SURVIVAL MODELLING, MISSING VALUES AND FRAILTY WITH APPLICATION TO CERVICAL CANCER DATA

NURADHIATHY BINTI ABD RAZAK

INSTITUTE OF GRADUATE STUDIES
UNIVERSITY OF MALAYA
KUALA LUMPUR

2016

SURVIVAL MODELLING, MISSING VALUES AND FRAILTY WITH APPLICATION TO CERVICAL CANCER DATA

NURADHIATHY BINTI ABD RAZAK

THESIS SUBMITTED IN FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

INSTITUTE OF GRADUATE STUDIES
UNIVERSITY OF MALAYA
KUALA LUMPUR

2016

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: **NURADHIATHY BINTI ABD RAZAK**

Registration/Matric No: **HHC090004**

Name of Degree: **DOCTOR OF PHILOSOPHY (SCIENCE)**

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

"SURVIVAL MODELLING, MISSING VALUES AND FRAILTY WITH APPLICATION TO CERVICAL CANCER DATA"

Field of Study: **MEDICAL STATISTICS**

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

ABSTRACT

Data of cervical cancer patients treated in Hospital Universiti Sains Malaysia are analysed using the Cox proportional hazards regression analysis to model the prognostic factors. Since there is a non-proportional hazards covariate, the analysis is extended to the stratified Cox model. Also, parametric survival models including the Weibull, lognormal and log-logistic models are performed on the data. Among these parametric models, Weibull is the best. Then, a stratified Weibull model is performed because the proportional hazards assumption is violated. A comparison between the stratified Cox and stratified Weibull models shows that the stratified Cox model gives a better fit.

Commonly, a complete case analysis is considered when there are missing values in a data set. This approach may reduce the sample size and power of the study. The performance of several methods for handling missing values is studied including the Expectation-Maximization (EM) algorithm by method of weight, hot deck, multiple imputation by chained equation with predictive mean matching (MICE-PMM) and complete case analysis methods for the Weibull data. The values are assumed missing at random (MAR). Simulation studies are performed, and the cervical cancer data is used for illustration. Overall, the EM algorithm by method of weight performs well compared to other methods.

In survival data, there may exist unmeasured factors that also influence the survival and cause heterogeneity among individuals. This unobserved random effect is known as frailty. This study also focuses on the test for detecting frailty in a positive stable Gompertz model. The Zhu’s score test (Zhu, 1998), modified score test and ln *s* based test (Sarker, 2002) may also be derived from such a model. Thus, this study investigates the tests properties, and found that the modified score test performs better than the other tests based on the convergence rate and power of the test via simulation.

ABSTRAK

Data pesakit kanser serviks yang dirawat di Hospital Universiti Sains Malaysia dianalisa menggunakan analisis model regresi bahaya berkadaran Cox untuk membina model faktor-faktor prognostik. Memandangkan terdapat kovariat bahaya tidak berkadaran, analisis ini dipanjangkan kepada model Cox berstrata. Disamping itu, model kemandirian parametrik termasuk model-model Weibull, lognormal dan log-logistik dijalankan ke atas data. Antara model-model parametrik ini, Weibull adalah yang terbaik. Kemudian, model Weibull berstrata dijalankan kerana andaian bahaya berkadaran tidak dipatuhi. Perbandingan di antara model-model Cox berstrata dan Weibull berstrata menunjukkan bahawa model Cox berstrata memberi kesesuian yang lebih baik.

Lazimnya, analisis kes lengkap dipertimbangkan apabila terdapat nilai-nilai lenyap dalam sesuatu set data. Pendekatan ini mungkin mengurangkan saiz sampel dan kuasa kajian. Prestasi beberapa kaedah untuk mengendalikan nilai-nilai lenyap dikaji termasuk kaedah-kaedah algortima pemaksimuman jangkaan (EM) menggunakan pemberat, dek panas, imputasi berganda oleh persamaan berantai dengan padanan min ramalan (MICE-PMM) dan analisis kes lengkap untuk data Weibull. Nilai-nilai diandaikan lenyap secara rawak (MAR). Kajian simulasi dijalankan dan data kanser serviks digunakan untuk ilustrasi. Keseluruhannya, kaedah algoritma EM menggunakan pemberat menunjukkan prestasi yang baik berbanding kaedah-kaedah yang lain.

Dalam data kemandirian, mungkin wujud faktor-faktor tidak diukur yang juga mempengaruhi kemandirian dan menyebabkan keheterogenan antara individu-individu. Kesan rawak yang tidak dilihat ini dikenali sebagai *frailty*. Kajian ini juga fokus kepada ujian untuk mengesan *frailty* dalam model positif stabil Gompertz. Ujian skor Zhu (Zhu, 1998), ujian skor terubah suai, dan ujian berasaskan $\ln s$ (Sarker, 2002) mungkin boleh juga diterbitkan daripada model tersebut. Oleh itu, kajian ini mengkaji sifat ujian-ujian tersebut, dan mendapati bahawa ujian skor terubah suai lebih baik daripada ujian-ujian lain berdasarkan kadar penumpuan dan kuasa ujian melalui simulasi.

ACKNOWLEDGEMENTS

I am most grateful to Allah SWT, the Most Merciful and Most Compassionate, for His Blessings and Guidance throughout this study.

I would like to express my sincere gratitude to my supervisors, Associate Professor Dr. Yong Zulina Zubairi and Dr. Rossita Mohamad Yunus for their guidance, patience, support, advices, and invaluable suggestions throughout the course of my study at the Univeristy of Malaya. Without their supervision and help, this study would never have been possible.

I thank my beloved parents, Encik Abd Razak Abdullah and Puan Rusnani Rashid for their love, patience, encouragements, and advices. Thanks to my sister, Nurikmaliathy, my brother, Ahmad Lukhfhy, and other family members who have always supported me throughout the period.

I am grateful to the Ministry of Higher Education Malaysia and University of Malaya (Grant No.: PS003/2010A) for providing the scholarship and also financial support. Special thank to my great friends: Fauzia, Safwati, Shazlin, Siti Zanariah, Siti Fatimah, Anum, Norhafizah, Ruhana, Anis, and Khairany, who have never stop supported me from the beginning of this Ph.D journey and shared their invaluable knowledge. Thank you very much also to those who have directly or indirectly contributed and helped me all the way through completing this Ph.D study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

β	coefficient of regression
$h(t)$	hazard function
$H(t)$	cumulative hazard function
k-nn	*k*-nearest neighbour
HR	hazard ratio
$S(t)$	survival function
se	standard error
t	survival time
TR	time ratio
ADC	adenocarcinoma
AFT	accelerated failure time
AIC	Akaike information criterion
CI	confidence interval
CT	computed tomography
DA	data augmentation
DM	distant metastasis
EM	expectation-maximization
EMB	expectation-maximization algorithm with a bootstrap
FIGO	International Federation of Gynecology and Obstetrics
HPE	histopathological examination
HPV	Human Papillomavirus
HUSM	Hospital Universiti Sains Malaysia
LML	log-minus-log

MCAR	missing completely at random
MAE	mean absolute error
MAR	missing at random
MICE	multiple imputation by chained equation
MLE	maximum likelihood estimate
MNAR	missing not at random
PMM	predictive mean matching
PVF	positive variance function
RMSE	root mean squared error
SCC	squamous cell carcinoma
SCJ	squamocolumnar junction
WHO	World Health Organization

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

Survival analysis is a collection of statistical procedures for data analysis for which the outcome variable is survival time (Kleinbaum & Klein, 2005). Survival time is the time measured from a well-defined starting point until the occurrence of an event of interest such as time measured from diagnosis of cancer until death (see Figure 1.1).

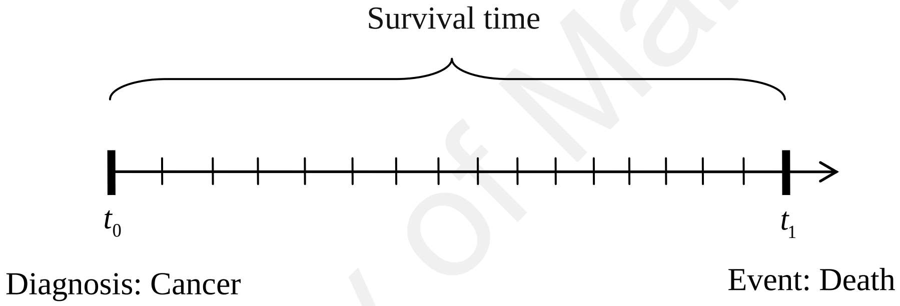


Figure 1.1: Diagram of time to event

One of the purposes of survival analysis is to estimate the survival function and hazard function. Furthermore, this analysis is useful for identifying factors that are significantly associated with the survival time or the risk of getting the event, which such factors are known as prognostic factors. Consequently, this analysis may provide valuable knowledge for medical practitioners such as the survival rate of patients, treatment progression, and factors contributing to either cure, recurrent or death after diagnosed with any particular disease. In addition, the findings are expected to help in the management of the disease such as in controlling and monitoring the prognostic factors in the population (Pruegsanusak *et al.*, 2012; Schneider *et al*., 2014).

There are three methods to analyse survival data, namely non-parametric, semi-parametric and parametric. Non-parametric analysis includes the Kaplan-Meier (or Product Limit) method that is used to estimate the survival probability. The survival difference between two or more groups is checked based on the log-rank test. Meanwhile, semi-parametric method that is the Cox proportional hazards regression analysis is often used to develop a prognostic model for any disease such as cancer (see Taib *et al*., 2008; Ghazali *et al*., 2010; El-Sherbieny *et al*., 2011; Wahidah *et al*., 2012; Schneider *et al*., 2014). When survival times follow any particular statistical distribution, parametric survival analysis is more suitable compared to the Cox proportional hazards regression analysis (see Wang *et al*., 2011; Zhu *et al*., 2011).

The Cox proportional hazards regression model is preferable than parametric survival models since less assumption is required. More assumptions need to be checked in order to apply parametric survival models such as identifying the appropriate statistical distribution for the data, checking the proportional hazards assumption, and accelerated failure time (AFT) assumption. Even Hosmer and Lemeshow (1999) highlighted that parametric survival models should be developed with caution. However, parametric survival analysis is more powerful (Lee & Wang, 2003) and may yield precise estimates (Klein & Moeschberger, 1997; Orbe *et al*., 2002) when a parametric model is chosen correctly. Study on the comparison of survival models has gained much attention in the past few years (see Sayehmiri *et al*., 2008; Ding *et al*. 2009; Aktürk Hayat *et al*., 2010; Grover *et al*., 2013). This type of study has been conducted using cancer data such as gastric cancer (Pourhoseingholi *et al*., 2007; Zhu *et al*., 2011), stomach cancer (Moghimi-Dehkordi *et al*., 2008), oral cancer (Köhler & Kowalski, 2012) and breast cancer (Pari Dayal *et al*., 2013).

In this study, data of cervical cancer patients treated in Hospital Universiti Sains Malaysia (HUSM) have been analysed. Patients who were diagnosed with cervical cancer between 1st July 1995 and 30th June 2007, and received at least one treatment for cervical cancer in HUSM were included in this study. The survival probability of these patients is estimated using the Kaplan-Meier method. The Cox proportional hazards regression analysis has been conducted to develop the prognostic model. As there is a non-proportional hazards covariate, the model has been extended to the stratified Cox model.

Also, this study is interested to identify a suitable parametric survival model for the aforementioned cervical cancer patients data. Therefore, the Weibull, log-logistic and lognormal models have been considered, and their performances are assessed based on the Akaike Information Criterion (AIC) statistic. The best parametric model is then compared to the stratified Cox model. In the comparison analysis, the non-proportional hazards covariate is also incorporated, and its importance has been emphasized. To our knowledge, parametric models have not been used extensively in the analysis of survival data in Malaysia. Furthermore, study on the comparison of survival models with non-proportional hazards covariate has not received enough attention yet.

Missing data are common to occur in many research studies. In medical studies, for instance, missing data are very difficult to be avoided especially when the studies involve retrieving information from any reported sources such as patients' medical record. Most standard statistical methods do not consider missing values in the analysis. Thus, the easiest option being applied by many researchers is to remove incomplete observations from the analysis. Such a method is known as a complete case analysis. However, this approach may reduce sample size, power of the study and also contribute to the loss of information. In addition, parameter estimates may be biased and inefficient especially when the amount of missing values is large (Barzi & Woodward, 2004). Therefore, treatment of missing values is necessary to avoid any devastating impact on the statistical inference especially due to the exclusion of subjects from the study.

Many referred literatures proposed techniques for handling missing covariate values in survival data. Jerez *et al*. (2006) applied mean substitution and hot deck methods on breast cancer survival data. Ibrahim and his co-workers (1998, 2001, 2004) developed Expectation-maximization (EM) type algorithm method for missing categorical and continuous covariate values for the Cox proportional hazards model. Meanwhile, Marshall *et al*. (2010a, 2010b) and Baade *et al*. (2015) applied various multiple imputation by chained equation (MICE) techniques for survival data in their studies.

The performance of missing data methods has been extensively investigated for the Cox proportional hazards regression model. However, little attention has been paid to missing data methods for the parametric survival model. Therefore, four methods namely the complete case analysis, EM algorithm by method of weight, hot deck imputation and multiple imputation by MICE-PMM have been considered in this study for handling missing covariate values in the parametric survival model. The findings of this study would be very beneficial given that the application of the parametric model on survival data has received much attention recently (see Köhler & Kowalski, 2012; Pari Dayal *et al*., 2013; Grover *et al*., 2013). The Weibull model is considered because it is the most common parametric model used for analysing survival data. These covariate values are assumed to be missing at random (MAR).

Most of the survival models assume that the hazard function is fixed, the individual's survival is independent of each other, and they have similar survival time distribution (Hosmer & Lemeshow, 1999). However, if there are other factors than covariates of interest which may also affect the survival and cause the population under study to be heterogeneous, the aforementioned assumptions may not be satisfied.

Heterogeneous population exists when there are individuals who are more likely to experience the event than other individuals in the group. For such data, the standard survival analysis is no longer appropriate to be applied because it may produce invalid inferences (Herring *et al*., 2002). A suitable survival model for this type of data is known as a frailty model or random effects model.

The frailty model considers heterogeneity caused by unmeasured or unobserved factors (Wienke, 2003). Vaupel *et al*. (1979) was the first who introduced the term frailty to describe an unobserved effect associated with each observation. Method for detecting frailty is very important since ignoring its effect may cause biased and inconsistent estimation (Andersen *et al*., 1999; Henderson & Oman, 1999). Commenges and Andersen (1995) derived a score test from the marginal partial likelihood of the Cox proportional hazards regression model. Meanwhile, Crowder and Kimber (1997) proposed a score test for the Weibull based model with gamma frailty. Zhu (1998) derived a Weibull based score test for a positive stable frailty that has infinite variance. Sarker (2002) extended Zhu's score test and proposed two new tests namely a modified score test and test based on $\ln s$.

In this study, tests for detecting frailty for a bivariate positive stable Gompertz model have been derived following the tests proposed by Zhu (1998) and Sarker (2002). Sarker (2002) pointed out that the null variances of the modified score test and the $\ln s$ based test would be different for the non-Weibull case with nuisance parameters. Thus, some modification on the variance estimations for the positive stable Gompertz model have been done in this study. Also, the properties of these tests and their performance have been evaluated based on the convergence rate and power of the tests.

1.2 Problem Statement

As the Cox proportional hazards regression model is very common in survival modelling, checking the proportional hazards assumption is very important. However, some studies tend to ignore this assumption, or if this assumption is violated, there is no proper treatment is considered. The Cox proportional hazards regression model is no longer valid to the model that violates the proportional hazards assumption. Thus, an extended Cox model is necessary to control for the non-proportional effect. In Malaysia, numerous studies have applied the Cox proportional hazards regression analysis in modelling prognostic factors. However, survival modelling with non-proportional hazards covariates has not been extensively applied especially using Malaysian data set. Hence, this study demonstrates the development of the stratified Cox model that considers the non-proportional hazards covariate using data of cancer patients in Malaysia.

Parametric survival models are less preferred than the Cox proportional hazards regression model because more assumptions are necessary to be verified in order to perform parametric survival analysis. However, when survival times fit a particular statistical distribution very well, that parametric model would give a powerful, precise and meaningful interpretation. At the same time, the model would be more informative than the Cox proportional hazards regression model. In Malaysia, survival modelling using parametric models is somewhat scarce. To our knowledge, there is no published study that discussed the performance of different types of survival model using Malaysian medical data, particularly cervical cancer data. Therefore, this study considers several parametric survival models and demonstrates the development of these models using data of cancer patients in Malaysia.

Many survival studies fail to address missing values problem in their studies and opt to exclude them from the analysis. Handling missing values is deemed necessary to avoid any misleading or devastating impact on the statistical inferences. There are numerous methods for handling missing values which have been developed. Many studies focused on investigating the performance of these methods for the Cox proportional hazards regression model since this model is commonly applied in survival analysis. Thus, several methods for handling missing values for parametric models are investigated in this study.

It is crucial to check the presence of frailty in a survival model since the effect may lead to bias parameter estimates when a standard survival analysis being applied. However, most developed frailty tests concentrated on a common survival model, in particular, the Cox and Weibull model. Frailty tests should be explored further for other types of distribution model. Therefore, this study derives frailty tests for a Gompertz distribution model and investigates the performance of the tests.

1.3 Objectives of the Study

The objectives of the study are as the following:

1. To develop a prognostic model using semi-parametric survival analysis for cervical cancer patients' data.
2. To propose a parametric survival model for cervical cancer patients' data.
3. To propose a feasible method of handling missing covariate values in a parametric survival model.
4. To derive a score test, modified score test and $\ln s$ based test for testing frailty in a positive stable Gompertz model and investigate the performance of the tests.

1.4 Scope and Limitation of the study

This study looks at survival modelling, missing values methods and frailty tests for survival data. The first part of this research work is to develop a prognostic model for cervical cancer patients' data using several types of survival model. The data set is obtained from the Hospital Universiti Sains Malaysia (HUSM). Patients who were diagnosed with cervical cancer between 1st July 1995 and 30th June 2007, and received at least one treatment related to cervical cancer in HUSM are included in the study. Non-parametric, semi-parametric and parametric survival analyses are performed on the data set. Also, the aim of this study is to propose a feasible method of handling missing covariate values in a parametric survival model. The Weibull AFT model with missing at random (MAR) categorical covariate values are considered. Simulation studies are performed to investigate the performance of the complete case analysis, EM algorithm by method of weight, hot deck imputation and multiple imputation by MICE-PMM for such a model. This study also investigates the frailty tests for the bivariate positive stable Gompertz model using the Zhu's score test (Zhu, 1998), modified score test and ln *s* based test (Sarker, 2002). The properties of these tests are studied. Simulation studies are conducted to investigate the performance of these tests based on the convergence rate and power of the study.

The main limitation of this study is that the results obtained are from a hospital-based data. It may not be the best model to describe the national database of cervical cancer patients in Malaysia. Nevertheless, the aim of this study is to demonstrate the proper analyses of survival data.

1.5 Significance of the Study

The findings from this study will be beneficial in the following ways:

1. This study may demonstrate a development of non-proportional hazards model using a real data set from a hospital in Malaysia.
2. This study may contribute to the knowledge of parametric survival analysis with an application to a real data set from a hospital in Malaysia.
3. This study may suggest an appropriate method for handling missing values in survival data which optimise the estimation of parameters when the model is the parametric model.
4. The research finding may help to determine the best method to detect frailty in a positive stable Gompertz model.

1.6 Organisation of the Thesis

Chapter 1 provides the background of the study, problem statement and outline of chapters in the thesis. Chapter 2 presents the literature review related to the study. Chapter 3 presents the Kaplan-Meier analysis, and the stratified Cox model for the data of cervical cancer patients treated in HUSM. Chapter 4 explores further the aforementioned data set using parametric models namely the Weibull, lognormal and log-logistic models. Besides, the best parametric model obtained is compared with the stratified Cox model that has been obtained in Chapter 3. Chapter 5 discusses on the performance of the complete case analysis, EM algorithm by method of weight, hot deck imputation and multiple imputation methods for handling missing covariate values in the parametric survival model. Chapter 6 focuses on the properties and performances of frailty tests for the bivariate positive stable Gompertz model. Chapter 7 gives a brief overview of the findings and suggests several further works that may be performed in the future.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Section 2.2 explains the background to survival analysis, the concept of censoring and also lists the survival functions that are used to describe the survival time data. Previous studies on survival and prognostic factors for cervical cancer are reviewed in Section 2.3. The non-parametric survival analysis is described in Section 2.4. Several types of survival regression model are described in Section 2.5, where the assumptions of each model and its verification methods are also given. Meanwhile, Section 2.6 describes missing data problems, missing values mechanisms and methods to handle missing values in survival data. The theoretical background of frailty models including the types of frailty model, frailty distributions and tests available for detecting frailty are given in Section 2.7.

2.2 Survival Analysis

Survival analysis is a collection of statistical procedures for data analysis for which the outcome variable is time that is measured from the origin until an event occurs (Kleinbaum & Klein, 2005). In conducting a survival study, the time origin, scale of measuring the time and event of interest must be clearly defined by the researcher (Cox & Oakes, 1984).

The time is defined as the duration (either in days, weeks, months or years) from the beginning of follow-up of an individual until the event of interest occurs. Also, it is known as survival time or failure time. Any designated experience of interest that may happen to an individual is known as event (Kleinbaum & Klein, 2005). This event may be a development of a disease, response to a treatment, relapse or death. Thus, survival time may be tumour-free time, the time from the start of treatment response, and time to death (Lee & Wang, 2003).

2.2.1 Censoring

A unique feature of survival analysis is its ability for handling censored observations. Censoring occurs when the end-point of interest has not been observed due to some causes. There are three conditions that may cause censoring to occur; (i) a subject does not experience the event by the time of the closure of the study, (ii) lost to follow-up during the study period or (iii) withdraws from the study due to some other reasons (such as adverse drug reaction or other competing risk) that makes further follow-up impossible (Clark *et al.*, 2003a; Kleinbaum & Klein, 2005). All these may cause the actual survival time of those individuals remain unknown.

There are several types of censoring such as right censoring, left censoring and interval censoring. This study only focuses on right-censored data since it is the most common type of censoring to occur. Right censoring occurs when a person's exact survival time is unknown at the right side of the follow-up period (Kleinbaum & Klein, 2005). Figure 2.1 illustrates the mechanism of right-censoring for eight individuals where "●" represents the time entering the study. Patients 1, 3, 4 and 6 experience the event of interest (×) within the study duration. Meanwhile, there are four censored observations (○), where patients 2 and 7 are lost to follow-up while 5 and 8 are alive at the end of study period.

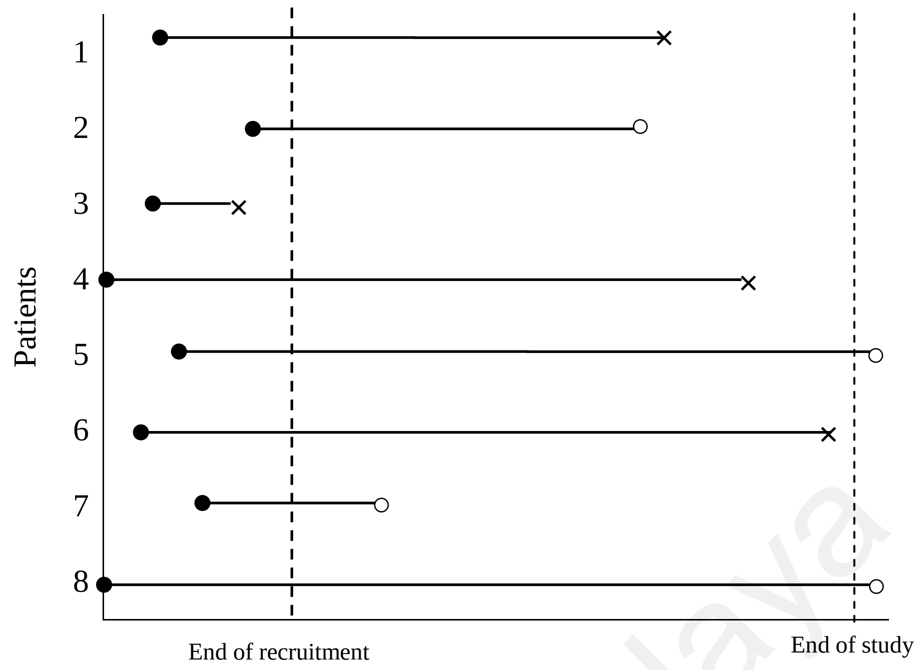


Figure 2.1: Study time for eight patients in a survival study

2.2.2 Survival Time Functions

The distribution of survival time may be characterised by three main functions, namely the probability density function, survival function, and hazard function. Let a non-negative continuous random variable T denotes the survival time that is measured from the time origin to an event of interest. Suppose that T has a density function that is given by

$$f\left(t\right)=\lim_{\Delta t\to 0}\left\{\frac{P\left(t\leq T<t+\Delta t\right)}{\Delta t}\right\}. \qquad (2.1)$$

The cumulative distribution function $F\left(t\right)$ is then given by

$$F\left(t\right)=P\left(T\leq t\right)=\int_{0}^{t}f\left(u\right)du, \qquad (2.2)$$

which gives the probability that the survival time is less than some value t (Collet, 2003). The survival function $S\left(t\right)$ may be obtained from the cumulative distribution function in (2.2) and is given by

$$S(t) = P(T > t) = 1 - F(t). \tag{2.3}$$

The survival function in (2.3) represents the probability of an individual survives longer than t.

Another important quantity that is commonly used to describe the risk or hazard of death at some time t is the hazard function $h(t)$. This function is defined as the probability that an individual get an event in an interval of t to $t + \Delta t$ given that the individual has survived up to that time t (Lee & Wang, 2003). This function may be written as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t}. \tag{2.4}$$

The survival function $S(t)$ may be derived if the hazard function $h(t)$ is known and vice versa. The relationship between these two functions may be expressed as follows

$$S(t) = \exp\{-H(t)\}, \tag{2.5}$$

where $H(t)$ is the cumulative hazard function that is given by

$$H(t) = \int_0^t h(u)\,du. \tag{2.6}$$

Therefore, the hazard function may also be written in term of a derivative involving $S(t)$ by the following expression

$$h(t) = -\frac{dS(t)/dt}{S(t)}. \tag{2.7}$$

2.3 Cervical Cancer

Cervical cancer occurs when cancer cell (malignant neoplasm) has developed at the cervix that is in the lower part of the uterus. One of the major causes of this abnormality growth is the Human papillomavirus (HPV) infection. This virus is predominantly transmitted through sexual intercourse. In addition to the infection of HPV, among other risk factors of cervical cancer are early age at first intercourse, multiple sexual partners, multiparity and smoking.

The cervix is divided into two parts; ectocervix and endocervix. The ectocervix is lined with stratified nonkeratinizing squamous epithelium while endocervix with a single layer of columnar epithelium. Squamocolumnar junction (SCJ) is a point at which squamous ectocervical epithelium and columnar endocervical epithelium met. Meanwhile, the transformation zone is the area of the transition point of the columnar cells into squamous cells of the ectocervix (see Figure 2.2). This area is the most vulnerable to the HPV infection, and the place where the normal cell change to pre-cancer (dysplasia) (Escobar *et al*., 2007). Thus, most of the cancer cell arise at this zone (Kavanagh *et al*., 2006).

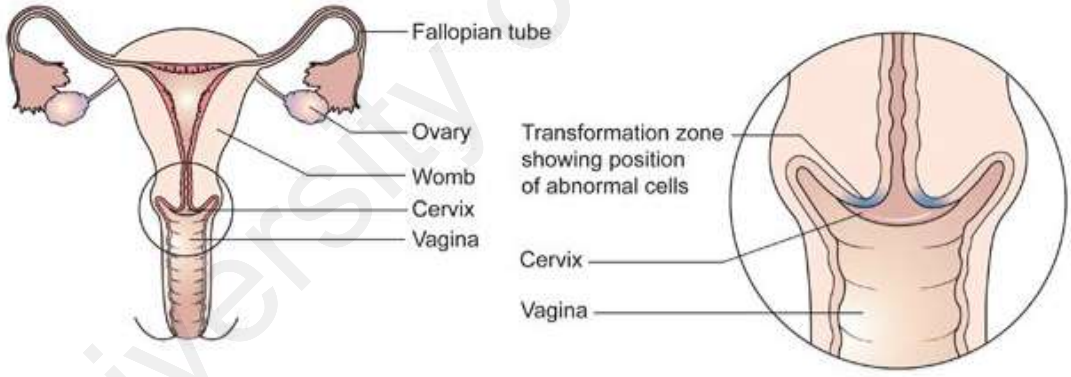


Figure 2.2: Diagram of transformation zone on the cervix
(Source: Cancer Research UK, 2014)

2.3.1 Stage of Cancer

The International Federation of Gynecology and Obstetrics (FIGO) system is used in staging the cervical cancer. The main consideration is the size of the tumour or its extension to pelvis and adjacent organs. Table 2.1 shows the classification of the disease and its descriptions.

Table 2.1: FIGO staging for cervical cancer (Source: Waggoner, 2003)

Stage	**Description**
Stage 0:	Carcinoma in situ, cervical intraepithelial neoplasis Grade III.
Stage I:	The carcinoma is strictly confined to the cervix (Extension to the corpus would be disregarded).
Ia	Invasive carcinoma which can be diagnosed only by microscopy. All macroscopically visible lesions – even with superficial invasion – are allotted to Stage Ib carcinomas. Invasion is limited to a measured stromal invasion with a maximal depth of 5.0 mm and a horizontal extension of not >7.0 mm Depth of invasion should not be >5.0 mm taken from the base of the epithelium of the original tissue should not change the stage allotment.
Ia1	Measured stromal invasion of not >3.0 mm in depth and extension of not >7.0 mm.
Ia2	Measured stromal invasion of >3.0 mm and not >5.0 mm with an extension of not >7.0 mm.
Ib	Clinically visible lesions limited to the cervix uteri or preclinical cancers greater than Stage Ia.
Ib1 Ib2	Clinically visible lesions not >4.0 cm. Clinically visible lesions >4.0 cm.
Stage II:	Cervical carcinoma invades beyond uterus but not to the pelvic wall or the lower third of the vagina.
IIa IIb	No obvious parametrial involvement. Obvious parametrial involvement.
Stage III:	The carcinoma has extended to the pelvic wall. On rectal examination, there is no cancer-free space between the tumour and the pelvic wall. The tumour involves the lower third of the vagina. All cases with hydronephrosis or nonfunctioning kidney are included unless they are known to be due to other causes.
IIIa	Tumour involves lower third of the vagina, with no extension to the pelvic wall.
IIIb	Extension to the pelvic wall and/or hydronephrosis or nonfunctioning kidney.
Stage IV:	The carcinoma has extended beyond the true pelvis or has involved (biopsy proven) the mucosa of the bladder or rectum. A bullous edema, as such, does not permit a case to be allotted to Stage IV.
IVa IVb	Spread of the growth to adjacent organs. Spread to distant organs.

2.3.2 Treatment

There are several factors that influence the choice of treatment such as the age, general condition of the patient, the stage of the tumour, and patient’s own preference (Radstone & Kunkler, 2003). At early stage of cancer, patients will be treated by radical hysterectomy and pelvic lymphadenectomy or alternatively combined external pelvic irradiation and brachytherapy with concomitant chemotherapy (Jensen *et al*., 2007). Meanwhile, patients with more advanced cancer will be given a combination radiotherapy and concomitant chemotherapy. Waggoner (2003) provided the types of treatment for cervical cancer patients according to the stage of cancer as given in Table 2.2.

Table 2.2: Treatment algorithm for cervical cancer (Source: Waggoner, 2003)

Stage	Clinical features	Treatment
IA1	Invasion 3·0 mm or less	If patient desires fertility, conisation of cervix. If she does not, simple hysterectomy (abdominal or vaginal).
IA2	With lymphatic space invasion	Hysterectomy with or without pelvic lymphadenectomy.
IB1	3·0-5·0 mm invasion, <7·0 mm lateral spread	Radical hysterectomy with pelvic lymphadenectomy. Radiotherapy.
IB2	Tumour 4 cm or less	Radical hysterectomy with pelvic lymphadenectomy plus chemoradiotherapy for poor prognostic surgical-pathological factors*. Radiotherapy.
IIA	Tumour bigger than 4 cm	Radical hysterectomy with pelvic lymphadenectomy plus chemoradiotherapy for poor prognostic surgical and pathological factors*. Chemoradiotherapy. Chemoradiotherapy plus adjuvant hysterectomy.
IIB	Upper-two-thirds vaginal involvement	Radical hysterectomy with pelvic lymphadenectomy. Chemoradiotherapy.
IIIA	With parametrial extension Lower-third vaginal involvement	Chemoradiotherapy. Chemoradiotherapy.
IVA	Local extension within pelvis	Chemoradiotherapy. Primary pelvic exenteration.
IVB	Distant metastases	Palliative chemotherapy. Chemoradiotherapy.

*Pelvic lymph-node metastases; large tumour; deep cervical stromal invasion; lymphovascular space invasion; positive vaginal or parametrial margins.

2.3.3 Histologic Type

World Health Organization (WHO) has classified cervical carcinoma into three main histological types: squamous cell carcinoma, adenocarcinoma and other epithelial tumours (Cheah & Looi, 1999). Two most common histologic types are squamous cell carcinoma and adenocarcinoma. The development of these two histologic types is highly associated with the infection of high-risk type of HPV (Walboomers *et al*., 1999). Squamous cell carcinoma develops from flat cells that cover the outer surface of the cervix. Meanwhile, adenocarcinoma is endocervical cancers originating from glandular epithelium. Other examples of histologic type that may be found are adenosquamous carcinoma, glassy cell carcinoma, adenoid basal carcinoma, and adenoid cystic carcinoma.

2.3.4 Metastasis

Cervical cancer may spread through direct local extension and also lymphatic. The cancer spreads directly to the vaginal mucosa, endometrial cavity, parametrial tissues and ligaments, pelvic side wall, bladder, and rectum (Moore-Higgs & Chafe, 2001). Also, pelvic and para-aortic lymph node metastases are one of the most significant prognostic factors of cervical cancer (Ho *et al*., 2004). In the worst case, patient may experience distant metastasis when the cancer spreads to any distant organs such as bladder, bones or lungs.

2.3.5 Survival and Prognostic Factors of Cervical Cancer

The survival of cervical cancer patients may vary by country. The five-year survival of cervical cancer in developed countries such as United States of America, Germany and Spain were higher than 60% (American Cancer Society, 2011). Meanwhile, the five-year survival exceeded 70% in Korea (Ahn *et al*., 2011; Shin *et al*., 2011; Woo *et al*., 2011), and 55% in Turkey (Eser, 2011). Flores-Luna *et al*. (2001) studied the survival of Mexican women and found that the overall five-year survival was 66.6%. In Asian countries like China (Xiang *et al*., 2011) and Thailand (Sumitsawan *et al*., 2011), the five-year survival exceeded 50%. Pomros *et al*. (2007) had done a study on cervical cancer patients treated with radiation therapy in

Srinagarind Hospital in Thailand and found that the five-year survival was 62.5%. Meanwhile, the five-year survival in least developed countries such as Gambia and Uganda was remarkably low, which was less than 25% (Sankaranarayanan *et al*., 2011).

Many studies reported the survival of cervical cancer patients based on the stage of the cancer. In general, the overall five-year survival nearly approaches 100% for patients diagnosed at stage IA and drops remarkably to almost 20% for stage IVB (Kyrgiou & Shafi, 2010). In Korea, a study found that the relative five-year survival rate according to stage were 94.2%, 69.7%, 38.9% and 21.1% for stage I, II, III and IV, respectively (Chung *et al*., 2006). Meanwhile, a hospital-based study in Indonesia obtained lower five-year survival, where for stage I was 50%, stage II was 40%, stage III was 20% and stage IV was 0% (Aziz, 2009).

Numerous studies have been done to determine factors affecting the survival of cervical cancer patients (Pomros *et al*., 2007; Ho *et al*., 2011; Seamon *et al*., 2011). There were various prognostic factors identified such as stage at diagnosis, age at diagnosis, lymph node involvement and tumour size (Brun *et al*., 2003; Acs & Gombos, 2006; Atahan *et al*., 2007; Dueňas-González *et al*., 2012). In most referred literatures, stage at diagnosis was frequently found as one of the significant factors affecting the prognosis of cervical cancer patients (see Grigienė *et al*., 2007; Zarchi *et al*., 2010; Katanyoo *et al*., 2012).

In Korea, a study of 44,182 patients who were diagnosed with cervical cancer between 1993 and 2002 found that stage at diagnosis and histologic type were important prognostic factors (Chung *et al*., 2006). Based on 479 surgical specimens obtained from radical abdominal hysterectomy, the statistical ranking of the significant prognostic factors were lymph node metastases, size of lymph node metastases, tumour volume, parametrial involvement and lymphatic space invasion (Pickel *et al*., 1997). Also, a study of 381 cervical cancer patients in Kentucky found that stage of cancer was the significant prognostic factor for overall survival (Seamon *et al*., 2011).

Besides tumour diameter and pelvic lymph node enlargement, Endo *et al*. (2015) also found that distant metastasis was significantly associated with poor outcomes in patients with cervical cancer in their study. Meanwhile, Xia *et al*. (2014) identified that parametrial invasion and pelvic node metastasis were the significant factors affecting the outcome of patients treated by radical hysterectomy and pelvic lymphadenectomy. There were also studies that looked at the survival differences based on ethnicity and races (see Patel *et al*., 2005; Bates *et al*; 2008; Redaniel *et al*., 2009; Coker *et al*., 2009; Priest *et al*., 2010). A population-based study in Singapore found that Malays had 33% higher excess hazard of death compared to Chinese in patients with localized cervical cancer (Wang *et al*., 2003).

2.4 Non-parametric Analysis

Survival function (also known as survival probability) denoted as $S(t)$ is defined as the probability that an individual survives from the time of origin to a specified future time t (Clark *et al*., 2003a). Kaplan-Meier (or product limit) method is used to estimate the survival probability at a given time t and provides graphical presentation of the survival distribution.

Suppose that there are *n* individuals with observed survival times $t_1, t_2, \ldots, t_n$ and there are *r* times of death amongst the individuals. The *r* rank ordered death times are $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$, where $t_{(j)}$ is the *j*th death time ($j = 1, 2, \ldots, r$). The estimated survival function at any time *t*, in the *k*th time interval from $t_{(k)}$ to $t_{(k+1)}$ ($k = 1, 2, \ldots, r$), is the estimated probability of surviving beyond $t_{(k)}$. This is the probability of surviving through the interval from $t_{(k)}$ to $t_{(k+1)}$, and all preceding intervals (Collet, 2003), which is obtained by the Kaplan-Meier estimate of the survival function from the equation

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right), \tag{2.8}$$

where n_j represents the number of individuals who are alive just before time $t_{(j)}$ and d_j denotes the observed number of deaths at $t_{(j)}$. The $100(1-\alpha)\%$ confidence interval for $S(t)$ is given by

$$\hat{S}(t) \pm z_{\alpha/2} se\left\{ \hat{S}(t) \right\}, \tag{2.9}$$

where α is the level of significance and $se\left\{ \hat{S}(t) \right\}$ is the standard error of the Kaplan-Meier estimate of the survival function obtained from Greenwood's formula that is given by

$$se\left\{\hat{S}(t)\right\} \approx \hat{S}(t)\left\{\sum_{j=1}^{k}\left(\frac{d_j}{n_j\left(n_j - d_j\right)}\right)\right\}^{\frac{1}{2}}. \tag{2.10}$$

The median survival time where the time beyond which 50% of the individuals in the population under study is expected to survive is given by (Collet, 2003)

$$\hat{t}(50) = \min\left\{t_{(j)} \mid \hat{S}\left(t_{(j)}\right) < 0.50\right\}. \tag{2.11}$$

The $100(1-\alpha)\%$ confidence interval of median survival time is given by

$$\hat{t}(50) \pm z_{\alpha/2}se\left\{\hat{t}(50)\right\}, \tag{2.12}$$

where

$$se\left\{\hat{t}(50)\right\} = \frac{1}{\hat{f}\left\{\hat{t}(50)\right\}}se\left[\hat{S}\left\{\hat{t}(50)\right\}\right], \tag{2.13}$$

$se\left[\hat{S}\left\{\hat{t}(50)\right\}\right]$ is obtained from Greenwood's formula in (2.10), and $\hat{f}\left\{\hat{t}(50)\right\}$ is the estimate of the probability function of the survival times at $t(50)$.

The plot of the Kaplan-Meier estimate of the survival function can be used to compare survival functions between two or more groups (Lee & Wang, 2003). If the curves are clearly separated, there is a possible difference in survival between two or more groups. However, if the survival pattern is similar or the curves cross each other, this plot suggests that there may be no difference in survival (Kleinbaum & Klein, 2005). Meanwhile, the log-rank test is used to check whether the Kaplan-Meier curves for two or more groups are statistically different (Kleinbaum & Klein, 2005). In this test, the values of observed (O_i) and expected (E_i) number of events are calculated for each independent variable group. The log-rank statistic is obtained by the following approximate formula

$$\chi^2 \approx \sum_{i}^{\text{no.of groups}} \frac{\left(O_i - E_i\right)^2}{E_i}. \tag{2.14}$$

The *p*-value given by comparing the log-rank statistic to the chi-square distribution with $G-1$ degree of freedom (where G is the number of groups) is an evidence to suggest a difference in survival between groups (Kleinbaum & Klein, 2005).

2.5 Survival Regression Model

There are two common approaches used in modelling the relationship between a set of explanatory variables of interest and survival experience. The first type of survival model is a proportional hazards model that is mainly used for modelling the risk or hazard whilst the second model is an accelerated failure time (AFT) model that is used to model the survival time. The difference between these two models is how the covariates act that is either multiplicatively on the hazard functions for the proportional hazards model or survival time for the AFT model.

2.5.1 Proportional Hazards Model

The most widely used model in survival analysis studies is the proportional hazards model. The hazards function of the proportional hazards model is given by

$$h(t)=h_0(t)\exp\left(\beta_1 x_1+\ldots+\beta_p x_p\right), \tag{2.15}$$

where t is the survival time and $h_0(t)$ is the baseline hazard function that is the hazard for a value of 0 of the covariates (Hougaard, 2000). Meanwhile, $\beta_1,\beta_1,\ldots,\beta_p$ are the regression coefficients of p covariates $x_1,x_2,\ldots,x_p$. The hazard function $h(t)$ must always be positive. The expression in (2.15) indicates that the effect of explanatory variables is multiplicative with respect to hazard. The baseline hazard function $h_0(t)$ involves t thus this function describes the hazard function changes as a function of t. Meanwhile, the second term in (2.15) that is $\exp\left(\beta_1 x_1+\ldots+\beta_p x_p\right)$ involves x which describes how the hazard function changes as a function of covariates (Hosmer & Lemeshow, 1999).

The assumption of the proportional hazards model is the hazard for one subject is proportional to the hazard of any other subject, and the proportionality is independent of time (Kleinbaum & Klein, 2005). Equivalently, the hazard ratio of any two individuals is assumed to be a time-independent constant (Lee & Wang, 2003). This assumption implies that the true survival functions of these two individuals are parallel (Collet, 2003).

Hazard ratio (*HR*) is defined as the ratio of the hazards function for two individuals, let say i and i', with covariate values denoted x_i and $x_{i'}$, respectively, which is computed as follows:

$$\begin{aligned} HR &= \frac{h_0(t)\exp(\beta x_i)}{h_0(t)\exp(\beta x_{i'})} \\ &= \frac{\exp(\beta x_i)}{\exp(\beta x_{i'})}. \end{aligned} \tag{2.16}$$

Suppose that $x_i = 1$ and $x_{i'} = 0$, then the hazard ratio for the subject i, relative to the subject i' is given by

$$\begin{aligned} HR &= \frac{\exp(\beta(1))}{\exp(\beta(0))} \\ &= \exp(\beta(1) - \beta(0)) \\ &= \exp(\beta). \end{aligned} \tag{2.17}$$

Given that the event of interest is defined as death due to a particular disease, the hazard ratio gives the risk or hazard of death at time *t*. When $HR > 1$, the hazard of death is greater for subject i, relative to the subject i'. Meanwhile, the hazard of death is smaller for subject i, relative to the subject i' if $HR < 1$. The interpretation of this quantity is meaningful and understandable especially for medical practitioners such as for describing the prognosis of patients. The $100(1-\alpha)\%$ confidence interval of the estimated hazard ratio (HR) is obtained from the equation

$$\exp\left[\hat{\beta}_j \pm z_{\alpha/2} se\left(\hat{\beta}_j\right)\right], \tag{2.18}$$

where $se\left(\hat{\beta}_j\right)$ is the estimated standard error of the coefficient.

2.5.2 Accelerated Failure Time Model

The other approach for modelling the effect of explanatory variables is through the accelerated failure time (AFT) model. This model assumes that the effect of covariates is multiplicative with respect to the time scale (Kleinbaum & Klein, 2005). The standard way to express the AFT model is in a log-linear form which is given by

$$\log T_i = \mu + \alpha_1 x_{1i} + \cdots + \alpha_p x_{pi} + \sigma\varepsilon_i, \tag{2.19}$$

where $\alpha_1, \ldots, \alpha_p$ are the unknown regression coefficients of the p explanatory variables $x_1, \ldots, x_p$, μ is the intercept, σ is the scale, and ε_i is a random error following a particular probability distribution (Collet, 2003). The hazard function of the AFT model is given by

$$h(t) = \exp(-\boldsymbol{\alpha}'\boldsymbol{x}) h_0\left(\exp(-\boldsymbol{\alpha}'\boldsymbol{x})t\right), \tag{2.20}$$

where $\boldsymbol{\alpha}'\boldsymbol{x} = \alpha_1 x_1 + \ldots + \alpha_p x_p$.

Another way to present the AFT model is through the relationship between the survival function of an individual $S_1(t)$ relative to the baseline survival function $S_0(t)$, which is given by

$$S_1(t) = S_0\left(\exp(-\alpha)t\right), \tag{2.21}$$

where $\exp(-\alpha)$ is called an acceleration factor. The effect of this quantity is either “accelerate” or “decelerate” the survival time (Hosmer & Lemeshow, 1999), where

$\exp(-\alpha) > 1$ speeds up the time to get an event and $\exp(-\alpha) < 1$ slows down the time to get the event.

The acceleration factor is also known as a time ratio (TR). Time ratio quantity is used for interpreting the AFT model like the hazard ratio in the proportional hazards model. The effect of the covariates on the survival time is clear and easy to understand as well as clinically meaningful (Hosmer & Lemeshow, 1999). The $100(1-\alpha)\%$ confidence interval for the time ratio is calculated by the following formula (Hosmer & Lemeshow, 1999)

$$\exp\left[\hat{\alpha}_j \pm z_{\alpha/2} se\left(\hat{\alpha}_j\right)\right], \tag{2.22}$$

where $se\left(\hat{\alpha}_j\right)$ denotes the estimated standard error of the coefficient $\hat{\alpha}_j$.

2.5.3 Semi-parametric Model

The proportional hazards model was first introduced by Cox (1972) that is known as the Cox proportional hazards regression model. This model is the most common model used in the analysis of survival data. This model is also known as a semi-parametric model because no particular form of the probability distribution is assumed for the survival time. Similarly, there is no assumption has been made on the actual form of the baseline hazard function $h_0(t)$ thus the model is more flexible and applicable. The hazard function of the Cox proportional hazards regression model is similar to (2.15) that is

$$h(t) = h_0(t)\exp\left(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right), \tag{2.23}$$

where the hazard function $h(t)$ is dependent on a set of p covariates $\left(x_1, x_2, .., x_p\right)$, where the impact is measured by the size of the respective coefficients $\left(\beta_1, \beta_2, \ldots, \beta_p\right)$. This expression may be interpreted as a unit increase in variable x_p corresponds to multiplication of the baseline hazard function $h_0(t)$ by the other factor $\exp(\beta_p)$ given that other covariates in the model are kept fixed (McShane & Simon, 2001).

In the Cox proportional hazards regression model, β-parameters are estimated by maximising the partial log-likelihood function. Of the n observed survival times, suppose that there are D uncensored times. Let $t_1 < t_2 < \ldots < t_D$ denote the ordered D distinct event times (no ties between the event times). The partial likelihood function of the Cox proportional hazards model may be written as

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp\left[\sum_{j=1}^{p} \beta_j x_{(i)j}\right]}{\sum_{l \in R(t_i)} \exp\left[\sum_{j=1}^{p} \beta_j x_{lj}\right]} \tag{2.24}$$

where $x_{(i)j}$ be the *j*th covariate associated with the individual whose failure time is t_i. Meanwhile, $R(t_i)$ consists of all individuals whose survival times are at least t_i.

It is called the partial likelihood function since the likelihood does not consider probabilities of all subjects. The likelihood function takes into account probabilities only for those patients who get the event, and does not explicitly consider probabilities for those patients who are censored (Kleinbaum & Klein, 2005). Commonly, the Newton-Raphson iterative method is used to obtain the maximum likelihood estimates of β-parameters. The $100(1-\alpha)\%$ confidence interval for the estimated $\hat{\beta}_{(j)}$ is given by

$$\hat{\beta}_{(j)} \pm z_{\alpha/2} se\left(\hat{\beta}_j\right). \tag{2.25}$$

where $se\left(\hat{\beta}_{(j)}\right)$ is the estimated standard error of the coefficients.

2.5.3.1 Proportional Hazards Assumption

In order to apply the Cox proportional hazards regression analysis, the proportional hazards assumption must be checked and fulfilled. The proportional hazards assumption implies that the hazard function of one individual is proportional to the hazard function of the other individual (Kleinbaum & Klein, 2005). In addition, the hazard curves for the groups should be proportional and parallel (Bradburn *et al*., 2003a). The proportional hazards assumption of the fitted model may be assessed based on:

(i) The scaled Schoenfeld residuals and global test.

The proportional hazards assumption for individual variable is tested based on the scaled Schoenfeld residuals. Meanwhile, the global Schoenfeld residuals test is used to assess for the overall model. The proportional hazards assumption is violated if the tests are statistically significant ($p\text{-value} < 0.05$). From Collet (2003), the Schoenfeld residual denotes r_{Pji} may be obtained as follows:

$$r_{Pji} = \delta_i \left\{ x_{ji} - \hat{a}_{ji} \right\}, \tag{2.26}$$

where x_{ji} is the value of the *j*th explanatory variable, $j = 1, \ldots, p$ for the *i*th subject in the study,

$$\hat{a}_{ji} = \frac{\sum_{l \in R(t_i)} x_{jl} \exp\left(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l\right)}{\sum_{l \in R(t_i)} \exp\left(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l\right)}, \tag{2.27}$$

and $R(t_i)$ is the set of all individuals at risk at the time t. Meanwhile, the scaled Schoenfeld residuals r^*_{Pji} are the components of the vector

$$\boldsymbol{r}^*_{Pi} = r \operatorname{var}\left(\hat{\boldsymbol{\beta}}\right) \boldsymbol{r}_{Pi}, \tag{2.28}$$

where r is the number of deaths among the n individuals, $\text{var}(\hat{\boldsymbol{\beta}})$ is the variance-covariance matrix of the parameter estimates, and $\boldsymbol{r}_{Pi} = (r_{P_{1i}}, r_{P_{2i}}, \ldots, r_{P_{pi}})'$ is the vector of Schoenfeld residuals for the *i*th subject.

(ii) The log-cumulative hazard plot

For categorical variables, the log-cumulative hazard functions against the logarithm of survival time plots (LML plot) are constructed and examined. The curves on the plot should not cross for the proportional hazards assumption to hold.

There are several options may be considered when the proportional hazards assumption is not satisfied. One may choose to model the non-proportionality by stratified model, partition the time axis, or time-dependent covariates model (Therneau & Grambsch, 2000).

2.5.3.2 Stratified Cox Model

In this study, the stratified Cox model is used to handle nonproportional hazards covariates. This model is the simplest option (Therneau & Grambsch, 2000) to model prognostic factors when there is categorical covariate that does not satisfy the proportional hazards assumption (Hosmer & Lemeshow, 1999). Other covariates that meet the proportional hazards assumption are included in the model, while the covariate that violates the assumption is controlled by stratification (Kleinbaum & Klein, 2005).

Let x_j be a non-proportional hazards covariate with s categories. Thus, subjects are stratified into s stratum where the hazard function for stratum s is given by

$$h_s(t) = h_{0s}(t)\exp\left(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right). \tag{2.29}$$

Meanwhile, other covariates are assumed to have proportional hazards within each level of stratification factor (Royston & Lambert, 2011). In this model, the baseline hazard is assumed to be different across levels of stratification variables. Thus, the fitted stratified model will yield different estimated survival curves for each stratum. However, the coefficients are constrained to be the same across strata.

