\beta_L LYMPHABS + \beta_1(LYMPHABS \times t) + \boldsymbol{\beta}^{'}\mathbf{x})\},$$

where x is the vector of all the fixed covariates (age, study arms, hemoglobin, body mass index, sex and SGOT) and $\boldsymbol{\beta}$ is the corresponding vector of the regression coefficient for the fixed covariates. The effect of LYMPHABS is $\exp(\beta_L + \beta_1 t)$. β_L can be interpreted as the effect of LYMPHABS at study enrollment. This model states that the effect of LYMPHABS is assumed to increase or decrease linearly in relation to time.

Alternatively, the proportional hazards assumption may hold at least approximately over short time periods instead of the entire time period. In this situation, hazard ratio of LYMPHABS may change at discrete intervals. We partitioned the time period into 5 sub-periods and created five binary, time-dependent covariates to represent them (Table 5.5). The first interval goes from 0 to half of a year; the second time interval goes from half to one year; the third time interval goes from 1 year to 2 years; the fourth interval goes from 2 to 3 years; and the last interval goes from 3 years onward. This model assumes that there are five different hazard ratios estimates in five intervals.

The Cox non-PH model is fitted as follows:

$$\begin{aligned} h(t, \mathbf{x}(t)) = h_0\,(t) \exp\{&\delta_L LYMPHABS + \delta_1(LYMPHABS \times g_1\,(t)) + \delta_2(LYMPHABS \times g_2\,(t)) \\ &+ \delta_3(LYMPHABS \times g_3\,(t)) + \delta_4(LYMPHABS \times g_4\,(t)) + \boldsymbol{\delta}^{'}\mathbf{x}\}. \end{aligned}$$

Period	δ	HR	P-value	95%CI
0-6 months	-1.09	0.34	<.0001	(0.23-0.49)
6-12 months	0.26	0.44	<.0001	(0.31-0.61)
1-2 years	0.53	0.57	<.0001	(0.43-0.76)
2-3 years	1.10	1.02	0.92	(0.76-1.36)
>3 years	1.16	1.07	0.83	(0.55-2.08)

HR= hazard ratio CI= confidence interval

Table 5.6: Time-dependent effect of absolute lymphocyte count (LYMPHABS) in five time intervals

where x is the vector of all the fixed covariates (age, study arms, hemoglobin, body mass index, sex and SGOT) and δ is the corresponding vector of the regression coefficient for the fixed covariates. The non-PH model allows the effect of LYMPHABS varies with time periods. The coefficients δ_1 to δ_4 denote the interaction effect between LYMPHABS and time. The effect of LYMPHABS in the first half year is given by $\exp(\delta_L)$. The effect of LYMPHABS in the subsequent period are estimated by $\exp(\beta_1 + \delta_i), i = 1, 2, 3, 4$. The results are presented in Table 5.6.

The result shows a significant effect of LYMPHABS below 2 years and a nonsignificant effect of LYMPHABS after 2 years. Hazard ratios in the first three time intervals are similar and hazard ratios for the last two time intervals are similar, we therefore separate the data into two time intervals. We use one Heaviside function $g(t)$, where

$$g(t) := \begin{cases} 0 & \text{if } t \leq 2 \\ 1 & \text{if } t > 2 \end{cases}.$$

The fitted model (model B) with time-dependent variables are

$$h(t, \mathbf{x}(t)) = \exp\{-0.73 LYMPHABS + 0.77(LYMPHABS \times g(t)) + \boldsymbol{\alpha}^{'} \boldsymbol{x}\}.$$

Tables 5.7 gives the results of model A: the Cox non-PH model with smooth time-dependent hazard ratio and model B: the Cox non-PH model with discrete time interval.

In model A, the hazard ratios for the effect of LYMPHABS in each time interval,

	Smooth Cox non-PH model				Piecewise Cox non-PH model			
Variables	β	HR	95%CI	P-value	β	HR	95%CI	P-value
PPD-positive								
Placebo		1				1		
6H	-0.25	0.78	(0.55,1.12)	0.18	-0.24	0.78	(0.55,1.12)	0.17
3HR	-0.13	0.88	(0.62,1.24)	0.47	-0.13	0.87	(0.62,1.24)	0.45
3HRZ	0.01	0.95	(0.72,1.49)	0.95	0.003	1.03	(0.71,1.43)	0.98
Anergic								
Placebo	0.26	1.29	(0.90,1.86)	0.17	0.26	1.29	(0.90,1.86)	0.17
6H	0.27	1.31	(0.93,1.86)	0.12	0.26	1.30	(0.92,1.94)	0.13
AGE	0.02	1.02	(0.93,1.86)	0.03	0.02	1.02	(1.00,1.03)	0.03
BMI								
19<BMI<=25		1				1		
BMI<=19	0.65	1.92	(1.48,2.48)	<.0001	0.67	1.95	(1.51,2.51)	<.0001
BMI>25	-0.62	0.54	(0.38,0.76)	0.0005	-0.62	0.54	(0.38,0.77)	0.0006
HGB	-0.32	0.73	(0.69,0.77)	<.0001	-0.32	0.73	(0.69,0.77)	<.0001
SEX	-0.54	0.58	(0.46,0.74)	<.0001	-0.54	0.58	(0.46,0.74)	0.0004
SGOTX	0.57	1.77	(1.31,2.39)	0.0002	0.58	1.78	(1.32,2.40)	0.0002
LYMPHABS	-1.08	0.34	(0.25,0.47)	<.0001	-0.73	0.48	(0.40,0.58)	<.0001
LYMPHABS*t	0.40	1.50	(1.27,1.76)	<.0001				
LYMPHABS*g(t)					0.77	2.23	(1.57,2.97)	<.0001
-2loglikelihood	4878.172				4879.781			

Note: LYMPHABS= Absolute lymphocyte count.

g(t) is the heaviside function, which is zero when time is less than or equal to 2 years and 1 when time is greater than 2 years.

Table 5.7: Cox models with time-dependent covariates

Period	β	HR	P-value	95%CI
0-2 years	-0.73	0.48	<.0001	(0.40-0.58)
>2 years	0.04	1.04	0.75	(0.80-1.35)

Table 5.8: Time-dependent effect of LYMPHABS in two time intervals

controlling all the other covariates is given as follows:

$$HR = \exp(\beta_L + \beta_1 t) = \exp\left(-1.08 + 0.4t\right).$$

$\beta_1 = 0.4$ is positive, which indicates that the effect of LYMPHABS increases linearly with time. When time = 1, $HR = \exp(-1.08 + 0.4 * 1) = 0.51$; when time = 2, $HR = \exp(-1.08 + 0.4 * 2) = 0.76$; when time = 3, $HR = \exp(-1.08 + 0.4 * 3) = 1.13$. In the first two years, HR less than one indicates that the AIDS progression hazard decreases as the value of the LYMPHABS increases. After two years, HR greater than 1 indicates the LYMPHABS is positively associated with the AIDS progression probability.

In model B, we define one cut point (2 years) on the time axis. The two hazard ratios are given by separately exponentiating each of the two estimated coefficients. When time is less than 2 years, $\widehat{HR} = \exp(-0.73) = 0.48$. When time is greater than 2 years, $\widehat{HR} = \exp(-0.73 + 0.77) = \exp(0.04) = 1.04$. We can obtain that the 95% confidence interval for the two hazard ratios are (0.40-0.58) and (0.80-1.35) manually based on the fitted model B. The 95% confidence interval in the first two years does not include one, which means the effect of LYMPHABS is statistically significant. The 95% confidence interval includes one after two years, which means that the effect of LYMPHABS is not statistically significant any more after two years. The estimated LYMPHABS effects are presented in Table 5.8. Therefore, the hazard ratio decreases by 1/2 when LYMPHABS increases by one unit in the first two years. After two years, there is no evidence that LYMPHABS has an effect on AIDS progression. The estimated treatment effects are similar in two models. The results are also very similar to the Cox PH model except for the effect of LYMPHABS.

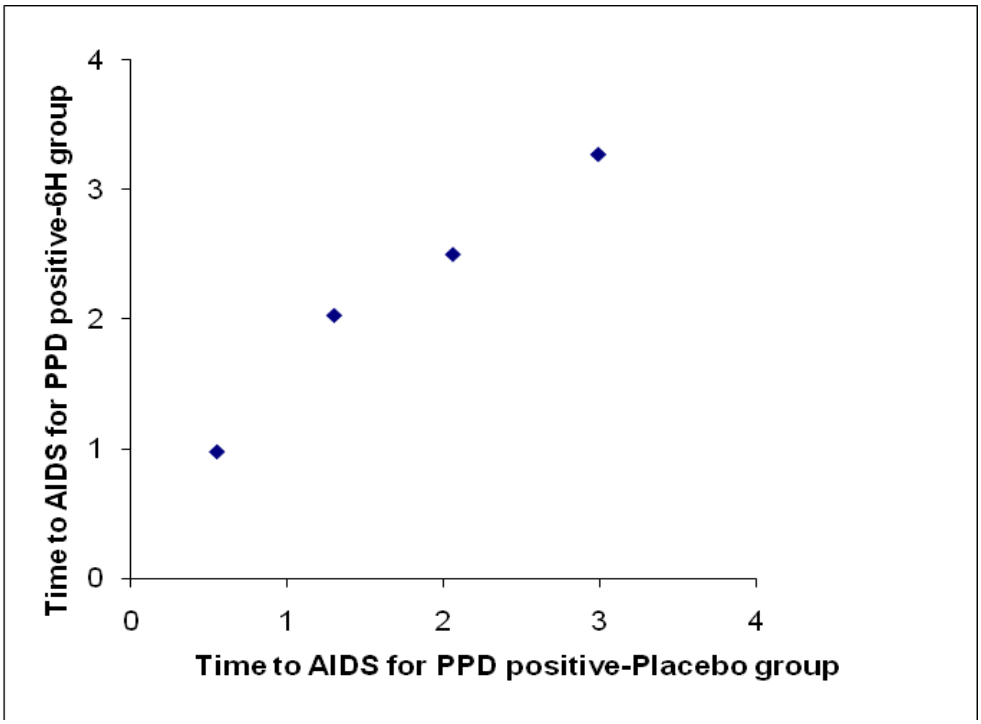

Figure 5.7: Q-Q plot for time to AIDS progression

5.3.4 AFT model

The accelerated failure time (AFT) model is another alternative of the Cox PH model when the PH assumption is violated. The AFT model can be used to express the magnitude of effect in a more accessible way in terms of difference between treatment in survival time. We fit the dataset using exponential, Weibull, log-logistic, log-normal and gamma AFT model. For each kind of model, we fit both the univariate and multivariate AFT model. In both univariate and multivariate AFT models, age, hemoglobin, body mass index, sex, SGOT, and absolute lymphocyte count are statistically significantly associated with disease progression to AIDS. No interactions are statistically significant in multivariate AFT models. The results from the different AFT models applied to the time to AIDS progression are presented in Table 5.9. There is no big difference for the estimations in different models.

The Q-Q plot (Section 4.2.7) is used to check the AFT assumption. The Q-Q plot in Figure 5.7 approximates well to a straight line from the origin indicating that the AFT model may provide an appropriate model.