The stratified model can be described as a no-interaction model where the model assumes that there is no interaction between any non-proportional hazards covariate (the stratified variable) with other variables in the model. This model, therefore, yields similar regression coefficients across the strata. Another type of stratified models is known as an interaction model that assumes the regression coefficients vary over the strata. The hazard function for such a model is

$$h_s(t) = h_{0s}(t)\exp\left(\beta_{1s}x_1 + \beta_{2s}x_2 + \ldots + \beta_{ps}x_p\right), \tag{2.30}$$

or may be written as

$$h_s(t) = h_{0s}(t)\exp\Big(\beta_1 x_1 + \ldots + \beta_p x_p + \beta_{(11)}\left(x_1 \times x_j\right) + \ldots + \beta_{(1p)}\left(x_p \times x_j\right)\Big). \tag{2.31}$$

The likelihood ratio test is used to compare both models. The no-interaction assumption is satisfied and sufficient if the likelihood ratio test is not statistically significant $(p\text{-value} > 0.05)$.

2.5.4 Parametric Survival Model

A parametric survival model assumes that survival times follow a particular probability distribution. Most of these models belong to the AFT model. However, some of these models are belong to both: the proportional hazards model and the AFT model such as the exponential and Weibull models. Unlike the Cox proportional hazards regression model, the MLEs of the parameter θ for the parametric models are obtained from the full likelihood function that is given by

$$L(\theta) = f(t) \times S(t). \tag{2.32}$$

Parametric survival models that have been considered throughout this study are described as follows:

(i) The Weibull Model

Suppose the survival times T of n individuals has a Weibull distribution with a scale parameter λ and shape parameter ϕ. The baseline hazard function is given by

$$h_0(t) = \lambda \phi t^{\phi-1}. \tag{2.33}$$

It follows that the hazard function of the Weibull proportional hazards model is given by

$$h(t) = \exp\left(\beta_1 x_1 + \ldots + \beta_p x_p\right) \lambda \phi t^{\phi-1}, \tag{2.34}$$

where $\beta_1, \beta_2, \ldots, \beta_p$ are the regression coefficients for p covariates $x_1, x_2, \ldots, x_p$. The shape of the hazard function depends on the value of ϕ. The hazard is constant and survival times has an exponential distribution if $\phi = 1$. Meanwhile, the hazard is increasing over time when $\phi > 0$ and decreasing when $\phi < 0$.

Meanwhile, for the Weibull AFT model the hazard function is defined by

$$h(t) = \exp(-\eta)\left\{h_0\left(t \exp(-\eta)\right)\right\}, \tag{2.35}$$

where $\eta = \alpha_1 x_1 + \cdots + \alpha_p x_p$ and $\alpha_1, \ldots, \alpha_p$ are the regression coefficients for p covariates $x_1, x_2, \ldots, x_p$. The hazard function of the Weibull AFT model is obtained by substituting for the baseline hazard from (2.33) in the hazard function in the equation (2.35) as follows:

$$\begin{aligned} h(t) &= e^{-\eta}\left\{h_0\left(te^{-\eta}\right)\right\}, \\ &= e^{-\eta}\left\{\lambda \phi \left(te^{-\eta}\right)^{\phi-1}\right\}, \\ &= \left(e^{-\eta}\right)^{\phi}\left\{\lambda \phi t^{\phi-1}\right\}, \end{aligned} \tag{2.36}$$

where $\phi = 1/\sigma$ and σ is a scale parameter.

A Weibull proportional hazards model is also an AFT model, yet requires different parameterisations. Therefore, when the proportional hazards assumption is satisfied, the AFT assumption also hold (Kleinbaum & Klein, 2005). From the equation (2.36), the corresponding β-parameters for the Weibull proportional hazards model in (2.34) may be obtained as

$$\beta_j = \frac{-\alpha_j}{\sigma}, \tag{2.37}$$

for $j = 1, \ldots, p$. The survival function may be expressed in the form of

$$S(t) = \exp\left(-\lambda \mathrm{t}^{\sigma^{-1}}\right). \tag{2.38}$$

The suitability of the Weibull model is assessed via the plot of the log negative plot of the survival estimates against the log of survival time. The assumption of the Weibull proportional hazards and AFT is reasonable when the plot gives parallel straight lines with slope ϕ (Kleinbaum & Klein, 2005).

(ii) The Log-logistic Model

A log-logistic model is an AFT model with two parameters ξ and κ. It is also known as a proportional odds model, where the odds ratio is assumed to remain constant over time (Kleinbaum & Klein, 2005). The hazard function for the log-logistic model is given by

$$h(t) = \frac{\xi \kappa t^{\kappa-1}}{1+\xi t^{\kappa}}. \tag{2.39}$$

The shape $\kappa > 1$ indicates that the hazard increases initially to a maximum point and then decreases, while $\kappa \leq 1$ describes that the hazard decreases monotonically with time. The survivor function may be expressed as

$$S(t) = \frac{1}{1+\xi t^{\kappa}}. \tag{2.40}$$

For the log-logistic model, if the proportional odds assumption holds, then AFT assumption also holds and vice versa. This assumption is evaluated graphically through the plot of $\ln\left(\hat{S}(t)\Big/\left(1-\hat{S}(t)\right)\right)$ against the log of survival time where $\hat{S}(t)$ is the survival estimates. The assumption is valid if the plot depicts straight and parallel lines (Kleinbaum & Klein, 2005). The log-logistic model is the only parametric model with both the proportional odds and the AFT representation (Klein & Moeschberger, 1997).

(iii) The Lognormal Model

Another common AFT model is known as a lognormal model. T has a lognormal distribution if $\log T$ has a normal distribution with mean μ and variance σ^2. The shape of the hazard function for this model is very similar to the log-logistic distribution, and these two models give very close results in most cases (Klein & Moeschberger, 1997; Kleinbaum & Klein, 2005). The survival function of the lognormal model is

$$S(t)=1-\Phi\left(\frac{\log t-\mu}{\sigma}\right), \tag{2.41}$$

where $\Phi(\bullet)$ is the standard normal cumulative density function. The lognormal assumption is checked based on the plot of $\Phi^{-1}\left\{1-\exp\left(-\hat{H}(t)\right)\right\}$ against the log of survival time, where the lines on the plot should be straight.

(iv) The Gompertz model

A Gompertz model is another type of a parametric proportional hazards model. Unlike the Weibull model, the Gompertz model is not an AFT model. The hazard function of this model may be expressed as follows

$$h(t) = ae^{bt}, \tag{2.42}$$

where a is a positive parameter and b is a shape parameter. The hazard function of the Gompertz model does not initially begin at 0 (Lee & Wang, 2003). The shape of the hazard function depends on the b value where the hazard is exponentially increasing with time when $b > 0$, and exponentially decreasing when $b < 0$. If $b = 0$, the hazard function in (2.42) become constant, a, thus the model is reduced to the exponential model. The survival function is defined by

$$S(t) = \exp\left\{-a\left(e^{bt} - 1\right)/b\right\}. \tag{2.43}$$

2.5.4.1 The Adequacy of the Parametric Model

In the parametric model, survival times are assumed to follow a specific statistical distribution. Thus, it is important to assess the suitability of the statistical distribution considered for describing the survival times (Bradburn *et al.*, 2003b). The parametric survival model is appropriate, if there is a linear relationship between the survival times and the cumulative hazard function (or a function of it) (Lee & Wang, 2003). Thus, a plot of the cumulative hazard function (or a function of it) versus the survival time (or a function of it) is constructed to check for the adequacy of the parametric model. The points on the plot should indicate approximately a straight line or a linear trend (Collet, 2003). It is worthwhile to note that, this graphical assessment is not used to identify a particular correct model, yet to reject any model that is not suitable to represent the survival times distribution (Klein & Moeschberger, 1997).

There are various ways to estimate the cumulative hazard function, $\hat{H}(t)$. Klein and Moeschberger (1997) suggested using a Nelson-Aalen estimator, while Collet (2003) and Kleinbaum and Klein (2005) suggested a transformation of the survivor function which is estimated using the Kaplan-Meier estimate. The results have not much different among these methods of estimation. Thus, in this study, the Nelson-Aelen estimator is used to estimate the cumulative hazard function. The cumulative hazard function is computed according to the following formula (Klein & Moeschberger, 1997):

$$\hat{H}(t) = \begin{cases} 0, & \text{if } t \le t_1, \\ \sum_{t \le t_1} \frac{d_i}{Y_i} & \text{if } t_1 \le t, \end{cases} \tag{2.44}$$

where d_i be the number of event and Y_i be the number of individuals who are at risk at time t_i.

The cumulative hazard function of the Weibull model is given by

$$H(t) = \lambda t^{\phi}, \tag{2.45}$$

and taking logarithms of $H(t)$ gives

$$\log H(t) = \log \lambda + \phi \log t. \tag{2.46}$$

If the assumption of Weibull is satisfied, the relationship between $\log \hat{H}(t)$ and $\log t$ should be linear and the log-cumulative hazard plot or the plot of $\log \hat{H}(t)$ versus $\log t$ should give an approximately straight line. As given in (2.46), the slope and intercept of the straight line is ϕ and $\log \lambda$, respectively. Table 2.3 lists the plots that may be used to assess the adequacy of other parametric models.

Table 2.3: Plots to check the suitability of exponential, log-logistic and lognormal distribution

Model	**Function of the cumulative hazard**	**Plot**
Exponential	$\hat{H}(t)$	$\hat{H}(t)$ versus t
Log-logistic	$\log\left\{\exp\left(\hat{H}(t)\right)-1\right\}$	$\log\left\{\exp\left(\hat{H}(t)\right)-1\right\}$ versus $\log t$
Lognormal	$\Phi^{-1}\left\{1-\exp\left(-\hat{H}(t)\right)\right\}$	$\Phi^{-1}\left\{1-\exp\left(-\hat{H}(t)\right)\right\}$ versus $\log t$

One of the approaches to select the best model among the parametric models under study is based on the Akaike's Information Criterion (AIC) statistic (Bradburn *et al*., 2003b). This statistic is appropriate for comparing the feasibility of different parametric models since these models are fitted similarly by maximum likelihood (Royston & Lambert, 2011). The formula is given by

$$AIC = -2(\text{log-likelihood}) + 2(c+k), \tag{2.47}$$

where c is the number of unknown parameters (coefficient regressions), and k are numbers of other parameters such as scale and shape for Weibull model. Therefore, k is equal to 2 for the Weibull, log-logistic and lognormal models (Klein & Moeschberger, 1997). A smaller value of the AIC indicates a better model.

2.5.4.2 Proportional Hazards Assumption for a Parametric Model

For the Weibull model, there is an additional step required to check for the proportional hazards assumption. If the proportional hazards assumption holds, then AFT assumption also holds (vice versa).

(i) Log-cumulative hazard plot: One way to assess the proportional hazards assumption for two or more levels of covariates in the Weibull model is by plotting the log-cumulative hazard plot. The curves must be parallel to support the proportional hazards assumption, and straight to support the Weibull assumption.

(ii) Likelihood ratio test: In the Weibull model, the assumption of the proportional hazards is violated if the shape parameters across the groups, l, for a particular variable are different. A likelihood ratio test may be used to check for the aforementioned assumption. Separate Weibull model that contains the same linear component is developed for each of the l groups, and these models should yield different values of the shape and scale parameters. The sum of the value of the statistic $-2\log\hat{L}$ for each of these models is obtained, which is denoted by $-2\log\hat{L}_1$. Then, a Weibull model which combines the l sets of data is fitted and this model is corresponding to the assumption of equal shape parameter. The $-2\log\hat{L}$ value for this model is computed and denoted by $-2\log\hat{L}_0$. The difference between $-2\log\hat{L}_0$ and $-2\log\hat{L}_1$ is compared using a chi-squared distribution with $l-1$ degrees of freedom. The difference between these two values is the change in $-2\log\hat{L}$ due to constraining the Weibull shape parameters to be equal. The proportional hazards assumption is not satisfied if the difference is significant (Collet, 2003).

2.5.4.3 Stratified Weibull Model

In the Weibull model, the proportional hazards assumption may be violated because the scale parameters (or the shape parameter) across any particular group are different. Therefore, to deal with this problem, a stratified model may be adopted. Such a model may relax the proportional hazards assumption (Kalbfleisch & Prentice, 1980). In a parametric stratified model, both intercept and any ancillary parameters such as scale in the Weibull model are allowed to vary for each level of the strata variable. Meanwhile, the coefficient regression for each covariate is assumed to be similar across strata. For the AFT model, the stratified model also allows the actual shape of the baseline survival function to vary with the strata (Cleves *et al*., 2010).

The log-linear form of the stratified Weibull regression model for an individual in the *s*th stratum may be written as

$$\log\left(T \mid X, S=s\right)=\mu_s+\alpha_1 x_1+\ldots+\alpha_p x_p+\sigma_s \varepsilon\,, \tag{2.48}$$

where μ_s and σ_s denote stratum-specific intercept and scale parameters. In (2.48), the regression coefficient α_j on the AFT scale are assumed to be stratum independent (Gu *et al.*, 2014).

2.5.5 Model Checking

The fitness of the fitted model and possible influential observations are assessed by examining the residuals plots; martingale, deviance and dfbeta. The martingale residuals plot is used to determine the correct functional form of covariates to be included in the model and assess any lack of fit of the model (Collet, 2003). The martingale residual is given by

$$r_{M_i}=\delta_i-\hat{H}_i\left(t_i\right), \tag{2.49}$$

where $\hat{H}_i\left(t_i\right)$ is the estimated cumulative hazard for the *i*th individual and δ_i is unity for uncensored case and zero otherwise. The deviance residuals plot is used for checking the overall model fitness and detecting the outliers. The residuals are

$$r_{D_i}=\operatorname{sgn}\left(r_{Mi}\right)\sqrt{-2\left\{r_{Mi}+\delta_i \log\left(\delta_i-r_{Mi}\right)\right\}}\;. \tag{2.50}$$

Meanwhile, any subject that has a strong influence on parameter estimates can be identified by examining the dfbeta residuals plots. These residuals approximate the change in the coefficient estimate for the *j*th covariate if the *i*th observation is removed from the model (Collet, 2003). The dfbeta residual is denoted by

$$\Delta_i \hat{\beta} \approx \hat{\beta}_j - \hat{\beta}_{j(t)}. \tag{2.51}$$

The Cox-snell residuals (Collet, 2003) is given by

$$r_{C_i} = \exp\left(\hat{\boldsymbol{\beta}} \boldsymbol{x}_i\right) \hat{H}_0\left(t_i\right), \tag{2.52}$$

where $\hat{H}_0\left(t_i\right)$ is an estimate of the baseline cumulative hazard function at time t_i. Also, this residuals are defined by

$$r_{C_i} = \hat{H}_i\left(t_i\right) = -\log \hat{S}_i\left(t_i\right), \tag{2.53}$$

where $\hat{H}_i\left(t_i\right)$ is the estimated cumulative hazard function and $\hat{S}_i\left(t_i\right)$ is the estimated survival function. For an AFT model, the estimated survival function is given by

$$\hat{S}_i\left(t\right) = S_{\varepsilon_i}\left(\frac{\log t - \hat{\mu} - \hat{\alpha}_1 x_{1i} - \hat{\alpha}_2 x_{2i} - \cdots - \hat{\alpha}_p x_{pi}}{\hat{\sigma}}\right), \tag{2.54}$$

where $S_{\varepsilon_i}\left(\varepsilon\right)$ is the survival function of ε_i in the AFT model, $\hat{\alpha}_j$ is the estimated coefficient, and $\hat{\mu}$, $\hat{\sigma}$ are the estimated value of μ and σ.

2.5.6 Some Studies on the Comparison of Survival Models

The most common survival model is the Cox proportional hazards regression model. This model has been widely applied in modelling prognostic factors especially in clinical studies, see Chemay *et al*. (2008), Abdul Razak *et al*. (2010), Mangantig *et al*. (2013), Yang *et al*. (2013) and Suh *et al*. (2013).

Kleinbaum and Klein (2005) have given several reasons for the popularity of the Cox proportional hazards regression model. In practice, it is doubtful to identify the correct parametric model that represents the survival data. However, one should not be worried from choosing a wrong parametric model when the analysis is conducted using the Cox proportional hazards regression model since this model does not depend on the assumption of the survival time distribution. If a survival model belongs to a particular parametric model such as the Weibull model, the results from the Cox proportional hazards regression model will be approximately close to the results of the Weibull model. In addition, the hazard ratio, hazard function and survival function can be estimated from the Cox proportional hazards model, even though the baseline hazard function is not specified.

Recently, study on the association between the explanatory variables and survival time using parametric survival models has gained much attention. Parametric models tend to produce more precise estimates if the models fit the data well (Klein & Moeschberger, 1997). Also, parametric models tend to give smaller standard error estimates for the quantity such as relative hazards and median survival time (Collet, 2003). Parameter estimates obtained from the model may completely specify the survival function and hazard function (Kleinbaum & Klein, 2005), and provide estimates of survival times (Hosmer & Lemeshow, 1999). Besides, the residuals may be simply computed from the differences between observed and predicted values of survival time (Hosmer & Lemeshow, 1999).

Numerous studies have been conducted to look at the performance of different types of survival model using medical data. Some of these studies investigated the performance between the semi-parametric model and parametric models (see Pourhoseingholi *et al*., 2007, 2009, 2011; Wang *et al*., 2011; Köhler & Kowalski, 2012; Hashemian *et al*., 2013). Meanwhile, some studies just compared the performance among parametric models (see Aktürk Hayat *et al*., 2010; Nakhaee & Law 2011; Pari Dayal *et al*., 2013). The exponential, Weibull, lognormal and log-logistic models are several type of parametric models those are commonly used in these studies.

Often, the Weibull model (Sayehmiri *et al*., 2008; Moghimi-Dekordi *et al*., 2008; Nakhaee & Law, 2011; Zhu *et al*., 2011) and log-normal model (Pourhoseingholi *et al*., 2007; Wang *et al*., 2011; Köhler & Kowalski, 2012) were found to be the best parametric model to represent the survival data understudied. Pourhoseingholi *et al*. (2011) conducted a study using the same group of patients as in Pourhoseingholi *et al*. (2007) by incorporating more parametric models and it turned out that the log-logistic was the best model. Even though Gompertz model was not commonly considered in such comparison studies, it was found to be the best survival model for breast cancer data in a study done by Aktürk Hayat *et al*. (2010).

Parametric model may be an alternative model when the proportional hazards assumption is violated, as being pointed by many referred literatures (Pourhoseingholi *et al.*, 2011; Ravangard *et al.*, 2011). However, some of these studies just mentioned about the assumption, but did not verify it for their models (Pourhoseingholi *et al*., 2011; Ravangard *et al*., 2011). Some studies such as Köhler and Kowalski (2012) and Bessell *et al*. (2012) found that their models violated the proportional hazards assumption, but did not perform any proper method to handle this problem. A good survival study should clarify the method used to check for the proportional hazards assumption for the Cox proportional hazards regression model (Mallett *et al*., 2010a). According to Altman *et al*. (1995) in their review of 132 papers, the proportional hazards assumption was verified in 5% of the 43 papers that used the Cox proportional hazards regression model. Meanwhile, in a systematic review study done by Mallett *et al*. (2010b) found that only 10 out of 47 studies on cancer survival tested this assumption. Bellera *et al*. (2010) summarised the methods for checking the proportional hazards assumption using graphical, testing or modelling approaches, together with working examples.

One important step in conducting the parametric analysis is to verify the appropriateness of the chosen statistical distribution. Zhu *et al*. (2011) presented the log-cumulative hazard plot against the log time to show that the Weibull model was suitable for modelling their data. Similarly, Moran *et al*. (2008) also checked the adequacy of the log-normal and log-logistic models by plotting log time against a linear function of the cumulative hazard rate. However, there are many studies that have not given any justification for using parametric models in their analyses (see Paillisse *et al*., 2005; Pourhoseingholi *et al*., 2009; Wang *et al*., 2011; Zare *et al*., 2012). Checking the suitability of parametric distribution is imperative to ensure that the parametric form of the survival distribution is correct (Royston, 2001).

In many studies, the Akaike Information Criterion (AIC) statistic was used to measure the performance and efficiency of survival models being compared. The AIC statistic may be appropriate for comparing parametric models such as in Grover *et al*. (2013). However, using this statistic to measure the performance between semi-parametric and parametric models as demonstrated by Bessell *et al*. (2012), Abidoun *et al*. (2012) and Hashemian *et al*. (2013) may be questionable. Royston and Lambert (2011) pointed out that the AIC value for semi-parametric and parametric models are incomparable because the Cox proportional hazards model is fitted by maximising the partial log-likelihood while parametric models are fitted by maximising the full log-likelihood.

Also, residuals plot may be used to find the best model such as using the plot of deviance residuals (Royston, 2001; Pari Dayal *et al*., 2013), martingale residuals (Köhler & Kowalski, 2012), Cox-Snell residuals (Ding *et al*., 2009; Ravangard *et al*. 2011) and normal-deviate residuals (Nardi & Schemper, 2003). In addition, the performance of these models may be checked based on the efficiency of parameter estimates using the measure of explained variation (Royston, 2001; Nardi & Schemper, 2003; Moran *et al*., 2008) or model fitness using the goodness of fit test (Sayehmiri *et al*., 2008; Nakhaee & Law, 2011).

2.6 Missing Values

Missing data occur when any or all of the variables of interest are unidentified or unobserved (Schafer, 1997). Data may be missing because of many reasons such as in a survey study, participants sometimes fail to respond to some items in the questionnaire given either intentionally or unintentionally. In many survival studies, most of the required information is retrieved from patients' medical records. However, some of these information may not be properly recorded or sometimes medical reports such as histopathological examination (HPE) report and Computed Tomography (CT) scan report may be lost. Meanwhile, in a clinical trial study, there might be a group of patients who neglected to visit the clinic or centre for further follow-up. Therefore, the progress of these patients may not be monitored, and their information on that particular clinic visit is not available. Figure 2.3 illustrates the multivariate data with missing covariate values.

	Variables				
Observations	y	x_1	x_2	...	x_p
1	y_1	x_{11}	**NA**	...	x_{1p}
2	y_2	**NA**	x_{22}		x_{2p}
.	y_3	x_{31}	x_{32}	...	x_{3p}
.	y_4	x_{41}	**NA**	...	**NA**
.	y_5	x_{51}	x_{52}	...	x_{5p}
.	y_6	x_{61}	x_{62}		x_{6p}
⋮	⋮	⋮	⋮	⋮	⋮
n	y_n	**NA**	x_{n2}	...	**NA**

Figure 2.3: Multivariate data set with missing values

2.6.1 Mechanism of Missing Data

The performance of missing data methods depends on the mechanism of missingness or how data are missing. Some of these methods have been developed based on the underlying assumption of a particular type of missing values mechanism such as in Lin and Ying (1993) study. If this assumption is violated, the method may yield biased parameter estimates (Allison, 2002). Thus, it is very crucial to understand the assumption of each mechanism of missing values. There are three mechanisms of missingness namely missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

Suppose that there are two random variables, x_1 and x_2, where some values in x_1 are missing and x_2 are all observed. The values in x_1 are said to be MCAR if the missingness probability of x_1 is independent of x_1 itself and x_2. In other words, the missing data values are a simple random sample of all data values (Schafer, 1997). The assumption of MCAR is very restrictive (Van Buuren, 2012) and often unreasonable in many real situations (Allison, 2002; Little & Rubin, 2002; Van Buuren, 2012).

The x_1 values are said to be MAR if the probability of being missing is independent of x_1 itself. However, the missingness of x_1 is related to other variable, x_2, which this variable may provide information about the missing values and a basis for imputation (Barzi & Woodward 2004). In this case, the appropriate approach needs to be considered to handle missing data. Missing completely at random is a special case of MAR (Schafer & Graham, 2002), yet MAR assumption is less restrictive than the MCAR (Schafer, 1997). Both MCAR and MAR are called ignorable nonresponse (Allison, 2002).

In contrast, the nonignorable nonresponse that is MNAR depends on both missing (x_1) and observed variables (x_2). This type of missingness requires an extra step where the missing data mechanism needs to be modelled in order to optimise the estimation of parameters (Allison, 2002). Herring *et al*. (2004) emphasised that the model for missing data mechanism should be specifically formulated prior to the analysis. Information that explains why data are missing must be gathered and utilised, so that the right model may be formulated. Leong *et al*. (2001) reported that using the complete case analysis for MNAR data have affected the parameter estimates and also the significant effect of important prognostic factors.

2.6.2 Missing Data Methods

Many studies have been conducted to identify the best approach to handle missing values in survival data. Different approaches are investigated either by single imputation, multiple imputation, or likelihood based methods. Imputation methods are more preferred as these approaches are easy to perform using existing methods that available in any statistical software.

In this study, four methods of handling missing values namely the complete case analysis, hot deck imputation, expectation-maximization (EM) algorithm and multiple imputation are considered. Hot deck imputation, EM algorithm and multiple imputation methods are chosen because the "imputation" of each missing value depends on other variables in the model. Consequently, as more variables are included in the model, the more the assumption of missing at random is likely to hold because the uncertainty associated with missingness is reduced (Schafer, 1997).

2.6.2.1 Complete Case Analysis

A common approach adopted by many researchers when there are missing values is the complete case analysis. This method removes any subject with missing values from the study. The complete case analysis is appropriate for MCAR data since the reduced sample is assumed to be a random subsample of the full data (Allison, 2002; Little & Rubin, 2002). In MAR data, parameter estimates based on this method are unbiased when the percentage of missing values is low (Barzi & Woodward, 2004; Marshall *et al*., 2010a, 2010b). Also, parameter estimates may be unbiased if the probability of missing in independent variable does not depend on the outcome variable (Allison, 2002).

2.6.2.2 Hot Deck Imputation

Hot deck imputation involves imputing missing values by values drawn from "similar" responding units in the sample (Little & Rubin, 2002). This method relies on the information of the next most similar case with completely observed variables. It is suitable for imputing categorical covariate values and preferable since it is simple, flexible and powerful for handling data with complex missing-data patterns (Little *et al*., 2008; Andridge & Little, 2010; Liao *et al*., 2014). Also, this method maintains the associations with variables in the data set (Barzi & Woodward, 2004) and the imputed values are proper measurement level of variables (Brown & Kros, 2003).

2.6.2.3 Expectation-Maximization (EM) Algorithm

The Expectation-Maximization (EM) algorithm is an iterative procedure that is useful for computing maximum likelihood estimates for various types of data in the presence of missing values. In missing values problem, the EM algorithm is preferred than other optimisation methods due its stability and simplicity (Schafer, 1997). The theory was first developed by Dempster *et al*. (1977). This method consists of two iterative steps: the expectation step (E-step) and maximization step (M-step).

Let X denotes a complete data set that contains p variables for n $\left(i=1,,\ldots n\right)$ observations. The probability density function of this complete data may be written as

$$f\left(X \mid \theta\right)=\prod_{i=1}^{n} f\left(x_i \mid \theta\right), \tag{2.55}$$

where $f\left(x_i \mid \theta\right)$ is the density function for the *i*th observation and θ is unknown parameters. Suppose that X contains two components: the missing component that is denoted as X_{mis}, and observed component that is denoted as X_{obs}. The probability density function for the incomplete data is given by

$$f\left(X \mid \theta\right)= f\left(X_{obs} \mid \theta\right) f\left(X_{mis} \mid X_{obs},\theta\right), \tag{2.56}$$

where $f\left(X_{obs} \mid \theta\right)$ is the density function of the observed data X_{obs} and $f\left(X_{mis} \mid X_{obs},\theta\right)$ is the density function of the missing data given the observed data. The corresponding log-likelihood function is given by

$$\ell\left(\theta \mid X\right)=\ell\left(\theta \mid X_{obs}\right)+\log f\left(X_{mis} \mid X_{obs},\theta\right), \tag{2.57}$$

where $\log f\left(X_{mis} \mid X_{obs},\theta\right)$ is called the predictive distribution of the missing data given θ. This term portrays the interdependence between X_{mis} and θ where X_{mis} contain information that is relevant to estimate the unknown parameters θ, while θ also help to estimate the possible values for X_{mis} (Schafer, 1997).

The term $\log f\left(X_{mis} \mid X_{obs},\theta\right)$ in (2.57) may not be computed since X_{mis} component is unobserved. The EM algorithm method solves this problem by replacing $\ell\left(\theta \mid X\right)$ in (2.57) by computing the conditional expectation of the log-likelihood given the observed data X_{obs} and a preliminary estimates of θ. In the E-step, the expected log-likelihood is denoted as $Q\left(\theta \mid \theta^{(k)}\right)$ and given by

$$Q\left(\theta \mid \theta^{(k)}\right)=E\left[\ell\left(X \mid \theta\right) \mid X_{obs},\theta\right]. \tag{2.58}$$

Meanwhile, in the M-step, $Q\left(\theta \mid \theta^{(k)}\right)$ is maximised to obtain parameter estimates of $\theta^{(k)}$. The E-step and M-step are repeated until the difference

$$Q\left(\theta^{(k+1)} \mid \theta^{(k)}\right) - Q\left(\theta^{k} \mid \theta^{(k-1)}\right), \tag{2.59}$$

changes by an arbitrarily small amount (McLachlan & Krishnan, 2008) or until the EM algorithm converges.

Ibrahim (1990) proposed the EM algorithm by method of weight to handle missing categorical covariate values in the generalized linear model. The study introduced a weighted complete data log-likelihood in the E-step. Let y_i be the outcome variable for the *i*th observation. It is assumed that covariates $\boldsymbol{x} = \left\{x_1, x_2, \ldots, x_p\right\}$ are random variables from a discrete distribution with finite range parameterised by the vector $\boldsymbol{\alpha} = \left(\alpha_1, \ldots, \alpha_r\right)$. Now, let $\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \boldsymbol{\alpha}\right)$, the complete data log-likelihood for *n* observations may be written as

$$\ell\left(\boldsymbol{\theta} \mid \boldsymbol{x}, y\right) = \sum_{i=1}^{n} \left\{\ell_{y_i \mid x_i}\left(\boldsymbol{\beta}\right) + \ell_{x_i}\left(\boldsymbol{\alpha}\right)\right\}, \tag{2.60}$$

where $\ell_{y_i \mid x_i}\left(\boldsymbol{\beta}\right)$ is the log-likelihood based on the conditional distribution of $y \mid \boldsymbol{x}$ and $\ell_{x_i}\left(\boldsymbol{\alpha}\right)$ is the contribution from the marginal distribution of $\boldsymbol{x}$. When there are some covariates values missing for the *i*th observation, $\boldsymbol{x}_i$ consist of $\boldsymbol{x}_{mis,i}$ (the missing components of $\boldsymbol{x}_i$) and $\boldsymbol{x}_{obs,i}$ (the observed components of $\boldsymbol{x}_i$). Thus, the E-step is given by

$$Q_i\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}\right) = E\left[\ell\left(\boldsymbol{\theta} \mid \boldsymbol{x}_i, y_i\right) \mid \boldsymbol{x}_{obs,i}, y_i, \boldsymbol{\theta}^{(k)}\right] = \sum_{x_{mis,i}} w_i^{(k)} \ell\left(\boldsymbol{\theta} \mid \boldsymbol{x}_i, y_i\right), \tag{2.61}$$

where $w_i^{(k)}$ are the weights corresponding to the incomplete observations. To obtain the maximum likelihood estimates of the parameter of interest, expression in (2.61) is maximised in the M-step, and these two steps are repeated until convergence.

Lipsitz and Ibrahim (1996a) extended Ibrahim's (1990) method to a parametric survival model with incomplete categorical covariate values. Let t_i denotes the survival time (the outcome variable), δ_i is a censoring indicator and $\boldsymbol{x} = \{x_1, x_2, \ldots, x_p\}$ is a set of p covariates for the ith subject. The complete data log-likelihood for the ith observation is

$$\ell(\theta \mid \boldsymbol{x}_i, t_i, \delta_i) \; = \ell(\boldsymbol{\beta}, \gamma \mid \boldsymbol{x}_i, t_i, \delta_i) + \log\left[p(\boldsymbol{x}_i \mid \boldsymbol{\alpha}) \right], \tag{2.62}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of the regression coefficients, γ is a scale parameter, $p(\boldsymbol{x}_i \mid \boldsymbol{\alpha})$ is the density of the covariates indexed by α. The expected log-likelihood in E-step is given by

$$Q_i\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}\right) = E\left[\ell(\theta \mid \boldsymbol{x}_i, t_i, \delta_i) \mid \boldsymbol{x}_{obs,i}, t_i, \delta_i, \boldsymbol{\theta}^{(k)} \right] = \sum_{x_{mis,i}} w_i^{(k)} \ell(\theta \mid \boldsymbol{x}_i, t_i, \delta_i). \tag{2.63}$$

This study also proposed saturated multinomial probabilities to model the $p(\boldsymbol{x}_i \mid \boldsymbol{\alpha})$. However, this approach led to a large number of nuisance parameters that need to be estimated when there were many missing covariates values, consequently affected the estimates of the parameters of interest $\boldsymbol{\beta}$. In their following study, Lipsitz and Ibrahim (1996b) solved this problem by proposing a conditional model for the covariate distribution that helped to reduce the number of nuisance parameters that needs to be estimated. Both of these studies directly applied their methods to a liver cancer data set, and compared the performance to the complete case analysis method.

Lipsitz and Ibrahim (1998) extended the idea of EM algorithm method to accommodate the problem of missing categorical covariates values in the Cox proportional hazards model. Unlike the parametric model, maximum likelihood estimates are obtained from the partial likehood function in the Cox model. However, Ibrahim’s (1990) weighted EM algorithm approach was not feasible since the partial likelihood function caused a problem in finding the solution. Hence, Lipsitz and Ibrahim (1998) proposed a Monte Carlo method that was similar to that of Wei and Tanner (1990) to overcome this problem. This algorithm was similar to the algorithm of EM.

Herring and Ibrahim (2001) studied the method for missing continuous and mixed covariate for the Cox proportional hazards model. This study proposed a solution to the computational burden that was addressed by Lipsitz and Ibrahim (1998) for missing categorical covariate so that the weighted EM algorithm (Ibrahim, 1990) method may be used to estimate the parameters. In addition, method developed by Ibrahim *et al.* (1999) for missing continuous and mixed covariates values was implemented in this study.

Some of the aforementioned methods for MAR data have been extended to handle MNAR data. Ibrahim *et al*. (1999) were the first to explore Ibrahim's (1990) method for missing categorical covariate, and Monte Carlo EM (Wei & Tanner, 1990) for missing continuous and mixed covariates in generalized linear model. Leong *et al*. (2001) adopted method that was proposed by Lipsitz and Ibrahim (1998) and applied to categorical MNAR data. Method that was proposed by Herring and Ibrahim (2001) was extended by Herring *et al*. (2004) for categorical, continuous and mixed MNAR data. The ability to handle different types of missing covariate values may be one of the advantages of this method over the one proposed by Leong *et al.* (2001). Both methods, Leong *et al.* (2001) and Herring *et al*. (2004) were developed for the Cox proportional hazards model.

Various extensions of EM algorithm methods that have been proposed earlier were adapted to accommodate missing covariate values in other types of survival model. For instance, Herring *et al*. (2002) developed a method based on Ibrahim *et al.* (1999) idea of estimating the parameters in a frailty model when some of the covariates were missing. Recently, Fonseca *et al.* (2013) applied the EM algorithm by method of weight to the cure rate survival model with missing categorical covariates. The effects of different cure fraction level on the parameter estimates were also studied.

Even though initially Ibrahim's (1990) method was adapted to parametric survival models (Lipsitz & Ibrahim, 1996a), the extension of this method has numerously focused on the Cox model. In addition, most of these studies compared only between the proposed EM algorithm method and the complete case analysis. Studies on the performance of EM algorithm method over other available methods for handling missing covariate values remain scarce especially for parametric survival models. Ibrahim *et al*. (2005) have done a comparison study among the EM algorithm by method of weight, multiple imputation, fully Bayesian and weighted estimating equations methods but for the generalized linear model.

2.6.2.4 Multiple Imputation

Multiple imputation creates more than one set of complete data where missing values are replaced with a set of plausible values (Clark *et al*., 2003b). Each missing value is substituted with different possible values, producing several sets of complete data, *m*, with the same observed values yet different imputed values. Then, all these *m* imputed data sets are analysed separately using a standard statistical analysis for a complete data. It is worthwhile to note that the differences of the imputed values and the parameter estimates from each analysis reflect the uncertainty about what value to impute (Van Buuren, 2012). All *m* parameter estimates that have been obtained from each analysis are pooled into one estimate $\tilde{\theta}$ following Rubin's rule. Figure 2.4 illustrates *m*=3 imputed data sets.

Let $\hat{Q}_i \quad (i=1,\ldots,m)$ be different parameters estimated from m data sets and these values are combined using the following formulae:

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i \, . \tag{2.64}$$

The total variance is

$$V = \bar{U} + \left(1 + \frac{1}{m}\right)B, \tag{2.65}$$

where $\bar{U}$ is the within imputation variance,

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m} U_i , \tag{2.66}$$

and B is the between-imputations variance,

$$B = \frac{1}{(m-1)}\sum_{i=1}^{m}\left[\hat{Q}_i - \bar{Q}\right]^2 . \tag{2.67}$$

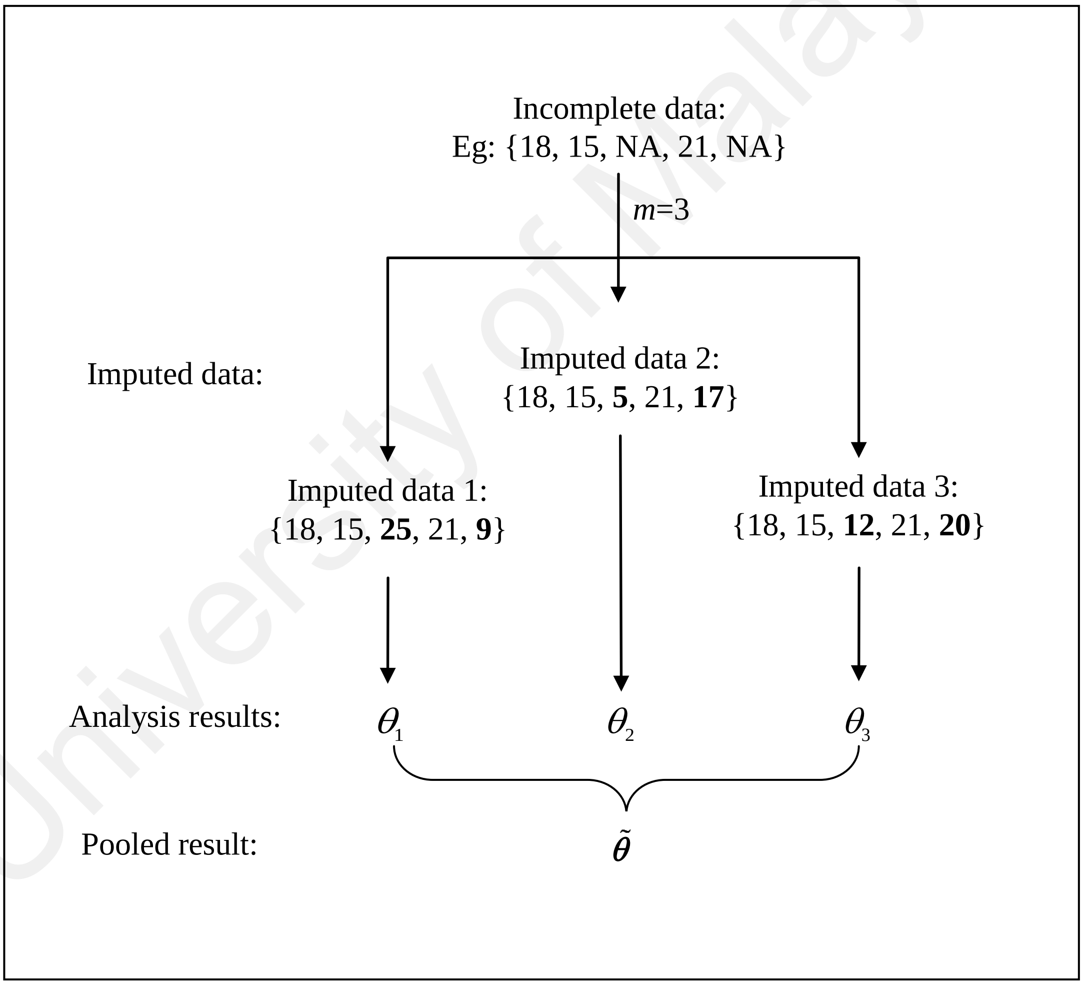


Figure 2.4: Basic steps in multiple imputation method

According to Schafer (1997), three to five imputed data sets would be enough to obtain good parameter estimates. Also, it depends on the percentage of missing values, where large amount of missing values require more imputation (Barzi & Woodward, 2004). Unlike the single imputation method, multiple imputation method considers the variability or uncertainty about the missing values by imputing each missing value with multiple possible values. Also, if a correct model is specified for the imputation, this method enhances the efficiency and validity of the parameter estimates (Little & Rubin, 2002).

Several survival studies reported that multiple imputation outperformed other missing data methods understudied (see in Alkan *et al*., 2013; Moghimbeigi *et al*., 2014). Various techniques of multiple imputation are available in statistical software packages for missing data. For instance, the R statistical software provides package for multiple imputation by chained equation (MICE), data augmentation (DA) and EM algorithm with a bootstrap (EMB) methods. Some of these methods may handle different types of variable. Marshall *et al*. (2010a, 2010b) compared the performance of several types of multiple imputation methods available in R software for handling missing values in the Cox proportional hazards regression model. Both of these studies found that, MICE with predictive mean matching (PMM) method performed the best among all methods considered. Also, according to Marshall *et al*. (2010a) the PMM method produced the least biased parameter estimates.

2.6.3 The Importance of Treating Missing Values

In most cases, subjects with missing covariate values will be removed from the analysis. However, this approach may reduce the sample size (Allison, 2002), and efficiency whenever uncensored subjects are excluded (Lin & Ying, 1993). Sometimes, the covariate itself will be excluded from the model if the percentage of missing values of that particular covariate is high. This may lead to a model misspecification (Herring & Ibrahim, 2001; Ibrahim *et al*., 2005) and reduce the power of study (Lin & Ying, 1993; Clark *et al*., 2003b). The complete case analysis still produces unbiased parameter estimates when data are MCAR. However, if the MCAR assumption is violated, this method may yield biased parameter estimates (Allison, 2002) and inefficient inferences (Little & Rubin, 2002). Therefore, proper method for handling missing values in a data set is very important to avoid any devastating impact on the statistical inferences. Moreover, most of the referred literatures considered large sample size in their studies. Hardt *et al*. (2013) pointing out the need of studying the impact of sample size on the performance of missing data methods. Thus, this study is interested to investigate the performance of missing data methods based on different sample sizes.

2.7 Frailty Model

Frailty model is a random effect model for survival data that considers heterogeneity caused by unmeasured or unobserved factors (Wienke, 2003). Frailty model can be divided into two types: univariate and multivariate frailty models. In univariate model, each observation has its own frailty value that causes the hazard functions to be different among them (Wienke, 2011). In multivariate frailty model, the assumption of independent survival times often violated when more than an individual in the study have similar characteristics. For instance, in a multicentre study, the survival experience of individuals from the same centre may not be similar to other individuals belong to a different centre. This may be due to different medical practices across the centers.

The frailty W is assumed to be identically and independently distributed random variables with a density function $G(w)$. Frailty may follow several statistical distributions such as gamma, inverse Gaussian, positive stable and positive variance function (PVF). The variance parameter σ^2 (if it exists) is defined as a measure of heterogeneity across the population in baseline risk (Wienke, 2011). Large σ^2 means the values of W are more dispersed, suggesting greater heterogeneity in the individual hazards. When σ^2 is small, the values of W is close to one.

2.7.1 Univariate Frailty Model

Individuals in a study population may differ, for instance, according to the effects of drug, a treatment, or the influence of various explanatory variables (Wienke, 2011). Each one has its own frailty, and those with higher frailty will experience an event earlier than those with lower frailty. In some cases, not all prognostic factors can be identified and included in the analysis. Some of these factors may be unobserved. Thus, there exist two sources of variability of survival time: one is variability accounted for observable factors, and the other is heterogeneity caused by unobserved factors.

Frailty W is a positive random variable that is assumed to follow a statistical distribution with probability density function $G(w)$. Given the survival time t, the hazard function for proportional hazards model conditional on the frailty is given by

$$h(t\,|\,W) = Wh_0(t), \tag{2.68}$$

where $h_0(t)$ is the baseline hazard function. In (2.68), W acts multiplicatively on the baseline hazard function. The baseline hazard is common to all the individuals whereas each individual has different frailty *W*. Thus, the larger the value of the frailty, the more likely an individual to experience an event as the hazard function also becomes large (Hosmer & Lemeshow, 1999). The conditional (individual) survivor function of t given W is given by

$$S(t \mid W = w) = \exp\left[-\int_0^t h(t \mid w)dt\right]$$

$$= \exp\left[-w\int_0^t h(t)\,dt\right]$$

$$= \exp\left[-wH_0(t)\right], \qquad (2.69)$$

where $H_0(t) = \int_0^t h_0(t)\,dt$ is the cumulative baseline hazard function. Data for individual level are not observable, thus the population survival function or unconditional survival function is considered. The unconditional survivor function of t is given by

$$S(t) = E_W\left\{\exp\left[-wH_o(t)\right]\right\}$$

$$= \int_w \exp\left[-wH_0(t)\right] \mathrm{dG}(w)$$

$$= \boldsymbol{L}\left[H_0(t)\right], \qquad (2.70)$$

where $\boldsymbol{L}\left[H_0(t)\right]$ is the Laplace transform of $H_0(t)$. The population survival function is the weighted mean of the conditional survivor function with weights given by the density function of the frailty distribution (Wienke, 2011). It is obtained from the conditional survival function by integrating out the frailty.

2.7.2 Multivariate Frailty Models

In multivariate frailty model, the frailty causes dependency among individuals in a cluster or a group. Frailty term is used to model the associations between the survival times. Under this model, it is assumed that the survival times for individuals within the same group or cluster correlate with each other. The dependence usually arises in multicentre studies (clustered sampling), genetic studies, longitudinal studies of recurrent events, and repeated measure studies.

Let T_{ij} denotes the survival time of individual i in the *j*th group, where $i=1,\ldots,n$ and $j=1,\ldots,p$. For the *i*th individual, conditional on the frailty w_i, survival times $T_i=\left(T_{i1},\ldots,T_{ip}\right)$ are independent with a hazard of the form

$$h(t\,|\,w)=w_i h_j\left(t_{ij}\right). \qquad (2.71)$$

The probability results can be studied for a single individual and, therefore, the index *i* below is omitted whenever possible. The joint conditional survivor function can be written as the following

$$\begin{aligned} S\left(t_1,\ldots,t_p\,|\,W=w\right) &= \exp\left[-\sum_{j=1}^{p}\int_0^{t_i} h_j\left(t_j\,|\,w\right)dt_j\right] \\ &= \exp\left[-w\sum_{j=1}^{p}\int_0^{t_i} h_j\left(t_j\right)dt_j\right] \\ &= \exp\left[-w\sum_{j=1}^{p}H_j\left(t_j\right)\right], \qquad (2.72) \end{aligned}$$

where $H_j=\int_0^{t_j} h_j\left(t_j\right)dt_j$ is the baseline cumulative hazard function for the *j*th group. Given the w_i with distribution function of $G(w)$, the unconditional joint survivor function can be written as

$$\begin{aligned} S\left(t_1,\ldots,t_p\,|\,W=w\right) &= \int_w \exp\left[-w\sum_{j=1}^{p}H_j\left(t_j\right)\right]dG(w) \\ &= L(s), \qquad (2.73) \end{aligned}$$

where $s=\sum_{j=1}^{p}H_j\left(t_j\right)$, and $L(s)$ denotes the Laplace transform of s.

One important concept in multivariate frailty model is a shared frailty model which had been introduced by Clayton (1978). The shared frailty model is a common type of multivariate frailty model that includes frailty to represent a characteristic whose values are shared or common among individuals in one group or cluster (Collet, 2003). Therefore, it creates dependence between survival times among individuals within the group (Hougaard, 2000). Positive dependence is observed whenever the variation of frailty variable is nonzero (Wienke, 2011). In this model, the frailty means a measure of relative risk that is a group share. Unlike the univariate frailty model, the frailty variable is associated with groups of individuals. The conditional survival function for the bivariate case is given by

$$S\left(t_1, t_2 \mid W = w\right) = S\left(t_1 \mid w\right) S\left(t_2 \mid w\right) = e^{-wH_0\left(t_1\right)} e^{-wH_0\left(t_2\right)}. \tag{2.74}$$

Suppose that there k groups of individuals with n_j individuals per group, $j = 1, \ldots, k$. For the proportional hazards model, the hazard of death at the time t for the *i*th individual, $i = 1, \ldots, n_j$, in the group *j*th is

$$h_{ij}\left(t\right) = \exp\left(\boldsymbol{\beta} \boldsymbol{x}_{ij} + \varsigma_j\right) h_0\left(t\right), \tag{2.75}$$

where x_{ij} is a vector of p explanatory variables and $h_0\left(t\right)$ is the baseline hazard function. The frailty, ς_j, is the random effect shared by subjects in the *k*th group, and the values are independent. It has a multiplicative relationship with the baseline hazard function. Meanwhile, the general AFT model that incorporates the shared frailty component is of the form

$$h_{ij}\left(t\right) = e^{-\eta_{ij}} h_0\left(t / e^{\eta_{ij}}\right), \tag{2.76}$$

where $\eta_{ij} = \alpha x_{ij} + \varsigma_j$.

The extension of the shared frailty model namely the correlated frailty model is another type of multivariate frailty model. For this model, only parts of the frailty are shared among individuals in a group. The conditional survival function in the bivariate case is given by

$$S\left(t_1, t_2 \mid W = w\right) = S\left(t_1 \mid w_1\right) S\left(t_2 \mid w_2\right) = e^{-w_1 H_0\left(t_1\right)} e^{-w_2 H_0\left(t_2\right)}, \tag{2.77}$$

where W_1 and W_2 are two correlated random variables. In the correlated frailty model, there is also an additional parameter for correlation between frailties in each group besides the parameter of the frailty. Unlike the shared frailty model, there are two associated random variables used to characterise the frailty effect for each cluster. For instance in twin data, different random variable is assigned to twin 1 and another to twin 2 (Wienke *et al*., 2005).

2.7.3 Statistical Distributions For Frailty

The frailty is a random variable W with a probability distribution function $G(w)$. Examples of frailty distribution are gamma distribution, positive stable, and inverse Gaussian.

2.7.3.1 Gamma Distribution

The most widely used frailty distribution is a gamma distribution. It has been applied in many researches due to the simplicity of the Laplace transform. Also, it is a flexible distribution depending on the value of the shape parameter. The probability density function of a gamma distributed random variable is given by (Wienke, 2011)

$$G(w) = w^{\alpha-1} e^{-\beta w} \beta^{\alpha} \frac{1}{\Gamma(\alpha)}, \tag{2.78}$$

where α is a shape parameter and β is a scale parameter. Given that $s = H(t)$, the unconditional survivor function $S_f(t)$ is

$$\begin{aligned} S_f(t) &= \int_0^{\infty} e^{-ws} w^{\alpha-1} e^{-\beta w} \frac{\beta^{\alpha}}{\Gamma(\alpha)} dw \\ &= \int_0^{\infty} e^{(-ws-\beta w)} w^{\alpha-1} \frac{\beta^{\alpha}}{\Gamma(\alpha)} dw \\ &= \int_0^{\infty} e^{-w(s+\beta)} w^{\alpha-1} \frac{1}{\Gamma(\alpha)} (\beta+s)^{\alpha} \frac{\beta^{\alpha}}{(\beta+s)^{\alpha}} dw. \end{aligned} \tag{2.79}$$

The last fraction in (2.79) does not depend on w and can be taken out of the integral. What is left is a gamma density with parameters α and $\beta+s$, and, therefore, integrates to one. This gives

$$S_f(t)=\frac{\beta^{\alpha}}{(\beta+s)^{\alpha}}. \tag{2.80}$$

When $\beta=1$, the unconditional survivor function is given by

$$S_f(t)=(1+s)^{-\alpha}. \tag{2.81}$$

2.7.3.2 Positive Stable Distribution

The probability density function of a positive stable distributed random variable is given by (Wienke, 2011)

$$G(w)=\frac{1}{\pi}\sum_{\kappa=1}^{\infty}(-1)^{\kappa+1}\frac{\Gamma(\kappa v+1)}{\kappa!}w^{-\kappa v-1}\sin(\kappa v\pi), \tag{2.82}$$

with $w\geq 0$ and $0<v\leq 1$. The Laplace transform of positive stable distribution is given by

$$L(s)=E\{\exp(-ws)\}=\exp(-s^{v}). \tag{2.83}$$

Therefore, the unconditional survivor function for the positive stable frailty model is given by

$$S_f(t)=\int_0^{\infty}e^{-ws}dG(w)=\exp(-s^{v}). \tag{2.84}$$

All moments of the positive stable distribution are infinite, thus the mean of the frailty is infinite and variance does not exists (Wienke, 2011).

2.7.3.3 Inverse Gaussian Distribution

The inverse Gaussian distribution was introduced by Hougaard (1984). The density function may be expressed in the form of (Wienke, 2011)

$$G(w)=\frac{\sqrt{\lambda}}{\sqrt{2\pi w^3}}\exp\left(-\frac{\lambda}{2\mu^2 w}(w-\mu)^2\right), \tag{2.85}$$

where λ and μ are positive parameters.