Variable	Exponential			Weibull				Log-logistic				Log-normal				Gamma			
	β	TR	95% CI	P-value	β	TR	95% CI	P-value	β	TR	95% CI	P-value	β	TR	95% CI	P-value	β	TR	95% CI	P-value
Intercept	-2.28			<.0001	-1.85			<.0001	-2.97			<.0001	-2.8			<.0001	-2.69			<.0001
PPD-positive																				
Placebo		1				1				1				1				1		
6H	0.23	1.26	(0.88,1.79)	0.21	0.18	1.20	(0.89,1.62)	0.22	0.21	1.23	(0.92,1.64)	0.16	0.22	1.25	(0.92,1.68)	0.15	0.21	1.23	(0.92,1.66)	0.16
3HR	0.15	1.16	(0.82,1.63)	0.4	0.12	1.13	(0.84,1.50)	0.42	0.14	1.15	(0.87,1.53)	0.32	0.19	1.21	(0.89,1.63)	0.21	0.17	1.19	(0.88,1.59)	0.26
3HRZ	0.01	1.01	(0.71,1.43)	0.95	-0.01	0.99	(0.74,1.33)	0.95	-0.01	0.99	(0.74,1.32)	0.94	-0.06	0.94	(0.70,1.27)	0.7	-0.04	0.96	(0.71,1.29)	0.80
Anergic																				
Placebo	-0.27	0.76	(0.53,1.10)	0.15	-0.24	0.79	(0.58,1.06)	0.12	-0.29	0.75	(0.55,1.02)	0.07	-0.27	0.76	(0.55,1.05)	0.1	-0.27	0.76	(0.55,1.04)	0.05
6H	-0.26	0.77	(0.55,1.09)	0.14	-0.22	0.80	(0.60,1.07)	0.13	-0.33	0.72	(0.54,0.97)	0.03	-0.03	0.97	(0.54,1.01)	0.05	-0.30	0.74	(0.54,1.00)	0.87
AGE	-0.02	0.98	(0.97,0.99)	0.02	-0.01	0.99	(0.97,0.99)	0.02	-0.01	0.99	(0.97,1.0009)	0.07	-0.01	0.99	(0.97,0.99)	0.1	-0.01	0.99	(0.97,1.00)	0.08
BMI																				
19<BMI<=25		1				1				1				1				1		
BMI<=19	-0.65	0.52	(0.40,0.67)	<.0001	-0.57	0.57	(0.45,0.70)	<.0001	-0.54	0.58	(0.46,0.73)	<.0001	-0.58	0.56	(0.44,0.73)	<.0001	-0.57	0.56	(0.44,0.72)	<.0001
BMI>25	0.58	1.79	(1.27,2.54)	<.0001	0.50	1.65	(1.24,2.21)	0.0007	0.44	1.56	(1.19,2.05)	0.002	0.48	1.62	(1.23,2.14)	0.0005	0.47	1.61	(1.22,2.11)	0.0007
HGB	0.32	1.38	(1.30,1.45)	<.0001	0.27	1.31	(1.25,1.38)	<.0001	0.32	1.38	(1.30,1.45)	<.0001	0.32	1.38	(1.30,1.46)	<.0001	0.32	1.37	(1.30,1.44)	<.0001
Absolute lymphocyte count	0.49	1.63	(1.39,1.90)	<.0001	0.42	1.53	(1.33,1.74)	<.0001	0.44	1.56	(1.37,1.77)	<.0001	0.41	1.51	(1.34,1.72)	<.0001	0.42	1.52	(1.34,1.73)	<.0001
SEX	0.53	1.70	(1.35,2.16)	<.0001	0.45	1.56	(1.28,1.91)	<.0001	0.65	1.91	(1.55,2.37)	<.0001	0.64	1.89	(1.52,2.36)	<.0001	0.62	1.86	(1.50,2.32)	0.0001
SGOT	-0.56	0.57	(0.42,0.77)	0.0002	-0.50	0.61	(0.47,0.78)	0.0001	-0.58	0.56	(0.43,0.73)	<.0001	-0.6	0.55	(0.41,0.73)	<.0001	-0.59	0.56	(0.42,0.73)	<.0001
Scale	1				0.83		(0.76,0.91)		0.71	2.03	(0.65,0.77)		1.36	3.90	(1.26,1.48)		1.23	3.42	(1.06,1.43)	
shape	1.00				1.20		(1.10,1.31)										0.24	1.27	(-0.03,0.51)	
Log-likelihood	-1031				-1096				-1080						-1081		-1080			

Table 5.9: Results from AFT models for time to AIDS progression

	No of parameters	Log-likelihood	Testing against the Gamma distribution	
Distribution	m	L	LR	df
Exponential	1	-1103.797	47.902	2
Weibull	2	-1095.900	32.108	1
Log-Normal	2	-1081.191	2.690	1
Gamma	3	-1079.846		
Log-logistic	2	-1079.740	Not nested	

Table 5.10: The log-likelihoods and likelihood ratio (LR) tests, for comparing alternative AFT models

Distribution	Log-likelihood	k	c	AIC
Exponential	-1103.797	12	1	2233.594
Weibull	-1095.900	12	2	2219.800
Log-Normal	-1081.191	12	2	2190.382
Gamma	-1079.846	12	3	2189.692
Log-Logistic	-1079.740	12	2	2187.480

Table 5.11: Akaike Information Criterion (AIC) in the AFT models

We compared all these AFT models using statistical criteria (likelihood ratio test and AIC). The nested AFT models can be compared using the likelihood ratio (LR) test. The exponential model, the Weibull model and the log-normal model are nested within the gamma model (Table 5.10).

According to the LR test, the log-normal model fits better. However, the LR test is not valid for comparing models that are not nested. In this case, we use AIC to compare the models (Table 5.11) (The smaller AIC is the better). The log-logistic AFT model appears to be an appropriate AFT model according to AIC compared with other AFT models, although it is only slightly better than log-normal or Gamma model. We also note that the Weibull and exponential model are poorer fits according to LR test and AIC. This provides more evidence that the PH assumption for this data is not appropriate.

Furthermore, we check the goodness of fit of the model using residual plots. The

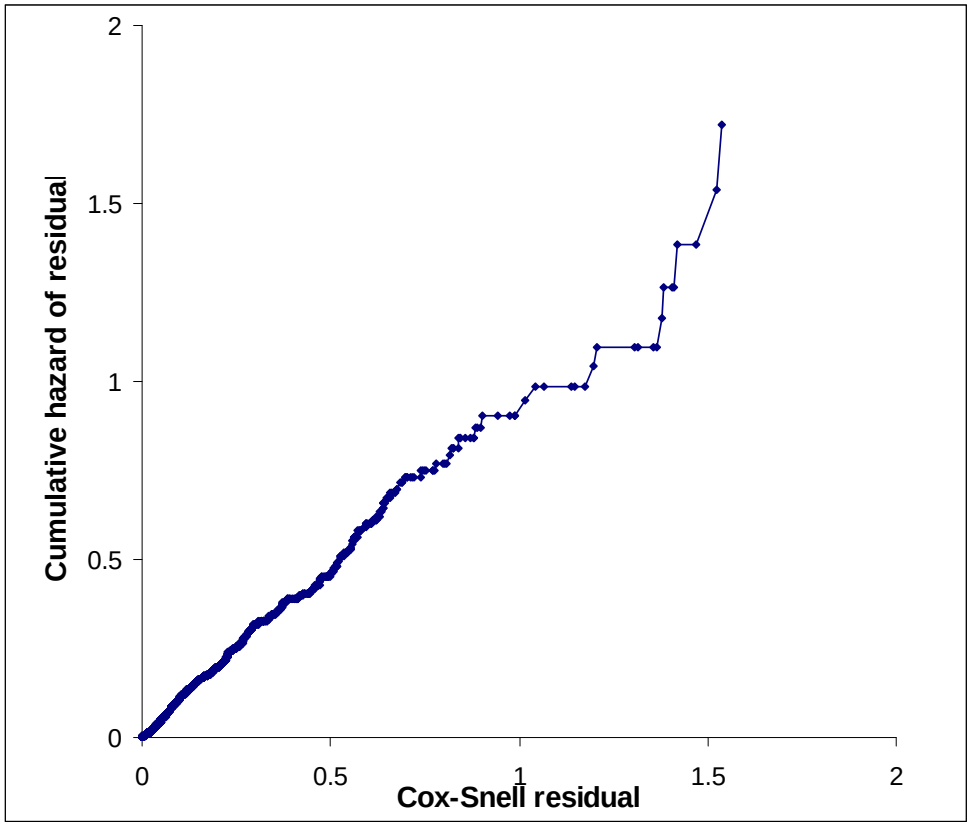

Figure 5.8: Cumulative hazard plot of the Cox-Snell residual for log-logistic AFT model

cumulative hazard plot of the Cox-Snell residuals in log-logistic model is presented in Figure 5.8. The plotted points lie on a line that has a unit slope and zero intercept. So there is no reason to doubt the suitability of this fitted log-logistic model. At last, we conclude that the log-logistic model is the best fitting the AFT model based on AIC criteria and residuals plot.

Under the log-logistic AFT model, in PPD-positive cohort, the estimated acceleration factor for an individual in 6H group, 3H group and 3HRZ group relative to an individual in placebo group is 1.23, 1.15, 0.99 respectively. This indicates that the effect of 6H, 3HR prolongs the time to AIDS progression, but the effect of 3HRZ speeds up the time to AIDS progression. However, they are not statistically significant. In the anergic cohort, the effect of placebo appears to slow down the time to AIDS progression but it is nonsignificant. We can calculate the acceleration factors and the corresponding confidence interval for every pair of groups manually. We can also obtain these by refitting the model in which different

dummy variables for treatment regimens are created.

The acceleration factor for age is 0.99, which indicates that the earlier AIDS progression and shorter survival time are more likely for the older persons. The time ratio for HGB and LYMPHABS above 1 implies that these variables prolong the time to AIDS progression as they increase. Men have longer survival time and AIDS progression time than women. The SGOT higher than 40 U/L in a HIV-infected patient speeds up AIDS disease progression and mortality than that less than 40 U/L. The patient with BMI less than 19 has shorter AIDS progression than patients with normal BMI, but the patients with BMI above 25 has longer AIDS progression than patients with BMI with normal BMI.

We now derive model-based predictions. From equation (4.13), the median survival time for the ith individual under the log-logistic model is given by $t\,(50) = \exp(\mu + \alpha x_i)$. For the individual (STUDYARMA$_i = 0$, $i = 2, ..., 6$, mean age = 30, BMIB$_2 = 0$, BMIB$_3 = 0$, mean HGB = 12.7, mean LYMPHABS = 2.27, sex = 0, SGOT = 0), the estimated median survival time for this individual in placebo group in PPD positive cohort is

$$\exp(-2.97 - 0.01 \times 30 + 0.32 \times 12.7 + 0.44 \times 2.27) = 6.01.$$

The estimated median survival time for a individual (STUDYARMA$_2 = 1$, STUDYARMA$_i = 0$, $i = 3, ..., 6$, mean age = 30, BMIB$_2 = 0$, BMIB$_3 = 0$, mean HGB = 12.7, mean LYMPHABS = 2.27, sex = 0, SGOT = 0) in 6H group in PPD positive cohort is

$$\exp(-2.97 + 0.21 \times 1 - 0.01 \times 30 + 0.32 \times 12.7 + 0.44 \times 2.27) = 7.41.$$

Note that the ratio of the two median time is the acceleration factor for 6H group compared with placebo group, which is 1.23.

Instead of predicting survival time, we often want to predict the probability of surviving to some specified time. For example, the five year survival probabilities for every individual in this data can be obtained. Table 5.12 shows the first 10 cases of this dataset. The last column (PROB) contains the five year survival probabilities based on the fitted log-logistic model.

Using equation (4.11), the fitted survival function for the ith individual is

$$\widehat{S}_i\,(t) = \left\{1 + t^{\frac{1}{\widehat{\sigma}}} \exp(\widehat{\eta}_i)\right\}^{-1} = \left\{1 + t^{\frac{1}{0.71}} \exp(\widehat{\eta}_i)\right\}^{-1},$$

where

$$\begin{aligned}\widehat{\eta}_i &= \frac{-\mu - \widehat{\alpha} x_i}{\widehat{\sigma}} \\ &= \frac{1}{0.71} 2.97 - 0.21 STUDYARMA_2 - 0.14 STUDYARMA_3 + 0.01 STUDYARMA_4 \\ &+ 0.29 STUDYARMA_5 + 0.33 STUDYARMA_6 + 0.01 AGE + 0.54 BMIB_2 \\ &- 0.44 BMIB_3 - 0.32 HGB - 0.44 LYMPHABS - 0.65 SEX + 0.58 SGOT\}.\end{aligned}$$

Using equation (4.12), the corresponding estimated hazard function for the ith individual is

$$\begin{aligned}\widehat{h}_i(t) &= \frac{1}{\widehat{\sigma} t}\{1 + t^{-\frac{1}{\widehat{\sigma}}} \exp(-\widehat{\eta}_i)\}^{-1} \\ &= \frac{1}{0.71t}\{1 + t^{-\frac{1}{0.71}} \exp(-\widehat{\eta}_i)\}.\end{aligned}$$

Note that the exponential and Weibull AFT models are also PH models. The signs of the coefficients in the AFT model are opposite to the signs for the PH model. The estimate of shape parameter in Weibull model is 1.2 which is greater than 1, and the 95% CI is (1.10,1.31) which does not cover the null value of 1. This suggests that the Weibull model may be better than the exponential model. Therefore we only compare Weibull AFT model and PH model here. Using equation (4.6), we can calculate the regression coefficients in Weibull PH model. The correspondence between the parameters of the Weibull PH and AFT models are presented in Table 5.13.

The estimated median survival time of the particular individual under the Weibull AFT model, from equation (4.8), is

$$t(50) = \exp\{\sigma \log(\log 2) + \mu + \alpha x_i\}.$$

The estimated median survival time for an individual in placebo group in PPD positive cohort after adjusting all the other covariates(STUDYARMA$_i$ = 0, $i = 2, ..., 6$, mean age = 30, BMIB$_2$ = 0, BMIB$_3$ = 0, mean HGB = 12.7, mean LYMPHABS = 2.27, sex = 0, SGOT = 0) is given by

$$\exp\{0.83 \log(\log 2) - 1.85 - 0.01 \times 30 + 0.27 \times 12.7 + 0.42 \times 2.27\} = 6.05.$$

The estimated median survival time for a individual in 6H group in PPD positive cohort is

$$\exp\{0.83 \log(\log 2) - 1.85 + 0.18 \times 1 - 0.01 \times 30 + 0.27 \times 12.7 + 0.42 \times 2.27\} = 7.24.$$

The ratio of the two medians is 1.20.