2.7.4 Tests for Detecting the Presence of Frailty

Kiefer (1984) proposed a score test for heterogeneity in exponential survival model and considered uncensored data. Lancaster (1985) modified the test using the unconditional variance and applied to the Weibull model. Later on, Blossfeld and Hamerle (1989) derived a score test following the test proposed by Lancaster (1985), yet using the conditional variance estimator suggested by Burdett *et al*. (1985). This study considered the Weibull model with censored observations.

Commenges and Andersen (1995) derived a score test from the marginal partial likelihood of the Cox proportional hazards regression model. The counting process arguments were used to obtain the asymptotic variance. The proposed test was studied in the case of individual frailty. Similar to Commenges and Andersen (1995), the basis of the tests proposed by Gray (1995) and Verweij *et al*. (1998) were martingale residuals. However, these studies considered the case of multivariate frailty. For instance, Gray (1995) focused on testing the institutional effect on study outcome and considered time-varying covariates in the model. Meanwhile, Andersen *et al*. (1999) applied the test proposed by Commenges and Andersen (1995) in a multicentre clinical trial study. This study compared the performance of frailty model with the common proportional hazards model. They found that ignoring the centre effect caused biased and inconsistent estimation. Sinha (2012) proposed a score test for detecting frailty for recurrent event data that was derived from Taylor series expansion of the likelihood function about the frailty mean. Apart from numerical methods, there were also graphical approaches for testing frailty in survival models such as in Viswanathan and Manatunga (2001), Economou and Caroni (2005), Economou and Caroni (2008) and Economou (2011). Also, an outlier test was proposed by Caroni and Kimber (2004) for detecting frailty in the Weibull model.

In parametric survival model, many frailty tests were developed based on the Weibull model. For instance, Crowder and Kimber (1997) proposed a score test for Weibull based model with gamma frailty. Zhu (1998) derived a Weibull based score test for infinite variance frailty, yet it has a slow convergence rate to the normal limit. Sarker (2002) improved this drawback by proposing two new score tests which were derived from Zhu’s score test namely modified score test and test based on ln s. The convergence to the normal limit for Sarker’s tests was faster than Zhu’s (1998) score test. Meanwhile, Bolfarine and Valença (2005) proposed score tests which have been derived from a Weibull AFT model. Besides the score test, Zhu (1998), Sarker (2002), Duchateau *et al*. (2002) and Claeskens (2008) also looked into the likelihood ratio test in their study. Thus, the aim of our study is to investigate the tests for frailty in a Gompertz survival time data since the model has not been extensively explored by other studies.

2.8 Summary

The theory of survival analysis such as the survival time functions and types of survival model are described in this chapter. Studies on survival and prognostic factors for cervical cancer have been reviewed. Also, studies on the comparison of survival models are discussed. Missing data problem in survival models are also studied, where missing data methods used in this study are described. Since most of the referred literatures on missing covariate values focused on the Cox proportional hazards model, this study interested to look at the performances of these methods on the parametric survival model. This chapter also explains the theory of frailty model which includes the types of frailty model, frailty distributions and tests for detecting frailty. Based on referred literatures, frailty test derived from the Gompertz distribution is limited. Thus, frailty tests for this type of survival data are derived and investigated in this study.

The methods described in this chapter are very relevant and often used in the analysis of survival data. The use of the Cox model is justified in determining the hazard ratio which is an important quantity in survival analysis as it gives a meaningful and understandable interpretation especially for describing patients' prognosis. The parametric survival models considered in this study are also appropriate as they are the most commonly used to model survival data. Meanwhile, methods for handling missing covariate values those have been considered in this study are also justified as they are the most commonly methods used in survival analysis. Also, the test for detecting frailty in survival data is of importance where further analyses are required especially when there is any unmeasured factor that may induce dependency among the survival times.

CHAPTER 3

NON-PARAMETRIC AND SEMI-PARAMETRIC ANALYSIS OF CERVICAL CANCER DATA

3.1 Introduction

The survival probability is estimated by the Kaplan-Meier method and the survival difference is checked based on the log-rank test. The relationship between the survival experience of patients and their socio-demographic and clinical characteristics is commonly studied based on the Cox proportional hazards regression model. This model identifies a set of covariates that is significantly influencing the hazard or the risk of getting an event. An important assumption of the Cox proportional hazards regression model is the proportional hazards assumption. When this assumption is violated, an extension of the proportional hazards model may be opted.

In this chapter, data of cervical cancer patients treated in Hospital Universiti Sains Malaysia (HUSM) are analysed to model the prognostic factors. Section 3.2 describes the data and variables that have been included in the analysis. In Section 3.3, results of the descriptive statistics are presented. Non-parametric methods using Kaplan-Meier estimate and log-rank test to check for the survival differences are described in Section 3.4. In Section 3.5, prognostic factors that were significantly affecting the risk of death of the cervical cancer patients treated in HUSM are identified using the Cox proportional hazards regression analysis. The proportional hazards assumption is verified, and the stratified Cox model is carried out since the proportional hazards assumption is violated. The results are presented in Section 3.5. Section 3.6 discusses all the results that have been obtained and the summary is given in Section 3.7.

3.2 Description of the Data Set

3.2.1 Source of Data

The cervical cancer data are obtained from Hospital Universiti Sains Malaysia (HUSM). The HUSM is located in Kubang Kerian, Kelantan, Malaysia. This hospital is a referral centre for the East Coast region of Malaysia. All those patients who were histopathologically and clinically diagnosed with cervical cancer between 1st July 1995 and 30th June 2007, and have received at least one treatment related to cervical cancer in HUSM are included in this study. Those who died due to other reason are excluded. Patients are followed until 31st December 2008. In all, 120 patients are included which among them 66 (55%) died and 54 (45%) alive. Ethical approval is obtained from the research ethics committee (see Appendix A).

3.2.2 Description of Variables

The outcome variable is survival time which is measured from the date of diagnosis with cervical cancer until date of death. Patient who survived beyond the study period is considered as censored observation. The independent variables are stage at diagnosis, ethnicity, histologic type, lymph node involvement, age at diagnosis, distant metastasis and primary treatment received. The stage of cancer follows the International Federation of Gynecology and Obstetrics (FIGO) system.

The number of patients diagnosed at stage IV is too small (5 patients) compared to all the other stages. Thus, it has been decided to combine both groups, stage III and stage IV, which yielded to 31 (25.8 %) patients. This variable is classified into three groups namely; stage I, stage II and stage III-IV. For the histologic type, cases are divided into squamous cell carcinoma and adenocarcinoma. Whenever the cancer has metastasized to the lymph node that is to pelvic, para-aortic or both, that patient is considered to have lymph node involvement. Patients are grouped according to their age at diagnosis namely; less than 40, 40-49 years, 50-59 years, and 60 and over. Ethnicity is classified according to Malay and non-Malay. The type of primary treatment received is divided into either surgery or non-surgery (chemotherapy and/or radiotherapy). These variables are described in Table 3.1.

Table 3.1: List of variables

Variables	**Description/Code**
Time	Survival time from diagnosis of cervical cancer to death (in months)

Status	Censored = 0
	Death = 1

Ethnicity	Non-Malay = 0,
	Malay =1

Lymph node involvement	Negative = 0
	Positive = 1

Histologic type	Squamous cell carcinoma (SCC) = 0, Adenocarcinoma (ADC) = 1
Age at diagnosis	≤ 39 years = 0, 40 – 49 = 1, 50 – 59 = 2, ≥ 60 = 3

Stage at diagnosis	I = 0,
	II = 1,
	III–IV = 2.

Primary treatment	Surgery = 0,
	Non-surgery = 1.

Distant metastasis	No = 0,
	Yes = 1.

3.3 Descriptive Statistics

The mean age at diagnosis of the cervical cancer patients is 49.73 ± 9.52. Majority of these patients are Malays (99[82.5%]) and the remaining are non-Malays (21[17.5%]). Exactly 31(25.8%) patients are diagnosed with lymph node involvement and 89(74.2%) are diagnosed without lymph node involvement. Meanwhile, the squamous cell carcinoma type constitutes about 77.5% (93 patients) of all histologic types. There are 35 (29.2%), 54 (45.0%) and 31 patients (25.8%) diagnosed in stage I, II, III-IV respectively. Of the 120 patients, there are 40 (33.3%) patients primarily treated with surgical treatment and 37(30.8%) patients have distant metastasis. The details are given in Table 3.2 below.

Table 3.2: Characteristics of patients with cervical cancer treated in HUSM

Characteristic	No. of cases (%)	Died (%)	Censored (%)
Ethnicity			
Non-Malay	21(17.5)	10(8.3)	11(9.2)
Malay	99(82.5)	56(46.7)	43(35.8)
Lymph node involvement			
Negative	89(74.2)	50(41.7)	39(32.5)
Positive	31(25.8)	16(13.3)	15(12.5)
Histologic type			
SCC[a]	93(77.5)	48(40.0)	45(37.5)
ADC[b]	27(22.5)	18(15.0)	9(7.5)
Age at diagnosis			
≤ 39 years	15(12.5)	11(9.2)	4(3.3)
40 – 49	46(38.3)	24(20.0)	22(18.3)
50 – 59	38(31.7)	21(17.5)	17(14.2)
≥ 60	21(17.5)	10(8.3)	11(9.2)
Stage at diagnosis			
I	35(29.2)	16(13.3)	19(15.8)
II	54(45.0)	28(23.3)	26(21.7)
III–IV	31(25.8)	22(18.3)	9(7.5)
Primary Treatment			
Surgery	40(33.3)	18(15.0)	22(18.3)
Non-surgery	80(66.7)	48(40.0)	32(26.7)
Distant Metastasis			
No	83(69.2)	39(32.5)	44(36.7)
Yes	37(30.8)	27(22.5)	10(8.3)

[a]Squamous cell carcinoma, [b]Adenocarcinoma

3.4 Non-parametric Analysis

The five year surival rate of the cervical cancer patients has been estimated using the Kaplan-Meier method. As defined by Lee and Wang (2003), the five-year survival rate is the cumulative proportion surviving at the end of the fifth year. Therefore, this study defines the five-year survival rate as the proportion of the cervical cancer patients who survived within five years after diagnosis.

The differences of the survival distribution between the groups for each covariate are further tested using the log-rank test. This test is performed using the function `survdiff` from the `survival` package (Therneau, 2014) in `R` software. The survival difference between the groups are statistically significant when the p-value is significant ($p\text{-value} < 0.05$). In addition, Kaplan-Meier survival curves are constructed for each variable to observe the difference in the survival distributions between two or more groups. Also, the median survival time which is the time when half of the patients died is computed.

The overall five-year survival estimates of the 120 cervical cancer patients treated in HUSM is 39.7% (95% CI: 30.7, 51.3). The median survival time is 40.8 (95% CI: 34.0, 62.0) months. Figure 3.1 presents the overall Kaplan-Meier estimates of the survivorship functions along with 95% confidence interval.

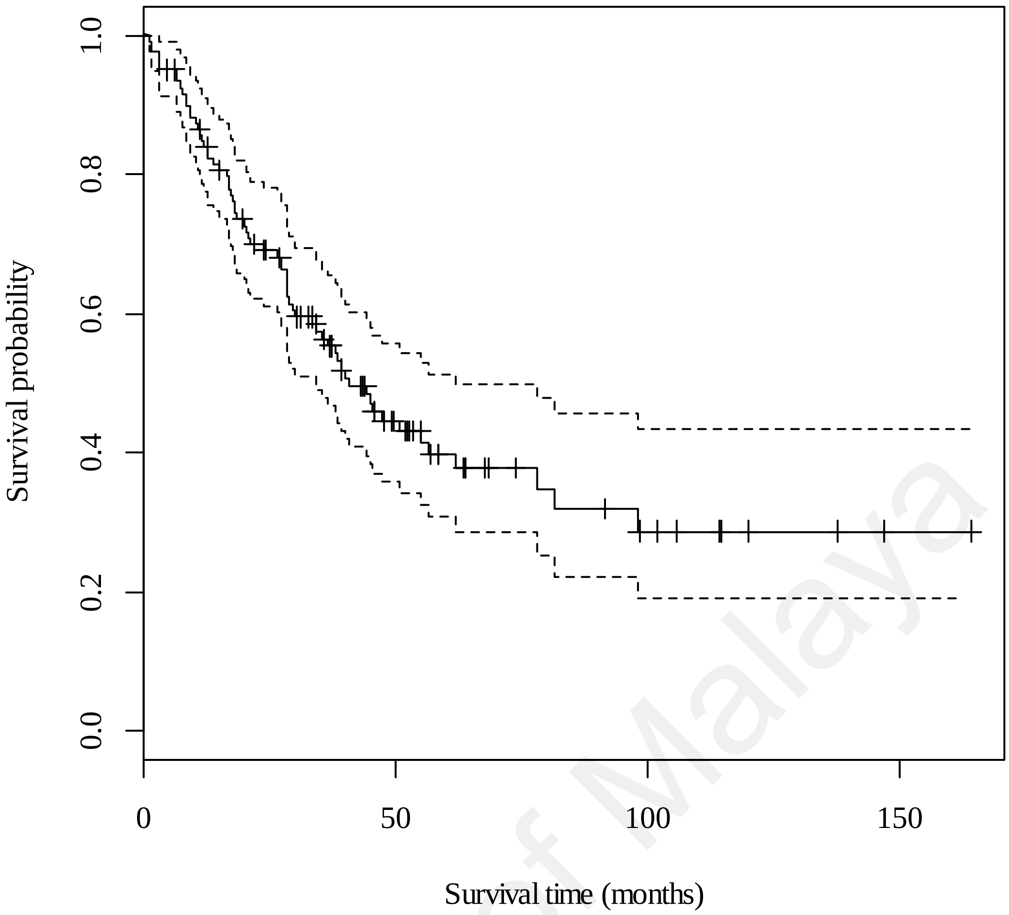


Figure 3.1: Kaplan-Meier estimate along with 95% confidence interval

The five-year survival rate for each factor is tabulated in Table 3.3. The log-rank test is used to compare the survival difference between the groups. The log-rank test and Kaplan-Meier survival curves may provide a preliminary idea of possible prognostic factors for further analysis. The log-rank test shows that there is significant difference in the survivorship function between the groups for variables stage, primary treatment and distant metastasis (see Table 3.3).

Table 3.3: Five-year survival according to patients' characteristics

Characteristic	Five-year survival (%)	95 % CI	χ^2 (df)	*p*-value (log-rank)
Ethnicity			0.2(1)	0.631
Non-Malay	33.0	12.7 – 86.0		
Malay	40.6	31.1 – 52.9		
Lymph node involvement			0.1(1)	0.762
Negative	36.6	26.6 – 50.4		
Positive	52.3	37.0 – 74.0		
Histologic type			1.4(1)	0.244
SCC[b]	41.2	30.9 – 54.8		
ADC[c]	35.2	19.9 – 62.3		
Age at diagnosis			3.3(3)	0.345
≤ 39 years	29.6	13.1 – 66.8		
40 – 49	44.6	30.6 – 65.0		
50 – 59	38.9	25.0 – 60.3		
≥ 60	42.1	23.6 – 75.1		
Stage at diagnosis			10.8(2)	0.005
I	54.7	38.7 – 77.2		
II	40.8	27.7 – 60.3		
III–IV	18.4	6.8 – 50.1		
Primary Treatment			5.1(1)	0.0242
Surgery	52.6	37.5 – 73.6		
Non-surgery	33.3	22.9 – 48.4		
Distant Metastasis			6.6(1)	0.0102
No	49.7	38.8 – 63.7		
Yes	16.4	6.6 – 40.8		

[a]log-rank test [b]Squamous cell carcinoma, [c]Adenocarcinoma

Kaplan-Meier curves with respect to the variables that significant in the log-rank test are constructed. Figure 3.2 illustrates the survival curves for the variable stage at diagnosis. Patients who are diagnosed at the latest stage (III-IV) are found to have the lowest survival compared to stage I and stage II. The survivorship function for stage III-IV lies below the other two groups (stage I and stage II) suggesting that this group has the least favorable survival experience.

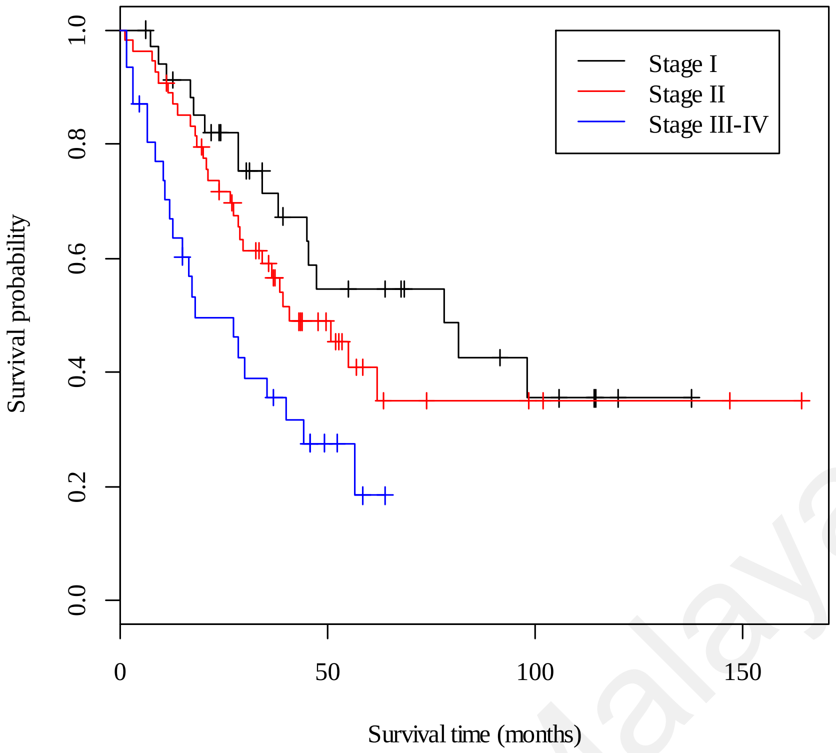


Figure 3.2: Kaplan-Meier survival curves for stage at diagnosis

Figure 3.3 shows that the curve for non-surgery group lies below the curve for the surgery group suggesting that the survival of patients in the surgery group is higher than the non-surgery group.

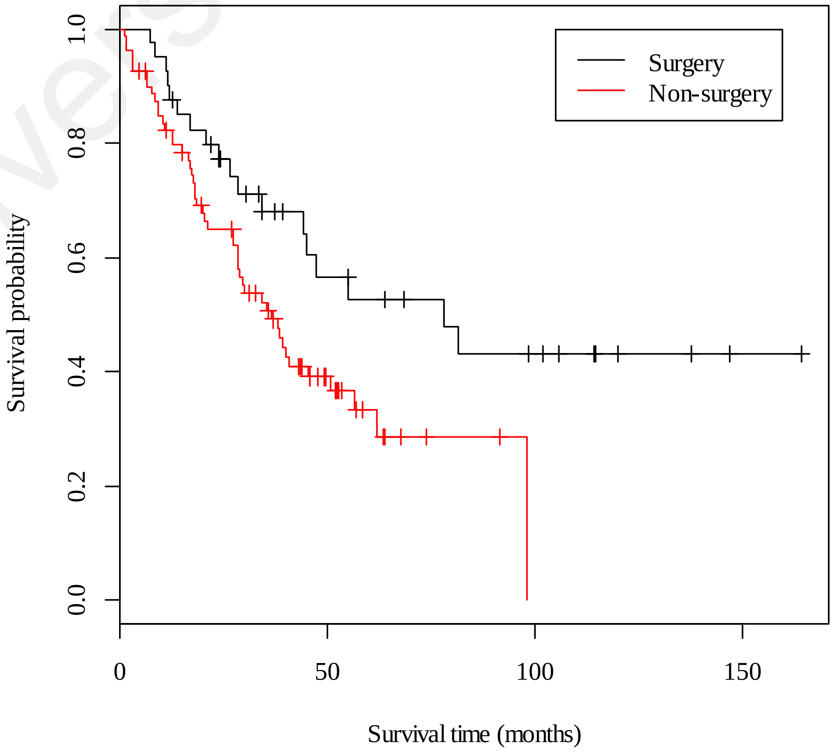


Figure 3.3: Kaplan-Meier survival curves for primary treatment

The Kaplan-Meier survival plot for the distant metastasis (see Figure 3.4) indicates that there is a clear separation between the lines for survival time beyond 25 months. The survivorship function for without distant metastasis group lies above the other group suggesting better survival experience.

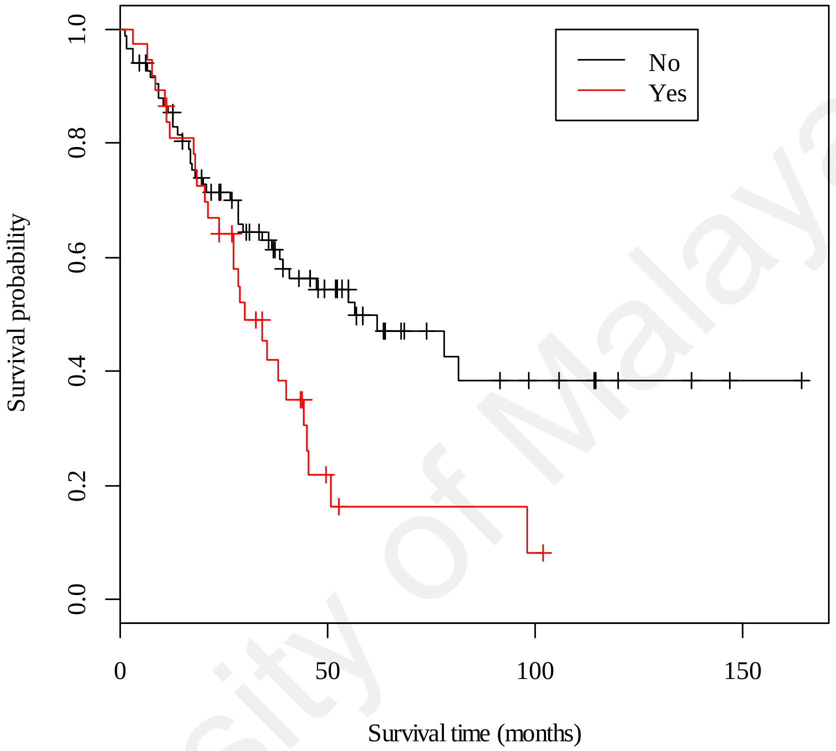


Figure 3.4: Kaplan-Meier survival curves for distant metastasis

3.5 Cox Proportional Hazards Regression Model

The prognostic model of the cervical cancer patients is commenced using the Cox proportional hazards regression model. The analysis is conducted to identify significant factors associated with the risk of death of these patients.

3.5.1 Proportional Hazards Assumption

The assessment of the proportional hazards assumption is a crucial step in modelling the Cox proportional hazards regression model. The numerical assessment is based on the scaled Schoenfeld residuals. This test verifies the proportional hazards assumption for each covariate. Table 3.4 shows that the proportional hazards assumption holds for all variables except that for distant metastasis (p-value < 0.05).

Table 3.4: The proportional hazards assumption test results for each independent variable

Variables	**rho**	χ^2	***p*-value**
Ethnicity	-0.0812	0.434	0.510
Lymph node involvement	-0.0170	0.019	0.892
Histologic type	0.0221	0.032	0.858
Age at diagnosis	NA	4.042	0.257
Stage at diagnosis	NA	2.102	0.350
Primary Treatment	0.0244	0.041	0.841
Distant metastasis	0.2580	4.600	0.032

For categorical independent variables, the proportionality assumption is further confirmed using a log-cumulative hazard functions against the logarithm of time plot (LML plot). The curves should be parallel so that the proportionality assumption holds (Lee & Wang, 2003). Figure 3.5 shows the LML plot for the distant metastasis variable. The plot supports the result from the scaled Schoenfeld residuals test, where the proportional hazards assumption for this variable is not satisfied. The two curves are parallel initially but afterwards the one representing the distant metastasis group crosses upwards the curve of the other group (without distant metastasis group).

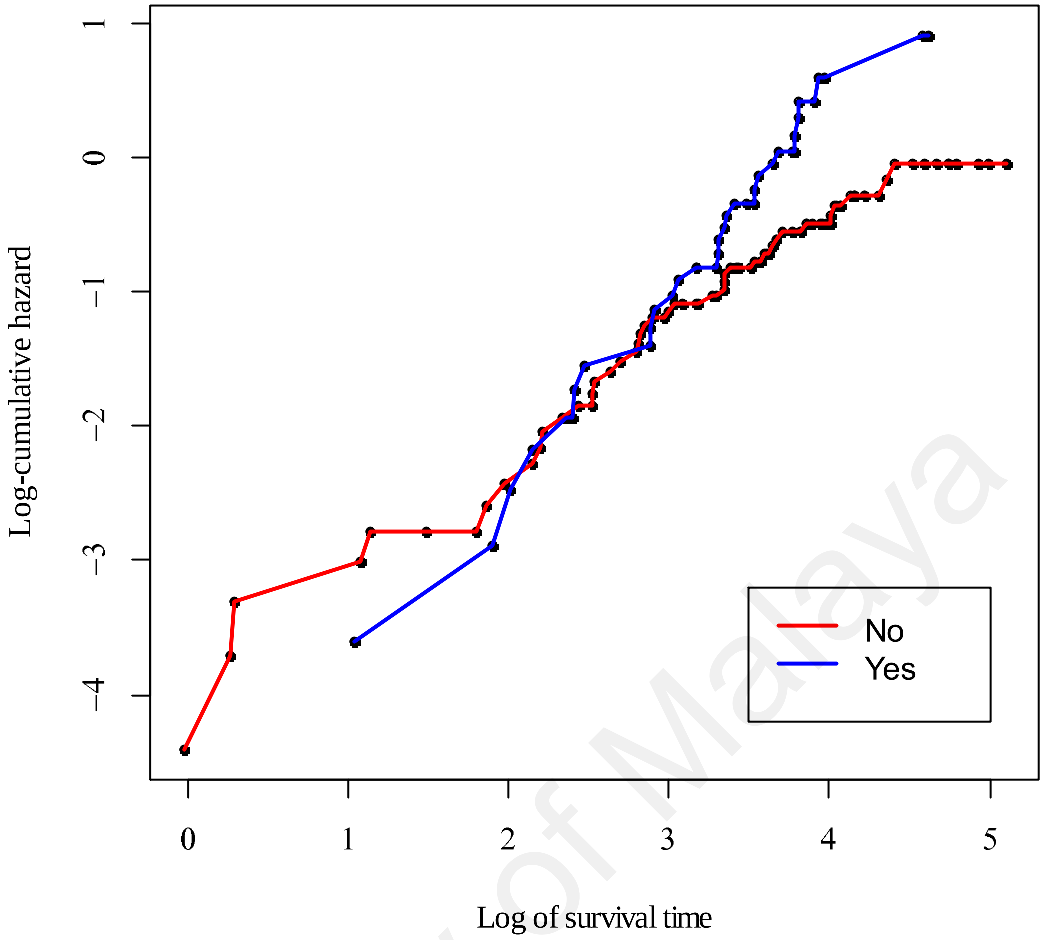


Figure 3.5: Log-cumulative hazard plot of distant metastasis

3.5.2 Univariate Cox Proportional Hazards Regression Analysis

In univariate analysis, each factor is analysed using the Cox proportional hazards regression model to identify the association between each covariate and the outcome individually. This analysis also may provide a preliminary idea on which variables have possible prognostic importance. As the distant metastasis variable does not satisfy the proportional hazards assumption, it is not appropriate to analyse this variable using the Cox proportional hazards model. Table 3.5 shows that stage at diagnosis and primary treatment are statistically significant at 5% level of significance.

Table 3.5: The univariate Cox proportional hazards regression model

Variables	**Coefficient**	**Crude HR (95% CI)**	**LR(df)**	***p*-value**
Ethnicity				
Non-Malay				
Malay	-0.1653	0.8476 (0.4312-1.666)	0.22(1)	0.6382
Lymph node involvement				
Negative				
Positive	-0.08687	0.9168 (0.5218-1.611)	0.09(1)	0.7609
Histologic type				
SCC[a]				
ADC[b]	0.3219	1.3797 (0.8013-2.376)	1.28(1)	0.2577
Age at diagnosis				
≤ 39 years				
40 – 49	-0.6247	0.5354 (0.2617-1.096)		
50 – 59	-0.4092	0.6642 (0.3194-1.381)		
≥ 60	-0.6155	0.5404 (0.2277-1.283)	2.97(3)	0.3962
Stage at diagnosis				
I				
II	0.3265	1.3860 (0.7448-2.579)		
III–IV	1.0176	2.7665 (1.4272-5.362)	9.6(2)	0.0080
Primary Treatment				
Surgery				
Non-surgery	0.6360	1.889 (1.077-3.312)	5.37(1)	0.0205

[a]Squamous cell carcinoma, [b]Adenocarcinoma

3.5.3 Multivariate Cox Proportional Hazards Regression Analysis

All variables are further analysed using the multivariate Cox proportional hazards regression analysis. Independent variables that significantly associated with the hazards of death of the cervical cancer patients under study are selected based on the stepwise selection method. This selection method consists of forward selection followed by backward elimination process, with $p\text{-value} < 0.05$ for variable entry, and $p\text{-value} > 0.10$ for variable removal. At this step, the preliminary main effects model is obtained.

In the multivariate analysis, histologic type, stage at diagnosis and distant metastasis are found to be statistically significant. This study has decided to combine the stage I and stage II groups since there is no significant difference between stage I and stage II (HR 1.413; 95% CI: 0.7579, 2.636; *p*-value=0.276). The likelihood ratio test also shows that there is no significant difference between model with three-level version and the collapsed two-level version of the stage at diagnosis variable. In addition, the binary variable yields a simpler model, and it has not changed the coefficients for any other variables in the model.

Interaction between the covariates in the preliminary main effects model is checked by adding the interaction term to the model. The interaction is not statistically significant at 5% level of significance. The preliminary final model is obtained after checking for the interaction. The results are presented in Table 3.6.

Table 3.6: The multivariate Cox proportional hazards regression model

Variables	**Coefficient**	**SE**	**Adjusted HR (95% CI)**	***p*-value**
Histologic type				
SCC[a]				
ADC[b]	0.5946	0.2897	1.8123(1.027-3.198)	0.0401
Stage at diagnosis				
I-II				
III–IV	0.8643	0.2700	2.3734(1.398-4.029)	0.0014
Distant Metastasis				
No				
Yes	0.7211	0.2636	2.0567(1.227-3.448)	0.0062

[a]Squamous cell carcinoma, [b]Adenocarcinoma

The proportional hazards assumption for each variable in the preliminary final model is tested based on the scaled Schoenfeld residuals. The global Schoenfeld residuals test is used to assess the assumption for the overall model. These tests are performed using the function `cox.zph` from the `survival` package (Therneau, 2014) in `R` software. The results are tabulated in Table 3.7. The global test indicates that the proportional hazards assumption for the overall model is violated. Meanwhile, the scaled Schoenfeld residuals test shows that the hazard for the distant metastasis variable is not proportional.

Table 3.7: The proportional hazards assumption test results for the preliminary final model

Variables	**rho**	χ^2	***p*-value**
Histologic type	0.0712	0.373	0.5416
Stage	-0.1434	1.336	0.2477
Distant metastasis	0.2934	6.591	0.0102
Global	NA	7.956	0.0469

3.5.4 Stratified Cox Model

As the distant metastasis violates the proportional hazards assumption, the stratified Cox model is applied. Only the stage at diagnosis and histologic type remain in the model, while the distant metastasis is included in the model as a stratification factor. Both no-interaction (see Table 3.8) and interaction models (see Table 3.9) are compared using the likelihood ratio test. The test indicates that the no-interaction assumption holds (*p*-value=0.0823). Thus, the no-interaction model in Table 3.8 is acceptable.

Table 3.8: The stratified Cox model (No-interaction model)

Variables	**Coefficient**	**SE**	**Adjusted HR (95% CI)**	***p*-value**
Histologic type	0.6529	0.2954	1.921 (1.077-3.428)	0.0271
Stage at diagnosis	0.9195	0.2744	2.508 (1.465-4.294)	0.0008

Log-likelihood = -222.6476

Table 3.9: The stratified Cox model (Interaction model)

Variables	**Coefficient**	**SE**	**Adjusted HR (95% CI)**	***p*-value**
Histologic type	1.3668	0.892	1.367 (0.6880-2.715)	0.3723
Stage at diagnosis	0.6760	0.3647	1.966(0.9619-4.018)	0.0638
Histologic type × distant metastasis	1.3700	0.6474	3.935(1.1064-13.997)	0.0343
Stage at diagnosis× distant metastasis	0.7773	0.5846	2.176(0.6918-6.842)	0.1837

Log-likelihood = -220.1503

3.5.5 Assessment of Model Adequacy

The fitness of the stratified Cox model (Table 3.8), and outliers are examined from the martingale, deviance and dfbeta residuals plots. Figure 3.6 illustrates the martingale residuals against survival time plot. The residuals scatter between -2 to 1, and there is no outlier seen in the plot. There are residuals that close to unity that correspond to patients who have shorter survival time than estimated by the model. Meanwhile, negative residuals indicate that patients have long survival time yet expected to die earlier. There is no indication of a lack of fit of the model.

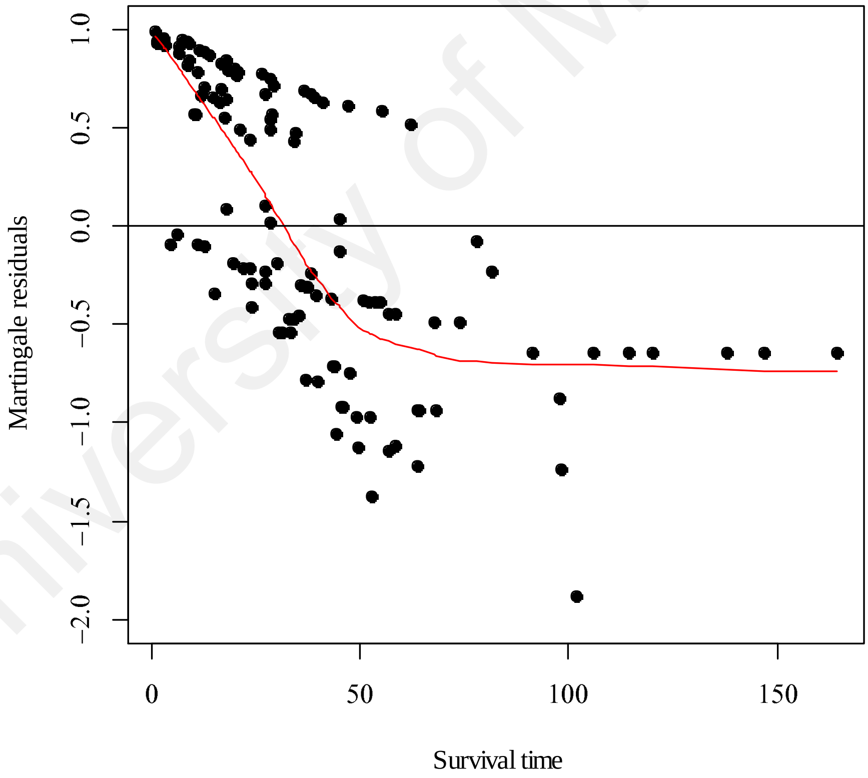


Figure 3.6: Plot of the martingale residuals against survival time for the stratified Cox model

The plot of the deviance residuals against survival times is examined to identify the presence of subjects who is poorly predicted by the model. Figure 3.7 shows that the residuals are roughly symmetrically distributed around zero, and the residuals ranged between -2 to 3 suggesting no wildly deviant observations.

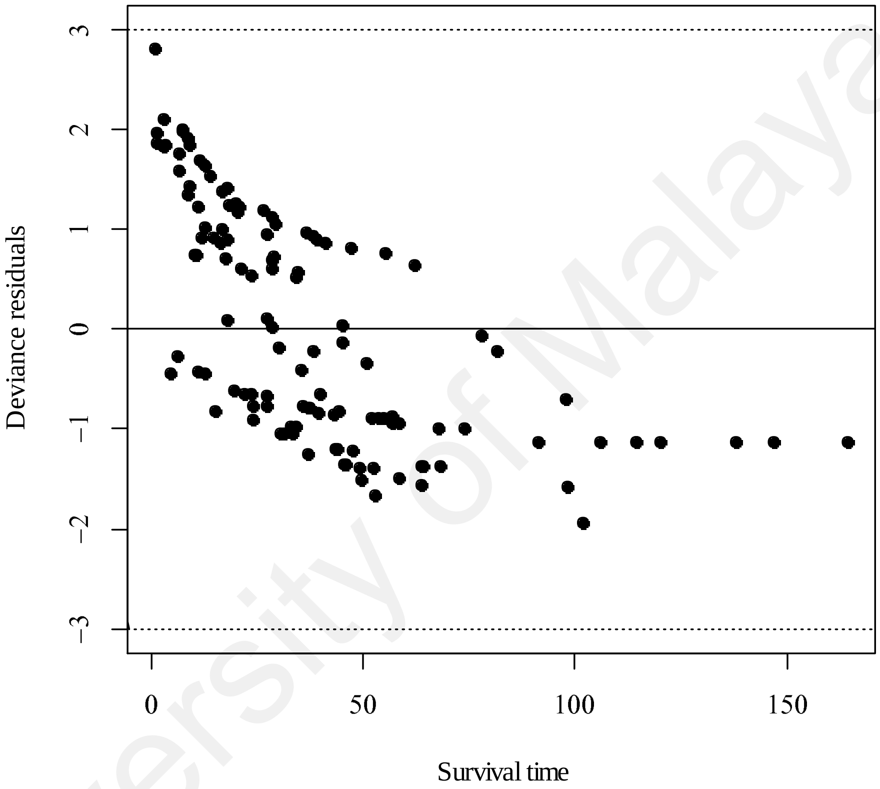


Figure 3.7: Plot of the deviance residuals against survival time for the stratified Cox model

In addition, the plot of delta-betas residuals for the histologic type (see Figure 3.8) and stage at diagnosis (see Figure 3.9) show that there is no influential observation since all residuals lie between -0.1 to 0.1, thus implying a good fit of the model.

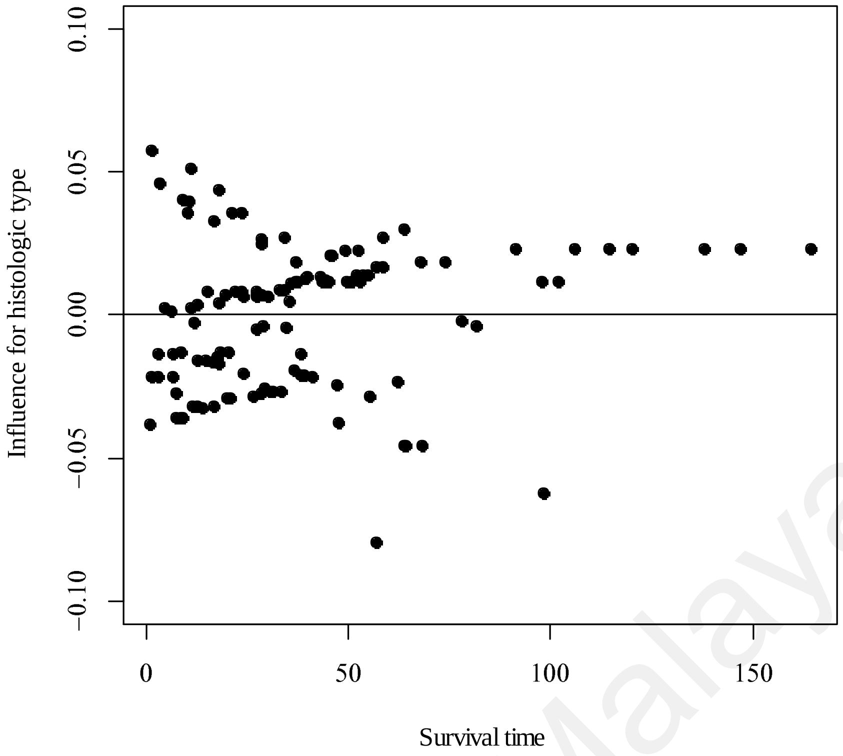


Figure 3.8: Plot of the delta-betas for histologic type against survival time for the stratified Cox model

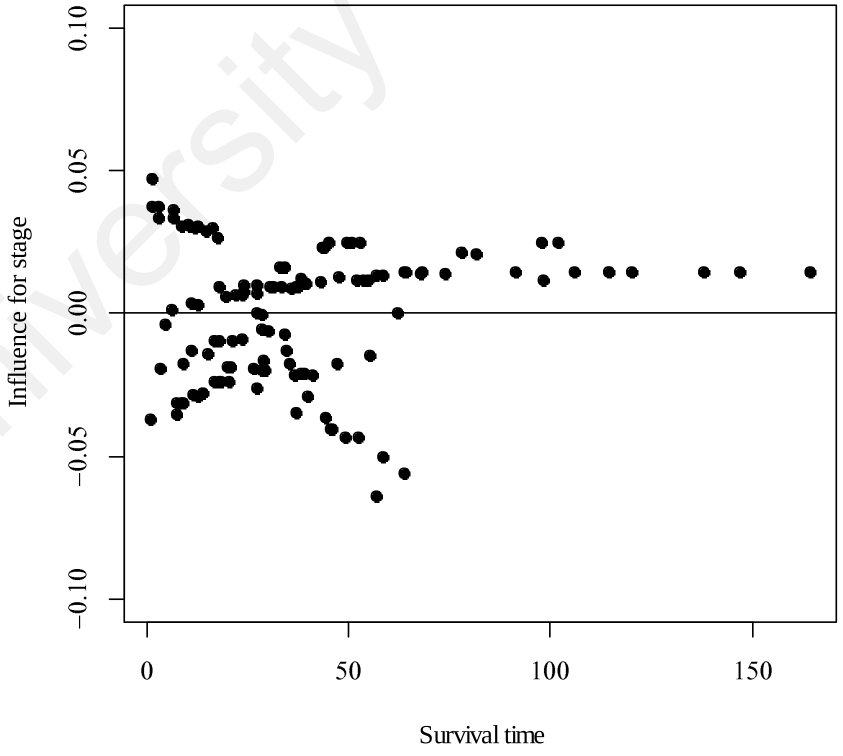


Figure 3.9: Plot of the delta-betas for stage at diagnosis against survival time for the stratified Cox model

3.5.6 Final Model

After checking for the model fitness, the final model is obtained. The result is given in Table 3.10. In this model, the histologic type and stage at diagnosis are found to be significant prognostic factors that affect the survival of cervical cancer patients under study. A patient in adenocarcinoma group has 1.921 times the hazard faced by patients in squamous cell carcinoma type. Meanwhile, patients who are diagnosed with stage III-IV are at 2.508 times the risk of death as those in stage I-II.

Table 3.10: The final stratified Cox model

Variables	**Coefficient**	**SE**	**Adjusted HR (95% CI)**	***p*-value**
Histologic type				
SCC[a]				
ADC[b]	0.6529	0.2954	1.921 (1.077-3.428)	0.0271
Stage				
I-II				
III–IV	0.9195	0.2744	2.508 (1.465-4.294)	0.0008

[a]Squamous cell carcinoma, [b]Adenocarcinoma

All analyses are conducted using R software version 3.0.3. Figure 3.10 summarises the procedures that have been performed in developing the prognostic model of the cervical cancer patients in this study.

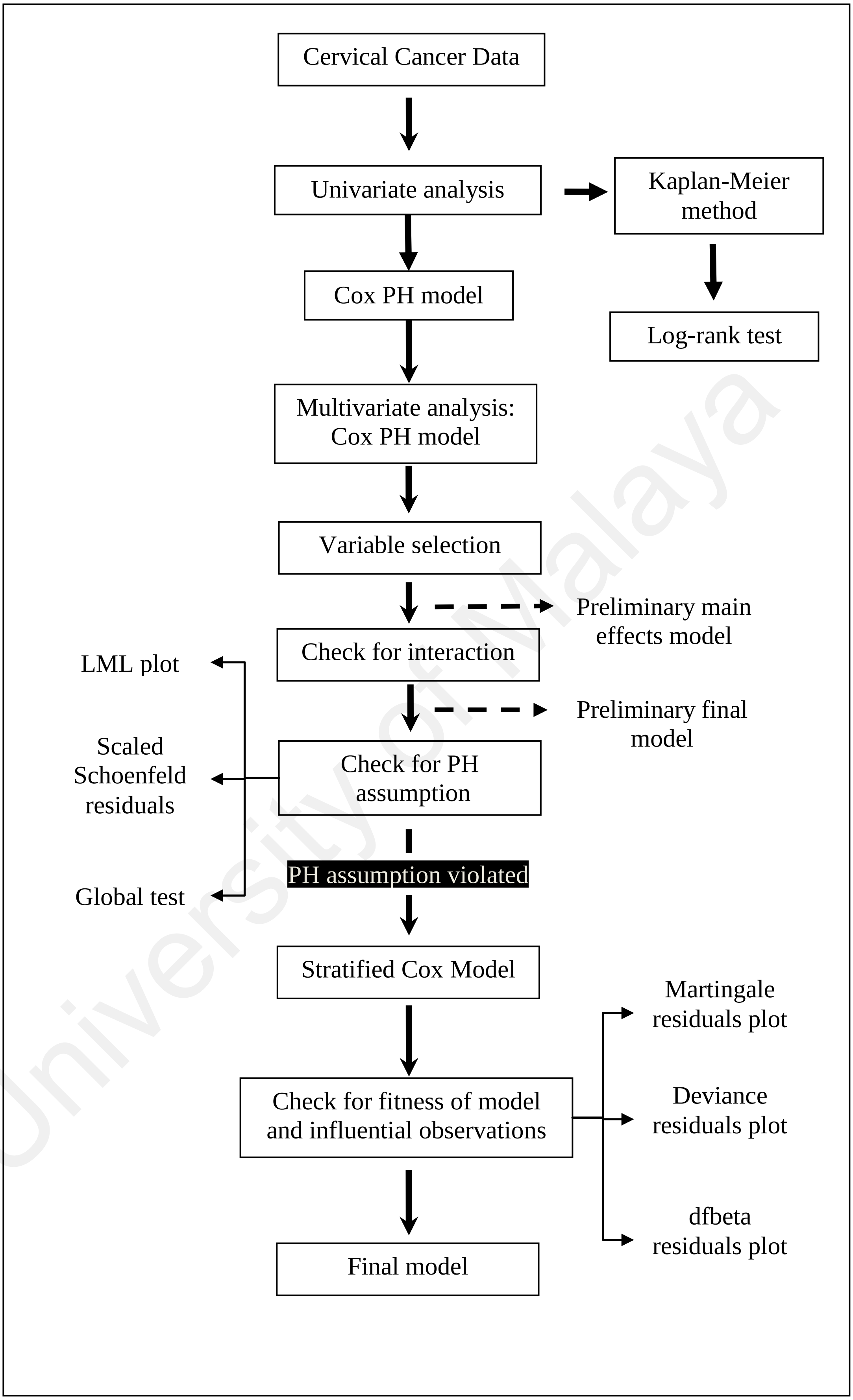


Figure 3.10: Steps involve in semi-parametric analysis of cervical cancer data

3.6 Discussion

The overall five-year survival of the 120 cervical cancer patients treated in HUSM is 39.7% with median survival time of 40.8 months. The finding of this study is almost similar to that of a study in Indonesia (Sirait *et al.*, 2003) and in India (Yeole *et al*., 2011) where the five-year survival is 40.3% and 42%, respectively. Meanwhile, the five-year survival of this study is slightly higher than that of a study in the Philippines, where the result was 34% (Laudico & Mapua, 2011).

The five-year survival of cervical cancer patients in this study is slightly lower than patients treated in the University of Malaya Medical Centre with a five-year survival of 50% (Wan Zamaniah *et al*., 2014). Similarly, in comparison with the finding of a study in Bulgaria (47.7%) that was done by Kostova *et al*. (2008), our result is low. Also, the five-year survival of our study is low compared to the overall five-year survival in other countries in Asia such as Hong Kong, the Republic of Korea and Singapore where the survival exceeded 65% (Sankaranarayanan *et al*., 2011). In addition, the average five-year survival in 23 European countries is 63% (Sant *et al*., 2009) showing better survival than that of patients in this study. Meanwhile, the median survival time of this study is nearly similar to that of the study in Indonesia which is 1208 days (Sirait *et al*., 2003).

The finding of this study shows that the five-year survival according to stage I, II and III-IV is 54.7%, 40.8% and 18.4% respectively. The log-rank test also shows that the survival is significantly different. This result is found to be consistent with those of other studies (Kostova *et al*., 2008; Seamon *et al*., 2011). Our finding is similar to other studies (Chen *et al*., 1999; Sirait *et al*., 2003; Chung *et al*., 2006) where the survival decreases as the stage of the disease increased. Kumari *et al*. (2010) also found that stage at diagnosis significantly influenced the prognosis of cervical cancer patients. However, their five-year survival obtained according to the stage was higher compared to our study. In fact, the survival of patients diagnosed at an advanced stage (Stage IV) was considerably high where the survival rate was 33%.

In our study, higher five-year survival is observed in patients treated with surgery compared to non-surgical treatment and the survival difference is statistically significant. Flores-Luna *et al*. (2001) found that patients who underwent surgical treatment had better survival (85.7%) than those who received radiotherapy (62.5%). Large proportion of individuals in the surgery group was diagnosed at an early stage. Therefore, longer survival time in this group was noted. Furthermore, the percentage of dying in surgical treatment group was lower than another treatment group.

The survival difference between with and without distant metastasis groups is statistically significant based on the log-rank test. The five-year survival of patients who have distant metastasis is lower than those who have no distant metastasis. This study also discovered that there is no significant difference in survival for variables ethnicity, lymph node involvement, age, and histologic type. In contrast, Yeh *et al*. (1999) reported a significant difference in survival of patients with lymph node and those without lymph node involvement. Garipagaoglu *et al*. (1999) and Flores-Luna *et al*. (2001) also reported that age did not influence the survival of cervical cancer in their study. Garipagaoglu *et al*. (1999) claimed that the survival difference was not observed due to a very small number of patients in younger age group (< 40 years). In contrast, Brun *et al*. (2003) reported an opposite finding. It was identified that the percentage of younger patients was large in their study. Meanwhile, Galic *et al*. (2012) found that the five-year survival of adenocarcinoma group was lower compared to the squamous cell carcinoma group for both early and advanced stage of cancer.

In this study, the stratified Cox model is performed as the proportional hazards assumption is violated for distant metastasis variable. Thus, this variable is stratified and considered in the model as a stratification factor. As a result, the significant prognostic factors associated with the survival of those 120 cervical cancer patients are histologic type and stage at diagnosis.

This study found that the prognosis of cervical cancer depends significantly on the stage at diagnosis. Patient who is diagnosed with advanced stages (stage III-IV) of cancer has higher risk of death than early stages, stage I-II. This finding is also concurred with findings from other researches (Rijke *et al*., 2002; Grigienė *et al*., 2007; Katanyoo *et al*., 2012; Douine *et al*., 2014). It is worthwhile to note that a study of 515 cervical cancer patients by Dueňas-González *et al*. (2012) showed a significant result for advance stage III & IV with adjusted hazard ratio of 1.54 (95% CI= 1.11-2.14), indicating that patients with advanced stage of disease had a 54% higher risk of progression or death at any time than earlier stage patients. Study in Thailand by Pomros *et al*. (2007) showed a significant result for stage III with adjusted HR of 1.65 (95% CI: 1.05; 2.59). However, stage IV was found not significant probably due to small number of patients in that group.

This study also found that the histologic type is significantly affecting the survival, as patients diagnosed with adenocarcinoma are identified to have higher risk of dying compared to squamous cell carcinoma, which are supported by previous findings of other researches. As an example, Galic *et al*. (2012) found that adenocarcinoma had significant negative impacts on the prognosis of cervical cancer patients studied. Furthermore, Yamauchi *et al*. (2014) claimed that adenocarcinoma was associated with a worse prognosis because those who were diagnosed with such a histologic type were detected later and at more advanced stages than the squamous cell carcinoma. In addition, this histologic type is often associated with HPV 18 (Kyrgiou & Shafi, 2010) and has poorer prognosis than other HPV type (Schwartz *et al*., 2001). Chung *et al*. (2006) and Atahan *et al*. (2007) also reported that histologic type was a significant prognostic factor of cervical cancer in their studies.

3.7 Summary

Data of cervical cancer patients treated in HUSM are analysed using the Cox proportional hazards regression model. From this model, it has been found that the stage at diagnosis, histologic type and distant metastasis are the significant prognostic factors that influence the risk of dying of these patients. However, the distant metastasis variable does not satisfy the proportional hazards assumption, thus the stratified Cox model is adopted. The stage at diagnosis and histologic type remain as the significant prognostic factors in the model. Meanwhile, the distant metastasis is considered as the stratification factor. Therefore, the findings indicate that cervical cancer patients treated at HUSM with stage III-IV with adenocarcinoma type are at the greatest risk of death from cervical cancer.

CHAPTER 4

PARAMETRIC ANALYSIS OF CERVICAL CANCER DATA

4.1 Introduction

Parametric survival models assume that survival times follow specific statistical distribution. This type of survival model may consist of a proportional hazards model and an accelerated failure time (AFT) model. The choice between these two models depends on the distribution of the survival times. The proportional hazards model involves modelling the hazards, while a major concern of the AFT model is on modelling the survival times.

Therneau and Grambsch (2000) listed several alternative methods to deal with nonproportional hazards data. Accelerated failure time model is one of the options that has been suggested since this model ignores the proportional hazards assumption (except for the exponential and Weibull models). The AFT model assumes that the predictors act multiplicatively on the survival times, which may be interpreted as the speed of progression of an individual along the time axis (Collet, 2003).

In this chapter, data of the 120 cervical cancer patients treated in HUSM that has been evaluated in Chapter 3 are further analysed using parametric survival models. The Weibull, log-logistic and lognormal models, are considered. The suitability of the aforementioned parametric models is checked in Section 4.2. In Section 4.3, the univariate and multivariate analyses are performed for each parametric model. All multivariate models are compared using Akaike information criterion (AIC) statistic to determine the best fitting model for the data. In Section 4.4, the best fit parametric survival model that has been obtained in Section 4.3 is compared with the stratified Cox model presented in Chapter 3. Discussions are given in Section 4.5, and summary of the findings is provided at the end of the chapter, in Section 4.6.

4.2 The Suitability of the Parametric Model

The suitability of the Weibull model is gauged based on the log-cumulative hazard plot. Figure 4.1 shows that the relationship between the $\log \hat{H}(t)$ and $\log t$ is approximately linear, suggesting that the Weibull assumption may be suitable.

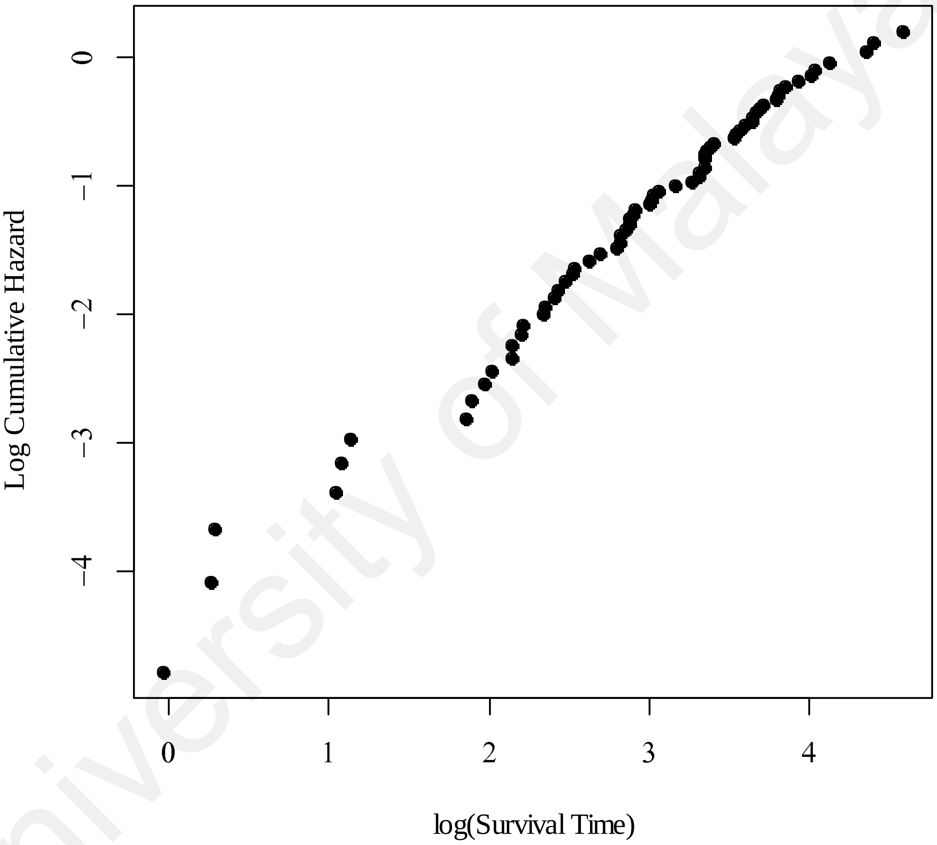


Figure 4.1: Log-cumulative hazard plot

The log-logistic model is also fitted to the data of the cervical cancer patients in this study. The suitability of the log-logistic model is assessed by a plot of the log-odds of survival against the log of survival time which is given in Figure 4.2. The plot indicates a remarkably straight line suggesting that the survival times may be appropriate to be modelled using the log-logistic distribution.

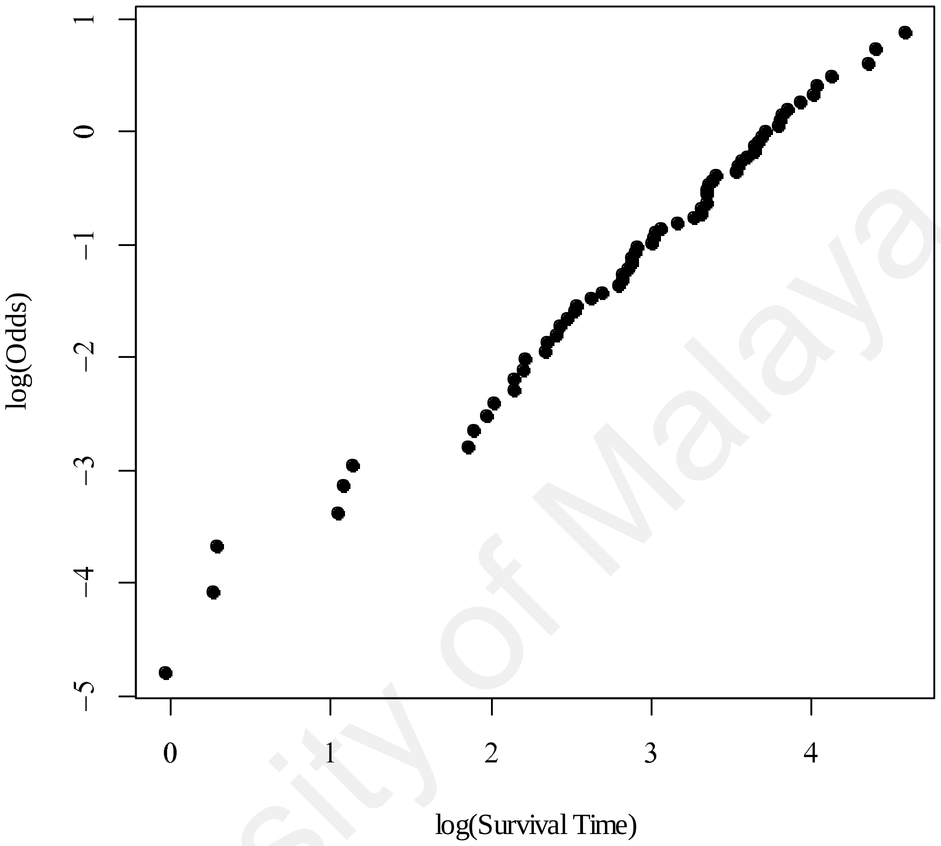


Figure 4.2: Log-odds of survival against the log of survival time plot

Figure 4.3 illustrates the plot of $\Phi^{-1}\left\{1-\exp\left(-\hat{H}(t)\right)\right\}$ against the log of survival time that is used to check for the suitability of the lognormal model. The plot gives a reasonably straight line suggesting that the survival times may also be appropriate to be modelled using the lognormal distribution.

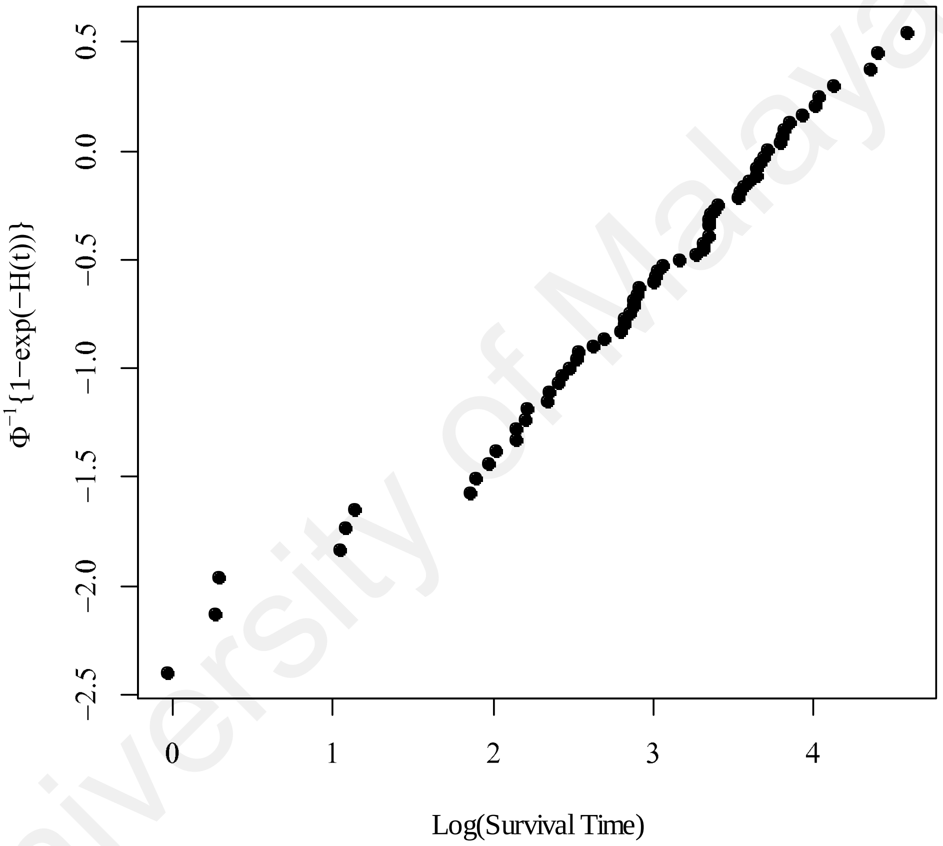


Figure 4.3: Plot of $\Phi^{-1}\left\{1-\exp\left(-\hat{H}(t)\right)\right\}$ against the log of survival time

4.3 Parametric Survival Regression Model

The accelerated failure time (AFT) model is considered to model the effect of stage at diagnosis, ethnicity, histologic type, lymph node involvement, age at diagnosis, distant metastasis and primary treatment on the survival times of the 120 cervical cancer patients treated in HUSM.

4.3.1 Univariate Parametric Survival Models

Univariate analysis is performed to obtain a preliminary idea of which factor that may be of prognostic value. Factors that are considered in these analyses are ethnicity, lymph node involvement, histologic type, age at diagnosis, primary treatment received, stage at diagnosis and distant metastasis. Each of these variables is analysed using the Weibull, log-logistic and lognormal models separately.

Results of the univariate analyses are presented in Table 4.1. In the case of the Weibull model, it has been found that variables with *p*-value less than 0.05 are stage at diagnosis, primary treatment and distant metastasis. Meanwhile, only the stage at diagnosis and primary treatment are statistically significant in the log-logistic and lognormal models.

Table 4.1: The univariate analyses of parametric models results

Variables	Weibull				Log-logistic				Lognormal			
	α[a]	TR[b]	95%CI[c]	*p*-value	*α*	TR	95% CI	*p*-value	*α*	TR	95%CI	*p*-value
Ethnicity												
Non-Malay												
Malay	0.21	1.24	0.62-2.48	0.551	0.08	1.09	0.53-2.24	0.821	0.12	1.12	0.52;2.42	0.769
Lymph node												
Negative												
Positive	0.14	1.14	0.64-2.05	0.652	0.12	1.13	0.61-2.08	0.698	0.18	1.20	0.63-2.28	0.580
Histologic												
SCC[d]												
ADC[e]	-0.40	0.67	0.38-1.18	0.166	-0.33	0.72	0.39-1.32	0.285	-0.36	0.70	0.36-1.34	0.280
Age												
≤ 39 years												
40 – 49	0.62	1.85	0.88--3.90	0.106	0.80	2.22	0.98-5.04	0.057	0.76	2.13	0.90-5.04	0.085
50 – 59	0.41	1.50	0.70-3.21	0.292	0.48	1.62	0.70-3.75	0.257	0.52	1.68	0.70-4.03	0.249
≥ 60	0.48	1.62	0.66-3.98	0.295	0.80	2.22	0.88-5.60	0.093	0.92	2.50	0.92-6.82	0.073
Stage												
I-II												
III-IV	-0.92	0.40	0.24-0.66	<0.0001	-0.95	0.39	0.22-0.68	0.001	-1.02	0.36	0.20-0.64	0.001
Treatment												
Surgery												
Non-surgery	-0.80	0.45	0.27-0.76	0.003	-0.69	0.50	0.28-0.88	0.017	-0.78	0.46	0.26-0.82	0.009
Metastasis												
No												
Yes	-0.73	0.48	0.29-0.79	0.004	-0.51	0.60	0.35-1.02	0.062	-0.46	0.63	0.35-1.15	0.131

α[a] = regression coefficient, TR[b]= time ratio, CI[c]= confidence interval, SCC[d]=Squamous cell carcinoma, ADC[e]=Adenocarcinoma

4.3.2 Multivariate Parametric Survival Models

In the multivariate analyses, the full model that contains all covariates is developed for each parametric survival model. Results for these full models are tabulated in Table 4.2. The histologic type, stage at diagnosis and distant metastasis are statistically significant $(p\text{-value} < 0.05)$ in the Weibull model. Meanwhile, for the log-logistic model, only the stage at diagnosis variable is statistically significant. In the case of the lognormal model, the significant prognostic factors are stage at diagnosis and age at diagnosis 60 years and older. It is worthwhile to note that, although the primary treatment is statistically significant in the univariate analyses for all models, the treatment effect is not statistically significant in the multivariate analyses.

Table 4.2: The multivariate analyses of full parametric models results

Variables	Weibull				Log-logistic				Lognormal			
	α[a]	TR[b]	95%CI[c]	*p*-value	*α*	TR	95% CI	*p*-value	*α*	TR	95%CI	*p*-value
Ethnicity												
Non-Malay												
Malay	-0.44	0.64	0.31-1.31	0.226	-0.53	0.59	0.25-1.36	0.216	-0.41	0.66	0.29-1.50	0.326
Lymph node												
Negative												
Positive	0.11	1.12	0.63-1.99	0.713	0.08	1.08	0.58-2.02	0.799	0.16	1.17	0.63-2.18	0.616
Histologic												
SCC[d]												
ADC[e]	-0.71	0.49	0.30-0.82	0.006	-0.54	0.58	0.33-1.04	0.067	-0.59	0.55	0.30-1.02	0.057
Age												
≤ 39 years												
40 – 49	0.48	1.62	0.77-3.41	0.206	0.59	1.80	0.78-4.14	0.169	0.75	2.12	0.91-4.91	0.081
50 – 59	0.35	1.42	0.69-2.95	0.34	0.36	1.43	0.62-3.29	0.404	0.52	1.68	0.72-3.93	0.229
≥ 60	0.55	1.73	0.76-3.96	0.193	0.74	2.10	0.86-5.10	0.105	1.00	2.72	1.07-6.88	0.035
Stage												
I-II												
III-IV	-0.86	0.42	0.25-0.72	0.002	-0.95	0.39	0.21-0.72	0.003	-0.98	0.38	0.20-0.70	0.002
Treatment												
Surgery												
Non-surgery	-0.49	0.61	0.36-1.05	0.077	-0.41	0.66	0.37-1.20	0.176	-0.44	0.64	0.35-1.16	0.143
Metastasis												
No												
Yes	-0.60	0.55	0.34-0.88	0.013	-0.35	0.70	0.41-1.19	0.187	-0.30	0.74	0.43-1.29	0.289

α[a] = regression coefficient, TR[b]= time ratio, CI[c]= confidence interval, SCC[d]=Squamous cell carcinoma, ADC[e]=Adenocarcinoma

4.3.3 Comparison of Parametric Survival Models

The Akaike Information Criterion (AIC) is used to compare and find the best fitted model among the Weibull, log-logistic and lognormal models. The AIC statistic is computed as follows:

$$AIC = -2\left(\text{log-likelihood}\right) + 2\left(c + k\right), \qquad (4.1)$$

where c is the number of unknown parameters (coefficient regressions), and k is the number of other parameters.

In determining the best fitted model by comparing models with reduced variables from backward, forward or stepwise selection methods may not be appropriate as different model (reduced variables) contains different number of significant variables. This means, comparison using AIC statistic from models with different number of variables may not be a proper approach (Lee and Wang, 2003). Thus, for this study, the AIC value is computed for each full parametric model (all variables) in Table 4.2, and these values are presented in Table 4.3. The smallest AIC statistic is observed for the Weibull model. The AIC value for the lognormal model is slightly higher than the log-logistic model. This result suggests that the Weibull model fits the data better compared to the log-logistic and lognormal models.

Table 4.3: The Akaike information criterion value for each parametric model

Model	Log-likelihood	AIC
Weibull	-332.2703	686.5406
Log-logistic	-333.1028	688.2056
Lognormal	-333.0527	688.1054

4.3.4 The Weibull Model

Further analysis is conducted for the Weibull model, where the variables which are significantly associated with the time to death of the cervical cancer patients are selected using the stepwise selection method. The histologic type, stage at diagnosis and distant metastasis are found to be significant (p-value < 0.05), and the result is given in Table 4.4. There is also no evidence of interaction among the variables in the model.

Table 4.4: The multivariate Weibull model

Variables	α^a	TR[b]	95% CI[c]	p-value
Histologic SCC[d] ADC[e]	-0.62	0.54	0.32-0.90	0.019
Stage I-II III-IV	-0.89	0.41	0.25-0.66	<0.0001
Distant metastasis No Yes	-0.74	0.48	0.30-0.76	0.002

α^a = regression coefficient, TR[b]= time ratio, CI[c]= confidence interval, SCC[d]=Squamous cell carcinoma, ADC[e]=Adenocarcinoma

The unique property of the Weibull model is that, if the proportional hazards assumption holds, then the AFT assumption also hold and vice versa (Kleinbaum & Klein, 2005). This means that a Weibull proportional hazards model is equivalent to a Weibull AFT model. It has been found that the proportional hazards assumption is violated for distant metastasis variable. Figure 4.4 illustrates the log-cumulative hazard plot for distant metastasis. The plot indicates that the two lines are straight, yet not parallel. This plot suggests that the Weibull assumption holds, but the proportional hazards may be violated (Kleinbaum & Klein, 2005).

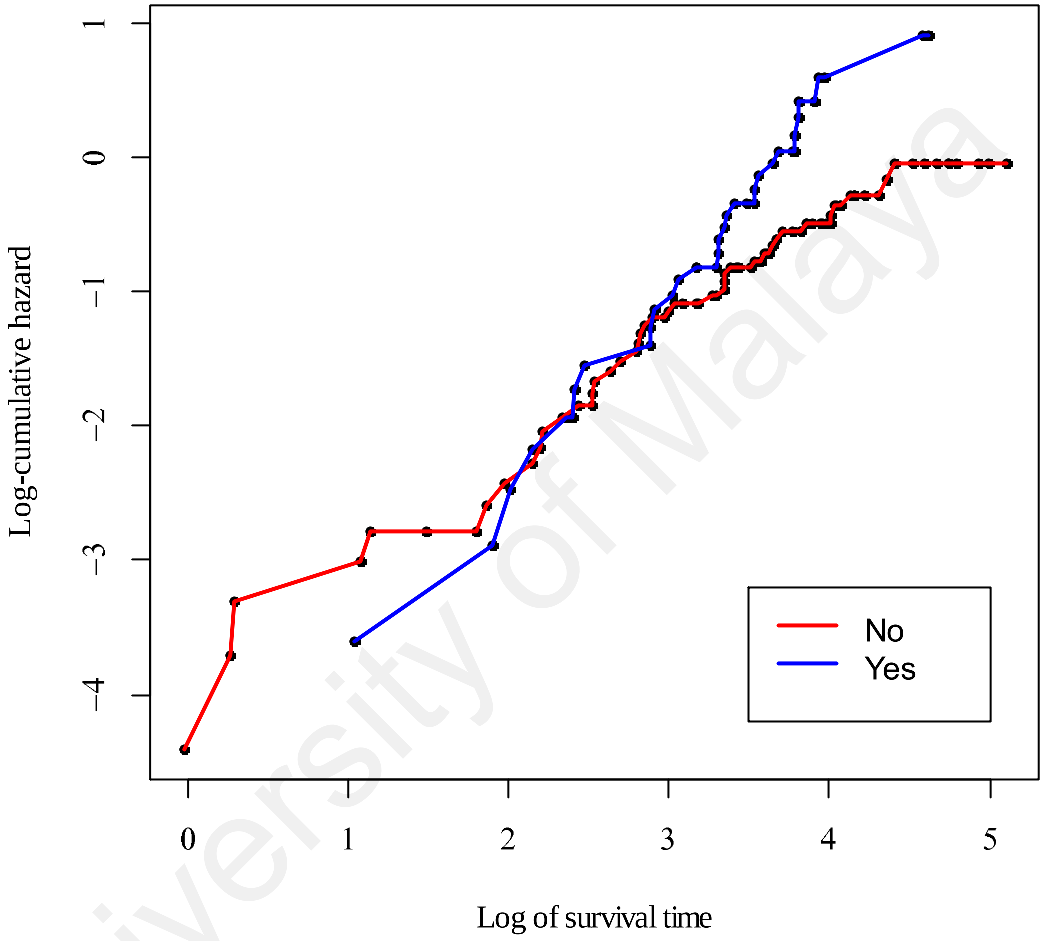


Figure 4.4: Log-cumulative hazard plot for distant metastasis

Apart from that, a likelihood ratio test is conducted to check whether the value of the scale parameter for the Weibull model with common linear component (histologic type and stage) for without distant metastasis group is similar to with distant metastasis group. If the scale parameter is similar, then the proportional hazards assumption is satisfied. From the test, it is confirmed that the two scale parameters are not equal (*p*-value=0.0003). This means the proportional hazards assumption is violated. Thus, a standard Weibull AFT model may not be suitable to model the data of the cervical cancer patients.

As the scale parameters are not identical between the model with and without distant metastasis, a stratified Weibull model is adopted. This model is developed by stratifying the distant metastasis variable while histologic type and stage at diagnosis remain as the covariates in the model. It is noted that, the stratification allows for a separate scale parameters for each distant metastasis group, yet the coefficients are assumed to be the same across the group. The stratified Weibull model is presented in Table 4.5. The estimated scale parameter for without metastasis (DM=0) group is 1.119, while for with distant metastasis (DM=1) group is 0.689.

Table 4.5: The stratified Weibull model

Variables	α^a	**TR**^{b}	**95% CI**^{c}	***p*-value**
Histologic type				
SCC^d				
ADC^e	-0.55	0.58	0.34-0.98	0.0429
Stage at diagnosis				
I-II				
III-IV	-0.98	0.37	0.23-0.60	<0.0001

α^a = regression coefficient, TR^b= time ratio, CI^c= confidence interval, SCC^d=Squamous cell carcinoma, ADC^e=Adenocarcinoma

4.3.5 Assessment of Model Adequacy

Fitness of the stratified Weibull model, outliers and influential observations are checked using the plot of martingale, deviance and delta-beta residuals graphically. Plots of the martingale residuals against the survival times are shown in Figure 4.5. According to Collet (2003), the martingale residuals for the parametric AFT model are not symmetrically distributed about zero, as observed in Figure 4.5. Figure 4.6 indicates that the deviance residuals lie between 3 to -2 suggesting no wildly deviant observations.

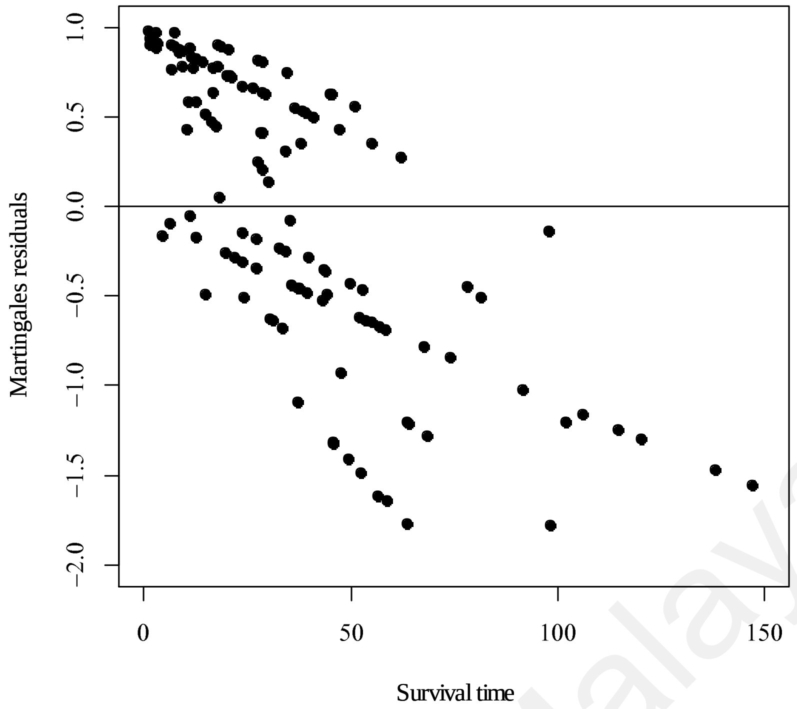


Figure 4.5: Plot of the martingale residuals against survival time

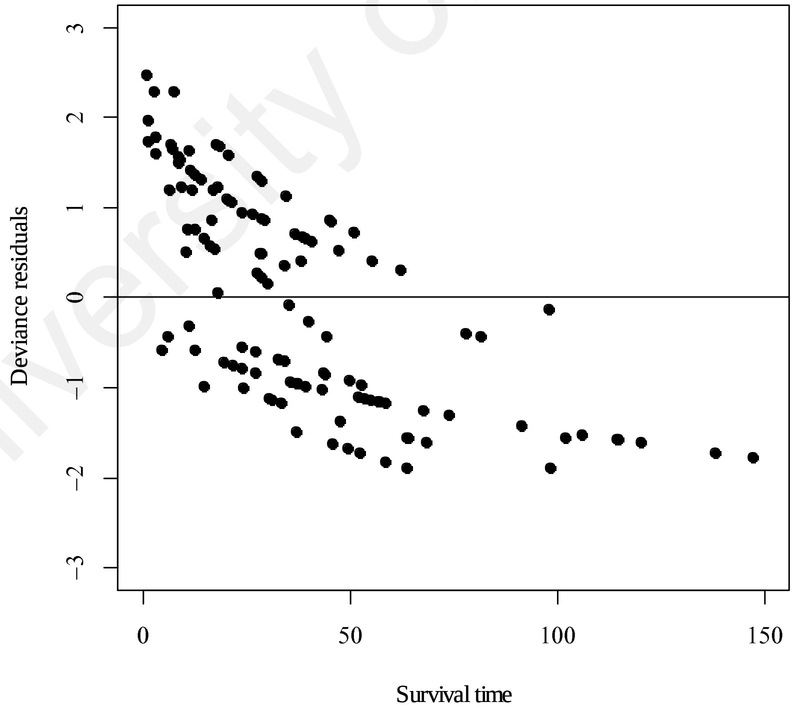


Figure 4.6: Plot of the deviance residuals against survival time for the stratified Weibull model

Figure 4.7 and Figure 4.8 show the delta-betas residuals plot against the survival times for the histologic type and stage at diagnosis, respectively. These plots show that there are no influential observations since all residuals lie within -0.1 to 0.1.

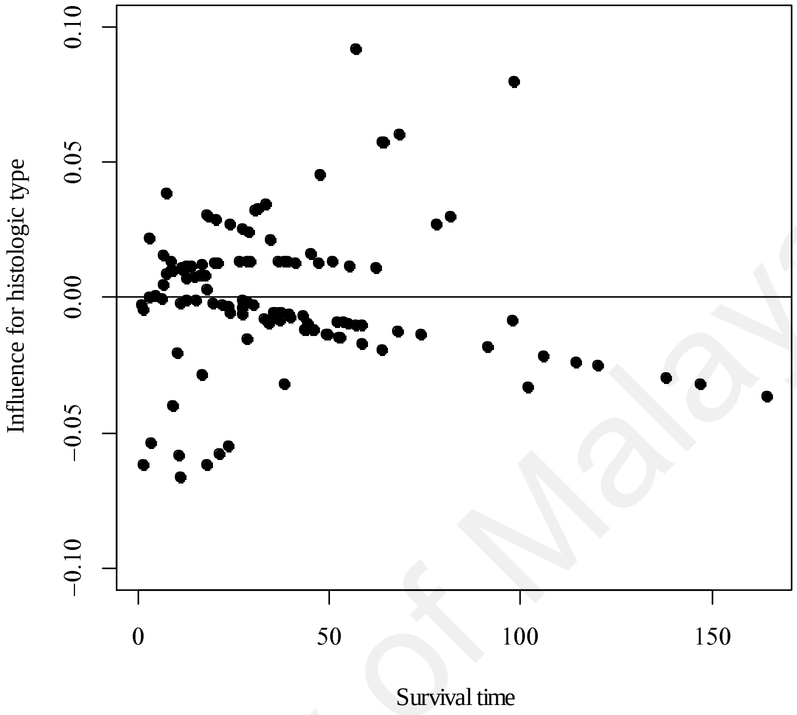


Figure 4.7: Plot of the delta-betas residuals for the histologic type for the stratified Weibull model

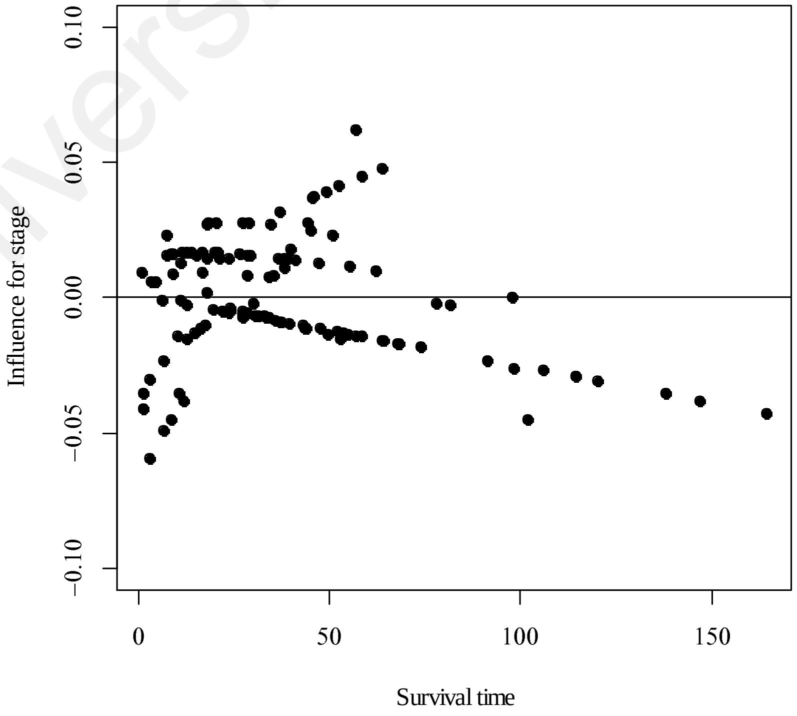


Figure 4.8: Plot of the delta-betas residuals for the stage at diagnosis for the stratified Weibull model

4.3.6 Final Model

Table 4.6 shows the final model obtained from the stratified Weibull model. The survival time for a patient who is diagnosed at stage III-IV is estimated to be 37% of that of a patient who was diagnosed with stage I-II. The estimated time ratio for the histologic type group is 0.58 which indicates that the earlier time to death is more likely for the patient with an adenocarcinoma type. Meanwhile, the scale parameter for without distant metastasis and with distant metastasis stratum is $\sigma_1 = 1.119$ and $\sigma_2 = 0.689$ respectively.

Table 4.6: The final stratified Weibull model

Variables	α^a	**SE**^{b}	**TR**^{c}**(95% CI**^{d}**)**	***p*-value**
Histologic type				
SCC^e				
ADC^f	-0.55	0.27	0.58 (0.34-0.98)	0.0429
Stage at diagnosis				
I-II				
III-IV	-0.98	0.24	0.37 (0.23-0.60)	<0.0001

α^a = regression coefficient, SE^b=standard error, TR^c= time ratio, CI^d= confidence interval, SCC^e=Squamous cell carcinoma, ADC^f=Adenocarcinoma

All analyses are conducted using the R software version 3.0.3. Figure 4.9 summarises the procedures that have been performed in identifying the best parametric model.

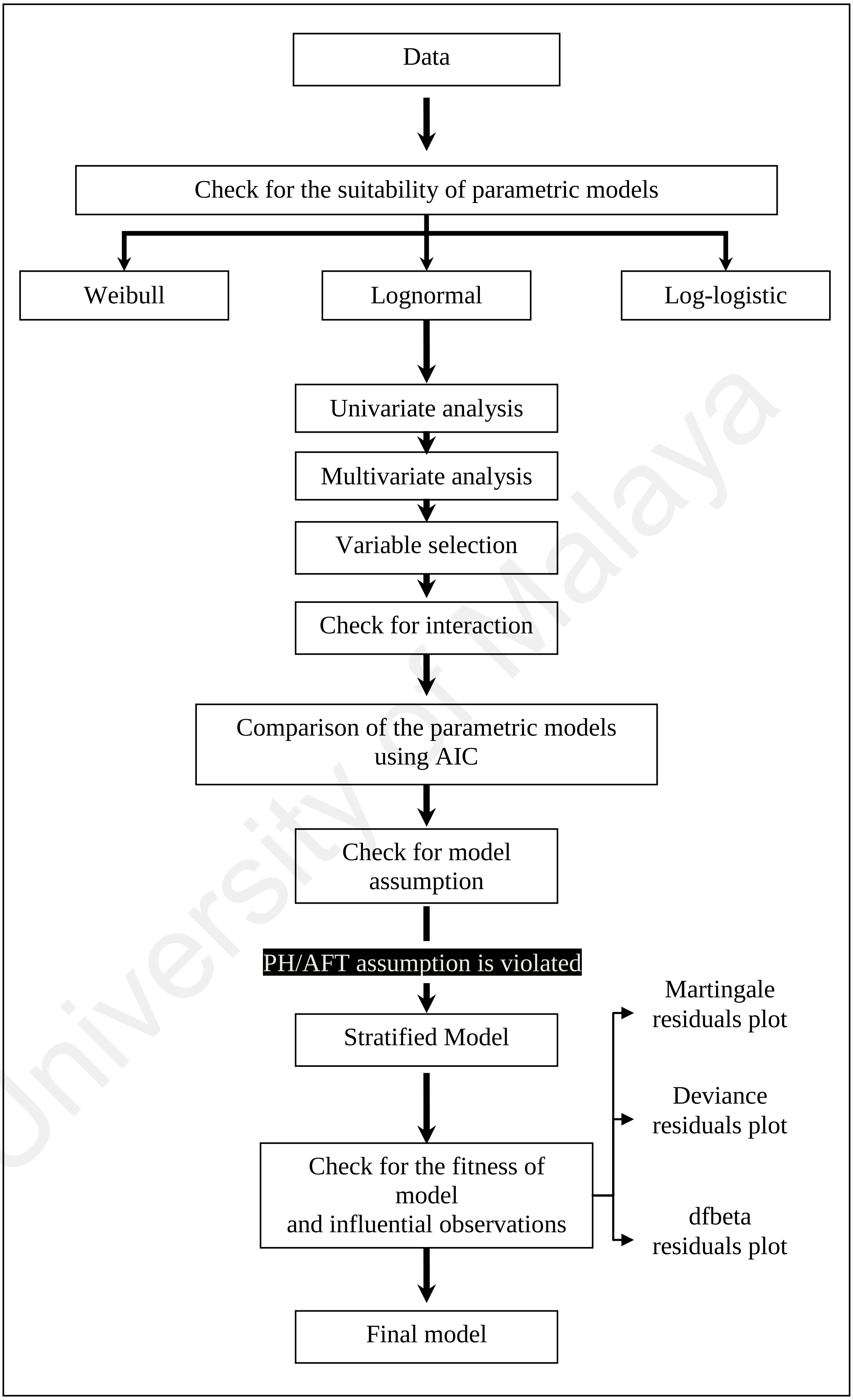

Data
Check for the suitability of parametric models
Weibull
Lognormal
Log-logistic
Univariate analysis
Multivariate analysis
Variable selection
Check for interaction
Comparison of the parametric models using AIC
Check for model assumption
PH/AFT assumption is violated
Stratified Model
Martingale residuals plot
Deviance residuals plot
Check for the fitness of model and influential observations
dfbeta residuals plot
Final model

Figure 4.9: Steps involve in parametric analyses of cervical cancer data

4.4 Comparison between the Stratified Cox and Stratified Weibull Models

The performance of the stratified Cox model and the stratified Weibull model are assessed further by examining the Cox-Snell residuals plots. Figure 4.10 and Figure 4.11 illustrate the Cox-Snell residuals plots for the stratified Weibull model for without metastasis and with distant metastasis stratum, respectively. The residuals for each group are computed separately as the scale parameters are different across the strata. Figure 4.10 shows that the line connecting the points is reasonably straight and nearly close to a straight line with zero intercept and slope equal to one. Meanwhile, the jagged line in Figure 4.11 is slightly deviates from the reference line.

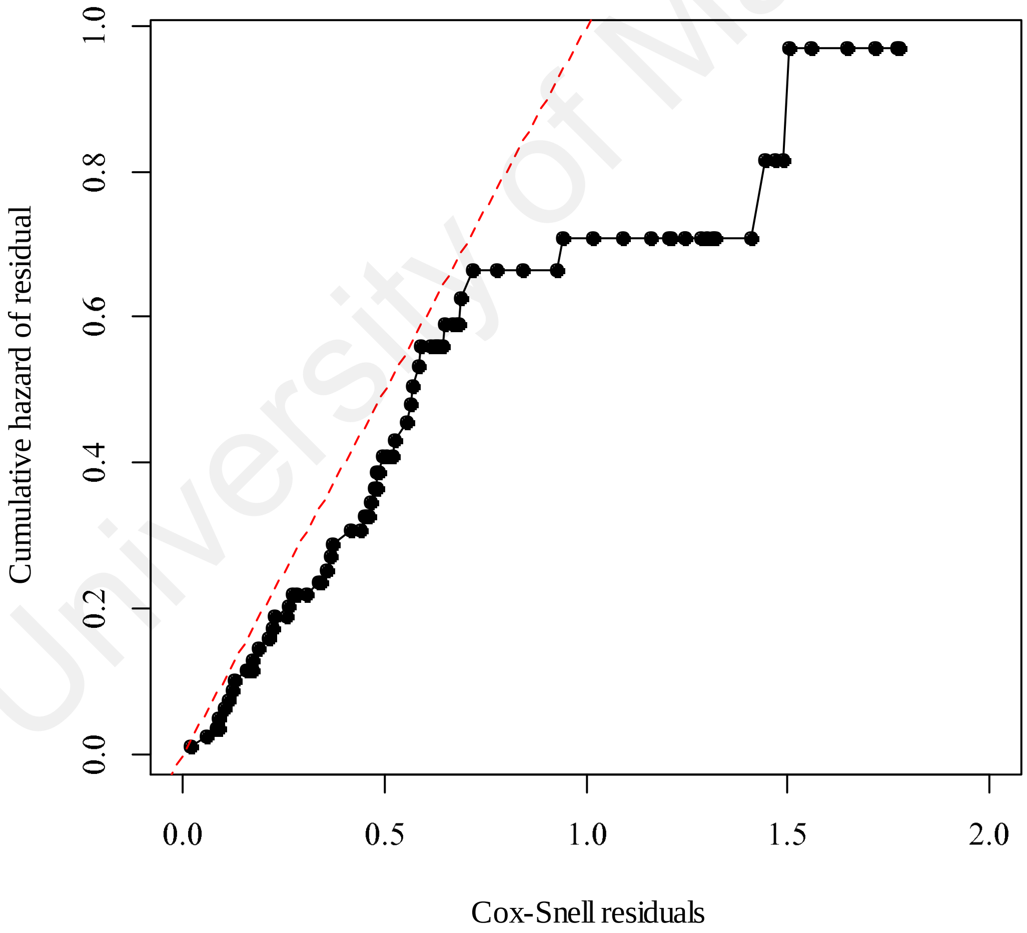


Figure 4.10: The Cox-Snell residuals plot for without distant metastasis stratum for the stratified Weibull model

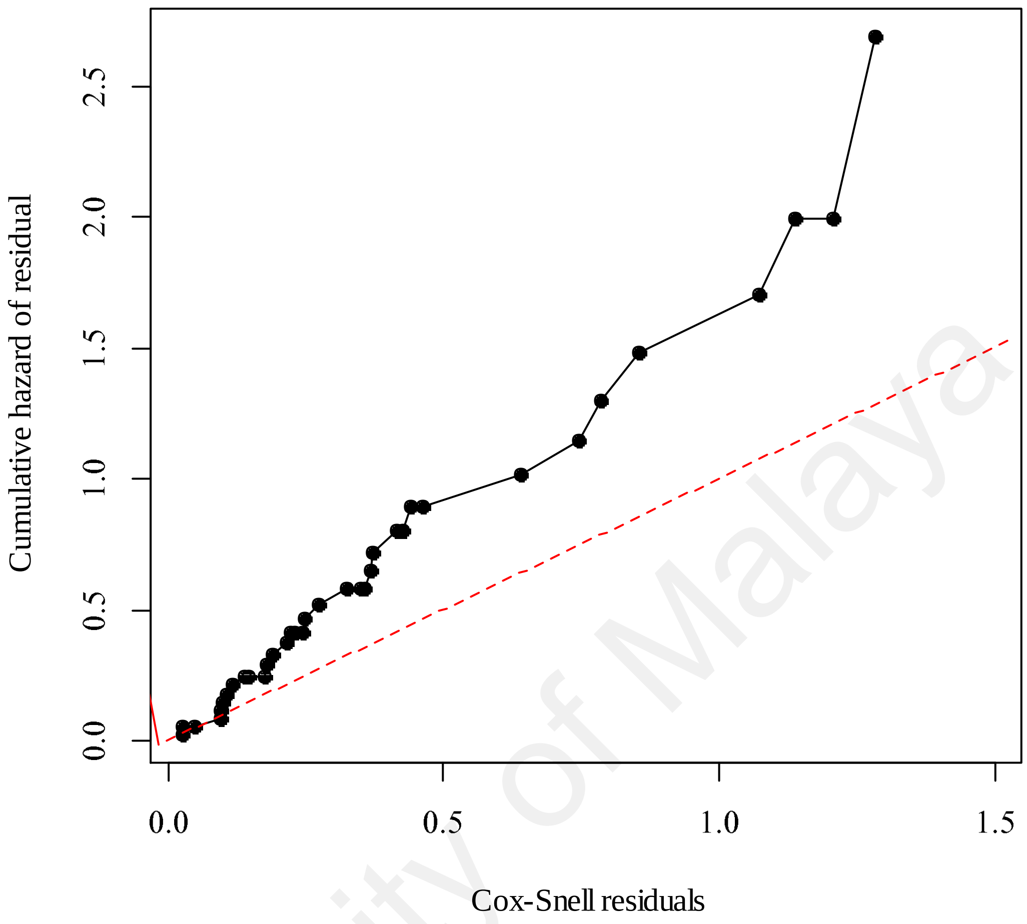


Figure 4.11: The Cox-Snell residuals plot for with distant metastasis stratum for the stratified Weibull model

The plot of the Cox-Snell residuals for the stratified Cox model is given in Figure 4.12. This figure shows that the majority of the plotted points lie on a 45-degree straight line through the origin suggesting that the stratified Cox model may fit the data relatively well. Therefore, based on these Cox-Snell residuals plots, it may be concluded that the stratified Cox model is the best model for the data of the 120 cervical cancer patients treated in HUSM compared to the stratified Weibull model.

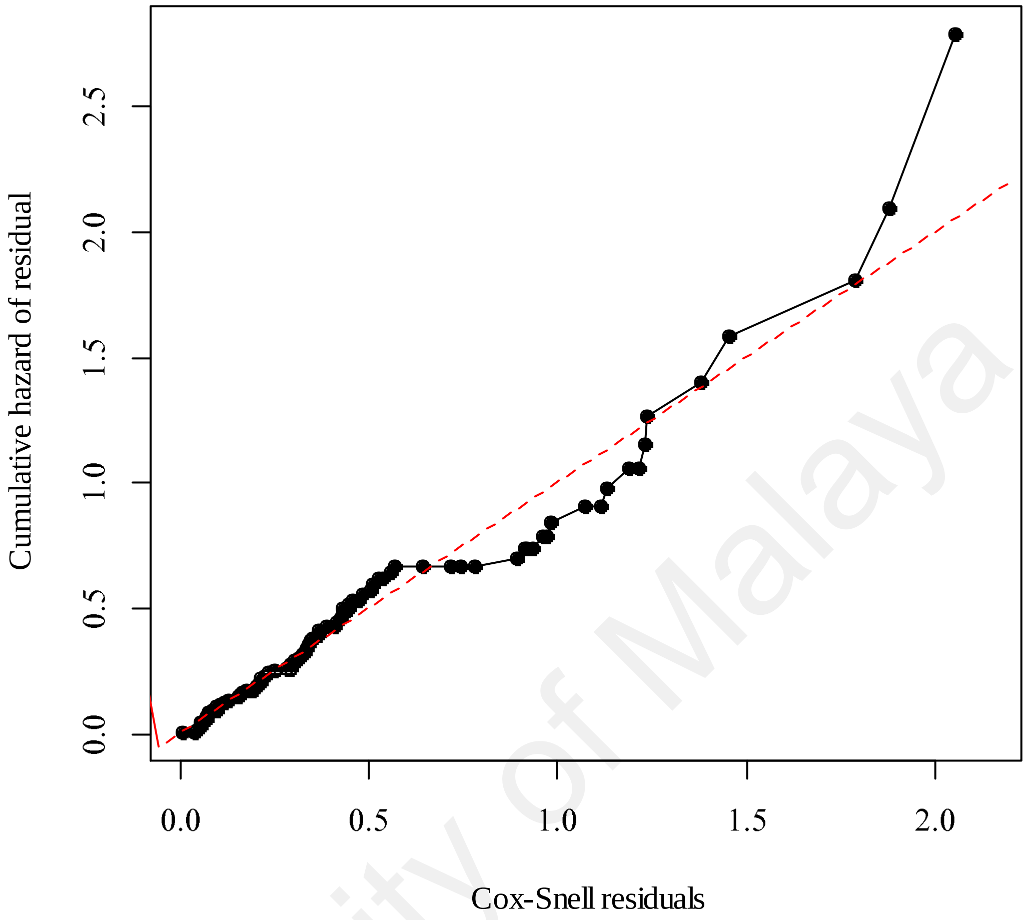


Figure 4.12: The Cox-Snell residuals plot for the stratified Cox model

4.5 Discussion

Data of the 120 cervical cancer patients treated in HUSM are analysed using parametric AFT models namely the Weibull, log-logistic and lognormal model. The suitability of the distributions is checked based on the plot of the cumulative hazard function (or a function of it) against the log of survival time. All plots give a reasonable straight line suggesting that the data may be appropriate to be modelled using the Weibull, lognormal and log-logistic model.

The stage at diagnosis and primary treatment are statistically significant in all univariate models. In addition, the distant metastasis variable is also significant in the univariate Weibull model. In the multivariate analysis, there are clear distinctions among different parametric models in the significant covariates. The stage at diagnosis, histologic type and distant metastasis are found to be significant factors in the Weibull model. For the lognormal model, the stage at diagnosis and age at diagnosis 60 years and older are significant. Meanwhile, only the stage at diagnosis variable is significant in the log-logistic model. The differences of the significant factors obtained from different models are also noted in Ravangard *et al*. (2011), Wang *et al*. (2011) and Köhler and Kowalski (2012).

It is worthwhile to note that, even though primary treatment variable is significant for all models in the univariate analysis, it is found to be not significant in any multivariate models. Meanwhile, stage at diagnosis may be regarded as the most important factor that affect the survival of cervical cancer patients since this variable is found statistically significant at all level of analyses (univariate and multivariate analysis) and type of models. Of these three parametric models, the Weibull model is the best fitted model because this model gives the smallest AIC value.

The development of the Weibull model is continued by checking the proportional hazards assumption for each significant variable identified in the multivariate model. It has been found that the distant metastasis variable does not satisfy the proportional hazards assumption, thus the AFT assumption is also violated. The violation of the proportional hazards assumption is confirmed by the likelihood ratio test since the scale parameters are different for without and with distant metastasis group.

Our study shows that only the Weibull model is able to detect the non-proportional hazards covariate, while other AFT models are not. In contrast, Köhler and Kowalski (2012) found that the non-proportional hazards covariates are significant in every AFT models tested in their study. Nardi and Schemper (2003) also emphasised that the AFT model tend to detect the significant effect of the non-proportional hazards covariate. However, Moran *et al*. (2008) supports our study by showing the less ability of the AFT model to detect the non-proportional covariates in comparison to the proportional hazards model.

Many studies pointed out that if the assumption of the proportional hazards is violated, the parametric model may be adopted (Pourhoseingholi *et al*., 2007; Moghimi-Dehkordi *et al*., 2008). This is not always true for the Weibull model because both assumptions, the proportional hazards and AFT, are used interchangeably. Qi (2009) reported that the performance of the Weibull model was poorer than other parametric models as the proportional hazards assumption was violated. Contradictory, Sayehmiri *et al*. (2008) observed that the Weibull model fitted the data well even though there was a non-proportional hazards covariate in the model, which is in parallel with our finding. However, they reported only the result without proposing any proper method to handle the non-proportional hazards covariate in the model.

In our study, a stratified model is proposed to relax the proportional hazards assumption of the Weibull model in which the distant metastasis variable became a stratification factor. Stage at diagnosis and histologic type remain as the covariates in the model. The final model of stratified Weibull indicates that shorter survival time are more likely for the patient who is diagnosed at stage III-IV than those in early stages (Stage I-II). Similarly, those who are diagnosed with adenocarcinoma had an earlier time to death compared to squamous carcinoma. The interpretation of an AFT model is easier and more meaningful for the clinician or medical practitioners to understand since it measures the direct effect of the covariates on the survival time (Hosmer & Lemeshow, 1999).

The set of covariates that are statistically significant in the Weibull model are also significant in the Cox model as presented in Chapter 3. Many studies are conducted to compare the performance of the Cox and parametric models. In fact, majority of these studies reported that the parametric model is better compared to the Cox model (Pourhoseingholi *et al.*, 2007, 2009, 2011; Wang *et al.*, 2011; Zhu *et al.*, 2011; Hashemian *et al.*, 2013). However, most of these studies rely on the AIC statistic to choose the best model which may be inappropriate for comparing between semi-parametric and parametric models (Bradburn *et al.*, 2003b; Royston & Lambert, 2011).

In this study, the Cox-Snell residuals plot is examined to check the performance of the stratified Weibull model and stratified Cox model. Ravangard *et al*. (2011), and Köhler and Kowalski (2012) are examples of study that used the Cox-Snell residual plot to compare the performance of semi-parametric and parametric models. Unlike the aforementioned studies, the Cox-Snell residuals plots show that the stratified Cox model fitted the cervical cancer data better than the stratified Weibull model. One possible reason for the poor fit of the parametric model is due to the percentage of censored observations in our study which is 45%. Nardi and Schemper (2003) revealed that the parametric model may fit the data well when the percentage of censored observations is less than 40 to 50 percent.

Similar to our study, Moran *et al*. (2008) also discovered that the Cox model with non-proportional hazards covariate is better than the parametric AFT models.

According to Moran *et al*. (2008), one of the reason for the contradictory findings of other studies is that the standard Cox proportional hazards model still being adopted even though there was a significant covariate that violated the proportional hazards in the model as demonstrated in Köhler and Kowalski (2012) study. However, with a proper treatment through the extended Cox's model that takes into account the non-proportional hazards effect may yield an acceptable result (Bellera *et al*., 2010).

4.6 Summary

The Weibull model fits the data of cervical cancer patients treated in HUSM better compared to the log-normal and log-logistic models. The stage at diagnosis, histologic type and distant metastasis are the significant factors in the Weibull model. However, since distant metastasis variable violates the proportional hazards assumption, the stratified Weibull model is applied. Also, this study is interested to identify the best model between the stratified Cox model that has been obtained from Chapter 3 and stratified Weibull model. Based on the Cox-Snell residuals plots, the stratified Cox model exhibits a better fit than the stratified Weibull model.

CHAPTER 5

MISSING VALUES IN PARAMETRIC SURVIVAL MODEL

5.1 Introduction

Missing covariate values are common in survival data. In this chapter, four methods for handling missing covariates values namely the complete case analysis, expectation-maximization (EM) algorithm by method of weight, hot deck and multiple imputation have been studied. These methods are investigated for the case of parametric model. The Weibull accelerated failure time (AFT) model is considered since this model is one of the most common parametric models being used in many survival studies. Also, this study focuses on missing categorical covariate values because most of the covariates in survival data are categorical. Data are assumed to be missing at random (MAR).