From equation (4.4), the estimated survival function for the ith individual is given by

$$\widehat{S}_i(t) = \exp\left\{-t^{\frac{1}{\widehat{\sigma}}}\widehat{\varsigma}_i\right\},$$

where

$$\begin{aligned}\widehat{\varsigma}_i &= \frac{-\mu - \widehat{\alpha}x_i}{\widehat{\sigma}} \\ &= \frac{1}{0.83}(1.85 - 0.18STUDYARMA_2 - 0.12STUDYARMA_3 + 0.01STUDYARMA_4 \\ &+ 0.24STUDYARMA_5 + 0.22STUDYARMA_6 + 0.01AGE + 0.57BMIB_2 \\ &- 0.50BMIB_3 - 0.27HGB - 0.42LYMPHABS - 0.45SEX + 0.50SGOT).\end{aligned}$$

From equation (4.7), the estimated hazard function for the ith individual is by

$$\begin{aligned}\widehat{h}_i(t) = \widehat{h}_i(t) &= \widehat{\sigma}^{-1}t^{\widehat{\sigma}^{-1}-1}\exp(\widehat{\varsigma}_i) \\ &= \frac{1}{0.83}t^{\frac{1}{0.83}-1}\exp(\widehat{\varsigma}_i). \qquad (5.1)\end{aligned}$$

We can also obtain this by the PH representation of the Weibull model. The estimated hazard function for the ith individual is

$$\begin{aligned}\widehat{h}_i(t) &= \lambda\gamma t^{\gamma-1}\exp(\boldsymbol{\beta}'\mathbf{x}_i) \\ &= -2.23 \times \frac{1}{0.83}t^{\frac{1}{0.83}-1}\exp(0.22STUDYARMA_2 - 0.14STUDYARMA_3 - 0.01STUDYARMA_4 \\ &- 0.29STUDYARMA_5 - 0.26STUDYARMA_6 - 0.02AGE - 0.69BMIB_2 + 0.60BMIB_3 \\ &+ 0.33HGB + 0.51LYMPHABS + 0.54SEX - 0.60SGOT).\end{aligned}$$

This turns out to be the same as equation (5.1).

We also fit the log-logistic AFT model for time to death and time to combination of AIDS and death. The results (see Table 5.14) for three kinds of events are very similar.

Obs	A2	A3	A4	A5	A6	AGE	B2	B3	HGB	LYMPHABS	SEX	SGOTX	death	t	prob
1	0	0	1	0	0	34	0	1	13.5	3.9	1	0	1	5	0.952
2	0	0	0	0	0	26	0	0	10.3	2	1	0	1	5	0.470
3	1	0	0	0	0	20	1	0	12.3	2.2	1	0	0	5	0.632
4	0	1	0	0	0	33	0	0	11.2	1.3	1	0	0	5	0.483
5	0	0	0	0	0	28	0	1	14.4	2.1	1	0	0	5	0.916
6	0	1	0	0	0	32	0	0	13.1	2.8	1	0	1	5	0.851
7	1	0	0	0	0	24	0	0	11	2.4	1	0	0	5	0.685
8	0	0	0	0	0	24	0	0	13.9	1.6	1	0	0	5	0.784
9	0	0	1	0	0	31	0	0	12.1	1.8	1	0	0	5	0.615
10	1	0	0	0	0	27	0	0	10.6	3.8	1	0	0	5	0.805

Table 5.12: Predicted 5 year survival probabilities for the first ten individuals based on log-logistic AFT model

5.3.5 Piecewise exponential model

All the AFT models we have considered assume that the hazard is a smooth function of time. The Cox model lacks the capability to test the shape of the hazard function. Alternatively, the underlying hazard function may be approximated by a step function. The piecewise exponential model (Section 4.1.2) can be used to describe both the effect of covariates and underlying hazard rate.

We first break up the whole follow-up period into five intervals of 1 year each and assume that the hazard is constant within each interval. We then create a new dataset with possibly multiple records for each person. And the time variable is coded as the length of time from the start of the interval until death. There is a certain arbitrariness that arises from the division of the observation period into intervals. To increase reliability in the result, we try other divisions and see if the results are stable. We reestimate the model with the division into 3 intervals of 2 years each. The likelihood ratio test is used to test the significant effect of interval by rerunning the model without variable intervals.

The results of the piecewise exponential model is in Table 5.15. For the five interval piecewise exponential model, the likelihood ratio test shows a significant effect of interval implying that the hazard is not constant over time. The coefficients for the four indicator variables are all compared with the first interval. The pattern of the coefficient estimates

	PH				AFT			
Variable	β	se	HR	95%CI	α	se	TR	95%CI
Intercept					-1.85	0.41		(-2.66,-1.05)
PPD-positive								
Placebo			1				1	
6H	-0.22	0.18	0.80	(-0.57,0.13)	0.18	0.15	1.20	(0.89,1.62)
3HR	-0.14	0.18	0.87	(-0.49,0.21)	0.12	0.15	1.13	(0.84,1.50)
3HRZ	0.01	0.18	1.01	(-0.34,0.36)	-0.01	0.15	0.99	(0.74,1.33)
Anergic	0.00							
Placebo	0.29	0.19	1.33	(-0.08,0.65)	-0.24		0.79	(0.58,1.06)
6H	0.26	0.18	1.30	(-0.08,0.61)	-0.22	0.16	0.80	(0.60,1.07)
AGE	0.02	0.01	1.02	(-0.04,0.005)	-0.01	0.01	0.99	(0.97,0.99)
BMI								
19<BMI<=25			1				1	
BMI<=19	0.69	0.14	1.99	(-0.96,0.42)	-0.57	0.11	0.57	(0.45,0.70)
BMI>25	-0.60	0.18	0.55	(0.24,0.97)	0.50	0.15	1.65	(1.24,2.21)
HGB	-0.33	0.02	0.72	(0.28,0.38)	0.27	0.03	1.31	(1.25,1.38)
LYMPHABS	-0.51	0.08	0.60	(0.35,0.67)	0.42	0.07	1.53	(1.33,1.74)
SEX	-0.54	0.12	0.58	(0.30,0.78)	0.45	0.10	1.56	(1.28,1.91)
SGOTX	0.60	0.16	1.82	(-0.91,-0.28)	-0.50	0.13	0.61	(0.47,0.78)
Scale	9.28	22.13		(-34.1,52.6)	0.83	0.04		(0.76,0.91)
Shape	1.20	0.05		(1.10,1.31)	1.20	0.05		(1.10,1.31)