Survival function of the Weibull AFT model and its maximum likelihood estimation for a complete data model are provided in Section 5.2. Next, Section 5.3 describes the methods of the complete case analysis, EM algorithm by method of weight, hot deck and multiple imputation. Meanwhile, simulation studies and the results are presented in Section 5.4. Then, Section 5.5 illustrates these methods on the data of cervical cancer patients treated in HUSM that has been evaluated in Chapter 3 and Chapter 4. The summary of the chapter is given in Section 5.6.

5.2 The Weibull AFT Model

Let Y_i be the survival time for subject i where $i = 1, \ldots, n$, and has a Weibull distribution with parameters λ and σ. The hazard function for the Weibull model is

$$h\left(y_i\right) = \lambda\sigma^{-1} y_i^{\sigma^{-1}-1}, \tag{5.1}$$

where $\lambda = \exp\left\{-\sigma^{-1}\left(\beta_0 + \sum_{j=1}^{p} \boldsymbol{\beta}_j \boldsymbol{x}_j\right)\right\}$ and σ is the scale parameter. The survival function is given by

$$S\left(y_i\right) = \exp\left(-\lambda y_i^{\sigma^{-1}}\right). \tag{5.2}$$

Commonly, the observation of Y is censored by a variable C so that the observable outcomes are the observed event time $T = \min\left(Y, C\right)$. The censoring indicator is denoted by δ_i, which $\delta_i = 1$ if the observed event is a failure $\left(Y_i \le C_i\right)$ and $\delta_i = 0$ otherwise $\left(Y_i > C_i\right)$. Also, let $\boldsymbol{x}_i = \left(x_{i1}, x_{i2}, \ldots, x_{ip}\right)'$ represents a $p \times 1$ vector of covariates associated with T_i, and $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \ldots, \beta_p\right)'$ be a $p \times 1$ vector of regression coefficients. In the case of noninformative censoring, the probability density function for $\left(t_i, \delta_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}, \sigma\right)$ is given by

$$p\left(t_i, \delta_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}, \sigma\right) = \left\{h\left(t_i, \delta_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}, \sigma\right)\right\}^{\delta_i} \left\{S\left(t_i, \delta_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}, \sigma\right)\right\}. \tag{5.3}$$

Hence, from the hazard function in (5.1) and survival function in (5.2), the probability density function for the Weibull AFT model is given by

$$\begin{aligned} p\left(t_i, \delta_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}, \sigma\right) = &\left\{\exp\left[-\sigma^{-1}\left(\beta_0 + \sum_{j=1}^{p} \boldsymbol{\beta}_j \boldsymbol{x}_j\right)\right]\sigma^{-1} t^{\sigma^{-1}-1}\right\}^{\delta_i} \\ &\times \exp\left\{-t^{\sigma^{-1}} \exp\left[-\sigma^{-1}\left(\beta_0 + \sum_{j=1}^{p} \boldsymbol{\beta}_j \boldsymbol{x}_j\right)\right]\right\}, \end{aligned} \tag{5.4}$$

while the log-likelihood function is given by

$$\ell\left(\boldsymbol{\beta},\sigma \mid \boldsymbol{x}_i,t_i,\delta_i\right)=\log\left[\left\{h\left(t_i,\delta_i \mid \boldsymbol{x}_i,\boldsymbol{\beta},\sigma\right)\right\}^{\delta_i}\left\{S\left(t_i,\delta_i \mid \boldsymbol{x}_i,\boldsymbol{\beta},\sigma\right)\right\}\right]$$

$$=\delta_i \log\left\{\exp\left[-\sigma^{-1}\left(\beta_0+\sum_{j=1}^{p}\boldsymbol{\beta}_j\boldsymbol{x}_j\right)\right]\sigma^{-1}t^{\sigma^{-1}-1}\right\} \qquad (5.5)$$

$$+\ \log\left[\exp\left\{-t^{\sigma^{-1}}\exp\left[-\sigma^{-1}\left(\beta_0+\sum_{j=1}^{p}\boldsymbol{\beta}_j\boldsymbol{x}_j\right)\right]\right\}\right].$$

When there are no missing covariate values, the maximum likelihood estimates of $\boldsymbol{\beta}$ and σ may be obtained by differentiating the log-likelihood function in (5.5) with respect to $\boldsymbol{\beta}$ and σ, setting the derivative equal to zero and evaluating them at $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$. The score equation for $\boldsymbol{\beta}$ is given by

$$\boldsymbol{u}_\beta\left(\hat{\boldsymbol{\beta}}\right)=\frac{\partial\ell\left(\boldsymbol{\beta},\sigma \mid \boldsymbol{x}_i,t_i,\delta_i\right)}{\partial\boldsymbol{\beta}}=0, \qquad (5.6)$$

and for the scale parameter σ is given by

$$\boldsymbol{u}_\beta\left(\hat{\sigma}\right)=\frac{\partial\ell\left(\boldsymbol{\beta},\sigma \mid \boldsymbol{x}_i,t_i,\delta_i\right)}{\partial\sigma}=0. \qquad (5.7)$$

As the solution may not be obtained in a closed form, numerical methods such as Newton-Raphson iteration method is often used to obtain the maximum likelihood estimates of $\boldsymbol{\beta}$ and σ.

5.3 Missing Data Methods

Four methods of handling missing covariate values are considered throughout this chapter namely the complete case analysis, EM algorithm by method of weight, hot deck imputation and multiple imputation.

5.3.1 Complete Case Analysis

Complete case analysis is the most common method opted in many survival studies. In this method, only individuals with complete information are included, while those individuals with incomplete information are excluded from a study.

5.3.2 Expectation-Maximization (EM) Algorithm by Method of Weight

The EM algorithm by method of weight that has been proposed by Lipsitz and Ibrahim (1996a) is considered in this study. When there are no missing values, the distribution of $\boldsymbol{x}_i$ is not needed for estimating $\boldsymbol{\beta}$ and σ. However, if some covariate values are missing, a model for the distribution of the covariates that are subject to missingness needs to be specified. This model is denoted by $p\left(\boldsymbol{x}_i \mid \boldsymbol{\alpha}\right)$, where $\boldsymbol{\alpha}$ is unknown parameters vector. Let $\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \sigma, \boldsymbol{\alpha}\right)$ be the vector of unknown parameters, and the complete data log-likelihood function is given by

$$\ell\left(\boldsymbol{\theta} \mid \boldsymbol{x}_i, t_i, \delta_i\right) = \ell\left(\boldsymbol{\beta}, \sigma \mid \boldsymbol{x}_i, t_i, \delta_i\right) + \log\left[p\left(\boldsymbol{x}_i \mid, \boldsymbol{\alpha}\right)\right], \tag{5.8}$$

where $\ell\left(\boldsymbol{\beta}, \sigma \mid \boldsymbol{x}_i, t_i, \delta_i\right)$ is defined in (5.5).

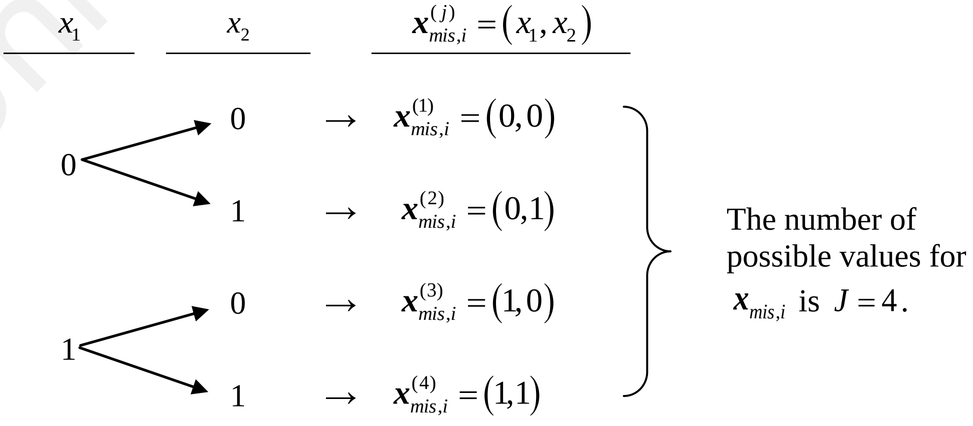
When some covariate values for subject i are missing, it is possible to write $\boldsymbol{x}_i = \left(\boldsymbol{x}_{mis,i}, \boldsymbol{x}_{obs,i}\right)$, where $\boldsymbol{x}_{mis,i}$ is the missing components of $\boldsymbol{x}_i$ that contains the unobserved covariates, while $\boldsymbol{x}_{obs,i}$ is the observed components that contains the completely observed covariates. The maximum likelihood estimates of $\boldsymbol{\theta}$ may be obtained by maximising the expected log-likelihood in (5.8) (Lipsitz & Ibrahim, 1996a). The E-step of the EM algorithm involves computing the conditional expectation for the $\ell\left(\boldsymbol{\theta} \mid \boldsymbol{x}_i, t_i, \delta_i\right)$ in (5.8) given the current estimate $\boldsymbol{\theta}^{(k)}$ and the observed data (Fonseca *et al*., 2013). The contribution of the *i*th subject to the expected log-likelihood is given by

$$Q_i\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}\right) = E\left[\ell\left(\boldsymbol{\theta} \mid \boldsymbol{x}_i, t_i, \delta_i\right) \mid \boldsymbol{x}_{obs,i}, t_i, \delta_i, \boldsymbol{\theta}^{(k)}\right], \tag{5.9}$$

where $\boldsymbol{\theta}^{(k)}$ denotes the estimates of $\boldsymbol{\theta}$ in the *k*th iteration.

Suppose that J is the number of possible values for $\boldsymbol{x}_{mis,i}$. If there are q missing categorical covariates, where $c_1, \ldots, c_q$ denotes the number of categories of each covariates, respectively, then the number of possible values for $\boldsymbol{x}_{mis,i}$ is $J = \prod_{i=1}^{q} c_i$. For instance, suppose $\boldsymbol{x}_{mis,i}$ consists of two covariates, x_1 and x_2. Both x_1 and x_2 are binary covariates {0,1}, hence the number of categories for x_1 and x_2 are $c_1 = 2$ and $c_2 = 2$, respectively. Therefore, the number of possible values for $\boldsymbol{x}_{mis,i}$ is $J = \prod_{i=1}^{2} c_i = c_1 \times c_2 = 2 \times 2 = 4$.

Possible values for :


x_1
x_2
$\boldsymbol{x}_{mis,i}^{(j)} = \left(x_1, x_2\right)$
0
0 → $\boldsymbol{x}_{mis,i}^{(1)} = (0,0)$
1 → $\boldsymbol{x}_{mis,i}^{(2)} = (0,1)$
1
0 → $\boldsymbol{x}_{mis,i}^{(3)} = (1,0)$
1 → $\boldsymbol{x}_{mis,i}^{(4)} = (1,1)$
The number of possible values for $\boldsymbol{x}_{mis,i}$ is $J = 4$.

Let $\boldsymbol{x}_i^{(j)} = \left(\boldsymbol{x}_{mis,i}^{(j)}, \boldsymbol{x}_{obs,i}\right)$ is the covariate vector with imputed values and observed values, where $j = 1, \ldots, J$. The conditional probability for the vector $\boldsymbol{x}_{mis,i}$ assuming the value $\boldsymbol{x}_{mis,i}^{(j)}$ is represented by $w_{ij}^{(k)}$ which is also known as weight. The $w_{ij}^{(k)}$ is given by

$$w_{ij}^{(k)} = p\left(\boldsymbol{x}_{mis,i}^{(j)} \mid \boldsymbol{x}_{obs,i}, t_i, \delta_i, \boldsymbol{\theta}^{(k)}\right) = \frac{p\left(t_i, \delta_i \mid \boldsymbol{x}_i^{(j)}, \boldsymbol{\beta}^{(k)}, \sigma^{(k)}\right) p\left(\boldsymbol{x}_i^{(j)} \middle| \alpha^{(k)}\right)}{\sum_{j=1}^{J} p\left(t_i, \delta_i \mid \boldsymbol{x}_i^{(j)}, \boldsymbol{\beta}^{(k)}, \sigma^{(k)}\right) p\left(\boldsymbol{x}_i^{(j)} \middle| \alpha^{(k)}\right)}. \quad (5.10)$$

It is worthwhile to note that $\sum_{j=1}^{J} w_{ij}^{(k)} = 1$ for all *k* and *i*. If all the covariates for the *i*th subject are observed, the weight $w_{ij}^{(k)} = 1$.

Hence, the E-step for the *i*th subject may be written as

$$\begin{aligned} Q_i\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}\right) &= \sum_{j=1}^{J} w_{ij}^{(k)} \ell\left(\boldsymbol{\theta} \mid \boldsymbol{x}_i, t_t, \delta_i\right) \\ &= \sum_{j=1}^{J} w_{ij}^{(k)} \ell\left(\boldsymbol{\beta}, \sigma \mid t_i, \delta_i, \boldsymbol{x}_i^{(j)}\right) + \sum_{j=1}^{J} w_{ij}^{(k)} \log\left[p\left(\boldsymbol{x}_i^{(j)} \mid \boldsymbol{\alpha}\right)\right], \end{aligned} \quad (5.11)$$

where $w_{ij}^{(k)}$ are the weights corresponding to the incomplete observations, or the conditional probability of a given missing data pattern indexed by *j*. Meanwhile, the E-step for all subjects may be expressed as

$$Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}\right) = \sum_{i=1}^{n}\sum_{j=1}^{J} w_{ij}^{(k)} \ell\left(\boldsymbol{\beta}, \sigma \mid t_i, \delta_i, \boldsymbol{x}_i^{(j)}\right) + \sum_{i=1}^{n}\sum_{j=1}^{J} w_{ij}^{(k)} \log\left[p\left(\boldsymbol{x}_i^{(j)} \mid \boldsymbol{\alpha}\right)\right]. \quad (5.12)$$

In the M-step, since the first term in (5.12) does not involve the parameter $\boldsymbol{\alpha}$ and the second term does not involve parameters $\boldsymbol{\beta}$ and σ, therefore $Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}\right)$ in (5.12) may be maximised separately with respect to $\boldsymbol{\beta}$ and σ,

$$Q\left(\boldsymbol{\beta}, \sigma \mid \boldsymbol{\theta}^{(k)}\right) = \sum_{i=1}^{n}\sum_{j=1}^{J} w_{ij}^{(k)} \ell\left(\boldsymbol{\beta}, \sigma \mid t_i, \delta_i, \boldsymbol{x}_i^{(j)}\right), \quad (5.13)$$

and with respect to $\boldsymbol{\alpha}$,

$$Q\left(\boldsymbol{\alpha} \mid \boldsymbol{\theta}^{(k)}\right) = \sum_{i=1}^{n} \sum_{j=1}^{J} w_{ij}^{(k)} \log\left[p\left(\boldsymbol{x}_i^{(j)} \mid \boldsymbol{\alpha} \right) \right]. \tag{5.14}$$

The iterative Newton-Raphson method is used in the maximization step. In general, the formulation of the maximum likelihood estimation using Newton-Raphson method is given as follows

$$\hat{\boldsymbol{\theta}}^{(s+1)} = \hat{\boldsymbol{\theta}}^{(s)} + \ddot{Q}\left(\hat{\boldsymbol{\theta}}^{(s)}\right)^{-1} \dot{Q}\left(\hat{\boldsymbol{\theta}}^{(s)}\right), \tag{5.15}$$

where $\dot{Q}(\boldsymbol{\theta})$ be the $q \times 1$ vector of the first derivatives of the expected log-likelihood function $Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}\right)$

$$\dot{Q}(\boldsymbol{\theta}) = \frac{\partial Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \sum_{j=1}^{J} w_{ij}^{(k)} \frac{\partial \ell\left(\boldsymbol{\theta} \mid \boldsymbol{x}_i^{(j)}, t_i, \delta_i\right)}{\partial \boldsymbol{\theta}}, \tag{5.16}$$

and $\ddot{Q}(\boldsymbol{\theta})$ be the $q \times q$ matrix of second derivatives of the expected log-likelihood function $Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}\right)$

$$\ddot{Q}(\boldsymbol{\theta}) = \frac{\partial^2 Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} = \sum_{i=1}^{n} \sum_{j=1}^{J} w_{ij}^{(k)} \frac{\partial^2 \ell\left(\boldsymbol{\theta} \mid \boldsymbol{x}_i^{(j)}, t_i, \delta_i\right)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}. \tag{5.17}$$

Since the maximization may be done separately, the standard statistical analysis for parametric survival model that allow weights to be assigned to each observation in the data set may be used.

5.3.2.1 Estimation of $\boldsymbol{\alpha}$

In survival modelling, the main interest is to obtain the parameter estimates of $\boldsymbol{\beta}$ and σ. Therefore, $\boldsymbol{\alpha}$ is considered as nuisance parameters since they are not parameters of interest. However, $\boldsymbol{\alpha}$ is very important in this EM algorithm procedure in order to estimate $\boldsymbol{\beta}$ and σ. Therefore, a model for the covariates distributions $p\left(\boldsymbol{x}_i \mid \boldsymbol{\alpha}\right)$ needs to be specified. In Lipsitz and Ibrahim (1996a) study, the saturated multinomial probability was used to model the covariates distributions. However, this approach was inefficient and computationally intensive especially in the case of a large number of nuisance parameters $\boldsymbol{\alpha}$ need to be estimated when the percentage of missing values is high.

As a result, Lipsitz and Ibrahim (1996b) suggested to model the distribution of the covariates as a product of one-dimensional conditional distributions. This method helps to reduce the number of nuisance parameters that need to be estimated in the M-step. The distribution of *p*-dimensional covariate vector $\boldsymbol{x}_i = \left(x_{i1}, x_{i2}, \ldots, x_{ip}\right)'$ may be written through a series of one-dimensional conditional distributions as follows

$$p\left(x_{i1}, \ldots, x_{ip} \mid \boldsymbol{\alpha}\right) = p\left(x_{ip} \mid x_{i1}, \ldots, x_{i(p-1)}, \boldsymbol{\alpha}_p\right) \times \ \ldots \times \ p\left(x_{i2} \mid x_{i1}, \boldsymbol{\alpha}_2\right) \times p\left(x_{i1} \mid \boldsymbol{\alpha}_1\right), \quad (5.18)$$

where $\boldsymbol{\alpha}_p$ is a vector of indexing parameters for the *p*th conditional distribution, and $\boldsymbol{\alpha} = \left(\alpha_1, \alpha_2, \ldots, \alpha_p\right)$. The model in (5.18) need to be specified only for covariates with missing values while observed covariates may be used as fixed regressor variables. To obtain a reduced model for dichotomous missing covariates, Lipsitz and Ibrahim (1996b) suggested to fit the logistic regression model, $p\left(x_{ij} \mid x_{i(j-1)}, \boldsymbol{\alpha}_j\right)$ in (5.18). Meanwhile, for the covariate with more than two levels, a multinomial logistic regression model may be used.

5.3.2.2 Estimation of Variance

The EM algorithm procedure does not give the right asymptotic covariance of the parameter estimates at convergence (Lipsitz & Ibrahim, 1996a). Therefore, the Louis (1982) method that was suggested by Lipsitz and Ibrahim (1996a) is used to compute the correct estimate of the asymptotic variance of the maximum likelihood estimates for $\boldsymbol{\theta}$. From Louis (1982) method, the observed information matrix is given by

$$I\left(\hat{\boldsymbol{\theta}}\right)=\ddot{Q}_i\left(\hat{\boldsymbol{\theta}}\right)-\sum_{i=1}^{n}\left[\sum_{j=1}^{J}w_{ij}S_i\left(\boldsymbol{x}_i,t_i,\delta_i,\hat{\boldsymbol{\theta}}\right)S_i\left(\boldsymbol{x}_i,t_i,\delta_i,\hat{\boldsymbol{\theta}}\right)'\right]+\sum_{i=1}^{n}\dot{Q}_i\left(\hat{\boldsymbol{\theta}}\right)\dot{Q}_i\left(\hat{\boldsymbol{\theta}}\right)', \quad (5.19)$$

where $\hat{\boldsymbol{\theta}}$ denotes the maximum likelihood estimates of $\boldsymbol{\theta}$ and

$$S_i\left(x_i,t_i,\delta_i,\hat{\boldsymbol{\theta}}\right)=\frac{\partial\ell\left(\hat{\boldsymbol{\theta}}\,|\,x_i,t_i,\delta_i\right)}{\partial\hat{\boldsymbol{\theta}}}. \quad (5.20)$$

The estimates of $\hat{\boldsymbol{\theta}}$ are obtained at the convergence of EM algorithm. Hence, an estimate of the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ may be obtained from the inverse of the information matrix $I\left(\hat{\boldsymbol{\theta}}\right)^{-1}$,

$$Var\left(\hat{\boldsymbol{\theta}}\right)=I\left(\hat{\boldsymbol{\theta}}\right)^{-1}. \quad (5.21)$$

The overall procedures of the EM algorithm by method of weight are summarised as follows:

a) The initial estimates of $\boldsymbol{\theta}^{(0)}=(\boldsymbol{\beta},\sigma,\boldsymbol{\alpha})$ are obtained from the complete case analysis.

b) E-step: At the $(k+1)$th EM iteration, the conditional probability or weight w_{ij} in (5.10) is computed using the parameter estimates at *k*th iteration of EM algorithm $\boldsymbol{\theta}^{(k)}$.

c) M-step: The estimates of $\boldsymbol{\theta}^{(k+1)}$ is obtained by maximising (5.12) using Newton-Raphson method as in (5.15).

d) The E-step and the M-step are repeated until convergence.

In this study, the programming code for the EM algorithm by method of weight has been developed in `R` statistical software. The function `survreg` from the `survival` package (Therneau, 2014) in `R` software is used for maximising $Q\left(\boldsymbol{\beta},\sigma\,|\,\boldsymbol{\theta}^{(k)}\right)$ in (5.13) in order to obtain $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$. Meanwhile, $\hat{\boldsymbol{\alpha}}$ is obtained by maximising $Q\left(\boldsymbol{\alpha} \mid \boldsymbol{\theta}^{(k)}\right)$ in (5.14) using the `maxLik` function (Henningsen & Toomet, 2011) in `R`.

5.3.3 Hot Deck Imputation

For this method, the function `imputation` from the `rminer` package (Cortez, 2013) in the `R` software has been considered. The imputed values are determined using the *k*-nearest neighbour (*k-nn*) method. Suppose that there is a data set that consists of n cases with two variables x and y. Let one value of the variable y is missing, and that missing term is denoted by y_j. Using the *k*-nearest neighbour method, each of other cases with complete data is checked and the missing value y_j is substituted by the value for the most similar case. This similar case is identified by finding the difference between the observed value x_j and the nearest neighbours of x_j, that is $x_k = \left(\ldots, x_{j-1}, x_{j+1}, \ldots\right)$ using the following formula

$$D = \left|x_k - x_j\right|. \tag{5.22}$$

The missing term y_j is substituted with the observed value y_k where the difference D between the observed items, x_k and x_j, is the smallest.

5.3.4 Multiple Imputation with MICE-PMM

The multiple imputation method begins with creating $m > 1$ imputed data sets, analysing the m imputed data sets separately and pooling the m parameter estimates into a single value. In this study, the multiple imputation by chained equation (MICE) with predictive mean matching (PMM) is considered. This method is performed using the package `mice` (Van Buuren & Groothuis-Oudshoorn, 2011) in the `R` software.

5.3.4.1 Multiple Imputation by Chained Equation

Multiple imputation by chained equation (MICE) method generates the imputed values from a set of imputation models which one imputation model is specified for each incomplete variable (Van Buuren & Groothuis-Oudshoorn, 2011). Since each variable with missing values is imputed using its own imputation model, this method able to handle different types of variable including continuous, binary, unordered and ordered categorical data (White *et al*. 2011; Royston & White, 2011). In the imputation model, all the variables those are supposed to be in the analysis are included in the imputation model in order to avoid bias (Schafer, 1997). Also, the outcome variable should be incorporated into the imputed model for imputing any missing covariate values (Moons *et al*., 2006).

Let the data be presented by the $n \times p$ matrix Y. The elements of Y is y_{ij} where $i = 1,\ldots,n$ and $j = 1,\ldots,p$. The jth column of Y is denoted as Y_j. Meanwhile, Y_{-j} is all columns of Y except Y_j. Missing component of Y_j is denoted by Y_j^{mis}, whilst Y_j^{obs} represents the observed component. The missing indicator R is the $n \times p$ binary response matrix, which its elements $r_{ij} = 1$ if y_{ij} is observed and if y_{ij} is missing, $r_{ij} = 0$. The unknown parameters of the imputation model are denoted as ψ_j. The MICE algorithm method specifies the imputation model on a variable-by-variable basis using a separate conditional distribution for each incomplete variable. The conditional distribution $P\left(Y_j^{mis} \mid Y_j^{obs}, Y_{-j}, R, \psi\right)$ is used to draw the imputed values for the Y_j^{mis}. The imputation model describing the conditional probabilities $P\left(Y_1^{mis} \mid Y_1^{obs}, Y_{-1}, R, \psi\right), \ldots, P\left(Y_j^{mis} \mid Y_j^{obs}, Y_{-j}, R, \psi\right)$ can be any appropriate regression model depending on the nature of the outcome variable. The MICE algorithm for imputation of multivariate missing data is described as follows (Van Buuren, 2012):

a) An imputation model $P\left(Y_j^{mis} \mid Y_j^{obs}, Y_{-j}, R\right)$ for variable Y_j $\left(j=1,\ldots,p\right)$ is specified.

b) For each j, starting imputations $\dot{Y}_j^0$ are filled in by random draws from Y_j^{obs}.

c) Repeat for $k=1,\ldots,K$:

d) Repeat for $j=1,\ldots,p$:

e) $\dot{Y}_{-j}^k = \left(\dot{Y}_1^k,\ldots,\dot{Y}_{j-1}^k,\dot{Y}_{j+1}^{k-1},\ldots,\dot{Y}_p^{k-1}\right)$ is defined as the currently complete data except Y_j.

f) $\dot{\psi}_j^k \sim P\left(\psi_j^k \mid Y_j^{obs}, \dot{Y}_{-j}^k, R\right)$ is drawn.

g) imputations $\dot{Y}_j^k \sim P\left(Y_j^{mis} \mid Y_j^{obs}, \dot{Y}_{-j}^k, R, \dot{\psi}_j^k\right)$ are drawn.

h) End repeat j.

i) End repeat k.

In the *k*th iteration of the algorithm, the imputed value is generated for the missing variable, then this imputed value is used for imputing the next variable. This process repeats until convergence is reached (Horton & Kleinman, 2007). According to Van Buuren (2012), the iteration number may be low, such as 5 and 10. Separate chains are performed to create multiple sets of complete data in parallel m times.

In this study, given that a survival data set contains four variables: survival time *t* that follows the Weibull distribution, censoring indicator δ, and two binary covariates x_1 and x_2. Let x_2 be the incomplete variable whilst all the others are completely observed. Since this study considers one incomplete variable that is x_2, only one cycle (*j*=1) of x_2 imputation involves in the kth iteration. Figure 5.1 portrays the process of multiple imputation with the MICE algorithm.

In this study, $m = 10$ complete data sets are generated. Since the Weibull AFT model is considered, each of these data sets is analysed separately by a standard parametric survival analysis in `R` using the function `survreg`. All the parameter estimates of the Weibull AFT model $\hat{\boldsymbol{\theta}} = (\boldsymbol{\beta}, \sigma)$ those have been estimated from m imputed data sets, $\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2, \ldots, \hat{\boldsymbol{\theta}}_m$, are pooled following Rubin's rule in (2.64) into one estimate $\hat{\boldsymbol{\theta}}$ using the function `pool` from the `R` package `mice` (Van Buuren *et al.*, 2014a).

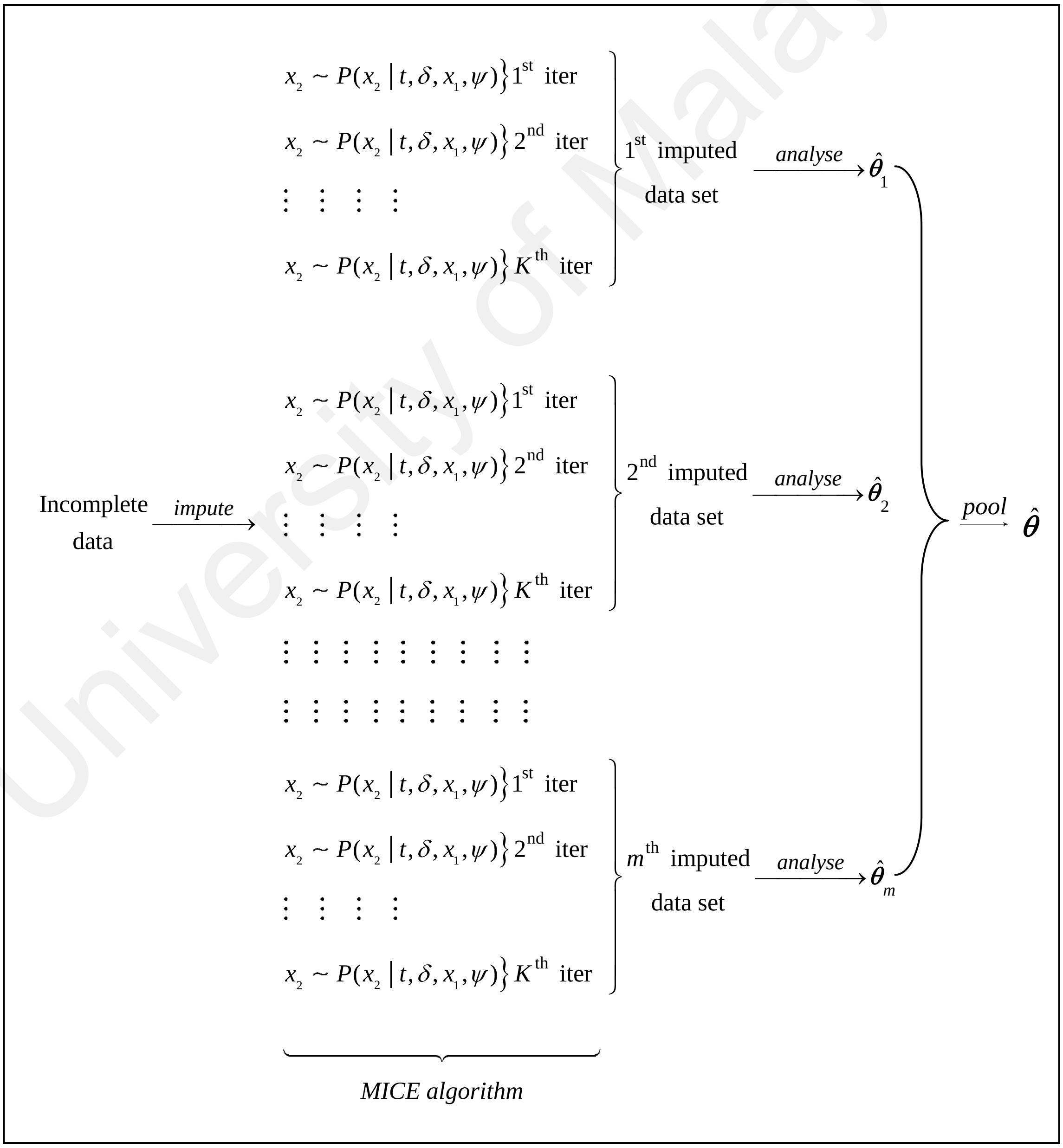

Incomplete data
impute
$x_2 \sim P(x_2 \mid t, \delta, x_1, \psi)\}$ 1st iter
$x_2 \sim P(x_2 \mid t, \delta, x_1, \psi)\}$ 2nd iter
$x_2 \sim P(x_2 \mid t, \delta, x_1, \psi)\}$ Kth iter
1st imputed data set
2nd imputed data set
mth imputed data set
analyse
$\hat{\boldsymbol{\theta}}_1$
$\hat{\boldsymbol{\theta}}_2$
$\hat{\boldsymbol{\theta}}_m$
pool
$\hat{\boldsymbol{\theta}}$
MICE algorithm

Figure 5.1: Multiple imputation by chained equation (MICE)

5.3.4.2 Predictive Mean Matching

The multiple imputation by chained equation (MICE) with predictive mean matching (PMM) has been considered in this study. The PMM method imputes each missing value of covariate x_j with the value that is sampled from the observed value of x_j. Based on the specified imputation model, the PMM method computes the predicted value for the missing value where all other variables serve as predictors. Then, a small number of candidate donors d from the observed data, where their predicted values $(\hat{x}_i)$ close to the predicted value for the missing value ($\hat{x}_j$) are selected. Each missing value is imputed with the observed value that is randomly drawn from these candidates. This method assumes that the distribution of imputed x_j follows the same distribution as the candidates' data (Van Buuren, 2012).

It is worthwhile to note that, the MICE-PMM is robust against misspecification of the imputation model (Van Buuren, 2012). White *et al*., (2011) provided an example which under misspecified model, the PMM method able to impute the missing values appropriately. Therefore, this method is suitable for any types of variable including categorical variable (Van Buuren, 2012; Van Buuren *et al*., 2014b). Also, the imputed values are realistic since the values drawn are within the range of the observed data (Van Buuren, 2012).

5.4 Simulation Study

Simulation procedures are carried out to investigate the performance of complete case analysis, EM algorithm by method of weight, hot deck imputation and multiple imputation using the MICE-PMM for data that is assumed missing at random (MAR).

In this simulation studies, survival times t_i are generated from a Weibull distribution with parameters $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = -1$, whilst the scale parameter

$\sigma = 2$. Survival times are randomly made censored yielding about 30% of censored observations in each data set. Also, two categorical independent covariates x_{i1} and x_{i2} are generated from a Bernoulli distribution where x_{i1} with a success probability of 0.6, and x_{i2} with success probability of

$$\frac{\exp\left(a_{20} + a_{21}x_{i1}\right)}{1+\exp\left(a_{20} + a_{21}x_{i1}\right)}, \tag{5.23}$$

given that $a_{20} = 1$ and $a_{21} = 1$.

Variable x_{i1} is always observed while x_{i2} is made MAR according to different percentage of missingness: 10%, 30% and 50%. When the mechanism of missingness of a covariate x is MAR, the probability of missing x values (conditional on the other observed covariates) does not depend on x or any other unobserved covariates. However, the missing probability may depend on the outcome variable and other observed covariates. Missing values are generated by specifying a missing indicator R_{i2}, where $R_{i2} = 1$ indicates that x_{i2} is missing and $R_{i2} = 0$ when x_{i2} is observed. The R_{i2} follows the logistic regression model, and the probability of x_{i2} to be missing is modelled as the following

$$p\left(R_{i2} = 1 \middle| \boldsymbol{\varphi}, x_{i1}, t_i^*\right) = \frac{\exp\left(\varphi_0 + \varphi_1 t_i^* + \varphi_2 x_{i1} + \varphi_3 x_{i1} t_i^*\right)}{1+\exp\left(\varphi_0 + \varphi_1 t_i^* + \varphi_2 x_{i1} + \varphi_3 x_{i1} t_i^*\right)}, \tag{5.24}$$

where

$$t_i^* = \frac{t - \mu_{t_i}}{\sigma_{t_i}}. \tag{5.25}$$

The values of $\boldsymbol{\varphi}$ for simulating 10%, 30% and 50% missing values for each data set are $\boldsymbol{\varphi}_{10\%} = \left(-2.5, -1.0, 0.2, 1.5\right)$, $\boldsymbol{\varphi}_{30\%} = \left(-1.12, -1.0, 0.2, 1.5\right)$ and $\boldsymbol{\varphi}_{50\%} = \left(-0.3, -1.0, 0.2, 1.5\right)$ respectively.

The simulation procedures are repeated 5000 times for different combination of sample size and the percentage of missing values. Sample size varies from 100, 300, and 500. The simulation mean, standard error (SE), mean absolute error (MAE) and root mean squared error (RMSE) are given by

$$\text{Mean, } \bar{\theta} = \frac{\sum_{k=1}^{5000} \hat{\theta}}{5000}, \tag{5.26}$$

$$\text{MAE} = \frac{\sum_{k=1}^{5000} |\hat{\theta} - \theta|}{5000}, \tag{5.27}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{k=1}^{5000} \left(\hat{\theta} - \theta\right)^2}{5000}}, \tag{5.28}$$

$$\text{SE} = \sqrt{\frac{\sum_{k=1}^{5000} \left(\hat{\theta} - \bar{\theta}\right)^2}{5000}}, \tag{5.29}$$

respectively, where $\hat{\theta}$ is the estimate of θ from the *k*th simulated sample. The `R` code for this simulation procedure is provided in Appendix B.

5.4.1 Simulation Results

Table 5.1 presents simulation results from the complete case analysis, EM algorithm by method of weight, hot deck and multiple imputation by MICE-PMM approaches for sample size $n = 100$. This table presents the mean, SE, MAE, and RMSE for parameter estimates based on 5000 generated samples.

With respect to the estimated values of SE, MAE, and RMSE of β_1 estimates, the performance of the complete case analysis is worse than that of other methods since all the aforementioned values of performance indicators are remarkably large for 30% and 50% missing value. Meanwhile, better performance may be seen after treating the missing values using multiple imputation method. Yet, when the percentage of missing values is 10%, the performance of EM algorithm by method of weight and hot deck methods are comparable to that of the multiple imputation method given that only small differences in the SE, MAE and RMSE values among these methods are observed.

The performance of hot deck method is the worst since the estimated values of SE, MAE, and RMSE of the coefficient β_2, the parameter corresponding to the missing covariate x_{i2}, are the largest, regardless the percentage of missing values in the data. Meanwhile, the complete case analysis performance is better than that of other methods when the percentage of missing values is 10%. However, as the percentage of missing values increases, the estimated values of SE, MAE and RMSE of β_2 also increases. When 30% of the data are missing, the multiple imputation method performs better. Meanwhile, the EM algorithm gives better results when the percentage of missing values is 50%.

It is worthwhile to note that the estimated SE, MAE and RMSE values of the scale parameter (σ) based on the complete case analysis are worse than those for other methods. Meanwhile, the other three methods perform almost similar. Small differences may be observed when the percentage of missing values is 50%. The estimates of σ obtained from the multiple imputation method are slightly better compared to the EM algorithm by method weight and hot deck imputation method.

Table 5.1: The simulation results for $n = 100$

Percentage of missing values	Indicator	Parameter	Complete Case	EM	Hot Deck	Multiple Imputation
10%	Mean	β_1	0.813	1.005	0.999	0.994
	SE		0.545	0.513	0.517	0.510
	MAE		0.459	0.409	0.411	0.406
	RMSE		0.576	0.513	0.517	0.510
	Mean	β_2	-0.937	-0.968	-0.956	-0.954
	SE		0.731	0.750	0.778	0.737
	MAE		0.575	0.588	0.609	0.579
	RMSE		0.733	0.750	0.779	0.738
	Mean	σ	1.935	1.960	1.960	1.959
	SE		0.198	0.189	0.190	0.189
	MAE		0.168	0.156	0.156	0.155
	RMSE		0.209	0.194	0.195	0.193
30%	Mean	β_1	0.570	1.007	0.991	0.975
	SE		0.629	0.523	0.529	0.508
	MAE		0.617	0.417	0.422	0.405
	RMSE		0.762	0.523	0.529	0.509
	Mean	β_2	-0.948	-0.974	-0.971	-0.956
	SE		0.981	0.967	1.058	0.886
	MAE		0.676	0.678	0.748	0.656
	RMSE		0.982	0.968	1.058	0.887
	Mean	σ	1.908	1.956	1.953	1.955
	SE		0.221	0.191	0.192	0.190
	MAE		0.195	0.157	0.159	0.157
	RMSE		0.240	0.196	0.197	0.195
50%	Mean	β_1	0.340	1.006	0.987	0.952
	SE		0.755	0.538	0.558	0.511
	MAE		0.817	0.427	0.442	0.409
	RMSE		1.002	0.538	0.558	0.513
	Mean	β_2	-1.085	-0.996	-1.110	-1.037
	SE		2.395	1.607	2.394	1.844
	MAE		0.925	0.818	1.025	0.825
	RMSE		2.396	1.607	2.396	1.844
	Mean	σ	1.889	1.952	1.947	1.952
	SE		0.257	0.193	0.196	0.191
	MAE		0.228	0.160	0.163	0.158
	RMSE		0.280	0.199	0.203	0.197

Table 5.2 presents simulation results for $n = 300$. For 10% missing values, the estimated values of SE, MAE, and RMSE of β_1 using the multiple imputation, EM algorithm by method of weight and hot deck method are almost similar. However, as the percentage of missing values increases, the estimated values of SE, MAE, and RMSE for the EM algorithm by method of weight followed by hot deck method are slightly higher than that of multiple imputation. Obvious differences may be observed from the complete case analysis as this method yields the largest SE, MAE and RMSE values and worsen as the percentage of missing values increases.

Based on β_2 estimates, the results show that the performance of EM algorithm method is better than all the others, except that when 10% of the data are missing where the performance of complete case analysis is better than EM algorithm. Largest values of the estimated SE, MAE and RMSE indicate that the hot deck estimations are the worst.

Table 5.2 also indicates that the SE, MAE and RMSE of σ are remarkably large suggesting that the complete case analysis performs less well. The performance of EM algorithm by method of weight, hot deck and multiple imputation methods for 10% and 30% missing covariate values are comparable. For 50% missing values, the estimated values of SE, MAE and RMSE are slightly higher based on the hot deck method than that of EM algorithm and multiple imputation. Meanwhile, the EM algorithm outperforms all the others.

Table 5.2: The simulation results for $n = 300$

Percentage of missing values	Indicator	Parameter	Complete Case	EM	Hot Deck	Multiple Imputation
10%	MEAN	β_1	0.821	1.000	0.997	0.993
	SE		0.309	0.291	0.293	0.290
	MAE		0.289	0.234	0.235	0.233
	RMSE		0.358	0.291	0.293	0.290
	MEAN	β_2	-0.931	-0.963	-0.956	-0.962
	SE		0.407	0.418	0.440	0.421
	MAE		0.329	0.333	0.351	0.336
	RMSE		0.413	0.420	0.442	0.422
	MEAN	σ	1.964	1.988	1.988	1.986
	SE		0.111	0.107	0.107	0.107
	MAE		0.093	0.086	0.086	0.086
	RMSE		0.116	0.107	0.108	0.108
30%	MEAN	β_1	0.591	1.003	0.996	0.982
	SE		0.354	0.295	0.299	0.290
	MAE		0.455	0.237	0.239	0.234
	RMSE		0.541	0.295	0.299	0.291
	MEAN	β_2	-0.931	-0.960	-0.950	-0.961
	SE		0.464	0.468	0.523	0.478
	MAE		0.373	0.373	0.418	0.383
	RMSE		0.469	0.469	0.525	0.480
	MEAN	σ	1.950	1.987	1.986	1.983
	SE		0.127	0.108	0.109	0.108
	MAE		0.110	0.087	0.088	0.087
	RMSE		0.136	0.108	0.110	0.109
50%	MEAN	β_1	0.380	1.007	0.993	0.968
	SE		0.424	0.299	0.307	0.291
	MAE		0.648	0.240	0.246	0.234
	RMSE		0.751	0.299	0.307	0.293
	MEAN	β_2	-0.981	-0.984	-0.993	-0.997
	SE		0.542	0.526	0.629	0.558
	MAE		0.432	0.420	0.502	0.447
	RMSE		0.543	0.526	0.629	0.558
	MEAN	σ	1.942	1.984	1.980	1.978
	SE		0.147	0.108	0.112	0.109
	MAE		0.127	0.087	0.091	0.089
	RMSE		0.158	0.110	0.114	0.112

Table 5.3 presents simulation results for $n = 500$. The estimation of β_1 from the complete case analysis are poor since the estimated SE, MAE and RMSE values are remarkably large especially when the percentage of missing values are 30% and 50%. Meanwhile, the performance of EM algorithm by method of weight is as good as that of the multiple imputation method even when the percentage of missing values is 50%.

The EM algorithm by method of weight gives the best estimates of β_2 since the SE, MAE and RMSE are the lowest. However, the complete case analysis performance is slightly better than EM algorithm when 10% of data are missing. The worst performance may be seen for the hot deck imputation method. Meanwhile, multiple imputation method performs slightly poor than the complete case analysis.

The complete case analysis yields the worst estimates of σ since the estimated values of SE, MAE, and RMSE are the largest. Meanwhile, the performance of EM algorithm by method of weight, multiple imputation and hot deck imputation methods are almost similar.

Table 5.3: The simulation results for $n = 500$

Percentage of missing values	Indicator	Parameter	Complete Case	EM	Hot Deck	Multiple Imputation
10%	MEAN	β_1	0.823	0.999	0.998	0.993
	SE		0.237	0.226	0.227	0.225
	MAE		0.240	0.180	0.181	0.179
	RMSE		0.296	0.226	0.227	0.226
	MEAN	β_2	-0.938	-0.971	-0.964	-0.977
	SE		0.313	0.321	0.339	0.327
	MAE		0.256	0.258	0.273	0.263
	RMSE		0.319	0.322	0.340	0.328
	MEAN	σ	1.971	1.993	1.993	1.991
	SE		0.089	0.086	0.086	0.086
	MAE		0.075	0.069	0.069	0.069
	RMSE		0.093	0.086	0.086	0.086
30%	MEAN	β_1	0.596	1.002	0.997	0.983
	SE		0.276	0.229	0.232	0.226
	MAE		0.423	0.182	0.184	0.180
	RMSE		0.489	0.229	0.232	0.226
	MEAN	β_2	-0.944	-0.973	-0.969	-0.987
	SE		0.356	0.358	0.406	0.383
	MAE		0.287	0.286	0.324	0.306
	RMSE		0.360	0.359	0.407	0.383
	MEAN	σ	1.958	1.993	1.991	1.988
	SE		0.099	0.086	0.087	0.086
	MAE		0.087	0.069	0.070	0.070
	RMSE		0.107	0.086	0.087	0.087
50%	MEAN	β_1	0.394	1.006	0.996	0.974
	SE		0.334	0.234	0.241	0.228
	MAE		0.617	0.185	0.191	0.182
	RMSE		0.692	0.234	0.241	0.230
	MEAN	β_2	-0.980	-0.981	-0.991	-1.014
	SE		0.424	0.410	0.486	0.456
	MAE		0.335	0.325	0.386	0.364
	RMSE		0.424	0.410	0.487	0.456
	MEAN	σ	1.953	1.991	1.988	1.984
	SE		0.117	0.087	0.089	0.088
	MAE		0.102	0.070	0.072	0.071
	RMSE		0.126	0.087	0.090	0.089

Table 5.1 to Table 5.3 show that as the percentage of missing values increasing, the estimated values of SE, MAE and RMSE of the parameter estimates are also increasing. On the other hand, the estimated SE, MAE, and RMSE values are decreasing as the sample size increasing. Overall, the EM algorithm method has shown the most favourable results in the simulation studies followed by the multiple imputation method. Based on the β_1 estimates, the performance of multiple imputation method is better than that of other methods. Meanwhile, the estimations of β_2 and σ from the EM algorithm by method of weight are better than all the others. Even though small differences might be observed between β_1 estimates obtained from these two methods, the performance indicators show that the values are almost similar. The complete case analysis yields poor estimates of β_1 and σ, while hot deck method gives the worst estimates for β_2. It is worthwhile to note that, when there are 10% missing values, the results from the complete case analysis is nearly as good as the other three methods.

5.5 Illustrative Example

Data of cervical cancer patients that has been evaluated in Chapter 3 and 4 is considered in this section. This data set consists of 120 cervical cancer patients treated in HUSM from 1997-2008. To illustrate the complete case analysis, EM algorithm by method of weight, hot deck and multiple imputation methods, only two covariates are considered. These two covariates are stage at diagnosis and distant metastasis. Variable stage at diagnosis consists of two groups, stage I-II and stage III-IV while distant metastasis is divided into with and without metastasis group. Variable distant metastasis is artificially set missing at random (MAR) with the percentage of 10%, 30% and 50%. Meanwhile, all values for the variable stage at diagnosis are observed. In addition, result from the analysis of complete data, denoted as FULL is also presented.

Table 5.4 gives the estimates of the coefficient regression for stage at diagnosis (β_1), distant metastasis (β_2) and the scale parameter (σ) from the complete data (model without missing values that is denoted as FULL), complete case analysis, EM algorithm by method of weight, hot deck and multiple imputation methods. The largest differences are observed in the estimates of the effect of stage at diagnosis based on the complete case analysis. The results indicate that as the percentage of missing values increasing, the estimates of β_1 from the complete case analysis deviate remarkably from the full model estimates. Also, it is noted that the estimated values of standard error from the complete case analysis are the largest among other methods. By looking at the estimates of β_1 and also its standard error from the EM algorithm by method of weight, this method performs well since the estimates are closer to that of the full model. Meanwhile, the estimates from multiple imputation method are slightly worse than that of the EM algorithm by method of weight.

By looking at the estimates of β_2, EM algorithm by method of weight gives the closest estimates to that of the full model. In contrast, the other three methods yield poor estimates especially when distant metastasis variable values are missing at 30% and 50%. The estimated values of standard error based on EM algorithm by method of weight are slightly larger than that of the hot deck and multiple imputation when the percentage of missing values are 30% and 50%. It is worthwhile to note that the standard error estimates from EM algorithm by method of weight are smaller than that of the complete case analysis. When 30% of distant metastasis values are missing, this variable remains statistically significant only after treating the missing values using the EM algorithm by method of weight and hot deck method. As the percentage of missing values increases to 50%, all methods fail to detect the significant effect of distant metastasis. However, EM algorithm by method of weight method shows slightly better performance as the p-value=0.084 is nearly to the level of significance $\alpha = 0.05$.

For the scale parameter σ, the closest estimates to the full model may be observed after treating the missing covariate values with EM algorithm by method of weight. The complete case analysis gives the worst estimates of σ, while the estimates based on the multiple imputation and hot deck methods are comparable.

It is found that EM algorithm by method of weight performs the best by looking at the estimates of the regression coefficients of stage at diagnosis (β_1), distant metastasis (β_2) and scale parameter (σ), although the standard error of β_2 is slightly larger than those of other methods, in particular the hot deck imputation and multiple imputation. In addition, as the percentage of missing value increases, the problem of detecting the prognostic effect of the covariate with missing values, which is distant metastasis is observed. This variable remains statistically significant in the model up to 30% missing values, and that only after using the EM algorithm by method of weight and hot deck method.

Table 5.4: Estimates for cervical cancer data with two variables

Percentage	Variable	Method	Estimates	SE	*p*-value
10%	Stage (β_1)	Full	-0.842	0.250	0.001
		CC	-1.074	0.268	<0.0001
		EM	-0.832	0.250	0.001
		HD	-0.826	0.257	0.001
		MI	-0.831	0.255	0.002
	Distant metastasis (β_2)	Full	-0.648	0.239	0.007
		CC	-0.612	0.258	0.018
		EM	-0.627	0.255	0.014
		HD	-0.515	0.245	0.035
		MI	-0.556	0.244	0.032
	Scale (σ)	Full	0.944		
		CC	0.939		
		EM	0.947		
		HD	0.960		
		MI	0.954		
30%	Stage (β_1)	Full	-0.842	0.250	0.001
		CC	-1.194	0.257	<0.0001
		EM	-0.833	0.252	0.001
		HD	-0.869	0.255	0.001
		MI	-0.857	0.256	0.001
	Distant metastasis (β_2)	Full	-0.648	0.239	0.007
		CC	-0.464	0.250	0.064
		EM	-0.600	0.278	0.031
		HD	-0.498	0.248	0.044
		MI	-0.462	0.252	0.091
	Scale (σ)	Full	0.944		
		CC	0.863		
		EM	0.952		
		HD	0.960		
		MI	0.962		
50%	Stage (β_1)	Full	-0.842	0.250	0.001
		CC	-1.564	0.371	<0.0001
		EM	-0.900	0.254	<0.0001
		HD	-0.922	0.256	<0.0001
		MI	-0.906	0.258	0.001
	Distant metastasis (β_2)	Full	-0.648	0.239	0.007
		CC	-0.438	0.357	0.220
		EM	-0.566	0.327	0.084
		HD	-0.396	0.245	0.105
		MI	-0.466	0.310	0.380
	Scale (σ)	Full	0.944		
		CC	0.968		
		EM	0.958		
		HD	0.963		
		MI	0.962		

5.6 Summary

In this chapter, the performance of the complete case analysis, EM algorithm by method of weight, hot deck and multiple imputation with MICE-PMM methods are investigated based on different sample size ($n = 100,\ 300 \text{ and } 500$) and different percentage of missing values ($10\%,\ 30,\ \text{and } 50\%$) on the parametric survival model namely the Weibull model. Data are assumed missing at random (MAR).

Only small differences are observed between the results obtained from the complete case analysis, EM algorithm by method of weight, hot deck imputation and multiple imputation with MICE-PMM method when percentage of missing values is 10%. Meanwhile, in a small sample, $n = 100$, the multiple imputation with MICE-PMM yields better results compared to the complete case analysis, EM algorithm by method of weight, and hot deck imputation method. However, as the sample size increases, EM algorithm by method of weight outperforms the other methods. In addition, when the percentage of missing values is as high as 50%, the EM algorithm by method of weight performs considerably well than all the others. In addition, the effects of variables in the final model may be remained statistically significant for small to moderate percentage of missing covariate values when EM algorithm by method of weight is applied.