Table 5.13: Comparison of Weibull PH and AFT model

	AIDS progression			Death			AIDS progression or death		
	β	95%CI	TR	β	95%CI	TR	β	95%CI	TR
PPD-positive									
Placebo			1			1			1
6H	0.21	(0.89,1.62)	1.23	0.19	(0.91-1.6)	1.21	0.23	(0.95,1.65)	1.26
3HR	0.14	(0.84,1.50)	1.15	0.12	(0.85,1.49)	1.13	0.18	(0.91,1.57)	1.20
3HRZ	-0.01	(0.74,1.33)	0.99	-0.02	(0.73,1.30)	0.98	0.02	(0.77,1.34)	1.02
Anergic									
Placebo			1			1			1
6H	-0.03	(0.71,1.32)	0.97	-0.02	(0.72,1.34)	0.98	0.0002	(0.74,1.35)	1.0002

Note: The time ratio (LR) is adjusted for age, body mass index, hemoglobin, absolute lymphocyte count, sex, and SGOT in the log-logistic AFT model.

Placebo is the reference group.

Table 5.14: The log-logistic AFT models for time to AIDS progression, death, and the combination of AIDS progression and death

is not monotonic. The estimated hazard increases from the first year to the second year, then decreases after two years. The Wald tests for the individual indicator variables show that the hazard of death in the fourth year is significantly lower than the first year. For the three interval piecewise exponential model, the likelihood ratio test shows a nonsignificant effect of interval. The pattern of the coefficients is not monotonic either. The estimated hazard increases from the first interval to the second interval, then decreases.

The estimates for other fixed covariates in two piecewise exponential models are virtually identical. But the likelihood ratio tests for the overall effect of interval are different. We have no confidence to support one piecewise exponential model, but they can give us some clue about the shape of the hazard function of survival time.

5.3.6 Conclusion

This study is based on a large number of participants from Uganda, where the prevalence of HIV infection and TB are very high. This study shows that the benefit of the TB preventive therapies to delay HIV disease progression to AIDS and death for HIV-infected adults is not confirmed, although they are effective in reducing the incidence of TB. Association of the TB preventive therapies with the AIDS progression is examined through the linkage

	Piecewise exponential model					
	Six intervals			Three intervals		
Variables	estimate	95%CI	P-value	estimate	95%CI	P-value
Intercept	-2.01	(-2.95,-1.06)	<.0001	-0.96	(-1.93,0.009)	0.05
PPD-positive						
Placebo						
6H	0.27	(-0.09,0.62)	0.14	0.26	(-0.10,0.61)	0.16
3HR	0.18	(-0.17,0.52)	0.32	0.17	(-0.18,0.51)	0.34
3HRZ	0.09	(-0.26,0.44)	0.61	0.04	(-0.31,0.39)	0.82
Anergic						
Placebo	-0.23	(-0.59,0.14)	0.22	-0.25	(-0.61,0.11)	0.18
6H	-0.19	(-0.54,0.15)	0.27	-0.20	(-0.55,0.14)	0.25
AGE	-0.02	(-0.03,-0.004)	0.01	-0.02	(-0.04,-0.005)	0.01
BMI						
19<BMI<=25						
BMI<=19	-0.64	(-0.89,-0.38)	<.0001	-0.71	(-0.97,-0.46)	<.0001
BMI>25	0.57	(0.23,0.92)	0.0012	0.62	(0.27,0.97)	0.0005
HGB	0.31	(0.26,0.37)	<.0001	0.24	(0.18,0.29)	<.0001
LYMPHABS	0.48	(0.32,0.63)	<.0001	0.53	(0.37,0.69)	<.0001
SEX	0.51	(0.28,0.75)	<.0001	0.40	(0.16,0.64)	0.001
SGOTX	-0.50	(-0.80,-0.20)	0.0011	-0.56	(-0.86,-0.26)	0.0002
j2	-0.12	(-0.36,0.12)	0.32	-0.21	(-0.44,0.02)	0.09
j3	-0.05	(-0.33,0.23)	0.73	-0.13	(-0.40,0.15)	0.36
j4	0.69	(0.16,1.23)	0.01			
j5	1.22	(-0.75,3.19)	0.23			

Table 5.15: Summary of the piecewise exponential models

of the signs and symptoms to replication of HIV virus.

A finding of the present study is the absence of protection of TB preventive therapies on AIDS progression, death and combined event of AIDS progression and death. This study presents similar estimates of risk for the covariates with the previous study with the baseline signs/ symptoms variables in the Cox PH model [39]. But the PH assumption does not hold for LYMPHABS in this analysis. To overcome this, time-dependent covariates are incorporated into the Cox model. We also use five different AFT models to fit the data. We find that the log-logistic AFT model fit better for this dataset. We provide the predicted hazard functions, predicted survival functions, median survival times and time ratios under the log-logistic AFT model.

In our study, age, anergic status, hemoglobin, body mass index, sex, SGOT and absolute lymphocytes count are significantly associated with the AIDS progression. The older person is prone to have shorter survival time and AIDS progression time. Men have longer survival time and AIDS progression time than women. The risks of AIDS progression, death and the combined event of AIDS progression and death are higher among anergic participants than among PPD-positive participants. HIV-infected participants with SGOT higher than 40 U/L have more rapid AIDS disease progression and mortality than those less than 40 U/L. The AIDS progression prolongs as hemoglobin increases. According to the Cox model with time-dependent variables, the predictive effect of absolute lymphocytes count clearly changes at about 2 years. Before 2 years, the hazard is less than one, which indicates that the risk of AIDS progression decreases as absolute lymphocyte count increases. After 2 years, the hazard is greater than one, which indicates that the risk of AIDS progression increases as absolute lymphocyte count increases. According to the log-logistic AFT model, LYMPHABS prolongs the time to AIDS progression as it increases.

The PH model is routinely applied to the analysis of survival data. The study considered here provides an example of a situation where PH model is inappropriate and where the AFT model provides a better description of the data. We have seen that the AFT model is a more valuable and realistic alternative to the PH model in some situations. Furthermore, the AFT model makes it possible for clinicians to interpret the treatment benefit in terms of an effect on expected duration of illness. To this content the AFT model may have explanatory advantage in that covariates have a direct effect on survival times rather on hazard functions as in the PH model.

Chapter 6

Discussion

The Cox PH model is the most widely used way of analyzing survival data in clinical research. In a review paper of survival analysis published in cancer journals [4], it was found that only five percent of all studies using the Cox PH models check PH assumption. However PH assumption is not always satisfied in the data. If this assumption does not hold there are various solutions to consider. One solution is to include the time-dependent variable for the predictors with non-proportional hazards. When this approach is used to account for a variable with non-proportionality, different results may be obtained from different choices of time-dependent variables. It is hard to choose between models. Alternatively we can use a model where we stratify on the non-proportional predictors. The stratified Cox model is not appropriate when the covariate with non-proportionality is continuous or of direct interests. And both ways are still based on comparison of hazards. The AFT model is an alternative method for the analysis of survival data even when hazards are not proportional. Based on asymptotic results, the AFT models should lead to more efficient parameter estimates than Cox model under certain circumferences [14], [45].

The Cox model expresses the multiplicative effect of covariates on the hazard. The AFT model provides an estimate of the median survival time ratios. The results from an AFT model are easier to interpret, more relevant to clinicians and provide a more appropriate description of survival data in many situations. The comparison of the Cox PH model and the AFT models are presented in Table 6.1.

In this thesis, we have analyzed the TB/HIV dataset using these alternative methods. This study provides an example of a situation where the PH model is inappropriate and where the AFT model provides a better description of the data. The PH assumption does not hold in this dataset. After fitting the Cox PH model, the goodness of fit of the model is assessed through residual plots. The PH model seems to display lack of fit in this example. In contrast, the AFT model provides an adequate description of the data. The family of

	Cox PH model	AFT model
Advantage	1) Widely used. 2) No assumption about the distribution for the survival time. 3) Survival curves can be estimated after adjusting for the explanatory variables. 4) Incorporation of time-dependent covariate is convenient using SAS software.	1) More informative. predicted hazard functions, predicted survival functions, median survival times and time ratios can be obtained. 2) The effect of covariate is to accelerate or delay the duration of illness by a constant amount (acceleration factor or time ratio). 3) The effect size is time ratio which is easier to interpret and more relevant to clinician.
Disadvantage	1) PH assumption must hold. 2) Effect size is hazard ratio which is less relevant to clinician.	1) Relatively unfamilar and rarely used. 2) AFT assumption must hold . 3) Need to specify the distribution of survival time, but an appropriate distribution may be difficult to indentify. 4) Incoporation of time-dependent covariate is not allowed using SAS software.

Table 6.1: Comparison of Cox PH model and AFT model

the AFT models containing the exponential AFT model, Weibull AFT model, log-logistic AFT model, log-normal AFT model, and gamma AFT model are applied to this dataset. We select the model that best describes the data. In addition, the example illustrates that the AFT model has a more realistic interpretation and provides more informative results as compared to PH model. Therefore, we suggest that using the Cox PH model may not be the optimum approach. The AFT model may provide an alternative method to fit some survival data.

One main disadvantage of using the AFT model is that the specific distribution of survival time is unknown in many cases. Further study of this data could attempt using a non-parametric version of the AFT model [57], which does not require the specification of the distribution can be applied in this dataset. The results from this model could then be compared with the standard AFT models and Cox PH models. In addition, further study can be carried out to evaluate the effects of practical cases such as large censoring.

Bibliography

[1] Ackah AA, Coulibaly D, Digbeu H, et al. Response to treatment, mortality, and CD4 lymphocyte counts in HIV-infected persons with tuberculosis in abidjan, cote d'ivoire. *Lancet 345* (1995), 607–610.

[2] Andersen, P. K., Borgan, O., Gaill, R. D., and Keiding, N. *Statistical Methods Based on Counting Processes*. Springer, New York, 1993.

[3] Anderson, P. K., Borgan, O., and Gill, R. D. Cox's regression model counting process: A large sample study. *Annals of Statistics 10* (1982), 1100–1120.

[4] Atman DG, De Stavola BL, Love SB, et al. Review of survival analysis published in cancer journals. *British Journal of Cancer 72* (1985), 511–518.

[5] Barlow, W. E., and Prentice, R. L. Residuals for relative risk regression. *Biometrika 75* (1988), 65–74.

[6] Berkson, J., and Gage, R. P. Calculation of survival rates for cancer. *Proceedings of the staff meetings of the Mayo clinic 25* (1950), 270–286.

[7] Breslow, N. E. Covariance analysis of censored survival data. *Biometrika 30* (1974), 89–100.

[8] Buckley, I. V., and James, I. Linear regression with censored data. *Biometrika* (1979), 429–436.

[9] Cain, K. C., and Lange, N. T. Approximate case influence for the proportional hazards regression model with censored data. *Biometrika 40* (1984), 493–499.

[10] Collett, D. *Modelling Survival Data in Medical Research*. Chapman and Hall, London, 2003.

[11] Cox, C. Delta method. In *Encyclopedia of Biostatistics* (1998), P. Armitage and T. Colton, Eds., vol. 2, pp. 1125–1127.

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

[13] Cox, D. R. Partial likelihood. *Biometrika 62* (1975), 269–276.

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

[15] Cox, D. R., and Snell, E. J. A general definition of residuals with discussion. *Journal of the Royal Statistical Society. Series B 30* (1968), 248–275.

[16] Crowder, J., and Hu, M. Covariance analysis of heart transplant survival data. *Journal of American Statistical Association 78* (1977), 27–36.

[17] Daley GL, Small PM, Schecter GF, et al. An outbreak of tuberculosis with accelerated progression among persons infected with the human immunodeficiency virus: an analysis using restriction-fragment-length polymorphism. *The New England Journal of Medicine 326* (1992), 231–235.

[18] Fleming, T. R., and Harrington, D. P. *Counting Processes and Survival Analysis*. Wiley, New York, 1991.

[19] Friedland, G., Churchyard, G. J., and Nardell, E. Tuberculosis and HIV coinfection: Current state of knowledge and research priorities. *Journal of Infectious Diseases 196* (2007), Suppl 1:S1–S3.

[20] Gehan, E. A. A generalized Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika 52* (1965), 203–223.

[21] Goletti D, Weissman D, Jackson RW, et al. Effect of mycobacterium tuberculosis on HIV replication. *J. Immunol. 157* (1996), 1271–1278.

[22] Gordin FM, Matts JP, Miller C, et al. A controlled trial of isoniazid in persons with anergy and human immunodefiency virus infection who are at high risk for tuberculosis. *The New England Journal of Medicine 337* (1997), 315–320.