As missing values are often encountered in survival data, this study may provide a suitable option for handling missing values particularly in parametric survival models. Amongst all, the EM algorithm by method of weight has shown a great potential in addressing the issue of missing values in parametric survival data analysis, especially when the sample size and the percentage of missing values are large.

CHAPTER 6

SCORE TESTS FOR DETECTING FRAILTY IN A BIVARIATE POSITIVE STABLE GOMPERTZ MODEL

6.1 Introduction

In a survival study, it is commonly assumed that survival times of a group of n individuals under study $t_1, t_2, \ldots, t_n$ are independent. However, there are circumstances where there exist unobserved or unmeasured factors that may induce dependency among the survival times. Frailty model is the best option to handle this type of data since the model takes into account the effect of these unobserved factors by introducing the frailty term into the model. Besides modelling the frailty, study on the tests for detecting the presence of frailty has received much attention.

Frailty may follow various types of statistical distribution, and one of the distributions is a positive stable distribution. Zhu (1998) derived a score test for detecting frailty based on the positive stable Weibull model. Later on, Sarker (2002) extended the study by deriving two new score based tests from the Zhus's score test, namely the modified score test and $\ln s$ based test. Sarker (2002) also found that amongst these three tests, the convergence rate of the Zhus's score test to the normal limit was remarkably slow. Meanwhile, the modified score test was preferable as the test converged faster to the normal limit and showed better performance based on the estimated power of the test.

These tests were proposed for a proportional hazards model with survival times follow Weibull distribution. It is worthwhile to note that a Gompertz distribution model is another type of survival models that adhere to this assumption which is better known as a proportional odds model. Thus, in this chapter, the behaviour of the aforementioned score tests is studied based on the Gompertz distribution model with positive stable frailty for a bivariate case $\left(p=2\right)$. The Gompertz distribution has many applications in medical field and biological sciences. In fact, this distribution was first introduced to model human mortality by Gompertz in 1825 (Collet, 2003). In addition, the Gompertz distribution will reduce to an exponential distribution if the shape parameter $b=0$.

This chapter is organised as follows. Section 6.2 describes the score based tests that have been proposed by Zhu (1998) and Sarker (2002). Section 6.3 presents the bivariate positive stable Gompertz model and its relation to the Zhu's score test, modified score test and $\ln s$ based test. The asymptotic properties of these score tests are presented in Section 6.4 for uncensored cases. The asymptotic variances for the modified score test and $\ln s$ based test for the uncensored case with nuisance parameters are derived. The asymptotic properties of these score tests for censored cases are described in Section 6.5. Section 6.6 explains the critical region of the tests. Results of the rate of convergence to the normal limit for these tests are presented in Section 6.7. In Section 6.8, the expressions for obtaining the critical values for any number of sample sizes are presented. The performance of these tests based on the estimated power is also evaluated in Section 6.9. Meanwhile, Section 6.10 presents the application of these tests on a simulated data set that follows the positive stable Gompertz model. Summary of the chapter is given in Section 6.11.

6.2 Derivation of the Score Tests from the Positive Weibull Model

The survival function $S(t)$ for a survival time variable $T > 0$ that follows a Weibull distribution is given by

$$S(t) = \exp\left(-\lambda t^{\phi}\right), \tag{6.1}$$

where λ is a scale parameter, and ϕ is a shape parameter. The cumulative hazard function of the Weibull distribution is given by

$$H(t) = \lambda t^{\phi}. \tag{6.2}$$

An approach for constructing the Weibull based frailty model is to consider a mixture, either continuous or discrete, of the Weibull distribution. From (2.70), given the frailty W, the unconditional survival function of T may be written as

$$S(t) = \int_0^{\infty} \exp\left(-w\lambda t^{\phi}\right) dG(w), \tag{6.3}$$

where $G(w)$ is the distribution function of the frailty W. Similarly, conditional on W, the unconditional survival function for the multivariate (*p*-variate) Weibull survival time $T = \left(T_1, T_2, \ldots, T_p\right)$ is given by

$$S\left(t_1, t_2, \ldots, t_p\right) = \int_0^{\infty} \exp\left(-ws_w\right) dG(w), \tag{6.4}$$

where $s_w = \sum_{j=1}^{p} \lambda_j t_j^{\phi_j}$.

Suppose that, W has a positive stable distribution with characteristic exponent v $\left(0 < v \leq 1\right)$. The Laplace transform of the positive stable distribution is

$$L(s) = E\left(\exp\left(-ws_w\right)\right) = \exp\left(-s_w^{v}\right). \tag{6.5}$$

Thus, the survival function for the multivariate positive stable Weibull model is obtained as Laplace transform

$$S\left(t_1, t_2, \ldots, t_p\right) = \int_0^{\infty} \exp\left(-ws_w\right) dG(w) = \exp\left(-s_w^{v}\right), \tag{6.6}$$

where $s_w = \sum_{j=1}^{p} \lambda_j t_j^{\phi_j}$.

Let T_1 and T_2 denote the two survival times, where the bivariate survival function $S\left(t_1,t_2\right)$ is given by

$$S\left(t_1,t_2\right) = \Pr\left(T_1 > t_1, T_2 > t_2\right) = 1 - F\left(t_1\right) - F\left(t_2\right) + F\left(t_1,t_2\right). \tag{6.7}$$

The relationship between the joint density function $f\left(t_1,t_2,\ldots,t_p\right)$ and the survival function $S\left(t_1,t_2,\ldots,t_p\right) = \Pr\left(T_1 > t_1, T_2 > t_2, \ldots, T_p > t_p\right)$ for survival times $T_1,T_2,\ldots,T_p$ is

$$f\left(t_1,t_2,\ldots,t_p\right) = \left(-1\right)^p \frac{\partial^p S\left(t_1,t_2,\ldots,t_p\right)}{\partial t_1, \partial t_2, \ldots, \partial t_p} = h_1 h_2 \ldots h_p \frac{\partial^p S_s\left(s_w\right)}{\partial s_w^p}, \tag{6.8}$$

where

$$h_j = \frac{\partial s_w}{\partial t_j} = \phi_j \lambda_j t_j^{\phi_j - 1}, \tag{6.9}$$

and

$$S_s\left(s_w\right) = \exp\left(-s_w^v\right). \tag{6.10}$$

Therefore, from (6.8) the density function for the bivariate $\left(j = 1,2\right)$ positive stable Weibull distribution may be written as follows

$$\begin{aligned} f\left(t_1,t_2\right) &= \left(-1\right)^2 \frac{\partial^2 S\left(t_1,t_2\right)}{\partial t_1 \partial t_2} \\ &= h_1 h_2 \left(v^2 s_w^{2v-2} - v\ \left(v-1\right) s_w^{v-2}\right) S_s\left(s_w\right) \\ &= \prod_{j=1}^{2} \lambda_j \phi_j t_j^{\phi_j - 1} \left(v^2 s_w^{2v-2} - v\ \left(v-1\right) s_w^{v-2}\right) \exp\left(-s_w^v\right), \end{aligned} \tag{6.11}$$

where $s_w = \lambda_1 t_1^{\phi_1} + \lambda_2 t_2^{\phi_2}$.

In a multivariate case ($p > 1$), the frailty causes positive association among the survival times $T_1,T_2,\ldots,T_p$. Under the positive stable Weibull model, when the characteristic exponent $v = 1$, the model degenerates to the independent Weibull model.

In other words, no frailty exists when $v=1$, and the components $T_1, T_2, \ldots, T_p$ are independent. Therefore, the null hypothesis for a positive stable frailty test is $H_0 : v = 1$, whilst the alternative hypothesis is $H_0 : 0 \leq v < 1$.

Zhu (1998) proposed a score test based on the positive stable Weibull model that was derived from the first derivative of the log-likelihood function of the model. Suppose that, there is no censored observation and the parameters of Weibull model λ_j and ϕ_j $(j=1,2)$ are known. The density function for a bivariate sample of size n $(t_{11}, t_{12}), \ldots, (t_{n1}, t_{n2})$ is given by

$$f(t_{i1}, t_{i2}) = \prod_{j=1}^{2} \lambda_j \phi_j t_j^{\phi_j - 1} \left(v^2 s_{wi}^{2v-2} - v \ (v-1) s_{wi}^{v-2} \right) \exp\left(-s_{wi}^{v}\right), \qquad (6.12)$$

where $i = 1, \ldots, n$. The log-likelihood function of the sample is given by

$$\ell_n(v) = -\sum_{i=1}^{n} s_{wi}^{v} + \sum_{i=1}^{n} \sum_{j=1}^{2} \ln\left(\lambda_j \phi_j t_{ij}^{\phi_j - 1}\right) + \sum_{i=1}^{n} \ln\left(v^2 s_{wi}^{2v-2} - v(v-1) \ s_{wi}^{v-2}\right). \qquad (6.13)$$

Hence, the first derivative of the log-likelihood with respect to v is

$$\frac{\partial \ell_n}{\partial v} = \sum_{i=1}^{n} \left(-s_{wi}^{v} \ln s_{wi} + \frac{1}{v} + \ln s_{wi} + \frac{s_{wi}^{v} + v s_{wi}^{v} \ln s_{wi} - 1}{v s_{wi}^{v} - v + 1} \right). \qquad (6.14)$$

The score statistic is obtained as the following

$$T_{(2)} = \frac{\partial \ell_n(v)}{\partial v} |_{v=1} = \sum_{i=1}^{n} \left(2 + 2 \ln s_{wi} - s_{wi} \ln s_{wi} - 1/s_{wi}\right), \qquad (6.15)$$

after substituting (6.14) for $v=1$ (under H_0).

Later on, Sarker (2002) proposed two score based tests namely the modified score test and $\ln s$ based test. The modified score test was derived from equation (6.15) by excluding the term $1/s_{wi}$ and the score statistic is given as follows

$$T_{(2)}^{*} = \sum_{i=1}^{n} \left(2 + 2 \ln s_{wi} - s_{wi} \ln s_{wi}\right). \qquad (6.16)$$

Meanwhile, the score statistic for the $\ln s$ based test is given by

$$T_{(2)}^{**} = \sum_{i=1}^{n} \ln s_{wi} \ . \tag{6.17}$$

6.3 Bivariate Positive Stable Gompertz Model

This study focuses on the bivariate positive stable Gompertz model $(p=2)$. The cumulative hazard function of the Gompertz distribution is given by

$$H(t) = a\left(e^{bt} - 1\right)\big/ b \, , \tag{6.18}$$

where a is a positive parameter and b is a shape parameter of the Gompertz model. The survival function for the bivariate positive stable Gompertz model is given by

$$S(t_1, t_2) = \exp\left(-s^v\right), \tag{6.19}$$

where $s = \sum_{j=1}^{2} a_j \left(e^{b_j t_j} - 1\right) / b_j$. The model degenerates to the basic Gompertz model when the exponent character $v = 1$, where there is no association between T_1 and T_2. The corresponding density function may be written from (6.8) as the following

$$f(t_1, t_2) = \prod_{j=1}^{2} a_j e^{b_j t_j} \left(v^2 s^{2v-2} - v \ (v-1) s^{v-2}\right) \exp\left(-s^v\right), \tag{6.20}$$

where s is defined as in (6.19). The log-likelihood function for uncensored bivariate sample from Gompertz distribution of size n $(t_{11}, t_{12}), \ldots, (t_{n1}, t_{n2})$ is given by

$$\ell_n(v) = -\sum_{i=1}^{n} s_i^v + \sum_{i=1}^{n} \sum_{j=1}^{2} \ln\left(a_j e^{b_j t_j}\right) + \sum_{i=1}^{n} \ln\left(v^2 s_i^{2v-2} - v \ (v-1) s_i^{v-2}\right), \tag{6.21}$$

,

where $s_i = a_1\left(e^{b_1 t_{i1}} - 1\right)/ b_1 + a_2\left(e^{b_2 t_{i2}} - 1\right)/ b_2$ and $i = 1, \ldots, n$. The first derivative of the log-likelihood function in (6.21) is given by

$$\frac{\partial \ell_n}{\partial v} = \sum_{i=1}^{n} \left(-s_i^v \ln s_i + \frac{1}{v} + \ln s_i + \frac{s_i^v + v s_i^v \ln s_i - 1}{v s_i^v - v + 1} \right), \tag{6.22}$$

which is similar to the first derivative of the log-likelihood of the positive Weibull model in (6.14). Hence, if the equation (6.22) is substituted for $v=1$, the similar score statistic of the Zhu's score test as in (6.15) is obtained as follows

$$T_{(2)}=\frac{\partial\ell_n\left(v\right)}{\partial v}|_{v=1}=\sum_{i=1}^{n}\left(2+2\ln s_i-s_i\ln s_i-1/s_i\right), \tag{6.23}$$

where $s_i=\sum_{j=1}^{2}a_j\left(e^{b_j t_{ij}}-1\right)/b_j$. Also, the other two tests proposed by Sarker (2002), the modified score

$$T_{(2)}^{*}=\sum_{i=1}^{n}\left(2+2\ln s_i-s_i\ln s_i\right), \tag{6.24}$$

and the $\ln s$ based test

$$T_{(2)}^{**}=\sum_{i=1}^{n}\ln s_i\ , \tag{6.25}$$

are derived from the log-likelihood function of the positive stable Gompertz model in the same manner as they were derived from the positive stable Weibull model. Thus, the Zhu's score test, modified score test and $\ln s$ based test may be applicable for detecting frailty in the positive stable Gompertz model.

6.4 Properties of the Score Tests for the Uncensored Case

6.4.1 The Asymptotic Null Properties of the Tests for the Uncensored Case without Nuisance Parameters

All these score statistics depend on the observations $\left(t_{i1},t_{i2}\right)$ where $i=1,\dots,n$ through the variables s_i which all have the same null distribution regardless the values of a_j and b_j. Hence, all these score statistics are the sum of independent, identically distributed variates. Such a property is useful for computing their asymptotic distributions.

Under the null hypothesis, $H_0 : v = 1$, and the assumption that a_j and b_j are known, the components t_{ij} $\left(j = 1, 2\right)$ of the *i*th observation are independently Gompertz distributed. If T is a random variable associated with the survival time of an individual and $S\left(t\right)$ is the corresponding survivor function, the random variable $Y = -\log S\left(T\right)$ has an exponential distribution with unit mean irrespective of the form $S\left(T\right)$ (Collet, 2003). Thus, $Y = H\left(t\right)$ because $H\left(t\right) = -\log S\left(t\right)$. Hence, the Gompertz cumulative hazard function, $a_j\left(e^{b_j t_{ij}} - 1\right)/b_j$ $\left(j = 1, 2\right)$ is distributed as a unit exponential.

The moment generating function for a unit exponential distribution variable Y is $M_Y\left(t\right) = \left(1-t\right)^{-1}$. The moment generating function of the sum of mutually independent random variables is the product of their moment generating functions. Here, since $s_i = \sum_{j=1}^{2} a_j\left(e^{b_j t_{ij}} - 1\right)/b_j$, therefore

$$M_{s_i}\left(t\right) = \left(1-t\right)^{-1} \times \left(1-t\right)^{-1} = \left(1-t\right)^{-2}. \qquad (6.26)$$

From the moment generating function in (6.26), $s_i = \sum_{j=1}^{2} a_j\left(e^{b_j t_{ij}} - 1\right)/b_j$ has gamma distribution with scale and shape parameters equal to 1 and 2, respectively. Hence, the density function of s_i is given by

$$f_0\left(s_i\right) = s_i e^{-s_i}. \qquad (6.27)$$

Expected values which may be derived from (6.27) are very useful for computing the null mean and variance of the score statistics. Those expected values are given by

$$
\begin{aligned}
&E\left(s_i^{-1}\right) = \int_0^\infty s_i^{-1} f_0\left(s_i\right) ds_i = \int_0^\infty e^{-s_i} ds_i = 1, \\
&E\left(\ln s_i\right) = \int_0^\infty \ln s_i f_0\left(s_i\right) ds_i = \Gamma'\left(2\right) = 1-\gamma, \\
&E\left(s_i \ln s_i\right) = \int_0^\infty s_i \ln s_i f_0\left(s_i\right) ds_i = \Gamma'\left(3\right) = 3-2\gamma, \\
&E\left(\ln^2 s_i\right) = \int_0^\infty \ln^2 s_i f_0\left(s_i\right) ds_i = \Gamma''\left(2\right), \\
&E\left(s_i \ln^2 s_i\right) = \int_0^\infty s_i \ln^2 s_i f_0\left(s_i\right) ds_i = \Gamma''\left(3\right), \\
&E\left(s_i^2 \ln^2 s_i\right) = \int_0^\infty s_i^2 \ln^2 s_i f_0\left(s_i\right) ds_i = \Gamma''\left(4\right), \\
&E\left(s_i^{-1} \ln s_i\right) = \int_0^\infty s_i^{-1} \ln s_i f_0\left(s_i\right) ds_i = \Gamma'\left(1\right) = -\gamma, \\
&E\left(s_i^{-2}\right) = \int_0^\infty s_i^{-2} f_0\left(s_i\right) ds_i = \int_0^\infty s_i^{-1} e^{-s_i} ds_i = \infty,
\end{aligned} \tag{6.28}
$$

where γ is Euler's constant and $\Gamma\left(\bullet\right)$ is the gamma function.

6.4.1.1 The Score Test

The mean of the score test $T_{(2)}$ in (6.23) may be written as

$$E\left(T_{(2)}\right) = \sum_{i=1}^{n} E\left(T_i\right), \tag{6.29}$$

where $T_i = 2 + 2\ln s_i - s_i \ln s_i - 1/s_i$. Using the expected values in (6.28), $E\left(T_i\right)$ may be obtained as the following

$$
\begin{aligned}
E\left(T_i\right) &= E\left(2 + 2\ln s_i - s_i \ln s_i - 1/s_i\right) \\
&= 2 + 2E\left(\ln s_i\right) - E\left(s_i \ln s_i\right) - E\left(1/s_i\right) \\
&= 2 + 2\left(1-\gamma\right) - \left(3-2\gamma\right) - 1 = 0 \\
&= 2\gamma - 2\gamma + 2 + 2 - 3 - 1 = 0 .
\end{aligned} \tag{6.30}
$$

Hence, from (6.29), the mean of $T_{(2)}$ under the null hypothesis is given by

$$E\left(T_{(2)}\right) = \sum_{i=1}^{n} E\left(T_i\right) = n \times 0 = 0 . \tag{6.31}$$

Meanwhile, the variance of $T_{(2)}$ may be expressed as

$$Var\left(T_{(2)}\right)=\sum_{i=1}^{n}E\left(T_i^2\right)-\sum_{i=1}^{n}E^2\left(T_i\right). \tag{6.32}$$

Under the null hypothesis, $E\left(T_i^2\right)$ may be obtained using (6.28) as

$$\begin{aligned}E\left(T_i^2\right)&=E\left(\left\{2+2\ln s_i-s_i\ln s_i-\frac{1}{s_i}\right\}^2\right)\\&=4+4E\left(\ln s_i\right)^2+E\left(s_i\ln s_i\right)^2+10E\left(\ln s_i\right)-4E\left(s_i\ln s_i\right)\\&\quad-4E\left(s_i\ln^2 s_i\right)-4E\left(1/s_i\right)-4E\left(\log s_i/s_i\right)+E\left(1/s_i^2\right).\end{aligned} \tag{6.33}$$

Each term in equation (6.33) has a finite value except that for $E\left(1/s_i^2\right)$ which is infinite. It follows that the variance of $T_{(2)}$ under the null hypothesis is

$$Var\left(T_{(2)}\right)=\infty. \tag{6.34}$$

Since the null variance of $T_{(2)}$ is infinite, the common central limit theorem argument is inapplicable to $T_{(2)}$. Instead, Zhu (1998) applied the central limit theorem for infinite variance (Feller, 1966) to obtain a standard normal test statistics as the following (see Zhu (1998) for the detail)

$$S_{(2)}=\frac{T_{(2)}}{\sqrt{\frac{1}{2}n\ln n}}\rightarrow N\left(0,1\right)\text{ as }n\rightarrow\infty. \tag{6.35}$$

6.4.1.2 The Modified Score Test

The mean of $T_{(2)}^*$ in (6.24) is

$$E\left(T_{(2)}^*\right)=\sum_{i=1}^{n}E\left(T_i^*\right), \tag{6.36}$$

where $T_i^*=2+2\ln s_i-s_i\ln s_i$. Thus, using the expected values in (6.28)

$$E\left(T_i^*\right)=E\left(2+2\ln s_i-s_i\ln s_i\right)$$

$$=2+2\left(1-\gamma\right)-\left(3-2\gamma\right)=1$$

$$=2\gamma-2\gamma+2+2-3=1. \tag{6.37}$$

and the mean of $T_{(2)}^*$ under the null hypothesis may be expressed as

$$E_0\left(T_{(2)}^*\right)=\sum_{i=1}^{n}E\left(T_i^*\right)=n\times 1=n\mu_0^*, \tag{6.38}$$

where μ_0^* is the null mean of $T_{(2)}^*$ for a single observation. The variance of $T_{(2)}^*$ under the null hypothesis may be expressed as

$$Var\left(T_{(2)}^*\right)=\sum_{i=1}^{n}E\left(T_i^{*2}\right)-\sum_{i=1}^{n}E^2\left(T_i^*\right), \tag{6.39}$$

where

$$E\left(T_i^{*2}\right)=E\left(\left(2+2\ln s_i-s_i\ln s_i\right)^2\right)$$

$$=4+4E\left(\ln s_i\right)^2+E\left(s_i\ln s_i\right)^2$$

$$+8E\left(\ln s_i\right)-4E\left(s_i\ln s_i\right)-4E\left(s_i\ln^2 s_i\right)$$

$$=4+\Gamma''\left(4\right)+4\Gamma''\left(2\right)-4\left(3-2\gamma\right)-4\Gamma''\left(3\right)+8\left(1-\gamma\right)$$

$$=2\gamma^2-6\gamma+\frac{\pi^2}{3}+4. \tag{6.40}$$

By substituting the Euler’s constant $\gamma=0.577215664$ and $\pi=3.1415926535$, the null variance of $T_{(2)}^*$ is

$$Var\left(T_{(2)}^*\right)=\sum_{i=1}^{n}E\left(T_i^{*2}\right)-\sum_{i=1}^{n}E^2\left(T_i^*\right)$$

$$=n\left(2\gamma^2-6\gamma+\frac{\pi^2}{3}+4\right)-n$$

$$=n\times\left(2\gamma^2-6\gamma+\frac{\pi^2}{3}+4-1\right)$$

$$=n\times 3.492929993$$

$$=n\sigma_0^{*2}, \tag{6.41}$$

where σ_0^{*2} is the null variance of $T_{(2)}^{*}$ for a single observation. By the central limit theorem, the standardised score statistic $S_{(2)}^{*}$ under the null hypothesis is given as the following

$$S_{(2)}^{*}=\frac{T_{(2)}^{*}-n\mu_0^{*}}{\sqrt{n\sigma_0^{*2}}}=\frac{T_{(2)}^{*}-n}{\sqrt{n\times 3.4929299993}}\rightarrow N\left(0,1\right)\text{ as }n\rightarrow\infty\,, \qquad (6.42)$$

where from (6.38), $\mu_0^{*}=1$, whilst $\sigma_0^{*2}=3.4929299993$ as given in (6.41).

6.4.1.3 The Test based on $\ln s$

The mean of $T_{(2)}^{**}$ under the null hypothesis is derived as follows

$$\begin{aligned}E\left(T_{(2)}^{**}\right)&=\sum_{i=1}^{n}E\left(\ln s_i\right)\\&=n\times\left(1-\gamma\right)\\&=n\times 0.4227843351\\&=n\mu_0^{**}\,,\end{aligned} \qquad (6.43)$$

where the Euler's constant $\gamma=0.577215664$. Also, the null variance of $T_{(2)}^{**}$ is

$$\begin{aligned}Var\left(T_{(2)}^{**}\right)&=\sum_{i=1}^{n}E\left(\ln^2 s_i\right)-\sum_{i=1}^{n}E^2\left(\ln s_i\right)\\&=n\times\left(\frac{\pi^2}{6}-1\right)\\&=n\times 0.6449340675\\&=n\sigma_0^{**2}\,,\end{aligned} \qquad (6.44)$$

where $\pi=3.1415926535$. Therefore, by the central limit theorem, the standardised score statistic for $T_{(2)}^{**}$ under the null hypothesis $S_{(2)}^{**}$ is

$$S_{(2)}^{**}=\frac{T_{(2)}^{**}-n\mu_0^{**}}{\sqrt{n\sigma_0^{**2}}}=\frac{\sum_{i=1}^{n}\ln s_i-n\times 0.4227843351}{\sqrt{n\times 0.6449340675}}\rightarrow N\left(0,1\right)\text{ as }n\rightarrow\infty\,, \qquad (6.45)$$

where from (6.43), $\mu_0^{**} = 0.4227843351$, whilst $\sigma_0^{**2} = 0.6449340675$ as given in (6.44).

6.4.2 The Non-null Case

The density function of s_i under the alternative hypothesis $H_1 : 0 < v < 1$ is given by

$$f_1(s_i) = \exp\left(-s_i^v\right)\left\{v^2 s_i^{2v-1} - v(v-1)s_i^{v-1}\right\}. \tag{6.46}$$

Some expected values those are obtained from (6.46) are needed for computing the non-null mean and variance of the tests. These expected values are given as follows:

$$\begin{aligned} E\left(s_i^{-1}\right) &= \int_0^\infty s_i^{-1} f_1(s_i)\, ds_i = -\infty, \\ E(\ln s_i) &= \int_0^\infty \ln s_i\, f_1(s_i)\, ds_i = 1 - \gamma / v, \\ E(s_i \ln s_i) &= \int_0^\infty s_i \ln s_i\, f_1(s_i)\, ds_i = v^2 \left[\left\{2\Psi(1/v) + 3v\right\}\Gamma(1/v)\right], \end{aligned} \tag{6.47}$$

where $\Psi(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$ is the di-gamma function.

The non-null mean of $T_{(2)}$ is given by

$$\begin{aligned} E\left(T_{(2)}\right) &= \sum_{i=1}^{n} E(T_i) \\ &= \sum_{i=1}^{n} E\left(2 + 2\ln s_i - s_i \ln s_i - 1/s_i\right) \\ &= 2 + 2(1 - \gamma / v) - E(s_i \ln s_i) - E(1/s_i) = -\infty . \end{aligned} \tag{6.48}$$

Also, Zhu (1998) showed that the non-null variance of $T_{(2)}$ is undefined since the term

$$E\left(s_i^{-2}\right) = \int_0^\infty s_i^{-2} f_1(s_i)\, ds_i = \infty . \tag{6.49}$$

The non-null mean and variance for $T_{(2)}^*$ and $T_{(2)}^{**}$ are given in Appendix C.

6.4.3 The Asymptotic Null Properties of the Tests for the Uncensored Case with Nuisance Parameters

In the case of parameters a_j and b_j are unknown, all the score statistics considered may be denoted as $\hat{T}_{(2)}$, $\hat{T}^{*}_{(2)}$ and $\hat{T}^{**}_{(2)}$. The score statistics $\hat{T}_{(2)}$, $\hat{T}^{*}_{(2)}$ and $\hat{T}^{**}_{(2)}$ are the same as $T_{(2)}$, $T^{*}_{(2)}$ and $T^{**}_{(2)}$, respectively, yet with a_j and b_j substituted by their corresponding maximum likelihood estimators (MLEs) under the null hypothesis. Thus, the score statistics for uncensored case with nuisance parameters are

$$\hat{T}_{(2)} = \sum_{i=1}^{n} \left(2 + 2\ln \hat{s}_i - \hat{s}_i \ln \hat{s}_i - 1/\hat{s}_i \right), \qquad (6.50)$$

$$\hat{T}^{*}_{(2)} = \sum_{i=1}^{n} \left(2 + 2\ln \hat{s}_i \text{ - } \hat{s}_i \ln \hat{s}_i \right), \qquad (6.51)$$

and

$$\hat{T}^{**}_{(2)} = \sum_{i=1}^{n} \ln \hat{s}_i \ , \qquad (6.52)$$

where $\hat{s}_i = \sum_{j=1}^{2} \hat{a}_j \left(e^{\hat{b}_j t_{ij}} - 1 \right) / \hat{b}_j$ and $\hat{a}_j$ and $\hat{b}_j$ are the MLEs of a_j, b_j under the null hypothesis.

The asymptotic mean of the test statistic for the case with nuisance parameters are similar to that of without nuisance parameters case. However, for the case with nuisance parameters, the asymptotic variances may reduce (Kimber, 1996; Crowder & Kimber, 1997). Sarker (2002) pointed out that the null variances for the non-Weibull distribution case may be different from his study finding when the nuisance parameters are replaced by their null MLEs. Therefore, in this study, the asymptotic variances for $\hat{T}^{*}_{(2)}$ and $\hat{T}^{**}_{(2)}$ for the positive stable Gompertz model are derived. The derivations of these asymptotic variances are based on the Pierce (1982) theorem. The asymptotic variance for $\hat{T}_{(2)}$ is not obtained using the Pierce (1982) results since the variance of $T_{(2)}$ is infinite.

6.4.3.1 The Modified Score Test $\hat{T}^{*}_{(2)}$

The variance of $\hat{T}^{*}_{(2)}$ in equation (6.51) under the null hypothesis is given by

$$Var\left(\hat{T}^{*}_{(2)}\right) = n\bar{\sigma}_0^{*2} = n\sigma_0^{*2} - B^T J_0^{-1} B_0 \, , \tag{6.53}$$

where $J_0 = E_0\left(-\partial^2 \ell_0 \big/ \partial\theta^2\right)$is the expected information matrix, and

$$B_0 = E\left(\partial T^{*}_{(2)} \big/ \partial\theta\right) = \sum_{i=1}^{n} E\left\{\left(2/s_i - \ln s_i - 1\right)\partial s_i / \partial\theta\right\}. \tag{6.54}$$

Under the null hypothesis, the log-likelihood function may be written as

$$\ell_0 = \sum_{i=1}^{n}\sum_{j=1}^{2}\left\{\ln a_j + b_i t_{ij}\right\} - \sum_{i=1}^{n} s_i \ . \tag{6.55}$$

The formulae for computing $B_0^T J_0^{-1} B_0$ are derived for $\hat{T}^{*}_{(2)}$ by considering $\ln a_{ij} = \beta_j$ $\left(j = 1, 2\right)$. The corresponding formulae are

$$B_0 = \begin{pmatrix} B_{0\beta} \\ B_{0b} \end{pmatrix}, \quad J_0 = \begin{pmatrix} J_{0\beta\beta} & J_{0\beta b} \\ J_{0b\beta} & J_{0bb} \end{pmatrix};$$

$$\begin{aligned}
\left\{B_{0\beta(2\times1)}\right\}_k &= \sum_{i=1}^{n} E_0\left\{\left(2/s_i - \ln s_i - 1\right)\partial s_i \big/ \partial\beta_k\right\} \\
&= \sum_{i=1}^{n}\int_0^{\infty}\int_0^{\infty}\left\{\left(2 / \sum_{j=1}^{2}\frac{e^{\beta_j}}{b_j}\left(e^{b_j t_{ij}} - 1\right)\right) - \ln\ \left(\sum_{j=1}^{2}\frac{e^{\beta_j}}{b_j}\left(e^{b_j t_{ij}} - 1\right)\right) - 1\right\}\frac{e^{\beta_k}}{b_k}\left(e^{b_k t_{ik}} - 1\right) \\
&\quad \times \prod_{j=1}^{2} e^{\beta_j} e^{b_j t_{ij}} \exp\left(-\frac{e^{\beta_j}}{b_j}\left(e^{b_j t_{ij}} - 1\right)\right) dt_{ij} \, ,
\end{aligned}$$

$$\left\{B_{0b(2\times1)}\right\}_k = \sum_{i=1}^{n} E_0\left\{\left(2/s_i - \ln s_i - 1\right)\ \partial s_i/\partial b_k\right\}$$

$$= \sum_{i=1}^{n}\int_0^{\infty}\int_0^{\infty}\left\{\left(2/\sum_{j=1}^{2}\frac{e^{\beta_j}}{b_j}\left(e^{b_jt_{ij}}-1\right)\right)-\ln\ \left(\sum_{j=1}^{2}\frac{e^{\beta_j}}{b_j}\left(e^{b_jt_{ij}}-1\right)\right)-1\right\}$$

$$\times\left(\frac{\left(b_je^{\beta_k}e^{b_kt_{ik}}t_{ik}\right)-e^{\beta_k}\left(e^{b_kt_{ik}}-1\right)}{b_k^2}\right)\prod_{j=1}^{2}e^{\beta_j}e^{b_jt_{ij}}\exp\left(-\frac{e^{\beta_j}}{b_j}\left(e^{b_jt_{ij}}-1\right)\right)dt_{ij},$$

$$\left\{J_{0\beta\beta(2\times2)}\right\}_{kl} = E_0\left(-\frac{\partial^2\ell_0}{\partial\beta_k\partial\beta_l}\right)$$

$$= E_0\left[-\frac{\partial^2}{\partial\beta_k\partial\beta_l}\left\{\sum_{i=1}^{n}\sum_{j=1}^{p}\left(\beta_j+b_jt_{ij}-\frac{e^{\beta_j}}{b_j}\left(e^{b_jt_{ij}}-1\right)\right)\right\}\right]$$

$$= E_0\left[\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{e^{\beta}}{b}\left(e^{bt}-1\right)\right]$$

$$= \delta_{kl}\sum_{i=1}^{n}\int_0^{\infty}\frac{e^{\beta}}{b}\left(e^{bt}-1\right)f\left(t_{ik}\right)dt_{ik}$$

$$= \delta_{kl}\sum_{i=1}^{n}\int_0^{\infty}\frac{e^{\beta}}{b}\left(e^{bt}-1\right)\ e^{\beta}e^{bt}\exp\left(-\frac{e^{\beta}}{b}\left(e^{bt}-1\right)\right)dt_{ik},$$

$$\left\{J_{0\beta b(2\times2)}\right\}_{kl} = E_0\left(-\frac{\partial^2 \ell_0}{\partial\beta_k \partial b_l}\right)$$

$$= E_0\left[-\frac{\partial^2}{\partial\beta_k\partial b_l}\left\{\sum_{i=1}^{n}\sum_{j=1}^{p}\left(\beta_j + b_j t_{ij} - \frac{e^{\beta_j}}{b_j}\left(e^{b_j t_{ij}}-1\right)\right)\right\}\right]$$

$$= E_0\left[\sum_{\mathrm{i}=1}^{n}\sum_{j=1}^{p}\left\{\frac{\left(b_j e^{\beta_j} e^{b_j t_{ij}} t_{ij}\right) - e^{\beta_j}\left(e^{b_j t_{ij}}-1\right)}{b_j^2}\right\}\right]$$

$$= \delta_{kl}\sum_{i=1}^{n}\int_0^\infty\left\{\frac{\left(b e^{\beta_k} e^{b_k t_{ik}} t_{ik}\right) - e^{\beta_k}\left(e^{b_k t_{ik}}-1\right)}{b_j^2}\right\} f\left(t_{ik}\right)\, dt_{ik}$$

$$= \delta_{kl}\sum_{i=1}^{n}\int_0^\infty\left\{\frac{\left(b e^{\beta_k} e^{b_k t_{ik}} t_{ik}\right) - e^{\beta_k}\left(e^{b_k t_{ik}}-1\right)}{b_j^2}\right\} e^{\beta_k} e^{b_k t_{ik}} \exp\left(-\frac{e^{\beta_k}}{b_k}\left(e^{b_k t_{ik}}-1\right)\right) dt_{ik},$$

$$\left\{J_{0bb(2\times2)}\right\}_{kl} = E_0\left[-\frac{\partial^2 \ell_0}{\partial b_k \partial b_l}\right]$$

$$= E_0\left[-\frac{\partial^2}{\partial b_k\partial b_l}\left\{\sum_{i=1}^{n}\sum_{j=1}^{p}\left(\beta_j + b_j t_{ij} - \frac{e^{\beta_j}}{b_j}\left(e^{b_j t_{ij}}-1\right)\right)\right\}\right]$$

$$= E_0\left[\sum_{i=1}^{n}\sum_{j=1}^{p}\left\{\frac{\left(b_j e^{\beta_j} e^{b_j t_{ij}} t_{ij}^2\right) - 2\left(e^{\beta_j} e^{b_j t_{ij}} t_{ij}\right)}{b_j^2} + \frac{2e^{\beta_j}\left(e^{b_j t_{ij}}-1\right)}{b_j^3}\right\}\right]$$

$$= \delta_{kl}\sum_{i=1}^{n}\int_0^\infty\left\{\frac{\left(b_k e^{\beta_k} e^{b_k t_{ik}} t_{ik}^2\right) - 2\left(e^{\beta_k} e^{b_k t_{ik}} t_{ik}\right)}{b_k^2} + \frac{2e^{\beta_k}\left(e^{b_k t_{ik}}-1\right)}{b_k^3}\right\} f\left(t_{ik}\right)\, dt_{ik}$$

$$= \delta_{kl}\sum_{i=1}^{n}\int_0^\infty\left\{\frac{\left(b_k e^{\beta_k} e^{b_k t_{ik}} t_{ik}^2\right) - 2\left(e^{\beta_k} e^{b_k t_{ik}} t_{ik}\right)}{b_k^2} + \frac{2e^{\beta_k}\left(e^{b_k t_{ik}}-1\right)}{b_k^3}\right\} e^{\beta_k} e^{b_k t_{ik}}$$

$$\times \exp\left(-\frac{e^{\beta_k}}{b_k}\left(e^{b_k t_{ik}}-1\right)\right) dt_{ik},$$

where the Euler's constant is $\gamma = 0.577215664$ and δ_{kl} is the Kronecker delta. Without loss of generality, the computation is simplified by letting the value of $\beta_j = 0 \ \left(a_j = 1\right)$

and $b_j = 1 \quad \left(j = 1, 2 \right)$. Then, the following values are obtained through numerical integration by *Mathematica* software version 10,

$$B_0 = \begin{bmatrix} n\left(-0.922784\right) \\ n\left(-0.922784\right) \\ n\left(-0.853955\right) \\ n\left(-0.853955\right) \end{bmatrix}, \tag{6.56}$$

$$J_0 = \begin{bmatrix} n & 0 & n\left(0.596347\right) & 0 \\ 0 & n & 0 & n\left(0.596347\right) \\ n\left(0.596347\right) & 0 & n\left(0.531931\right) & 0 \\ 0 & n\left(0.596347\right) & 0 & n\left(0.531931\right) \end{bmatrix}. \tag{6.57}$$

Thus, from (6.56) and (6.57), $B_0^T J_0^{-1} B_0$ may be obtained as

$$B_0^T J_0^{-1} B_0 = n \times 2.749073, \tag{6.58}$$

after substituting for $\beta_j = 0$ and $b_j = 1$. Hence, using $\sigma_0^{*2} = 3.4929299993$ from (6.41), and $B_0^T J_0^{-1} B_0$ in (6.58), the variance of $\hat{T}_{(2)}^*$ under the null hypothesis is given by

$$\begin{aligned} Var\left(\hat{T}_{(2)}^*\right) = n\bar{\sigma}_0^{*2} &= n\sigma_0^{*2} - B^T J_0^{-1} B_0 \\ &= \left(n \times 3.492929993\right) - \left(n \times 2.749073\right) \\ &= n \times \left(3.492929993 - 2.749073\right) \\ &= n \times 0.7438566. \end{aligned} \tag{6.59}$$

Following the central limit theorem, using (6.38) and (6.59), the standardised score statistic $\hat{S}_{(2)}^*$ under the H_0 is asymptotically distributed as standard normal, that is given by

$$\hat{S}_{(2)}^* = \frac{\hat{T}_{(2)}^* - n}{\sqrt{n\bar{\sigma}_0^{*2}}} = \frac{\hat{T}_{(2)}^* - n}{\sqrt{n \times 0.7438566}} \rightarrow N\left(0,1\right) \text{ as } n \rightarrow \infty. \tag{6.60}$$

The ratio of the null variance of $T_{(2)}^{*}$ in (6.41) to the null variance of $\hat{T}_{(2)}^{*}$ in (6.59) is

$$\sigma_0^{*2} \big/ \bar{\sigma}_0^{*2} = 3.492929993 \big/ 0.7438566 = 4.696 . \tag{6.61}$$

This explains that the asymptotic null variance of $T_{(2)}^{*}$ is four times as large as that of $\hat{T}_{(2)}^{*}$. Thus, the use of $\hat{T}_{(2)}^{*}$ as if it has the same null distribution as $T_{(2)}^{*}$ would lead to a conservative test.

6.4.3.2 The Test based on ln s $\hat{T}_{(2)}^{**}$

The null variance of $\hat{T}_{(2)}^{**}$ is given by

$$Var\left(\hat{T}_{(2)}^{**}\right) = n\bar{\sigma}_0^{**2} = n\sigma_0^{**2} - B^T J_0^{-1} B_0 , \tag{6.62}$$

where $J_0 = E_0\left(-\partial^2 \ell_0 \big/ \partial\theta^2\right)$ is the expected information matrix in (6.57), and

$$B_0 = E_0\left(\partial T_{(2)}^{**} \big/ \partial\theta\right) = \sum_{i=1}^{n} E_0 \left\{ 1\big/ s_i \, \partial s_i \big/ \partial\theta \right\} . \tag{6.63}$$

$$\left\{B_{0\beta(2\times1)}\right\}_k = \sum_{i=1}^{n} E_0 \left\{ \left(1/s_i\right) \partial s_i / \partial\beta_k \right\}$$

$$= \sum_{i=1}^{n} \int_0^\infty \int_0^\infty \left\{ \left(1/ \sum_{j=1}^{2} \frac{e^{\beta_j}}{b_j} \left(e^{b_j t_{ij}} - 1\right) \right) \right\} \frac{e^{\beta_k}}{b_k} \left(e^{b_k t_{ik}} - 1\right)$$

$$\times \prod_{j=1}^{2} e^{\beta_j} e^{b_j t_{ij}} \exp\left(-\frac{e^{\beta_j}}{b_j} \left(e^{b_j t_{ij}} - 1\right) \right) dt_{ij},$$

$$\left\{B_{0b(2\times1)}\right\}_k = \sum_{i=1}^{n} E_0 \left\{ \left(1\big/ s_i\right) \partial s_i \big/ \partial b_k \right\}$$

$$= \sum_{i=1}^{n} \int_0^\infty \int_0^\infty \left\{ \left(1/ \sum_{j=1}^{2} \frac{e^{\beta_j}}{b_j} \left(e^{b_j t_{ij}} - 1\right) \right) \right\} \left(\frac{\left(b_j e^{\beta_k} e^{b_k t_{ik}} t_{ik}\right) - e^{\beta_k} \left(e^{b_k t_{ik}} - 1\right)}{b_k^2} \right)$$

$$\times \prod_{j=1}^{2} e^{\beta_j} e^{b_j t_{ij}} \exp\left(-\frac{e^{\beta_j}}{b_j} \left(e^{b_j t_{ij}} - 1\right) \right) dt_{ij} .$$

By substituting for $\beta_j = 0$ and $b_j = 1$, B_0 is obtained as follows

$$B_0 = \begin{bmatrix} n(0.5) \\ n(0.5) \\ n(0.218945) \\ n(0.218945) \end{bmatrix}. \tag{6.64}$$

Using (6.57) and (6.64),

$$B_0^T J_0^{-1} B_0 = n \times\ 0.5712094, \tag{6.65}$$

and the null variance of $\hat{T}_{(2)}^{**}$ is given by

$$\begin{aligned} Var\left(\hat{T}_{(2)}^{**}\right) &= n\bar{\sigma}_0^{**2} = n\sigma_0^{**2} - B^T J_0^{-1} B_0 \\ &= (n \times 0.6449340675) - (n \times 0.5712094) \\ &= n \times (0.6449340675 - 0.5712094) \\ &= n \times\ 0.07372465, \end{aligned} \tag{6.66}$$

where from (6.44), $\sigma_0^{**2} = 0.6449340675$ and from (6.65), $B_0^T J_0^{-1} B_0 = 0.5712094$.

By the central limit theorem, the standardised score statistic $\hat{S}_{(2)}^{**}$ under the null hypothesis is

$$\hat{S}_{(2)}^{**} = \frac{\hat{T}_{(2)}^{**} - n\mu_0^{**}}{\sqrt{n\bar{\sigma}_0^{**2}}} = \frac{\sum_{i=1}^{n} \ln \hat{s}_i - n \times 0.4227843351}{\sqrt{n \times\ 0.07372465}} \to N(0,1) \text{ as } n \to \infty. \tag{6.67}$$

The ratio of the null variance of $T_{(2)}^{**}$ in (6.44) to the null variance of $\hat{T}_{(2)}^{**}$ (6.66) is given by

$$\sigma_0^{**2} / \bar{\sigma}_0^{**2} = 0.6449340675 / \ 0.07372465 = \ 8.748. \tag{6.68}$$

The use of $\hat{T}_{(2)}^{**}$ as if it has the same null distribution as $T_{(2)}^{**}$ would lead to a very conservative test as the asymptotic null variance of $T_{(2)}^{**}$ is about eight times larger than that of $\hat{T}_{(2)}^{**}$.

6.5 Properties of the Score Tests for the Censored Case

6.5.1 Censored Case without Nuisance Parameters

Consider now the case that T_j $\left(j=1,2\right)$ might be right censored and the Gompertz parameters a_j and b_j are known. Supposed that T_1 and T_2 are censored at fixed times c_1 and c_2, respectively. There are four possibilities may be observed for a particular individual with index i as follows (Zhu, 1998):

a) The likelihood contribution from $\left(t_{i1},t_{i2}\right)$, where t_{i1} and t_{i2} are uncensored is

$$L_i\left(v;t_{i1},t_{i2}\right)=\frac{\partial^2 S\left(t_{i1},t_{i2}\right)}{\partial t_{i1}\partial t_{i2}}=\prod_{j=1}^{2}a_j e^{b_j t_{ij}}\left(v^2 s_i^{2v-2}-v\left(v-1\right)s_i^{v-2}\right)\exp\left(-s_i^v\right), \tag{6.69}$$

and its corresponding contribution to the score statistic under the null hypothesis is

$$T_i\left(t_{i1},t_{i2}\right)=\frac{\partial \ell_i\left(v;t_{i1},t_{i2}\right)}{\partial v}\Big|_{v=1}=2+2\ln s_i-s_i\ln s_i-\frac{1}{s_i}, \tag{6.70}$$

where $s_i=a_1\left(e^{b_1 t_{i1}}-1\right)/b_1+a_2\left(e^{b_2 t_{i2}}-1\right)/b_2$.

b) The likelihood contribution from $\left(t_{i1},t_{i2}\right)$, where t_{i1} is censored and t_{i2} is observed is

$$L_i\left(v;c_1,t_{i2}\right)=\frac{\partial S\left(c_1,t_{i2}\right)}{\partial t_{i2}}=a_2 e^{b_{2j}t_{i2}}\left(v s_i^{v-1}\right)\exp\left(-s_i^v\right), \tag{6.71}$$

and its corresponding contribution to the score statistic under the null hypothesis is

$$T_i\left(c_1,t_{i2}\right)=\frac{\partial \ell_i\left(v;c_1,t_{i2}\right)}{\partial v}\Big|_{v=1}=1+\ln s_i-s_i\ln s_i, \tag{6.72}$$

where $s_i=a_1\left(e^{b_1 c_1}-1\right)/b_1+a_2\left(e^{b_2 t_{i2}}-1\right)/b_2$.

c) The likelihood contribution from $\left(t_{i1},t_{i2}\right)$, where t_{i1} is observed and t_{i2} is censored is

$$L_i\left(v;t_{i1},c_2\right)=\frac{\partial S\left(t_{i1},c_2\right)}{\partial t_{i1}}=a_1 e^{b_{1j}t_{i1}}\left(vs_i^{v-1}\right)\exp\left(-s_i^v\right), \qquad (6.73)$$

and its corresponding contribution to the score statistic under the null hypothesis

$$T_i\left(t_{i1},c_2\right)=\frac{\partial \ell_i\left(v;t_{i1},c_2\right)}{\partial v}\Big|_{v=1}=1+\ln s_i - s_i \ln s_i, \qquad (6.74)$$

where $s_i = a_1\left(e^{b_1 t_{i1}}-1\right)/b_1 + a_2\left(e^{b_2 c_2}-1\right)/b_2$.

d) The likelihood contribution from $\left(t_{i1},t_{i2}\right)$, where t_{i1} and t_{i2} are censored is

$$L\left(v\,|\,c_1,c_2\right)=S\left(c_1,c_2\right)=\exp\left(-s_i^v\right), \qquad (6.75)$$

and its corresponding contribution to the score function is

$$T_i\left(c_1,c_2\right)=\frac{\partial \ell_i\left(v\,|\,c_1,c_2\right)}{\partial v}\Big|_{v=1}=-s_i \ln s_i, \qquad (6.76)$$

where $s_i = a_1\left(e^{b_1 c_1}-1\right)/b_1 + a_2\left(e^{b_2 c_2}-1\right)/b_2$.

Summarising the aforementioned information in (6.70) to (6.76), the Zhu’s score test, modified score test and $\ln s$ based test for the bivariate model with censored observations is given by

$$T_{(2),c}=\sum_{i=1}^{n}\left\{I_i\left(1+\ln s_i\right)-s_i \ln s_i - \frac{I_i\left(I_i-1\right)}{2s_i}\right\}, \qquad (6.77)$$

$$T_{(2),c}^{*}=\sum_{i=1}^{n}\left\{I_i\left(1+\ln s_i\right)-s_i \ln s_i\right\}, \qquad (6.78)$$

and

$$T_{(2),c}^{**}=\sum_{i=1}^{n}\left\{I_i \ln s_i\right\}, \qquad (6.79)$$

respectively, where $s_i = a_1\left(e^{b_1 t_{i1}}-1\right)/b_1 + a_2\left(e^{b_2 t_{i2}}-1\right)/b_2$ and I_i $\left(i=1,2,\ldots,n\right)$ is an indicator variable defined as follows

$$I_i = \begin{cases} 0 & \text{if } t_{i1} \text{ and } t_{i2} \text{ are both censored,} \\ 1 & \text{if exactly one of } t_{i1} \text{ and } t_{i2} \text{ is censored}, \\ 2 & \text{if } t_{i1} \text{ and } t_{i2} \text{ are both uncensored.} \end{cases} \tag{6.80}$$

6.5.1.1 The Asymptotic Null Properties of $T_{(2),c}$

Zhu (1998) showed that under the null hypothesis

$$E\left(T_{(2),c}\right) = nE\left(I_i\left(1+\ln s_i\right) - s_i \ln s_i - \frac{I\left(I-1\right)}{2s_i}\right) = 0\,. \tag{6.81}$$

Meanwhile, $Var\left(T_{(2),c}\right) = \infty$ as the last term $I_i\left(I_i-1\right)/2s_i$ in (6.77) comes from (6.70) that has infinite variance. Thus, Zhu (1998) used the non-regular normalisation as in (6.35), that is given by

$$S_{(2),c} = \frac{T_{(2),c}}{\sqrt{\frac{1}{2}n\ln n}} \to N\left(0,1\right) \text{ as } n \to \infty\,. \tag{6.82}$$

6.5.1.2 The Asymptotic Null Properties of $T^*_{(2),c}$

For calculating the mean of $T^*_{(2),c}$ under the null hypothesis, the expression of $T^*_{(2),c}$ is specified as follows

$$T^*_{(2),c} = \sum_{i=2}^{n} g_i\left(t_{i1}, t_{i2}\right), \tag{6.83}$$

where

$$g_i\left(t_{i1}, t_{i2}\right) = I\left(1+\ln s_i\right) - s_i \ln s_i\,, \tag{6.84}$$

The subscript i of $g_i\left(t_{i1}, t_{i2}\right)$ is omitted in the following discussion since the modified score function depends on s_i which are identically distributed. Hence, since

$E\left(T_{(2),c}\right)=0$ from (6.81) and the component of t_j is right censored at c_j $\left(j=1,2\right)$, the expected value of $T^*_{(2),c}$ with respect to a joint density $f\left(t_1,t_2\right)$ is given by

$$E\left(T^*_{(2),c}\right)=nE\left(\frac{I\left(I-1\right)}{2s}\right)=n\int_0^{c_2}\int_0^{c_1}\frac{1}{s}\,f\left(t_1,t_2\right)\,dt_1dt_2\,. \tag{6.85}$$

Meanwhile, the variance of $T^*_{(2),c}$ is given by

$$Var\left(T^*_{(2),c}\right)=n\left\{E\left(g^2\left(t_1,t_2\right)\right)-E^2\left(g\left(t_1,t_2\right)\right)\right\}. \tag{6.86}$$

Let $z_j=a_j\left(e^{b_jt_j}-1\right)/\,b_j$ and $d_j=a_j\left(e^{b_jc_j}-1\right)/\,b_j$ $\left(j=1,2\right)$. Under the null hypothesis and the assumption that a_j and b_j are known, z_1 and z_2 are independently distributed as unit exponential variables (Sarker, 2002) with density function

$$f\left(z_j\right)=\exp\left(-z_j\right)\text{ for }z_j\le d_j\left(j=1,2\right). \tag{6.87}$$

Consider the special case, when there are no covariates and censoring points d_j for each component of t_j are equal $\left(d_1=d_2=d\right)$. Thus, from (6.85) the null mean of $T^*_{(2),c}$ is

$$\begin{aligned}E\left(T^*_{(2),c}\right)&=n\int_0^{c_2}\int_0^{c_1}\frac{1}{s}f\left(t_1,t_2\right)dt_1dt_2\\&=n\int_0^{d}\int_0^{d}\frac{e^{-\left(z_1+z_2\right)}}{z_1+z_2}\,dz_1dz_2\\&=n\left(1+e^{-2d}-2e^{-d}+2d\Gamma\left(0,d\right)-2d\Gamma\left(0,2d\right)\right)\\&=n\mu^*_{0,c}\,, \end{aligned}\tag{6.88}$$

where $\Gamma\left(n,d\right)=\int_d^{\infty}e^{-z}z^{n-1}dz$ is the incomplete gamma function. For calculating the variance of $T^*_{(2),c}$ in (6.86), $E\left(g^2\left(t_{i1},t_{i2}\right)\right)$ is obtained as follows

$$E\left(g^2\left(t_{i1},t_{i2}\right)\right)=\int_0^{c_2}\int_0^{c_1}g^2\left(t_1,t_2\right)f\left(t_1,t_2\right)\,dt_1dt_2+\int_0^{c_1}\int_{c_2}^{\infty}g^2\left(t_1,c_2\right)f\left(t_1,t_2\right)\,dt_2dt_1$$

$$+\int_0^{c_2}\int_{c_1}^{\infty} g^2\left(c_1,t_2\right) f\left(t_1,t_2\right)\, dt_1dt_2 + g^2\left(c_1,c_2\right)S\left(t_1,t_2\right). \tag{6.89}$$

As given in Sarker (2002), each part in (6.89) may be obtained from the following:

$$\begin{aligned}
EG_1 &= \int_0^{c_2}\int_0^{c_1} g^2\left(t_1,t_2\right) f\left(t_1,t_2\right) dt_1dt_2 \\
&= \int_0^{d}\int_0^{d} \left\{2+2\ln\left(z_1+z_2\right)-\left(z_1+z_2\right)\ln\left(z_1+z_2\right)\right\}^2 e^{-\left(z_1+z_2\right)} dz_1dz_2 \\
&= -8de^{-2d}\ln\left(2\right)-8de^{-2d}\ln\left(d\right)-8d\Gamma\left(0,2d\right)+8d\ln\left(d\right)e^{-d}+8d\Gamma\left(0,d\right) \\
&\quad +\int_0^{d} r\left(4-4r+r^2\right)\ln^2\left(r\right)e^{-r}dr+\int_d^{2d}\left(2d-r\right)\left(4-4r+r^2\right)\ln^2\left(r\right)e^{-r}dr,
\end{aligned}$$

$$\begin{aligned}
EG_2 &= \int_0^{c_1}\int_{c_2}^{\infty} g^2\left(t_1,c_2\right) f\left(t_1,t_2\right) dt_2dt_1 \\
&= \int_0^{d}\int_d^{\infty} \left\{1+\ln\left(z_1+d\right)-\left(z_1+d\right)\ln\left(z_1+d\right)\right\}^2 e^{-\left(z_1+z_2\right)} dz_2dz_1 \\
&= \int_0^{d} \left\{1+\ln\left(z_1+d\right)-\left(z_1+d\right)\ln\left(z_1+d\right)\right\}^2 e^{-\left(z_1+d\right)} dz_1 \\
&= e^{-2d}+4de^{-2d}\ln\left(2\right)+4de^{-2d}\ln\left(d\right)-e^{-d}-2d\ln\left(d\right)e^{-d} \\
&\quad +\int_d^{2d}\left(1-2r+r^2\right)\ln^2\left(r\right)e^{-r}dr,
\end{aligned}$$

$$\begin{aligned}
EG_3 &= \int_0^{c_2}\int_{c_1}^{\infty} g^2\left(c_1,t_2\right) f\left(t_1,t_2\right) dt_1dt_2 \\
&= \int_0^{d}\int_d^{\infty} \left\{1+\ln\left(d+z_2\right)-\left(d+z_2\right)\ln\left(d+z_2\right)\right\}^2 e^{-\left(z_1+z_2\right)} dz_1dz_2 \\
&= \int_0^{d} \left\{1+\ln\left(d+z_2\right)-\left(d+z_2\right)\ln\left(d+z_2\right)\right\}^2 e^{-\left(d+z_2\right)} dz_2 \\
&= e^{-2d}+4de^{-2d}\ln\left(2\right)+4de^{-2d}\ln\left(d\right)-e^{-d}-2d\ln\left(d\right)e^{-d} \\
&\quad +\int_d^{2d}\left(1-2r+r^2\right)\ln^2\left(r\right)\, e^{-r}dr,
\end{aligned}$$

and

$$EG_4 = g^2\left(c_1,c_2\right)S\left(c_1,c_2\right) = 4d^2\ln^2\left(2d\right)e^{-2d}.$$

Thus, the null variance of $T^*_{(2),c}$ is obtained as follows:

$$\begin{aligned}Var\left(T^{*}_{(2),c}\right) &= n\left[E\left(g^{2}\left(t_{1},t_{2}\right)\right)-E^{2}\left(g\left(t_{1},t_{2}\right)\right)\right] \\ &= n\left[\left(EG_{1}+EG_{2}+EG_{3}+EG_{4}\right)-E^{2}\left(g\left(t_{1},t_{2}\right)\right)\right] \\ &= n\Big[4d\ln\left(d\right)e^{-d}-1+4d^{2}\ln^{2}\left(2d\right)e^{-2d}+8d^{2}\Gamma\left(0,d\right)\Gamma\left(0,2d\right)+2e^{-d}-4e^{-2d} \\ &\quad +4d\Gamma\left(0,d\right)-4d\Gamma\left(0,2d\right)-4d^{2}\Gamma^{2}\left(0,d\right)-4d^{2}\Gamma^{2}\left(0,2d\right)+8de^{-d}\Gamma\left(0,d\right) \\ &\quad -8de^{-d}\Gamma\left(0,2d\right)-4de^{-2d}\Gamma\left(0,d\right)+4de^{-2d}\Gamma\left(0,2d\right)+4e^{-3d}-e^{-4d} \\ &\quad +\int_{d}^{2d}\left\{2d\left(4-4r+r^{2}\right)+\left(2-8r+6r^{2}-r^{3}\right)\right\}\ln^{2}\left(r\right)e^{-r}dr \\ &\quad +\int_{0}^{d}r\left(4-4r+r^{2}\right)\ln^{2}\left(r\right)e^{-r}dr\Big] \\ &= n\sigma^{*2}_{0,c}, \end{aligned} \qquad (6.90)$$

through numerical integration by *Mathematica* software.

6.5.2 Censored Case with Nuisance Parameters

The score statistics for censored cases with nuisance parameters are obtained by replacing the a_j and b_j in (6.77), (6.78) and (6.79) by their null MLEs, $\hat{a}_j$ and $\hat{b}_j$. The score statistics are given as follows

$$\hat{T}_{(2),c}=\sum_{i=1}^{n}\left\{I_i\left(1+\ln\hat{s}_i\right)-\hat{s}_i\ln\hat{s}_i-\frac{I_i\left(I_i-1\right)}{2\hat{s}_i}\right\}, \qquad (6.91)$$

$$\hat{T}^{*}_{(2),c}=\sum_{i=1}^{n}\left\{I_i\left(1+\ln\hat{s}_i\right)-\hat{s}_i\ln\hat{s}_i\right\}, \qquad (6.92)$$

and

$$\hat{T}^{**}_{(2),c}=\sum_{i=1}^{n}\left\{I_i\ln\hat{s}_i\right\}, \qquad (6.93)$$

where $\hat{s}_i=\hat{a}_1\left(e^{\hat{b}_1 t_{i1}}-1\right)/\hat{b}_1+\hat{a}_2\left(e^{\hat{b}_2 t_{i2}}-1\right)/\hat{b}_2$, $\hat{a}_j$ and $\hat{b}_j$ are MLEs of a_j and b_j, respectively under H_0 and I_i is defined in (6.80). For censored observations, t_{i1} and t_{i2} are replaced by censoring times c_{i1} and c_{i2}, respectively.

The null mean of $\hat{T}^{*}_{(2),c}$ is the same as asymptotic null mean of $T^{*}_{(2),c}$ but the null variance is different. Since finding the null variance of $\hat{T}^{*}_{(2),c}$ using the Pierce (1982) theorem might be a daunting task, a simulation study is conducted instead. For $n = 500$, 10000 values of $\hat{T}^{*}_{(2),c}$ are generated for each censoring point, where the censoring points lie between 0.2 and 20. This simulation procedure is repeated ten times in order to produce stable variances estimates. The averages of the estimated variances are regressed on the corresponding censoring points, d. As a result, it is found that the variance of $\hat{T}^{*}_{(2),c}$ may be approximated by the following quadratic equation:

$$\bar{\sigma}^{*2}_{0,c} \approx \left[0.745569 - 0.370864e^{-d} - 0.285308e^{-2d}\right]. \tag{6.94}$$

Practically, only the fixed censoring time c_j $\left(j = 1, 2\right)$ is known, yet not the value of a_j and b_j. Even the censoring time for both components are equal, where $c_1 = c_2 = c,$ the censoring point $\hat{d}_j = \hat{a}_j\left(e^{\hat{b}_j t_j} - 1\right)/\hat{b}_j$ might be different since it depends on the estimation of the nuisance parameters, $\hat{a}_j$ and $\hat{b}_j$. Therefore, it is worthwhile to note that, the equation of variance in (6.94) has very limited application.

6.6 The Critical Region

All the score tests considered in this chapter are one-tailed test. Negative value of the difference between the non-null mean and null mean gives a lower tailed test and vice versa. The non-null mean of $T_{(2)}$ in (6.47) is $-\infty$, thus obviously the test is a lower tailed test with the critical region of an asymptotically α-level test is $\left\{S_{(2)} < -z_{\alpha}\right\}$, where $S_{(2)} = T_{(2)} / \sqrt{\left(1/2\right) n \ln n}$. Also, Sarker (2002) verified the modified score test and the $\ln s$ based test are lower tailed tests. The critical region of an asymptotically

$\alpha-$level test for $T^{*}_{(2)}$ is $\left\{S^{*}_{(2)} < -z_{\alpha}\right\}$ where $S^{*}_{(2)} = \dfrac{T^{*}_{(2)} - n\mu^{*}_{0}}{\sqrt{n\sigma^{*2}_{0}}}$, while $\left\{S^{**}_{(2)} < -z_{\alpha}\right\}$ for $T^{**}_{(2)}$, where $S^{**}_{(2)} = \dfrac{T^{**}_{(2)} - n\mu^{**}_{0}}{\sqrt{n\sigma^{**2}_{0}}}$.

6.7 Evaluation of the Convergence Rates

Simulation procedures are carried out to investigate the rate of convergence to the normal limit for the Zhu's score test, modified score test and $\ln s$ based test for the bivariate positive stable Gompertz model. All simulations are done using the statistical computing software R version 3.0.3.

6.7.1 Convergence Rate for the Uncensored Case

6.7.1.1 Without Nuisance Parameters Case

A simulation study is conducted to examine the convergence rate for all score based tests $T_{(2)}$, $T^{*}_{(2)}$ and $T^{**}_{(2)}$ for no censored case and parameters a_j and b_j are known. The steps of the simulation are described as follows:

(i) Random variables t_{i1} and t_{i2} $\left(i = 1, 2, \ldots, n\right)$ from two independent Gompertz distribution with parameters $a_j = 1$ and $b_j = 1$ $\left(j = 1, 2\right)$ are generated by calling the `rgompertz` function from the `flexsurv` package (Jackson, 2014) in the R software.

(ii) The random variable s_i is computed based on the relation $s_i = \left(e^{t_{i1}} - 1\right) + \left(e^{t_{i2}} - 1\right)$ after substituting for $a_j = 1$ and $b_j = 1$.

(iii) Then, $T_{(2)}$, $T^{*}_{(2)}$ and $T^{**}_{(2)}$ are computed using the formulae given in (6.15), (6.16) and (6.17).

(iv) $S_{(2)}$ is obtained from the formula given in (6.35), whilst $S_{(2)}^{*}$ and $S_{(2)}^{**}$ by normalising the score statistics $T_{(2)}^{*}$ and $T_{(2)}^{**}$ as in (6.42) and (6.45), respectively.

(v) The mean quantiles with its respective standard deviation are determined at the point $\alpha = 0.01,\ 0.025,\ 0.05,\ 0.10$.

Sample size *n* is ranged from 50 to 10000. The simulations are repeated 10000 times for each *n*. Besides, the simulations are done ten times for each *n* to estimate the stability of the simulated results. The R code for this simulation procedure is provided in Appendix D. Table 6.1, Table 6.2 and Table 6.3 present the mean quantiles and corresponding standard deviation of $S_{(2)}$, $S_{(2)}^{*}$ and $S_{(2)}^{**}$, respectively. For comparison, the quantiles of the standard normal distribution are listed at the bottom line of these tables. The results show that the rate of convergence of $S_{(2)}^{**}$ to the normal limit is faster than $S_{(2)}^{*}$. On the other hand, the rate of convergence for $S_{(2)}$ is the slowest and does not reach the normal limit even though the sample size reaches 10000.

Table 6.1: Standardised critical values of Zhu's score test $S_{(2)}$ (Uncensored case without nuisance parameters)

n	$\sqrt{\frac{1}{2}n\log n}$	Quantiles (SD)			
		0.10	**0.05**	**0.025**	**0.01**
20	5.47	-2.698(0.03)	-3.952(0.09)	-5.272(0.19)	-7.504(0.44)
50	9.89	-2.435(0.04)	-3.487(0.08)	-4.629(0.12)	-6.396(0.25)
100	15.17	-2.323(0.06)	-3.307(0.08)	-4.351(0.11)	-6.110(0.23)
500	39.42	-2.097(0.03)	-2.929(0.03)	-3.761(0.06)	-5.124(0.22)
1000	58.77	-2.022(0.03)	-2.796(0.04)	-3.591(0.07)	-4.870(0.26)
5000	145.92	-1.924(0.03)	-2.617(0.03)	-3.336(0.07)	-4.498(0.19)
10000	214.57	-1.869(0.03)	-2.548(0.03)	-3.222(0.06)	-4.360(0.12)
∞		-1.28	-1.64	-1.96	-2.33

Table 6.2: Normalised critical values of the modified score test $S^{*}_{(2)}$ (Uncensored case without nuisance parameters)

n	$\sqrt{n}\sigma^{*}_{0}$	Quantiles (SD)			
		0.10	**0.05**	**0.025**	**0.01**
20	8.36	-1.326(0.02)	-1.836(0.03)	-2.314(0.03)	-2.924(0.07)
50	13.22	-1.323(0.02)	-1.786(0.02)	-2.205(0.03)	-2.701(0.06)
100	18.69	-1.321(0.02)	-1.751(0.03)	-2.135(0.04)	-2.604(0.04)
500	41.79	-1.299(0.02)	-1.688(0.03)	-2.038(0.04)	-2.460(0.07)
1000	59.10	-1.297(0.01)	-1.680(0.02)	-2.020(0.01)	-2.416(0.03)
5000	132.15	-1.281(0.01)	-1.658(0.02)	-1.983(0.03)	-2.375(0.04)
10000	186.89	-1.272(0.03)	-1.636(0.03)	-1.965(0.04)	-2.344(0.04)
∞		-1.28	-1.64	-1.96	-2.33

Table 6.3: Normalised critical values of the $\ln s$ based test $S^{**}_{(2)}$ (Uncensored case without nuisance parameters)

n	$\sqrt{n}\sigma^{**}_{0}$	Quantiles (SD)			
		0.10	**0.05**	**0.025**	**0.01**
20	3.59	-1.296(0.02)	-1.695(0.02)	-2.038(0.03)	-2.440(0.05)
50	5.68	-1.287(0.01)	-1.664(0.01)	-1.995(0.02)	-2.378(0.05)
100	8.03	-1.307(0.02)	-1.686(0.03)	-2.018(0.02)	-2.407(0.05)
500	17.96	-1.280(0.02)	-1.655(0.02)	-1.974(0.03)	-2.356(0.05)
1000	25.40	-1.285(0.02)	-1.658(0.02)	-1.986(0.03)	-2.357(0.03)
5000	56.79	-1.289(0.01)	-1.657(0.02)	-1.971(0.02)	-2.344(0.03)
10000	80.31	-1.283(0.02)	-1.643(0.02)	-1.964(0.02)	-2.331(0.03)
∞		-1.28	-1.64	-1.96	-2.33

6.7.1.2 With Nuisance Parameters Case

An additional step of maximum likelihood estimation that is for estimating the MLEs of a_j and b_j under the null hypothesis is included in the simulation procedure for the case in which the Gompertz parameters, a_j and b_j, are unknown. Complete steps for this simulation study are described as follows:

(i) Random variables t_{i1} and t_{i2} $(i=1,2,\ldots,n)$ from two independent Gompertz distribution with parameters $a_j=1$ and $b_j=1$ $(j=1,2)$ are generated by calling the `rgompertz` function.

(ii) The MLEs of a_j and b_j, $\hat{a}_j$ and $\hat{b}_j$, respectively, are estimated from sample t_{ij} using the Newton-Raphson iteration method by calling the `maxLik` function（Henningsen & Toomet, 2011) from the software `R`.

(iii) The random variable $\hat{s}_i$ is computed by

$$\hat{s}_i=\left(\frac{\hat{a}_1\left(e^{\hat{b}_1 t_{i1}}-1\right)}{\hat{b}_1}+\frac{\hat{a}_2\left(e^{\hat{b}_2 t_{i2}}-1\right)}{\hat{b}_2}\right). \tag{6.95}$$

(iv) Score statistics $\hat{T}_{(2)}$, $\hat{T}^{*}_{(2)}$ and $\hat{T}^{**}_{(2)}$ are computed using (6.50) to (6.52).

(v) $\hat{S}_{(2)}$ is obtained from (6.35) and $\hat{S}^{*}_{(2)}$ and $\hat{S}^{**}_{(2)}$ by normalising the test statistics $\hat{T}^{*}_{(2)}$ and $\hat{T}^{**}_{(2)}$ using the equation (6.60) and (6.67).

(vi) The mean quantiles with its respective standard deviation are determined at the point $\alpha=0.01,\ 0.025,\ 0.05,\ 0.10$.

Sample size is ranged from 20 to 10000, and the simulations are repeated 10000 times for each *n*. Simulations are carried out ten times for each *n* to estimate the stability of the simulated results.

The mean quantiles for normalised critical values of $\hat{S}_{(2)}$, $\hat{S}^{*}_{(2)}$ and $\hat{S}^{**}_{(2)}$ are tabulated in Table 6.4, Table 6.5, and Table 6.6, respectively, together with the standard deviations. Standard normal quantiles are presented at the bottom of the table. Table 6.4 shows that $\hat{S}_{(2)}$ does not converge to the normal limit even for sample size 10000. Meanwhile, the rate of convergence of $\hat{S}^{**}_{(2)}$ is slightly slower than $\hat{S}^{*}_{(2)}$.

Table 6.4: Standardised critical values of Zhu's score test $\hat{S}_{(2)}$ (Uncensored case with nuisance parameters)

n	$\sqrt{\frac{1}{2} n \log n}$	Quantiles (SD)			
		0.10	**0.05**	**0.025**	**0.01**
20	5.47	-2.112(0.05)	-3.183(0.09)	-4.509(0.15)	-6.849(0.26)
50	9.89	-1.894(0.03)	-2.812(0.03)	-3.922(0.09)	-5.924(0.22)
100	15.17	-1.804(0.03)	-2.653(0.07)	-3.629(0.11)	-5.410(0.25)
500	39.42	-1.685(0.02)	-2.422(0.04)	-3.270(0.08)	-4.670(0.12)
1000	58.77	-1.626(0.03)	-2.322(0.05)	-3.131(0.06)	-4.464(0.19)
5000	145.92	-1.592(0.02)	-2.238(0.04)	-2.929(0.06)	-4.123(0.15)
10000	214.57	-1.557(0.02)	-2.164(0.04)	-2.831(0.06)	-3.965(0.12)
∞		-1.28	-1.64	-1.96	-2.33

Table 6.5: Normalised critical values of the modified score test $\hat{S}^{*}_{(2)}$ (Uncensored case with nuisance parameters)

n	$\sqrt{n}\bar{\sigma}^{*}_{0}$	Quantiles (SD)			
		0.10	**0.05**	**0.025**	**0.01**
20	3.86	-1.445(0.03)	-1.866(0.03)	-2.237(0.03)	-2.672(0.05)
50	6.10	-1.365(0.01)	-1.765(0.02)	-2.115(0.03)	-2.516(0.04)
100	8.62	-1.339(0.02)	-1.727(0.03)	-2.066(0.04)	-2.465(0.06)
500	19.29	-1.307(0.01)	-1.683(0.02)	-1.999(0.02)	-2.390(0.03)
1000	27.27	-1.301(0.02)	-1.670(0.02)	-1.994(0.02)	-2.371(0.03)
5000	60.99	-1.285(0.01)	-1.657(0.02)	-1.980(0.02)	-2.337(0.03)
10000	86.25	-1.281(0.02)	-1.647(0.02)	-1.958(0.04)	-2.331(0.06)
∞		-1.28	-1.64	-1.96	-2.33

Table 6.6: Normalised critical values of the $\ln s$ based test of $\hat{S}^{**}_{(2)}$
(Uncensored case with nuisance parameters)

n	$\sqrt{n}\bar{\sigma}_0^{**}$	Quantiles (SD)			
		0.10	**0.05**	**0.025**	**0.01**
20	1.21	-1.477(0.03)	-1.953(0.04)	-2.375(0.04)	-2.897(0.06)
50	1.92	-1.397(0.02)	-1.822(0.02)	-2.215(0.03)	-2.671(0.05)
100	2.72	-1.357(0.01)	-1.762(0.03)	-2.128(0.04)	-2.574(0.05)
500	6.07	-1.317(0.01)	-1.706(0.02)	-2.048(0.02)	-2.431(0.03)
1000	8.59	-1.303(0.02)	-1.689(0.01)	-2.017(0.02)	-2.402(0.04)
5000	19.20	-1.296(0.01)	-1.666(0.02)	-1.990(0.02)	-2.359(0.03)
10000	27.15	-1.286(0.02)	-1.656(0.02)	-1.971(0.03)	-2.329(0.05)
∞		-1.28	-1.64	-1.96	-2.33

6.7.2 Convergence Rate for the Censored Case

Convergence rate for cases with censored observation is also investigated through simulations. It is worthwhile to note that, for censored case, the results of the estimated critical values have limited practical application. Thus, this study only focuses on the Zhu's score test and modified score test.

6.7.2.1 Without Nuisance Parameters Case

It is assumed that T_{i1} and T_{i2} are censored at fixed times c_1 and c_2, respectively. For censored case and the parameters a_j and b_j are known, the population of censored observations ε has the relationship with the marginal cumulative hazard $H_1(c_1)$ and $H_2(c_2)$ as follows

$$\begin{aligned}\varepsilon &= P\left(T_1 > c_1\right) + P\left(T_2 > c_2\right) - P\left(T_1 > c_1, T_2 > c_2\right) \\ &= \exp\left(-\frac{a_1\left(e^{b_1c_1}-1\right)}{b_1}\right) + \exp\left(-\frac{a_2\left(e^{b_2c_2}-1\right)}{b_2}\right) - \exp\left(-\frac{a_1\left(e^{b_1c_1}-1\right)}{b_1} - \frac{a_2\left(e^{b_2c_2}-1\right)}{b_2}\right) \\ &= 1-\left(1-e^{-d_1}\right)\left(1-e^{-d_2}\right), \qquad (6.96)\end{aligned}$$

where $H_j(c_j) = d_j$ and $e^{-H_j} = e^{-d_j}$ is the population proportion of censored observations for the component j. The full steps to investigate the rate of convergence of these score tests are given as the following

(i) Random variables t_{i1} and t_{i2} $\left(i = 1, 2, \ldots, n\right)$ from two independent Gompertz distribution with parameters $a_j = 1$ and $b_j = 1$ $\left(j = 1, 2\right)$ are generated by calling the `rgompertz` function.

(ii) The censoring time $c_j = 1.035$ is chosen so that about 30% of pairs are censored at least in one component of T_j according to (6.96). Any value of T_{ij} that is greater than $c_j = 1.035$ is replaced by 1.035.

(iii) The random variable s_i is computed.

(iv) The score statistics $T_{(2),c}$ and $T^*_{(2),c}$ are computed from equation (6.77) and (6.78), respectively.

(v) $S_{(2),c}$ is obtained from (6.82) and $S^*_{(2),c}$ is obtained by normalising the test statistics $T^*_{(2),c}$ with the mean and variance given in (6.88) and (6.90), respectively.

Table 6.7 presents the mean quantile for the Zhu's score test and its corresponding standard deviation. Meanwhile, Table 6.8 presents the mean quantile for the modified score test and its corresponding standard deviation. As expected $S^*_{(2),c}$ converge faster to the standard normal limit than $S_{(2),c}$.

Table 6.7: Normalised critical values of the Zhu's score test $S_{(2),c}$ (30% censoring and without nuisance parameters case)

n	$\sqrt{\frac{1}{2} n \log n}$	Quantiles (SD)			
		0.10	0.05	0.025	0.01
20	5.47	-2.320(0.03)	-3.465(0.05)	-4.743(0.09)	-7.002(0.14)
50	9.89	-2.151(0.05)	-3.137(0.09)	-4.298(0.10)	-6.262(0.24)
100	15.17	-2.040(0.03)	-2.953(0.04)	-3.960(0.08)	-5.727(0.25)
500	39.42	-1.897(0.03)	-2.661(0.06)	-3.475(0.07)	-4.886(0.15)
1000	58.77	-1.876(0.04)	-2.630(0.04)	-3.429(0.06)	-4.768(0.23)
5000	145.92	-1.769(0.02)	-2.414(0.03)	-3.110(0.06)	-4.237(0.14)
10000	214.57	-1.739(0.03)	-2.387(0.05)	-3.052(0.09)	-4.084(0.23)
∞		-1.28	-1.64	-1.96	-2.33

Table 6.8: Normalised critical values of the modified score test $S^*_{(2),c}$ (30% censoring and without nuisance parameters case)

n	$n\mu^*_{0,c}$	$\sqrt{n}\sigma^*_{0,c}$	Quantiles (SD)			
			0.10	0.05	0.025	0.01
20	17.95	6.81	-1.317(0.02)	-1.744(0.02)	-2.129(0.02)	-2.595(0.05)
50	44.87	10.76	-1.302(0.02)	-1.705(0.03)	-2.068(0.03)	-2.472(0.05)
100	89.74	15.22	-1.292(0.02)	-1.677(0.01)	-2.020(0.03)	-2.440(0.05)
500	448.71	34.04	-1.291(0.01)	-1.671(0.02)	-1.999(0.03)	-2.369(0.04)
1000	897.42	48.14	-1.288(0.02)	-1.654(0.02)	-1.986(0.03)	-2.371(0.04)
5000	4487.08	107.63	-1.282(0.02)	-1.651(0.02)	-1.970(0.02)	-2.350(0.04)
10000	8974.15	152.22	-1.283(0.02)	-1.649(0.02)	-1.959(0.03)	-2.331(0.02)
∞			-1.28	-1.64	-1.96	-2.33

6.7.2.2 With Nuisance Parameters Case

In this case, the similar simulation procedure for the censored case with nuisance parameters is performed. In addition, the maximum likelihood estimation method is included in order to obtain the MLEs of a and b using the Newton-Raphson iterative method by calling the `maxLik` function (Henningsen & Toomet, 2011) in the software `R`. The mean and variance are calculated based on the equation in (6.88) and (6.94), respectively. Meanwhile, the test statistics $\hat{T}_{(2),c}$ and $\hat{T}^*_{(2),c}$ are calculated from the formula in equation (6.91) and (6.92), respectively. The mean quantile of $\hat{S}_{(2),c}$ and $\hat{S}^*_{(2),c}$ and its corresponding standard deviation are tabulated in Table 6.9 and Table 6.10, respectively. The results indicate that the convergence rate of $\hat{S}_{(2),c}$ is slower than $\hat{S}^*_{(2),c}$.

Table 6.9: Normalised critical values of the Zhu's score test $\hat{S}_{(2),c}$ (30% censoring and with nuisance parameters case)

n	$\sqrt{\frac{1}{2} n \log n}$	Quantiles (SD)			
		0.10	**0.05**	**0.025**	**0.01**
20	5.47	-1.731(0.04)	-2.691(0.08)	-3.879(0.12)	-6.036(0.32)
50	9.89	-1.706(0.04)	-2.583(0.05)	-3.622(0.09)	-5.462(0.19)
100	15.17	-1.666(0.04)	-2.469(0.06)	-3.435(0.09)	-5.158(0.18)
500	39.42	-1.594(0.05)	-2.311(0.07)	-3.136(0.09)	-4.562(0.22)
1000	58.77	-1.576(0.03)	-2.248(0.03)	-3.020(0.08)	-4.422(0.22)
∞		-1.28	-1.64	-1.96	-2.33

Table 6.10: Normalised critical values of the modified score test $\hat{S}^*_{(2),c}$ (30% censoring and with nuisance parameters case)

n	$\sqrt{n}\bar{\sigma}^*_{0,c}$	Quantiles (SD)			
		0.10	**0.05**	**0.025**	**0.01**
20	3.75	-1.299(0.02)	-1.672(0.02)	-2.006(0.03)	-2.423(0.05)
50	5.87	-1.298(0.02)	-1.678(0.03)	-2.007(0.03)	-2.376(0.04)
100	8.25	-1.289(0.01)	-1.662(0.02)	-1.987(0.02)	-2.347(0.04)
500	18.41	-1.278(0.02)	-1.642(0.02)	-1.961(0.04)	-2.320(0.04)
1000	26.06	-1.272(0.02)	-1.635(0.03)	-1.948(0.03)	-2.298(0.06)
∞		-1.28	-1.64	-1.96	-2.33

6.8 Critical Values Estimation

In this section, simulation studies are conducted to produce reliable expressions for estimating the critical values that are applicable even to cases with small sample size. In practice, nuisance parameters are usually unknown. Therefore, in this section, only critical values for the case with nuisance parameters are considered.

The standardised critical values for the Zhu's score test, modified score test and $\ln s$ based test for uncensored case with nuisance parameters are estimated for different values of sample size, ranged between $20 \le n \le 10000$ for quantiles $\alpha = 0.01, 0.025, 0.05$ and 0.10. The steps for generating the standardised critical values follow the method described in Section 6.6. The estimated standardised critical values are then smoothed by calling the `nls` function in software `R`. The estimated regression coefficient of the fitted regression models for the Zhu's score test is tabulated in Table 6.11, for the modified score test in Table 6.12, and for the $\ln s$ based test in Table 6.13. Meanwhile, Figure 6.1, Figure 6.2 and Figure 6.3 indicate that the regression models fit the estimated standardised critical values very well.

Sarker (2002) pointed out that the estimation of critical values for censored cases with unknown parameters is extremely cumbersome. In this case, the test statistics depend on the Gompertz parameters and censoring time c through the censoring point $d_j = a_j(e^{b_j c_1} - 1)/b_j$ $\left(j = 1, 2\right)$. In practice, only on the censoring time, c, are known. Hence, d_j must be estimated as the term depends on the null maximum likelihood estimates (MLEs) of the Gompertz parameters. Thus, this study has decided not to pursue the investigation on the estimation of critical values for this case.

Table 6.11: Estimated expressions for calculating the standardised critical values of $\hat{S}_{(2)}$ for the case with nuisance parameters

Statistic	α	Expression	RSE[1]
$\hat{S}_{(2)}$	0.10	$-1.4988-1.3303/\sqrt[3]{n}-0.2413/\sqrt{n}$	0.0135
	0.05	$-2.0436-3.0252/\sqrt[3]{n}+0.2680/\sqrt{n}$	0.0198
	0.025	$-2.6068-5.4266/\sqrt[3]{n}+1.0764/\sqrt{n}$	0.0349
	0.01	$-3.5103-10.5617/\sqrt[3]{n}+3.2903/\sqrt{n}$	0.0799

[1] Estimated residual standard error

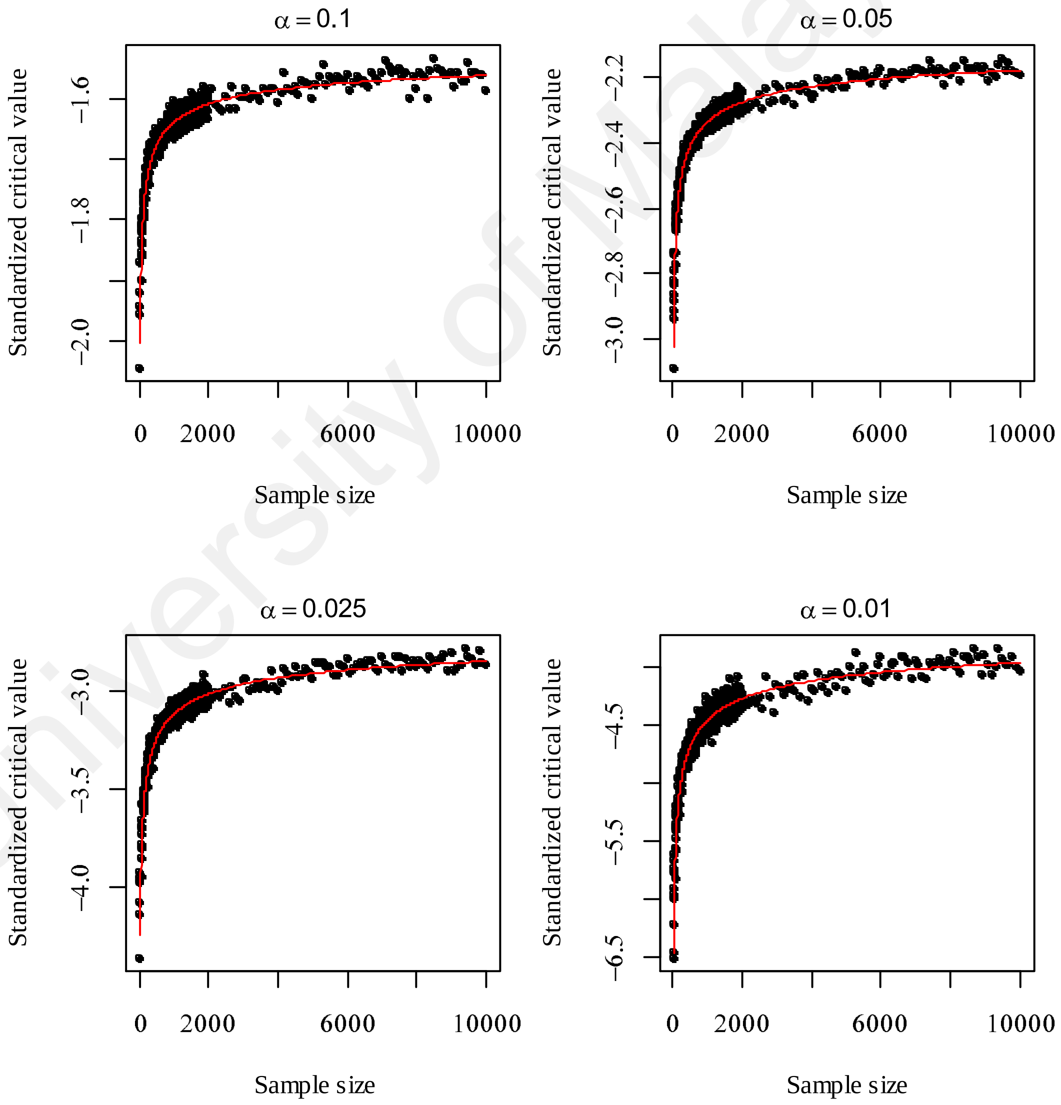


Figure 6.1: Estimated standardised critical values of the bivariate Zhu's score test statistics $\hat{T}_{(2)}$ with superimposed fit of the equation $C_\alpha = \alpha_0 + \alpha_1 / \sqrt[3]{n} + \alpha_2 / \sqrt{n}$

Table 6.12: Estimated expressions for calculating the standardised critical values of $\hat{S}^*_{(2)}$ for the case with nuisance parameters

Statistic	α	Expression	RSE[1]
$\hat{S}^*_{(2)}$	0.10	$-1.2818-0.4935/\sqrt{n}-0.7675/n$	0.0080
	0.05	$-1.6441-0.7706/\sqrt{n}-0.5154\ /n$	0.0098
	0.025	$-1.9588-1.0476/\sqrt{n}-0.3030/n$	0.0119
	0.01	$-2.3214-1.4912/\sqrt{n}+0.2777/n$	0.0166

[1] Estimated residual standard error

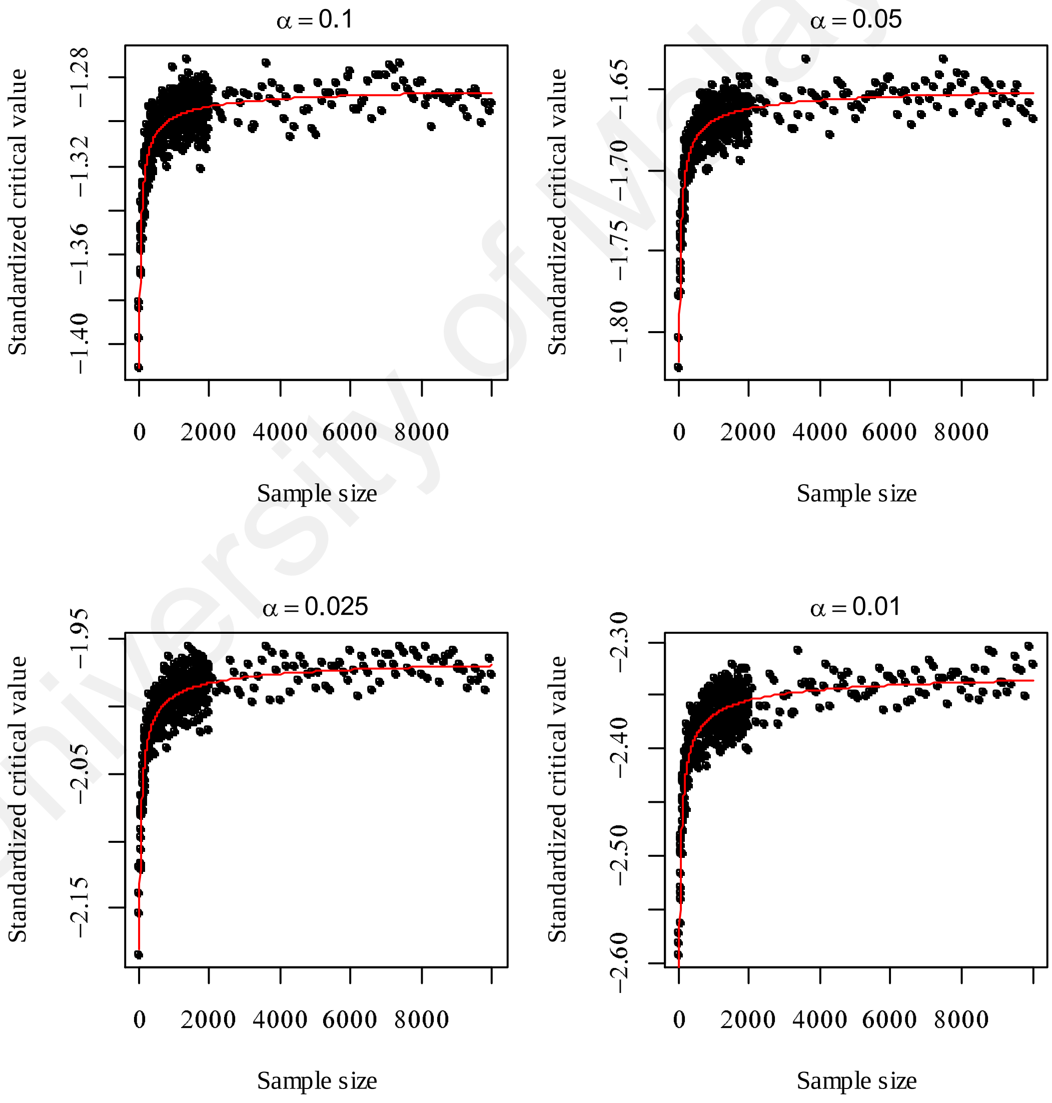


Figure 6.2: Estimated standardised critical values of the bivariate modified score test statistics $\hat{T}^*_{(2)}$ with superimposed fit of the equation $C_\alpha = \alpha_0 + \alpha_1/\sqrt{n} + \alpha_2/n$

Table 6.13: Estimated expressions for calculating the standardised critical values of $\hat{S}^{**}_{(2)}$ for the case with nuisance parameters

Statistic	α	Expression	RSE[1]
$\hat{S}^{**}_{(2)}$	0.10	$-1.2815-0.8654/\sqrt{n}-0.3965/n$	0.0078
	0.05	$-1.6447-1.2993/\sqrt{n}+0.3878\ /n$	0.0100
	0.025	$-1.9572-1.8380/\sqrt{n}-0.6334/n$	0.0129
	0.01	$-2.3217-2.5274/\sqrt{n}+0.4706/n$	0.0164

[1] Estimated residual standard error

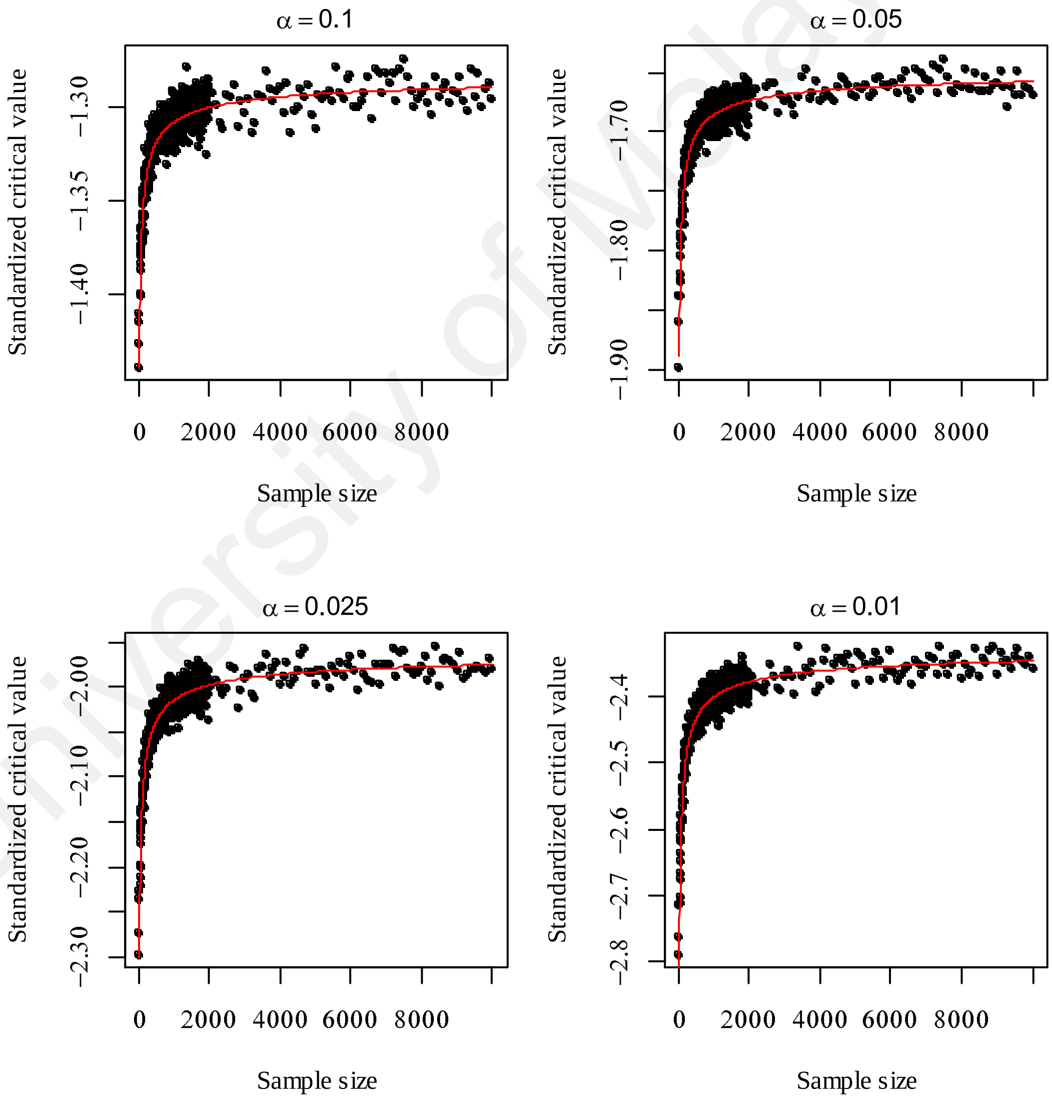


Figure 6.3: Estimated standardised critical values of the bivariate lns based test statistics $\hat{T}^{**}_{(2)}$ with superimposed fit of the equation $C_\alpha = \alpha_0 + \alpha_1/\sqrt{n} + \alpha_2/n$

6.9 Evaluation of Power

The power of a statistical test is the probability that the test correctly rejects a false null hypothesis (Daniel, 2005). The higher the power, the better the test compared to others. In this section, the power of the Zhu's score tests, modified score test and $\ln s$ based test are evaluated for the bivariate positive stable Gompertz model based on different sample sizes, amount of censoring, and amounts of frailty.

Simulation procedures are performed to estimate the power of the score statistics (a) $T_{(2)}$, $T_{(2)}^{*}$ and $T_{(2)}^{**}$ for uncensored cases without nuisance parameters, (b) $\hat{T}_{(2)}$, $\hat{T}_{(2)}^{*}$ and $\hat{T}_{(2)}^{**}$ for uncensored cases with nuisance parameters, (c) $T_{(2),c}$, $T_{(2),c}^{*}$ and $T_{(2),c}^{**}$ for censored case without nuisance parameters based on the pre-determined critical values obtained in Section 6.7.

6.9.1 Uncensored Case without Nuisance Parameters

For uncensored case without nuisance parameters, the steps of the simulation procedure are given as follows:

(i) Two random variables x_{i1} and x_{i2} $\left(i=1,2,\ldots,n\right)$ are generated from two independent Gompertz distribution with parameters $a_j=1$ and $b_j=1$ $\left(j=1,2\right)$ by calling the `rgompertz` function.

(ii) A positive stable random variable w_i is generated by calling the function of `stabledist` (Wuertz *et al.*, 2013) in `R`.

(iii) Then, positive stable Gompertz random variables t_{i1} and t_{i2} are obtained from $t_{ij}=x_{ij}/w_i$.

(iv) The test statistics $T_{(2)}$, $T_{(2)}^{*}$ and $T_{(2)}^{**}$ are computed.

(v) The test statistics the are obtained in step (iv) are compared to pre-determined critical values at 5% from Table 6.1 to Table 6.3.

(vi) The null hypothesis is rejected when the observed values of the test are lower than its corresponding critical values.

(vii) Power of the tests which is equal to the percentage of rejections in a repeated sampling is estimated.

All simulations procedures are performed for each combination of characteristic exponent $0.50 \le v \le 1.00$ and n $(n = 20,\ 50$ and $100)$. Simulations are repeated for 2000 times. Table 6.14 shows that the power estimates of the Zhu's score test and modified score test are comparable, with the Zhu's score test yields slightly higher power estimates for $0.50 \le v \le 0.90$. Meanwhile, the power estimates for the $\ln s$ based test is remarkably low compared to the other two tests.

Table 6.14: Estimated powers (%) of $T_{(2)}$, $T^{*}_{(2)}$ and $T^{**}_{(2)}$ at the 5% level of significance for $PS(v)$ frailty (Uncensored case without nuisance parameters)

n	v	$T_{(2)}$	$T^{*}_{(2)}$	$T^{**}_{(2)}$
20	0.50	100.00	99.99	33.54
	0.60	99.98	99.94	35.67
	0.70	99.70	99.22	30.73
	0.80	96.07	94.10	23.13
	0.90	63.17	62.49	13.67
	0.95	29.05	29.76	9.35
	0.99	7.78	8.04	5.28
	1.00	4.68	4.73	4.61
50	0.50	100.00	100.00	33.04
	0.60	100.00	100.00	38.36
	0.70	100.00	100.00	36.88
	0.80	99.95	99.91	28.32
	0.90	90.24	89.42	16.82
	0.95	48.40	49.58	10.52
	0.99	10.83	10.01	6.01
	1.00	5.26	5.11	5.03
100	0.50	100.00	100.00	30.89
	0.60	100.00	100.00	41.44
	0.70	100.00	100.00	42.50
	0.80	100.00	100.00	32.43
	0.90	99.08	99.21	18.13
	0.95	70.41	72.60	10.94
	0.99	12.39	12.29	5.78
	1.00	4.57	4.68	4.53

6.9.2 Uncensored Case with Nuisance Parameters

For this case, the similar simulation procedure for the case without nuisance parameters is performed, except that the method of maximum likelihood estimation is included by calling the `maxLik` function (Henningsen & Toomet, 2011). The pre-determined critical values in Table 6.4 to Table 6.6 are used. The estimated powers of the tests are tabulated in Table 6.15. The results indicate that the power for detecting the presence of frailty in the positive stable Gompertz for all the three tests is reasonable.

The power of the modified score test is the highest followed by the $\ln s$ based test especially for $0.50 \le v \le 0.90$.

Table 6.15: Estimated powers (%) of $\hat{T}_{(2)}$, $\hat{T}^{*}_{(2)}$ and $\hat{T}^{**}_{(2)}$ at the 5% level of significance for $PS(v)$ frailty (Uncensored case with nuisance parameters)

n	v	$\hat{T}_{(2)}$	$\hat{T}^{*}_{(2)}$	$\hat{T}^{**}_{(2)}$
20	0.50	99.53	99.89	99.87
	0.60	96.37	98.45	97.95
	0.70	83.35	90.22	88.35
	0.80	57.30	65.51	63.53
	0.90	27.62	30.05	29.70
	0.95	15.77	15.74	15.85
	0.99	6.42	6.14	6.15
	1.00	4.37	4.55	4.38
50	0.50	100.00	100.00	100.00
	0.60	100.00	100.00	100.00
	0.70	98.46	99.67	99.51
	0.80	86.42	92.75	91.84
	0.90	48.59	54.10	53.90
	0.95	25.86	25.16	26.42
	0.99	8.13	8.00	8.08
	1.00	5.26	5.56	5.51
100	0.50	100.00	100.00	100.00
	0.60	100.00	100.00	100.00
	0.70	99.99	100.00	100.00
	0.80	98.04	99.53	99.31
	0.90	68.33	75.51	75.57
	0.95	34.30	35.79	36.77
	0.99	9.74	8.63	9.23
	1.00	4.81	4.77	4.98

6.9.3 Censored Case without Nuisance Parameters

For censored case, only the power of the tests for the case without nuisance parameters are investigated. The power of the Zhu's score test and modified score test for three different cases of fixed censoring time are investigated. The censoring time are chosen which (a) $c_1 = c_2 = 1.38$, where 10% pairs are censored in at least one component, (b) $c_1 = 0.80,\ c_2 = \infty$, where 30% pairs are censored only in one component and (c) $c_1 = c_2 = 1.035$, where 30% pairs are censored in at least one component. The results are tabulated in Table 6.16. The results indicate that the power estimates of the Zhu's score test and modified score test are comparable, with slightly higher power for the Zhu's score test. In addition, the power estimates for both tests are the lowest for larger amount of censoring cases that is when 30% pairs are censored at least in one component of the survival times.