[23] Greenwood, M. The natural duration of cancer. *Reports on Public Health and Medical Subjects 33* (1926), 1–26.

[24] Holford, T. R. Life tables with concomitant information. *Biometrika 32* (1976), 587–597.

[25] Holford, T. R. The analysis of rates and of survivorship using log-linear models. *Biometrika 36* (1980), 299–305.

[26] IBRAHIM, J. G., CHEN, M.-H., and SINHA, D. *Bayesian Survival Analysis*. Springer-Verlag, New York, 2001.

[27] Johansen, S. The product limit estimate as a maximum likelihood estimate. *Scandinavian Journal of Statistics 5* (1978), 195–199.

[28] Johnson JL, Okwera A, Hom DL, Mayanja H, Kityo CM, et al. Duration of efficacy of treatment of latent tuberculosis infection in HIV-infected adults. *AIDS. 15* (2001), 2137–2147.

[29] Joint United Nations Programme on HIV/AIDS (UNAIDS)/World Health Organization. *AIDS epidemic update: December 2007*. Geneva, 2007.

[30] JP, N., MC, R., and A., K. HIV-associated tuberculosis in developing countries: epidemiology and strategies for prevention. *Tubercle and Lung Disease 73* (1992), 311–321.

[31] Kalbfleisch, J. D., and Prentice, R. L. Marginal likelihoods based on Cox's regression and life model. *Biometrika 60* (1973), 267–278.

[32] Kalbfleisch, J. D., and Prentice, R. L. *The statistical Analysis of Failure Time Data*, 2nd ed. Wiley, New York, 2002.

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

[34] Klein, J. P., and Moeschberger, M. L. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, 1997.

[35] Klembaum, D. G. *Survival Analysis: A Self learning text*. Springer, New York, 1996.

[36] Laird, N., and Olivier, D. Covariance analysis of censored survival data using log-linear analysis technique. *Journal of American Statistical Association 76* (1981), 231–240.

[37] Lawless, J. F. *Statistical Models and Methods for Lifetime Data Analysis*. Wiley, New York, 1982.

[38] Lee, E. T., and Wang, J. W. *Statistical Methods for Survival Data Analysis*, 3rd ed. Wiley, New York, 2003.

[39] Lim, HJ., Okwera, A., Mayanja-Kizza, H., Ellner, JJ., Mugerwan, RD., and Whalen, CC. Effect of tuberculosis preventive therapy on HIV disease progression and survival in HIV-infected adults. *HIV Clinical Trials 7* (2006), 172–183.

[40] Mantel, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Report 50* (1966), 163–170.

[41] Maria A. Qurgley, Alwyn Mwinga, et al. Long-term of preventive therapy for tuberculosis in a cohort of HIV-infected zambian adults. *AIDS. 15* (2001), 215–222.

[42] Mocroft AJ, Johnson MA, Sabin CA, Lipman M, Elford J, et al. Staging system for clinical aids patients. *Lancet 346* (1995), 12–17.

[43] Moreno S, Baraia-Etxaburu J, Bouza E, et al. Risk for developing tuberculosis among anergic patients infected with HIV. *Annals of Internal Medicine 119* (1993), 194–198.

[44] Murray, J. F., and Gursed, d. HIV infection and tuberculosis. *Respiration 57* (1990), 210–220.

[45] Oakes, D. The asymtotic information in censored survival data. *Biometrika 64* (1977), 441–448.

[46] Palella, F. J., Delaney, K. M., Moorman, A. C., Loveless, M. O., Fuhrer, J., Satten, G. A., Aschman, D. J., Holmberg, S. D., and The HIV Outpatient Study Investigators. Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. *The New England Journal of Medicine 338*, 13 (1998), 853–860.

[47] Pape JW, Jean SS, Ho JL, et al. Effect of isoniazid prophylaxis on incidence of active tuberculosis and progression of HIV infection. *Lancet 342* (1993), 268–272.

[48] Schoenfeld, D. Partial residuals for the proportional hazards regression model. *Biometrika 69* (1982), 239–241.

[49] Selwyn PA, Hartel D, Lewis VA, et al. A prospective study of the risk of tuberculosis among intravenous drug users with human immunodeficiency virus infection. *The New England Journal of Medicine 320* (1989), 545–550.

[50] Selwyn PA, Sckell BM, Alcabes P, Friedland GH, Klein RS, Schoenbaum EE. High risk of active tuberculosis in HIV-infected drug users with cutaneous anergy. *JAMA 268* (1992), 504–509.

[51] The White House Office of National AIDS Policy. *HIV and TB: Tuberculosis.* Ottawa, 1995.

[52] The White House Office of National AIDS Policy. *Report on the presidential mission 30 on children orphaned by AIDS in sub-Saharan Africa: Findings and plan of action.* Washington, DC, 1999.

[53] Therneau, T. M., Grambsch, P. M., and Fleming, T. R. Martingale-based residuals for survival models. *Biometrika 77* (1990), 147–160.

[54] Toossi Z, Mayanja-Kizza H, Hirsch GS, et al. Impact of tuberculosis (tb) on HIV-1 activity in dually infected patients. *Glin. Exp. Immunol. 123* (2001), 233–238.

[55] Tuner BJ, Markson LE, Mckee L, et al. The aids-defining diagnosis and subsequent complications: a survival-based severity index. *Journal of Acquired Immune Deficiency Syndromes 4* (1991), 1059–1071.

[56] Wadhawen D, Hira S, Mwansa N, et al. Preventive tuberculosis chemotherapy with isoniazid among patients infected with HIV-1. *IX International Conference on AIDS, Berlin, June 1993*.

[57] Wei, L. J. The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Statistics in Medicine 11* (1992), 1871–1879.

[58] Whalen CC, Johnson JL, Okwera A, Hom DL, Huebner R, et al. A trial of three regimens to prevent tuberculosis in ugandan adults infected with the human immunodefficiency virus. *The New England Journal of Medicine 337* (1997), 801–808.

[59] World Health Organization (WHO). Acquired immune defficiency syndrome (aids): Interim proposal for a who staging system for HIV-1 infection and disease. *WHO Weekly Epidemiol Record 65* (1990), 221–228.

[60] World Health Organization (WHO). Preventive therapy against tuberculosis in people living with HIV. *WHO Weekly Epidemiol Record 74* (1999), 385–398.

[61] World Health Organization (WHO). *Global tuberculosis control: surveillance, planning, financing.* Geneva, 2007.