Table 6.16: Estimated powers (%) of $T_{(2),c}$ and $T^*_{(2),c}$ at the 5% level of significance

n	***v***	$T_{(2),c}$	$T^*_{(2),c}$	$T_{(2),c}$	$T^*_{(2),c}$	$T_{(2),c}$	$T^*_{(2),c}$
		$c_1 = c_2 = 1.38$		$c_1 = 0.80,\ c_2 = \infty$		$c_1 = c_2 = 1.035$	
20	0.50	100.00	99.97	100.00	99.97	99.92	99.93
	0.60	99.93	99.80	99.93	99.85	99.31	99.21
	0.70	98.63	97.99	98.83	97.87	94.42	93.84
	0.80	88.12	87.21	90.15	88.91	76.81	74.55
	0.90	51.00	50.40	53.37	53.95	40.10	36.55
	0.95	24.93	24.87	25.16	27.22	20.90	18.03
	0.99	8.79	8.65	8.51	10.03	7.95	7.07
	1.00	5.82	5.84	5.44	6.80	4.96	5.03
50	0.50	100.00	100.00	100.00	100.00	100.00	100.00
	0.60	100.00	100.00	100.00	100.00	100.00	100.00
	0.70	100.00	100.00	100.00	100.00	99.97	99.97
	0.80	99.57	99.26	99.71	99.47	97.06	97.13
	0.90	79.38	77.88	81.04	81.70	66.92	64.36
	0.95	41.02	38.20	41.02	44.28	32.76	28.64
	0.99	11.23	8.92	10.60	12.40	8.76	7.59
	1.00	6.09	5.04	5.82	7.37	4.99	5.01
100	0.50	100.00	100.00	100.00	100.00	100.00	100.00
	0.60	100.00	100.00	100.00	100.00	100.00	100.00
	0.70	100.00	100.00	100.00	100.00	100.00	100.00
	0.80	100.00	100.00	100.00	100.00	99.98	99.95
	0.90	96.14	94.92	95.62	96.22	87.58	86.64
	0.95	59.57	54.04	60.01	65.11	47.28	43.71
	0.99	12.83	8.69	12.64	15.62	11.16	9.14
	1.00	5.41	3.98	5.64	8.04	5.08	4.90

6.10 Illustrative Example

6.10.1 Simulated Data

It has been found that the data set of cervical cancer patients considered in the previous chapters contains no frailty component in them. In addition, a real data set that may fit the positive stable Gompertz model is not available. Therefore, for illustration, a simulated data set that follows the positive stable Gompertz model is used. A data set with $n = 55$ is generated which consist of two random variables from two independent Gompertz distribution with parameters $a = 0.0001$ and $b = 0.1$. Meanwhile, a random variable from the positive stable distribution is generated with an exponent characteristic of $v = 0.68$. All three score based tests $\hat{T}_{(2)}$, $\hat{T}^{*}_{(2)}$ and $\hat{T}^{**}_{(2)}$ are considered. The critical values for each score statistics at the level of significance $\alpha = 0.05$ are obtained based on the expressions in Table 6.11 to Table 6.13. The results are tabulated in Table 6.17. All the tests show evidence that there is significant frailty effects in the simulated data set at 5% level of significance.

Table 6.17: The results of simulated data ($n = 55$) using the score based tests for bivariate positive stable frailty

Test statistic	Observed value	Critical values $C_{0.05}$	Significant frailty
$\hat{S}_{(2)}$	-32.38	-2.80	Yes
$\hat{S}^{*}_{(2)}$	-7.16	-1.76	Yes
$\hat{S}^{**}_{(2)}$	-6.93	-1.81	Yes

6.11 Summary

In this study, the Zhu’s score test, modified score test and $\ln s$ based test are derived for the positive stable Gompertz model. The asymptotic properties of these tests have been investigated for four different cases, the uncensored case without nuisance parameters, uncensored case with nuisance parameters, censored case without nuisance parameters, and censored case with nuisance parameters. Also, the new asymptotic variances for with nuisance parameters case are derived.

The performance of these tests are studied based on the rate of convergence and the power of the tests. As anticipated, the rate of convergence of the Zhus’s score test to the normal limit is the slowest amongst other tests. However, the estimated power of the test is considerably well. The results also suggest that the Zhu’s score test is less sensitive to censoring as its power is slightly higher than the modified test in such case.

The convergence rate of the modified score test and $\ln s$ based test is much faster than the Zhu’s score test. Meanwhile, the power estimates of the modified score tests are close to the power of the Zhus’s score test. In contrast, the $\ln s$ based test has lower power especially for the uncensored case without nuisance parameters. Based on the convergence rate and power of the tests, the results reveal that the modified score test performs better under all cases considered in this study.

CHAPTER 7

CONCLUSIONS

7.1 Summary

This study looked at some problems related to survival data analysis. Most survival analysis studies tend to ignore the possible implications of such problems on the study findings. There are three main problems which have been addressed in this study; a nonproportional hazard, missing values and frailty.

In Chapter 3, data of 120 cervical cancer patients treated in HUSM were analysed. This study found that the overall five-year survival of these patients was low. Factors considered in the analyses were stage at diagnosis, ethnicity, histologic type, lymph node involvement, age at diagnosis, distant metastasis and primary treatment received. Of all variables, it was found that survival difference was statistically significant for the stage at diagnosis, primary treatment, and distant metastasis variables.

Also, the relationship between the prognostic factors of cervical cancer and the hazard of dying due to the cancer for these patients was investigated. From the Cox proportional hazards regression analysis, it was found that the stage at diagnosis, histologic type and distant metastasis were significantly influenced the risk of dying for cervical cancer patients studied. However, since the proportional hazards assumption for the distant metastasis variable was violated, the stratified Cox model was considered. Consequently, stage at diagnosis and histologic type remained as significant prognostic factors associated with the hazard of dying of these patients after stratified for the distant metastasis variable. Women, who were diagnosed at an advanced stage (III-IV), were having a 2-fold greater risk of dying than patients with stage I-II. Patients who were diagnosed with histologic type of adenocarcinoma have poorer prognosis compared to those in squamous cell carcinoma group.

In Chapter 4, data of cervical cancer patients were further analysed using the Weibull, log-logistic and lognormal models. Based on the plots of the cumulative hazard function (or a function of it) against the log of survival time, these parametric survival models were suitable for modelling the data set. Stage at diagnosis, histologic type and distant metastasis variables were statistically significant for the Weibull model, whilst only the stage at diagnosis was significant in the log-logistic model. Meanwhile, the stage at diagnosis and age at diagnosis 60 years and older were significant in the lognormal model. Thus, the most important factor that affected the progress of the cancer for these patients was stage at diagnosis since it was found statistically significant at all level of analysis (univariate or multivariate) and all type of parametric models.

Amongst these three models, the Weibull model was the best-fitted model since the AIC value was the smallest. Then, fitness of the Weibull model was checked further. The likelihood ratio test indicated that the scale parameters were different for without distant metastasis and with distant metastasis groups. Thus, there was evidence that this variable did not satisfy the proportional hazards assumption. A stratified Weibull model was proposed, where the distant metastasis became the stratification factor. The final model indicated that shorter survival time were more likely for the patient who was diagnosed at stage III-IV than those in early stages (Stage I-II). Similarly, those who was diagnosed with adenocarcinoma had an earlier time to death compared to squamous carcinoma.

Also, the performances of the stratified Cox model and stratified Weibull model have been compared. Both of these models contained the same explanatory variables.

From the plots of Cox-Snell residuals, it was found that the stratified Cox model fitted the cervical cancer data better than the stratified Weibull model.

In Chapter 5, the performances of four missing data methods have been studied for the parametric survival model. The complete case analysis, EM algorithm by method of weight, hot deck and multiple imputation with MICE-PMM methods were considered for handling missing categorical covariate values in the Weibull AFT model. This study focused on MAR data. It was found that when the percentage of missing values was small, the complete case analysis method was acceptable. However, as the percentage of missing values increases, the complete case analysis yielded the worst estimates compared with other methods. Meanwhile, for smaller sample size, multiple imputation with MICE-PMM performed better, followed by the EM algorithm by method of weight and hot deck. As the sample size increases, the EM algorithm by method of weight gave better parameter estimates. Similarly, when the percentage of missing value increases, the EM algorithm by method of weight outperformed the other methods.

Also, these methods were applied to the data of cervical cancer patients treated in HUSM. Amongst all, the parameter estimates from the EM algorithm were closer to the full model which had no missing values. The results from the EM algorithm method also showed that the significant effect of the variables in the final model was remained when the percentage of missing values were between small to moderate values. Based on the results obtained, it was found that the performance of the EM algorithm by method of weight was the best for handling missing categorical covariate values in the parametric survival model.

Another problem in survival data analysis is the existence of frailty. Zhu (1998) proposed a score test for detecting the presence of positive stable frailty in a Weibull model. Then, Sarker (2002) extended this score test and derived two new score tests namely a modified score test and $\ln s$ based test. In Chapter 6, similar tests were derived and studied for the bivariate positive stable Gompertz model. Four cases have been considered; uncensored case without nuisance parameters, uncensored case with nuisance parameters, censored case without nuisance parameters, and censored case with nuisance parameters. Also, the new asymptotic variances for cases with nuisance parameters were derived from the positive stable Gompertz model.

The results from the simulation studies showed that the rate of convergence of the Zhus's score test to the normal limit was the slowest, yet the estimated power of the test was considerably well. The convergence rate of the modified score test and $\ln s$ based test were faster than the Zhu's score test. Meanwhile, the power estimates of the modified score tests was comparable to the power of the Zhus's score test. In contrast, the power estimation for the $\ln s$ based test was the lowest especially for uncensored case with known parameters. Overall, the modified score test performed better in all cases based on the convergence rate and power of the test. Unfortunately, the real data set that may follow the positive stable Gompertz model was not available. Thus, for illustration, a simulated data set has been used.

7.2 Contributions

This study has contributed to survival data analysis in the following ways:

1. This study has demonstrated the development of parametric survival models for Malaysian real data set using several types of parametric survival models. Also, this study has demonstrated on how to identify the best survival model to represent the data.
2. This study has demonstrated the development of non-proportional hazards model for Malaysian cancer data set using the stratified Cox model and stratified Weibull model.
3. This study has investigated the performance of complete case analysis, EM algorithm by method of weight, hot deck and multiple imputation with MICE-PMM method for handling missing covariate values in a parametric survival models. It has been found that the best method is the EM algorithm by method of weight. This study is imperative since the Cox proportional hazards model with missing values has often been given more attention in many published studies than the parametric model. In addition, this study has illustrated these methods for the data of cervical cancer patients treated in HUSM.
4. This study has derived the Zhu's score test, modified score test and $\ln s$ based test for the positive stable Gompertz model. Also, the new asymptotic variances for the case with nuisance parameters were derived. The performances of these tests have been investigated and this study found that the modified score test perform better than the other two tests.

7.3 Future Work

Further research may be executed based on the analyses that have been done in our study. Several studies that are possible to work on are listed as the following:

(i) Modelling the cervical cancer data using the flexible parametric model analysis

Another type of model that may be applied for analysing the cervical cancer data is a flexible parametric model. Such a model is more flexible than the standard parametric model and suitable for handling non-proportional hazards covariates.

(ii) Missing not at random (MNAR) data

This study only considered parametric model with missing at random (MAR) covariate. Therefore, further investigation on the performance of the missing data methods studied may be done for missing not at random (MNAR) data. Such a case has not widely studied especially for the parametric survival model. In addition, one may also consider cases with a mixture of different missingness mechanism such as MCAR, MAR and MNAR.

(iii) Missing continuous covariates

In practice, most survival data involve categorical covariates. Therefore, this study only focused on missing values for categorical covariate. However, sometimes continuous variables are more useful and important thus further investigation may be done for this type of covariate with missing values. There are several studies that proposed the methods to handle missing continuous covariates, yet these studies mostly focused on the Cox proportional hazards model. Thus, study on this method for handling missing continuous data in the parametric model may be informative.

(iv) Testing frailty for the multivariate $(p > 2)$ case

The study on the properties of all three score tests considered in Chapter 6 focused on the bivariate positive stable Gompertz model. Therefore, it is worthwhile to discuss further the properties of these score tests for multivariate $(p > 2)$ cases. Besides, more intensive investigation of these score tests for the positive stable Gompertz model in the presence of covariates information either with missing values or without missing values may be necessary.

(v) Power analysis of the positive stable Gompertz model for censored case with nuisance parameters

One may continue the investigation on the power of all three score tests for censored case with nuisance parameters using the method of bootstrap that had been proposed by Sarker (2002). In addition, the power of these score tests may be further evaluated under misspecified frailty.

REFERENCES

Abdul Razak, A., Saddki, N., Naing, N. N., and Abdullah, N. (2010). Oral cancer survival among malay patients in Hospital Universiti Sains Malaysia, Kelantan. *Asian Pacific Journal of Cancer Prevention*, 11: 187-191.

Abiodun, A. A. (2012). Survival analysis of mortality data among elderly patients in university of Ilorin teaching hospital Ilorin, Nigeria. *Scientia Africana*, 11(1): 14-24.

Acs, G., and Gombos, Z. (2006). Prognostic factors and new methods in cervical carcinoma. *Pathology Case review*, 11(3): 130-139.

Ahn, Y. O., and Shin, M. H. (2011). Cancer survival in Seoul, Republic of Korea, 1993-1997. In Sankaranarayanan R and Swaminathan R (Eds.), *Cancer survival in Africa, Asia, the Caribbean and Central America*. IARC Scientific Publications No. 162, Lyon, France: IARC. Retrieved from http://survcan.iarc.fr/survival/chap22.pdf.

Aktürk Hayat, E., Suner, A., Uyar, B., Dursun, Ö., Orman, M. N., and Kitapçioğlu G. (2010). Comparison of five survival models: breast cancer registry data from Ege University Cancer Research Center. *Turkiye Klinikleri Journal of Medical Sciences*, 30(5): 1665-1674.

Alkan, N., Terzi, Y., Cengiz, M. A., and Alkan, B. B. (2013). Comparison of missing data analysis methods in Cox proportional hazard models. *Turkiye Klinikleri Journal of Biostatistics*, 5(2):49-54.

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage Publications.

Altman, D. G., de Stavola, B. L., Love, S. B., and Stepniewska, K. A. (1995). Review of survival analyses published in cancer journals. *British Journal of Cancer*, 72(2): 511–518.

American Cancer Society (2011). *Global Cancer Facts & Figures* (2nd Eds). Atlanta: American Cancer Society.

Andersen, P. K., Klein, J. P., and Zhang, M. (1999). Testing for centre effects in multi-centre survival studies: a Monte Carlo comparison of fixed and random effects tests. *Statistics In Medicine*, 18: 1489-1500.

Andridge, R. R., and Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical* Review, 78(1): 40-64. doi:10.1111/j.1751-5823.2010.00103.x.

Atahan, I. L., Onal, C., Ozyar, E., Yiliz, F., Selek, U., and Kose, F. (2007). Long-term outcome and prognostic factors in patients with cervical carcinoma: a retrospective study. *International Journal of Gynecological Cancer*, 17: 833-842.

Aziz, M. F. (2009). Gynecological cancer in Indonesia. *Journal of Gynecologic Oncology*, 20: 8-10.

Baade, P. D., Royston, P., Youl, P. H. Weinstock, M. A., Geller, A., and Aitken, J. F. (2015). Prognostic survival model for people diagnosed with invasive cutaneous melanoma. *BMC Cancer*, 15:27. doi: 10.1186/s12885-015-1024-4

Barzi, F., and Woodward, M. (2004). Imputations of missing values in practice: results from imputations of serum cholestrol in 28 cohort studies. *American Journal of Epidemiology*, 160: 34-45.

Bates, J. H., Hofer, B. M., and Parikh-Patel, A. (2008). Cervical cancer incidence, mortality, and survival among Asian subgroups in California, 1990-2004. *Cancer*, 113(10 Suppl): 2955-2963.

Bellera, C. A., MacGrogan, G., Debled, M., de Lara, C. T., Brouste, V., and Mathoulin-Pélissier, S. (2010). Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC Medical Research Methodology*, 10: 20. doi:10.1186/1471-2288-10-20.

Bessell, E. M., Bouliotis, G., Armstrong, S., Baddeley, J., Haynes, A. P., O’Connor, S., … Bradley, M. (2012). Long-term survival after treatment for Hodgkin’s disease (1973–2002): improved survival with successive 10-year cohorts. *British Journal of Cancer*, 107: 531-536.

Blossfeld, H-P., and Hamerle, A. (1989). Unobserved heterogeneity in hazard rate models: a test and an illustration from a study of career mobility. *Quality & Quantity*, 23: 129-141.

Bolfarine, H., and Valença, D. M. (2005), Testing homogeneity in Weibull-regression models. *Biometrical Journal*, 47(5): 707-720. doi: 10.1002/bimj.200410064.

Bradburn, M. J., Clark, T. G., Love, S. B., and Altman, D. G. (2003a). Survival analysis part II: multivariate data analysis - An introduction to concepts and methods. *British Journal of Cancer*, 89: 431-436.

Bradburn, M. J., Clark, T. G., Love, S. B., and Altman, D. G. (2003b). Survival Analysis Part III: multivariate data analysis - choosing a model and assessing its adequacy and fit. *British Journal of Cancer*, 89: 605-611.

Brown, M., and Kros, J. (2003). Data mining and the impact of missing data. *Industrial Management & Data Systems*, 103(8): 611-621.

Brun, J. L., Stoven-Camou, D., Trouette, R., Lopez, M., Chene, G., and Hocké, C. (2003). Survival and prognosis of women with invasive cervical cancer according to age. *Gynecologic Oncology*, 91: 395-401.

Burdett, K., Kiefer N. M., and Sharma S. (1985). Layoffs and duration dependence in a model of turnover. *Journal of Econometric*, 28: 51-69.

Cancer Research UK. (2014). *Treatment if you have abnormal cervical cells*. Retrieved from:http://www.cancerresearchuk.org/about-cancer/type/cervical-cancer/smears/treatment-if-you-have-abnormal-cervical-cells.

Caroni C. and Kimber, A. (2004). Detection of frailty in weibull lifetime data using outlier tests. *Journal of Statistical Computation and Simulation*, 74(1): 15-23.

Cheah, P. L., and Looi, L. M. (1999). Carcinoma of the uterine cervix: a review of its pathology and commentary on the problem in Malaysians. *Malaysian Journal of Pathology*, 21(1): 1-15.

Chemay, N. K., Naing, N. N., Rahman, M. N. G., and Bachok, N. (2008). Prognostic factors of prostate cancer patients at Hospital Universiti Sains Malaysia. *International Medical Journal*, 15(3): 225-231.

Chen, R. J., Lin, Y. H., Chen, C. A., Huang, S. C., Chow, S. N., and Hsieh, C.Y. (1999). Influence of histologic type and age on survival rates for invasive cervical carcinoma in Taiwan. *Gynecologic Oncology*, 73(2): 184-190.

Chung, H. H., Jang, M. J., Jung, K. W., Won, Y. J., Shin, H. R., Kim, J. W., and Lee, H. P. (2006). Cervical cancer incidence and survival in Korea: 1993-2002. *International Journal of Gynecological Cancer*, 16(5): 1833-1838.

Clark, T. G., Bradburn, M. J., Love, S. B. and Altman, D. G. (2003a). Survival analysis part I: basic concepts and first analyses. *British Journal of Cancer*, 89: 232-238.

Clark, T. G., Bradburn, M. J., Love, S. B. and Altman, D.G. (2003b). Survival analysis part IV: further concepts and methods in survival analysis. *British Journal of Cancer*, 89: 781-786.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65: 141-151.

Claeskens, G., Nguti, R., and Janssen, P. (2008). One-sided tests in shared frailty models. *Test*, 17: 69–82.

Cleves, M., Gould, W., Gutierrez, R. G., and Marchenko, Y. V. (2010). *An Introduction to Survival Analysis Using Stata* (3rd ed.). College Station, TX: Stata Press.

Coker, A. L., DeSimone, C. P., Eggleston, K. S., White. A. L., and Williams, M., (2009). Ethnic disparities in cervical cancer survival among Texas women. *Journal Of Women's Health*, 18(10): 1577-1583.

Collet D. (2003). *Modelling survival data in medical research* (2nd ed.). London: Chapman & Hall/CRC.

Commenges, D., and Andersen, P. K. (1995). Score test of homogeneity for survival data. *Lifetime Data Analysis*, 1: 145-156.

Cortez, P. (2013). *rminer: Data Mining Classification and Regression Methods*. R package version 1.3.1. Retrieved from http://CRAN.R-project.org/package=rminer.

Cox, D. R., and Oakes D. (1984). *Analysis of survival data*. London: Chapman & Hall.

Crowder, M., and Kimber, A. (1997). A score test for the multivariate Burr and other Weibull mixture distributions. *Scandinavian Journal of Statistics*, 24(3): 419-432.

Daniel, W. W. (2005). *Biostatistics: a foundation for analysis in the health sciences* (8th ed.). Hoboken, New Jersey : John Wiley & Sons.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B* (*Methodological*), 39: 1-38.

Ding, S., Soong, S. J, Lin, H. Y., Desmond, R., and Balch, C. M. (2009). Parametric modeling of localized melanoma prognosis and outcome. *Journal of Biopharmaceutical Statistics*, 19: 732-747.

Douine, M., Roue, T., Fior, A., Adenis, A., Thomas, N., Nacher, M. (2014). Survival of patients with invasive cervical cancer in French Guiana, 2003–2008. *International Journal of Gynecology and Obstetrics,* 125(2): 166–167.

Duchateau, L., Janssen, P., Lindsey, P., Legrand, C., Ngutic, R., and Sylvester, R. (2002). The shared frailty model andthe power for heterogeneity tests in multicenter trials. *Computational Statistics & Data Analysis*, 40: 603-620.

Dueňas-González, A., Orlando, M., Zhou, Y., Quinlivan, M., and Barraclough, H. (2012). Efficacy in high burden locally advanced cervical cancer with concurrent gemcitabine and cisplatin chemoradiotherapy plus adjuvant gemcitabine and cisplatin: prognostic and predictive factors and the impact of disease stage on outcomes from a prospective randomized phase III trial. *Gynecologic Oncology,* 126(3): 334-340. doi: 10.1016/j.ygyno.2012.06.011.

Economou, P. and Caroni C. (2005). Graphical tests for the assumption of gamma and inverse gaussian frailty distributions. *Lifetime Data Analysis,* 11: 565–582,

Economou P. and Caroni C. (2008). Graphical tests for the frailty distribution in the shared frailty model. *Communications in Statistics-Simulation and Computation,* 37: 978–992.

Economou, P. (2011). On model selection in the case of nested distributions - An application to frailty models. *Statistical Methodology*, 8: 172-184.

El-Sherbieny, E., Rashwan, H., Lubis, S. H., and Choi, V. J. (2011). Prognostic Factors in Patients with nasopharyngeal carcinoma treated in Hospital Kuala Lumpur. *Asian Pacific Journal of Cancer Prevention*, 12(7): 1739-1743.

Endo, D., Todo, Y., Okamoto, K., Minobe, S., Kato, H., Nishiyama. N. (2015). Prognostic factors for patients with cervical cancer treated with concurrent chemoradiotherapy: a retrospective analysis in a Japanese cohort. *Journal of Gynecologic Oncology*, 26(1):12-18.

Escobar, P. F., Chiesa-Vottero, A. and Michener, C. M. (2007). Diagnosis, workup, and management of preinvasive lesions of the cervix. In Sokol, A. I. and Sokol, E. R. (Eds.), *General gynecology: the requisites in obstetrics and gynecology* (pp. 429-457). Philadelphia: Mosby Elsevier.

Eser, S. (2011). Cancer survival in Izmir, Turkey, 1995−1997. In Sankaranarayan R and Swaminathan R (Eds.), *Cancer survival in Africa, Asia, the Caribbean and Central America*, IARC Scientific Publications No 162, Lyon, France: IARC. Retrieved from http://survcan.iarc.fr/survival/chap29.pdf.

Feller, W. (1966). *An introduction to probability theory and its applications*, Vol 2. New York: John Wiley.

Flores-Luna, L., Salazar-Martinez, E., Escudero-De los Rios, P., Gonzalez-Lira, G., Zamora-Muñoz, S., and Lazcano-Ponce, E. (2001). Prognostic factors related to cervical cancer survival in Mexican women. *International Journal of Gynecology and Obstetrics*, 75(1): 33-42.

Fonseca, R. S., Valença, D. M., and Bolfarine, H. (2013). Cure rate survival models with missing covariates: a simulation study. *Journal of Statistical Computation and Simulation*, 83(1): 97-113.

Galic, V., Herzog, T. J., Lewin, S. N., Neugut, A.I., Burke, W. M., Lu, Y. S., … Wright, J. D. (2012). Prognostic significance of adenocarcinoma histology in women with cervical cancer. *Gynecologic Oncology*, 125(2): 287-291.

Garipagaoglu, M., Yalvac, S., Kose, M. F., Tulunay, G., Kayikcioglu, F., Çakmak, A., … Hayran, M. (1999). Treatment results and prognostic factors in inoperable carcinoma of the cervix treated with external plus high dose brachytherapy. *Cancer Letters*, 136: 17-26.

Ghazali, A. K., Musa, K. I., Naing, N. N., and Mahmood, Z. (2010). Prognostic factors in patients with colorectal cancer at Hospital Universiti Sains Malaysia. *Asian Journal of Surgery*, 33: 127-133.

Gray, R. J. (1995). Tests for variation over groups in survival. *Journal of the American Statistical Association*, 90(429): 198-203.

Grigienė, R., Valuckas, K. P., Aleknavičius, E., Kurtinaitis, J., and Letautienė, S. R. (2007). The value of prognostic factors for uterine cervical cancer patients treated with irradiation alone. *BMC Cancer*, 7(234): 1-9.

Grover, G., Sabharwal, A., and Mittal, J. (2013). An application of gamma generalized linear model for estimation of survival function of diabetic nephropathy patients. *International Journal of Statistics in Medical Research*, 2: 209-219.

Gu, X., Shapiro, D., Hughes, M. D., and Balasubramanian, R. (2014). Stratified Weibull regression model for interval-censored data. *The R Journal*, 6(1): 31-40.

Hardt, J., Herke, M., Brian, T., and Laubach, W. (2013). Multiple imputation of missing data: a simulation study on a binary response. *Open Journal of Statistics*, 3: 370-378. doi: http://dx.doi.org/10.4236/ojs.2013.35043.

Hashemian, A. H., Beiranvand, B., Rezae, M., and Reissi, D. (2013). A comparison between Cox regression and parametric methods in analyzing kidney transplant survival. *World Applied Sciences Journal*, 26(4): 502-507.

Henderson, R., and Oman, P. (1999). Effect of frailty on marginal regression estimates in survival analysis. *Journal of the Royal Statistical Society Series B* (*Statistical Methodology*), 61(2): 367-379.

Henningsen, A., and Toomet, O. (2011). maxLik: A package for maximum likelihood estimation in R. *Computational Statistics*, 26(3): 443-458. doi: 10.1007/s00180-010-0217-1

Herring, A. H., and Ibrahim, J. G. (2001). Likelihood-based methods for missing covariates in the Cox proportional hazards model. *Journal of the American Statistical Association*, 96(453): 292-302.

Herring, A. H., Ibrahim, J. G., and Lipsitz, S. R. (2002). Frailty models with missing covariates. *Biometrics*, 58(1): 98-109.

Herring, A. H., Ibrahim, J. G., and Lipsitz, S. R. (2004). Non-ignorable missing covariate data in survival analysis: a case-study of an international breast cancer study group. *Journal of the Royal Statistical Society. Series C* (*Applied Statistics*), 53: 293-310.

Ho, C. M., Chien, T. Y., Huang, S. H., Wu, C. J., Shih, B. Y., and Chang, S. C. (2004). Multivariate analysis of the prognostic factors and outcomes in early cervical cancer patients undergoing radical hysterectomy. *Gynecologic Oncology*, 93: 458-464.

Ho, K. C., Wang, C. C., Qiu, J. T., Lai, C. H., Hong, J. H., Huang, Y. T., … Yen, T. C. (2011). Identification of prognostic factors in patients with cervical cancer and supraclavicular lymph node recurrence. *Gynecologic Oncology*, 123(2): 253-256.

Horton, N. J. and Kleinman, K. P. (2007). Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 61(1): 79-90. doi: 10.1198/000313007X172556

Hosmer, D. W., and Lemeshow, S. (1999). *Applied survival analysis: regression modeling of time to event data*. New York: John Wiley & Sons.

Hougaard, P. (1984). "Life Table Methods for Heterogeneous Populations: Distributions Describing the Heterogeneity". *Biometrika*, 71: 75-83.

Hougaard, P. (2000). *Analysis of multivariate survival data*. New York: Springer Verlag.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85: 765-769.

Ibrahim, J. G., Lipsitz, S. R., and Chen, M. H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society*. *Series B* (*Statistical Methodology*), 61(1): 173-190.

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R., and Herring, A. H. (2005). Missing-data methods for generalized linear models: a comparative review. *Journal of the American Statistical Association*, 100(469): 332-346.

Jackson, C. (2014). *flexsurv: Flexible parametric survival and multi-state models*. R package version 0.3. Retrieved from http://CRAN.R-project.org/package=flexsurv.

Jensen, P.T. (2007). Gynaecological cancer and sexual functioning: does treatment modality have an impact? *Sexologies*, 16(4): 279-285.

Jerez, J. M., Molina, I., Subirats, J. L., and Franco, L., (2006). *Missing data imputation in breast cancer prognosis: Proceedings of the 24th IASTES International Multi-Conference Biomedical Engineering* (pp.323-328). Innsbruck, Austria.

Kalbfleisch, J. D, and Prentice, R. L. (1980) *The statistical analysis of failure time data*. New York: John Wiley.

Katanyoo, K., Sanguanrungsirikul, S., Manusirivithaya, S. (2012). Comparison of treatment outcomes between squamous cell carcinoma and adenocarcinoma in locally advanced cervical cancer. *Gynecologic Oncology*, 125(2): 292-296.

Kavanagh, J. J., Phan, A., Tangjitgamol, S. and Ramirez, P. T. (2006). Tumor of the uterine cervix. In Kantarjian, H. M., Wolff, R. A. and Koller, C. A. (Eds.), *MD Anderson manual of medical oncology* (pp. 602-652). New York: McGraw-Hill.

Kiefer, N. M. (1984). A simple test for heterogeneity in exponential models of duration. *Journal of Labor Economics*, 2(4): 539-549.

Kimber, A. C. (1996). A Weibull-based score test for heterogeneity. *Lifetime Data Analysis*, 2: 63-71.

Klein, J. P., and Moeschberger, M. L. (1997). *Survival analysis: techniques for censored and truncated data*. New York : Springer-Verlag.

Kleinbaum, D.G., and Klein, M. (2005). *Survival analysis: a self-learning text* (2nd ed.). New York: Springer.

Köhler, H. F., and Kowalski, L. P. (2012). A critical appraisal of different survival techniques in oral cancer patients. *European Archives of Oto-Rhino-Laryngology*, 269: 295-301.

Kostova, P., Zlatkov, V., and Danon, S. (2008). Five-year overall survival and prognostic factors among patients with cervical cancer in Bulgaria. *Journal of the Balkan Union of Oncology*, 13(3): 363-368.

Kumari, K.G., Sudhakar, G., Ramesh, M., Kalpana, V. L., and Paddaiah, G. (2010). Prognostic factors in cervical cancer: a hospital-based retrospective study from Visakhapatnam City, Andhra Pradesh. *Journal of Life Sciences*, 2(2): 99-105.

Kyrgiou, M., and Shafi, M. I. (2010). Invasive cancer of the cervix. *Obstetrics, Gynaecology and Reproductive Medicine*, 20(5): 147-154.

Lancaster, T. (1985). Generalised residuals and heterogeneous duration models with applications to the Weibull model. *Journal of Econometrics*, 28: 155-169.

Laudico, A., and Mapua, C. (2011). Cancer survival in Manila Philippines, 1994-1995. In Sankaranarayan R and Swaminathan R (Eds.), *Cancer survival in Africa, Asia, the Caribbean and Central America*, IARC Scientific Publications No 162, Lyon, France: IARC. Retrieved from http://survcan.iarc.fr/survival/chap18.pdf.

Lee, E. T., and Wang, J. W. (2003). *Statistical methods for survival data analysis* (3rd ed.). Hoboken, New Jersey: Wiley-Interscience.

Leong, T., Lipsitz, S. R. and Ibrahim, J. G. (2001). Incomplete covariates in the Cox model with applications to biological marker. *Journal of the Royal Statistical Society. Series C* (*Applied Statistics*), 50(4): 467-484.

Liao, S. G., Lin, Y., Kang, D. D., Chandra, D., Bon, J., Kaminski, N., … Tseng, G. C. (2014). Missing value imputation in high-dimensional phenomic data: imputable or not, and how?. *BMC Bioinformatics*, 15:346.

Lin, D. Y. and Ying, Z. (1993). Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association*, 88(424): 1341-1349.

Lipsitz, S. R., and Ibrahim, J. G. (1996a). Using the EM-algorithm for survival data with incomplete categorical covariates. *Lifetime Data Analysis*, 2(1): 5-14

Lipsitz, S. R., and Ibrahim, J. G. (1996b). A conditional model for incomplete covariates in parametric regression models. *Biometrika*, 83(4): 916-922.

Lipsitz, S. R., and Ibrahim, J. G. (1998). Estimating equations with incomplete categorical covariates in the Cox model. *Biometrics*, 54(3): 1002-1013.

Little, R. J., Yosef, M., Cain, K. C., and Nan, B. and Harlow, S. D. (2008). A hot-deck multiple imputation procedure for gaps in longitudinal data on recurrent events. *Statistics in Medicine*, 27:103-120. doi: 10.1002/sim.2939.

Little, R. J. A., and Rubin D. B. (2002). *Statistical Analysis with missing data* (2nd ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.

Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 44: 226-233.

Mallett, S., Royston, P., Dutton, S., Waters, R., and Altman, D. G. (2010a). Reporting methods in studies developing prognostic models in cancer: a review. *BMC Medicine*, 8:20.

Mallett, S., Royston, P., Waters, R., Dutton, S., and Altman, D.G. (2010b). Reporting performance of prognostic models in cancer: a review, *BMC Medicine*, 8:21.

Mangantig, E., Naing, N. N., Norsa'adah, B., and Azlan, Husin. (2013). Survival and prognostic factors in Malaysian acute myeloid leukemia patients after allogeneic haematopoietic stem cell transplantation. *International Journal of Hematology*, 98: 197–205.

Marshall, A., Altman, D. G., and Holder, R. L. (2010a). Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. *BMC Medical Research Methodology*, 10(112): 1-10.

Marshall, A., Altman, D.G., Royston, P., and Holder R. L. (2010b). Comparison of techniques for handling missing covariate data within prognostic modeling studies: a simulation study. *BMC Medical Research Methodology*, 10(7): 1-16.

McLachlan, G. J., and Krishnan, T. (2008). *The EM Algorithm and Extensions* (2nd ed.). New York: Wiley-Interscience.

McShane, L. M., and Simon, R. (2001). *Prognostic factors in cancer* (2nd ed.). New York: John Wiley & Sons, Inc.

Moghimbeigi, A., Tapak, L., Roshanaei, G., and Mahjub. H., (2014). Survival analysis of gastric cancer patients with incomplete data. *Journal of Gastric Cancer*, 14(4):259-265. doi:http://dx.doi.org/10.5230/jgc.2014.14.4.259.

Moghimi-Dehkordi, B., Safaee, A., Pourhoseingholi, M. A., Fatemi, R., Tabeie, Z., and Zali, M. R. (2008). Statistical comparison of survival models for analysis of cancer data. *Asian Pacific Journal of Cancer Prevention,* 9: 417-420.

Moons, K. G. M., Donders, R. A. R. T., Stijnen, T., & Harrell Jr, F. E. (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology*, 59: 1092-1101.

Moore-Higgs, G. J., and Chafe, S. M. (2001). *Outcomes in radiation therapy multidisciplinary management*, Sudbury: Jones and Bartlett Publishers, Inc.

Moran, J. L., Bersten, A. D., Solomon, P. J., Edibam, C., Hunt, T. (2008). Modelling survival in acute severe illness: Cox versus accelerated failure time models. *Journal of Evaluation in Clinical Practice*, 14: 83-93.

Nakhaee, F., and Law, M. (2011). Parametric modelling of survival following HIV and AIDS in the era of highly active antiretroviral therapy: data from Australia. *Eastern Mediterranean Health Journal*, 17(3): 231-237.

Nardi, A., and Schemper, M. (2003). Comparing Cox and parametric models in clinical studies. *Statistics in Medicine*, 22: 3597-3610.

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

Paillisse, C., Lacomblez, L., Dib, M., Bensimon, G., Garcia-Acosta, S., and Meininger, V. (2005). Prognostic factors for survival in amyotrophic lateral sclerosis patients treated with riluzole. *Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders*, 6: 37-44.

Pari Dayal, L., Leo Alexander, T., Ponnuraja, C., Ranganathan Rama, Venkatesan, P. (2013). Modelling of time to event breast cancer data using accelerated failure time (AFT) in South India women. *Global Research Analysis*, 2(5): 192-194.

Patel, D. A., Barnholtz-Sloan, J. S., Patel, M. K., Malone, J. J. M., Chuba, P. J., and Schwartz, K. (2005). A population-based study of racial and ethnic differences in survival among women with invasive cervical cancer: analysis of surveillance, epidemiology, and end results data. *Gynecologic Oncology*, 97(2): 550-558.

Pickel, H., Haas, J., and Lahousen, M. (1997). Prognostic factors in cervical cancer. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 71(2): 209-213.

Pierce, D. A. (1982). The asymptotic effect of substituting estimators for parameters in certain type of statistics, *Annals of Statistics*, 10: 475-478.

Pomros, P., Sriamporn, S., Tangvoraphonkchai, V., Kamsa-Ard, S., and Poomphakwaen, K. (2007). Factors affecting survival of cervical cancer patients treated at the radiation unit of Srinagarind Hospital, Khon Kaen University, Thailand. *Asian Pacific Journal of Cancer Prevention*, 8: 297-300.

Pourhoseingholi, M. A., Hajizadeh, E., Moghimi-Dehkordi, B., Safaee, A., Abadi, A., and Zali, M. R. (2007). Comparing Cox regression and parametric models for survival of patients with gastric carcinoma. *Asian Pacific Journal of Cancer Prevention*, 8: 412-416.

Pourhoseingholi, M. A., Moghimi-Dehkordi, B., Safaee, A., Hajizadeh, E., Solhpour, A., and Zali, M. R. (2009). Prognostic factors in gastric cancer using log-normal censored regression model. *Indian Journal of Medical Research*, 129(3): 262-267.

Pourhoseingholi, M. A., Pourhoseingholi, A., Vahedi, M., Moghimi-Dehkordi, B., Safaee, A., Ashtari, S., and Zali, M. R. (2011). Alternative for Cox regression: parametric model to analysis the survival of cancer patients. *Iranian Journal of Cancer Prevention*, 4(1): 1-9.

Priest, P., Sadler, L., Sykes, P., Marshall, R., Peters, J., and Crengle, S. (2010). Determinants of inequalities in cervical cancer stage at diagnosis and survival in New Zealand. *Cancer Causes Control*, 21: 209-214.

Pruegsanusak, K., Peeravut, S., Leelamanit, V., Sinkijcharoenchai, W., Jongsatitpaiboon, J., Phungrassami, T., … P. Thongsuksai. (2012). Survival and Prognostic Factors of Different Sites of Head and Neck Cancer: An Analysis from Thailand. *Asian Pacific Journal of Cancer Prevention*. 13: 885-890. doi:http://dx.doi.org/10.7314/APJCP.2012.13.3.885.

Qi, J. (2009). *Comparison of proportional hazards and accelerated failure time models* (Master thesis). Department of Mathematics and Statistics, University of Saskatchewan.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Radstone, D., and Kunkler I. H. (2003). Cervix, body of uterus, ovary, vagina, vulva, gestational trophoblastic tumours. In Bomford, C. K, Kunkler, I. H, Miller, H., and Walter, J. (Eds.), *Walter and Miller's textbook of Radiotherapy Radiation Physics, Therapy, and Oncology* (6th ed., pp. 465-486). Edinburgh: Churchill Livingstone.

Ravangard, R., Arab, M., Rashidian, A., Akbarisari, A., Zare, A., and Zeraati, H. (2011). Comparison of the results of Cox proportional hazards model and parametric models in the study of length of stay in a tertiary teaching hospital in Tehran, Iran. *Acta Medica Iranica*, 49(10): 650-658.

Redaniel, M. T., Laudico, A., Mirasol-Lumague, M. R., Gondos, A., Uy, G. L., Toral, J. A., … Brenner, H. (2009). Ethnicity and health care in cervical cancer survival: comparisons between a Filipino resident population, Filipino-Americans, and Caucasians. *Cancer Epidemiology, Biomarkers & Prevention*, 18(8): 2228-2234. doi: 10.1158/1055-9965.EPI-09-0317.

Rijke, J. M. D., Putten, H. W. H. M. V. D., Lutgens, L. C. H. W., Voogd, A. C., Kruitwagen, R. F. P. M., Dijck, J. A. A. M. V., and Schouten, L. J. (2002). Age-specific differences in treatment and survival of patients with cervical cancer in the Southeast of the Netherlands, 1986–1996. *European Journal of Cancer*, 38: 2041–2047.

Royston, P. (2001). The lognormal distribution as a model for survival time in cancer, with an emphasis on prognostic factors. *Statistica Neerlandica*, 55(1): 89-104.

Royston, P., and Lambert, P. C. (2011). *Flexible Parametric Survival Analysis using Stata: beyond the Cox model*. College Station, Texas: Stata Press.

Royston, P. and White, I. R. (2011). Multiple imputation by chained equations (MICE): implementation in Stata. *Journal of Statistical Software*, 45(4): 1-20.

Sankaranarayanan, R., Swaminathan, R., Jayant, K., and Brenner, H. (2011). An overview of cancer survival in Africa, Asia, the Caribbean and Central America: the case for investment in cancer health services. In Sankaranarayan R and Swaminathan R (Eds.), *Cancer survival in Africa, Asia, the Caribbean and Central America*, IARC Scientific Publications No 162, Lyon, France: IARC. Retrieved from http://survcan.iarc.fr/survival/chap32.pdf.

Sant, M., Allemani, C., Santaquilanib, M., Knijnb, A., Marchesia, F., Capocacciac, R., and the Eurocare Working Group Eurocare-4. (2009). Survival of cancer patients diagnosed in 1995–1999. Results and commentary. *European Journal of Cancer*. 45: 931-991.

Sarker, M. S. J. (2002). *Tests for Weibull based proportional hazards frailty models* (PhD thesis). Department of Mathematics & Statistics, University of Surrey.

Sayehmiri, K., Eshraghian, M. R., Mohammad, K., Alimoghaddam, K., Foroushani, A. R., Zeraati, H., … Ghavamzadeh, A. (2008). Prognostic factors of survival time after hematopoietic stem cell transplant in acute lymphoblastic leukemia patients: Cox proportional hazard versus accelerated failure time models. *Journal of Experimental & Clinical Cancer Research*, 27:74. doi: 10.1186/1756-9966-27-74.

Schafer J. L., and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, 7(2): 147-177.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Schneider, I. J. C., Flores, M. E., Nickel, D. A. Martins, L. G. T., and Traebert, J. (2014). Survival rates of patients with cancer of the lip, mouth and pharynx: a cohort study of 10 years. *Revista Brasileira de Epidemiologia*, 17(3): 680-691. doi: 10.1590/1809-4503201400030009

Schwartz, S. M., Daling, J. R., Shera, K. A., Madeleine, M. M., McKnight, B., Galloway, D. A., … McDougall, J. K. (2001). Human papillomavirus and prognosis of invasive cervical cancer: a population-based study. *Journal of Clinical Oncology*, 19(7): 1906-1915.

Seamon, L. G., Tarrant, R. L., Fleming, S. T., Vanderpool, R. C., Pachtman, S., Podzielinski, I., … DeSimone, C. P. (2011). Cervical cancer survival for patients referred to a tertiary care center in Kentucky. *Gynecologic Oncology*, 123(3): 565-570.

Shin, H. R., Lee, D. H., Lee, S. Y., Lee, J. T., Park, H. K., Rha, S. H., … Kong, H. J. (2011). Cancer survival in Busan, Republic of Korea, 1996-2001. In Sankaranarayan R and Swaminathan R (Eds.), *Cancer survival in Africa, Asia, the Caribbean and Central America*, IARC Scientific Publications No 162, Lyon, France: IARC. Retrieved from http://survcan.iarc.fr/survival/chap20.pdf.

Sinha, S. K. (2012). The use of score tests for frailty variance components in recurrent event data. *Journal of Biometrics and Biostatistics*, S4-003. doi:10.4172/2155-6180.S4-003.

Sirait, A. M., Soetiarto, F., and Oemiatil, R. (2003). Survival rate of cervical cancer patients in Dharmais Cancer Hospital, Jakarta. *Buletin Penelitian Kesehatan*, 31: 13-24.

Suh, D. H., Kim, T. H., Kim, J. W., Kim, S. Y., Kim, H. S., Lee, T. S., … Song, Y. S. (2013). Improvements to the FIGO staging for ovarian cancer: reconsideration of lymphatic spread and intraoperative tumor rupture. *Journal of Gynecologic Oncology*, 24(4): 352-358.

Sumitsawan, Y., Srisukho, S., Sastraruji, A., Chaisaengkhum, U., Maneesai, P., and Waisri, N. (2011). Cancer survival in Chiang Mai, Thailand, 1993−1997. In Sankaranarayan R and Swaminathan R (Eds.), *Cancer survival in Africa, Asia, the Caribbean and Central America*, IARC Scientific Publications No 162, Lyon, France: IARC. Retrieved from http://survcan.iarc.fr/survival/chap25.pdf.

Taib, N. A., Yip, C. H., and Mohamed, I. (2008). Survival analysis of Malaysian women with breast cancer: results from the University of Malaya Medical Centre. *Asian Pacific Journal of Cancer Prevention*, 9(2): 197-202.

Therneau, T. M., and Grambsch P. M. (2000). *Modeling survival data: extending the Cox model*. New York: Springer-Verlag.

Therneau T. M. (2014). *survival: Survival Analysis*. R package version 2.37.7. Retrieved from http://CRAN.R-project.org/package=survival.

Van Buuren, S., and Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software,* 45(3): 1-67.

Van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Boca Raton, Florida: CRC Press.

Van Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L., and Jolani, S. (2014a). *mice: Multivariate Imputation by Chained Equations*. R package version 2.21. Retrieved from http://cran.r- project.org/package=mice.

Van Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L., and Jolani, S. (2014b). *mice: Multivariate Imputation by Chained Equations*. R package version 2.22. Retrieved from http://cran.r-project.org/web/packages/mice/mice.pdf.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3): 439-454.

Verweij, P. J. M., van Houwelingen, H. C., and Stijnen, T. (1998). A goodness-of-fit test for Cox's proportional hazards model based on martingale residuals. *Biometrics,* 54(4): 1517-1526.

Viswanathan, B., and Manatunga, A. K. (2001). Diagnostic plots for assessing the frailty distribution in multivariate survival data. *Lifetime Data Analysis,* 7: 143–155.

Waggoner, S. E. (2003). Cervical cancer. *The Lancet,* 361 (9376): 2217-2225.

Wahidah, T., Khattak, M. N., Wan Arfah, N., and Naing, N. N. (2012). Prognostic factors of osteosarcoma patients in Hospital Universiti Sains Malaysia. *International Medical Journal,* 19(2): 150-153.

Walboomers, J. M. M., Jacobs, M. V., Manos, M. M., Bosch, F. X., Kummer, J. A., Shah, K. V., … Muñoz, N. (1999). Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *The Journal of Pathology*, 189(1): 12-19.

Wan Zamaniah, W. I., Mastura, M. Y., Phua, C. E., Adlinda, A., Marniza, S., and Rozita, A. M. (2014). Definitive Concurrent Chemoradiotherapy in Cervical Cancer - a University of Malaya Medical Centre Experience. *Asian Pacific Journal of Cancer Prevention*, 15(20): 8987-8992.

Wang, H., Chia, K. S., Du, W. B., Lee, J., Sankaranarayanan, R., Sankila, R., … Lee, H. P. (2003). Population-based survival for cervical cancer in Singapore, 1968-1992. *American Journal of Obstetrics & Gynecology*, 188(2): 324-329.

Wang, B. B., Liu, C. G., Lu, P., Latengbaolide, A., and Lu, Y. (2011). Log-normal censored regression model detecting prognostic factors in gastric cancer: a study of 3018 cases. *World Journal of Gastroenterology*, 17(23): 2867-2872.

Wei, G. C. G., and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411): 699-704.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, 30: 377-399.

Wienke, A. (2003). Frailty models. Max Planck Institute for Demographic Research (MPIDR) Working Paper WP 2003-032. Retrieved from http://www.demogr.mpg.de/papers/working/wp-2003-032.pdf.

Wienke, A., Arbeev, K. G., Locatelli, I., and Yashin, A. I. (2005). A comparison of different bivariate correlated frailty models and estimation strategies. *Mathematical Biosciences*, 198: 1-13.

Wienke, A. (2011). *Frailty models in survival analysis*. Boca Raton, Florida: CRC Press.

Woo, Z. H., Hong, Y. C., Kim, W. C., and Pu, Y. K. (2011). Cancer survival in Incheon, Republic of Korea, 1997−2001. In Sankaranarayan R and Swaminathan R (Eds.), *Cancer survival in Africa, Asia, the Caribbean and Central America*, IARC Scientific Publications No 162, Lyon, France: IARC. Retrieved from http://survcan.iarc.fr/survival/chap21.pdf.

Wuertz, D., Maechler, M., and Rmetrics core team members. (2013). *stabledist: Stable Distribution Functions*. R package version 0.6.6. Retrieved from http://CRAN.R-project.org/package=stabledist.

Xia, X., Xu, H., Wang, Z., Liu, R., Hu, T., and Li, S. (2014). Analysis of prognostic factors affecting the outcome of stage Ib-IIb cervical cancer treated by radical hysterectomy and pelvic lymphadenectomy. *American Journal of Clinical Oncology*, doi: 10.1097/COC.0000000000000100.

Xiang Y. B., Jin F, and Gao YT (2011). Cancer survival in Shanghai, China, 1992-1995. In Sankaranarayan R and Swaminathan R (Eds.), *Cancer survival in Africa, Asia, the Caribbean and Central America*, IARC Scientific Publications No 162, Lyon, France: IARC. Retrieved from http://survcan.iarc.fr/survival/chap7.pdf.

Yamauchi, M., Fukuda, T., Wada, T., Kawanishi, M., Imai, K., Hashiguchi, Y., … Sumi, T. (2014). Comparison of Outcomes Between Squamous Cell Carcinoma and Adenocarcinoma in Patients with Surgically Treated Stage I-II Cervical Cancer. *Molecular And Clinical Oncology*, 2: 518-524.

Yang, L. Y., Jia, X. B., Li, N. W., Chen, C., Liu, Y., and Wang H. J. (2013). Comprehensive clinic-pathological characteristics of cervical cancer in Southwestern China and the clinical significance of histological type and lymph node metastases in young patients. *PLoS ONE*, 8(10): e75849. doi:10.1371/journal.pone.0075849. eCollection 2013.

Yeh, S. A., Leung, S. W., Wang, C. J., and Chen, H. C. (1999). Postoperative radiotherapy in early stage carcinoma of the uterine cervix: treatment results and prognostic factors. *Gynecologic Oncology*,72: 10-15.

Yeole, B. B., Kurkure, A. P., and Sunny, L. (2011). Cancer survival in Mumbai (Bombay), India, 1992-1999. In Sankaranarayan R and Swaminathan R (Eds.), *Cancer survival in Africa, Asia, the Caribbean and Central America*, IARC Scientific Publications No 162, Lyon, France: IARC. Retrieved from http://survcan.iarc.fr/survival/chap16.pdf.

Zarchi, M. K., Akhavan, A., Fallahzadeh, H., Gholami, H., Dehghani, A., and Teimoori, S. (2010). Outcome of cervical cancer in Iranian patients according to tumor histology, stage of disease and therapy. *Asian Pacific Journal of Cancer Prevention,* 11: 1289-1291.

Zare, N., Doostfatemeh, M., and Rezaianzadeh, A. (2012). Modeling of breast cancer prognostic factors using a parametric log-logistic model in Fars Province, Southern Iran. *Asian Pacific Journal of Cancer Prevention*, 13: 1533-1537.

Zhu, C. Q. (1998). *Statistical methods for Weibull based random effects models* (Phd thesis). Department of Mathematics & Statistics, University of Surrey.

Zhu, H. P., Xia, X., Yu, C. H., Adnan, A., Liu, S. F., and Du, Y. K, (2011). Application of Weibull model for survival of patients with gastric cancer. *BMC Gastroenterology,* 11: 1-6.

LIST OF PUBLICATIONS AND PAPERS PRESENTED

List of Publications

1. Razak N. A., Khattak M. N., Zubairi Y. Z., Naing N. N., and Zaki N. M. (2013). Estimating the five-year survival of cervical cancer patients treated in Hospital Universiti Sains Malaysia. *Asian Pacific Journal of Cancer Prevention*, 14(2): 825-828.

2. Razak, N. A., Khattak, M. N., Naing, N. N., Zaki, N. M., and Zubairi, Y. Z. (2013). Survival analysis of cervical cancer patients treated in Hospital Universiti Sains Malaysia. *International Medical Journal*, 20(6): 707-710.

3. Razak, N. A., Zubairi Y. Z., and Yunus R.M. Imputing missing values in modelling the PM_{10} concentrations, *Sains Malaysiana*. 43(10): 1601–1609.

4. Juhan, N., Razak, N. A., Zubairi, Y. Z., Khattak, M. N., and Naing, N. N. (2013). Survey of patients with cervical cancer in Hospital Universiti Sains Malaysia: survival data analysis with time-dependent covariate. *Iranian Journal of Public Health*, 42(9): 980-987.

List of Conferences

1. 2010 Universiti Brunei Darussalam 1st Graduate Science Student Research Conference. University of Brunei Darussalam, Brunei. (2010).
   Title: *Prognostic factors influencing the survival of cervical cancer patients treated in Hospital Universiti Sains Malaysia*.

2. Seminar Kebangsaan Pascasiswazah Statistik ISM 1, University of Malaya. (2011).
   Title: *Comparison of methods for handling missing data*.

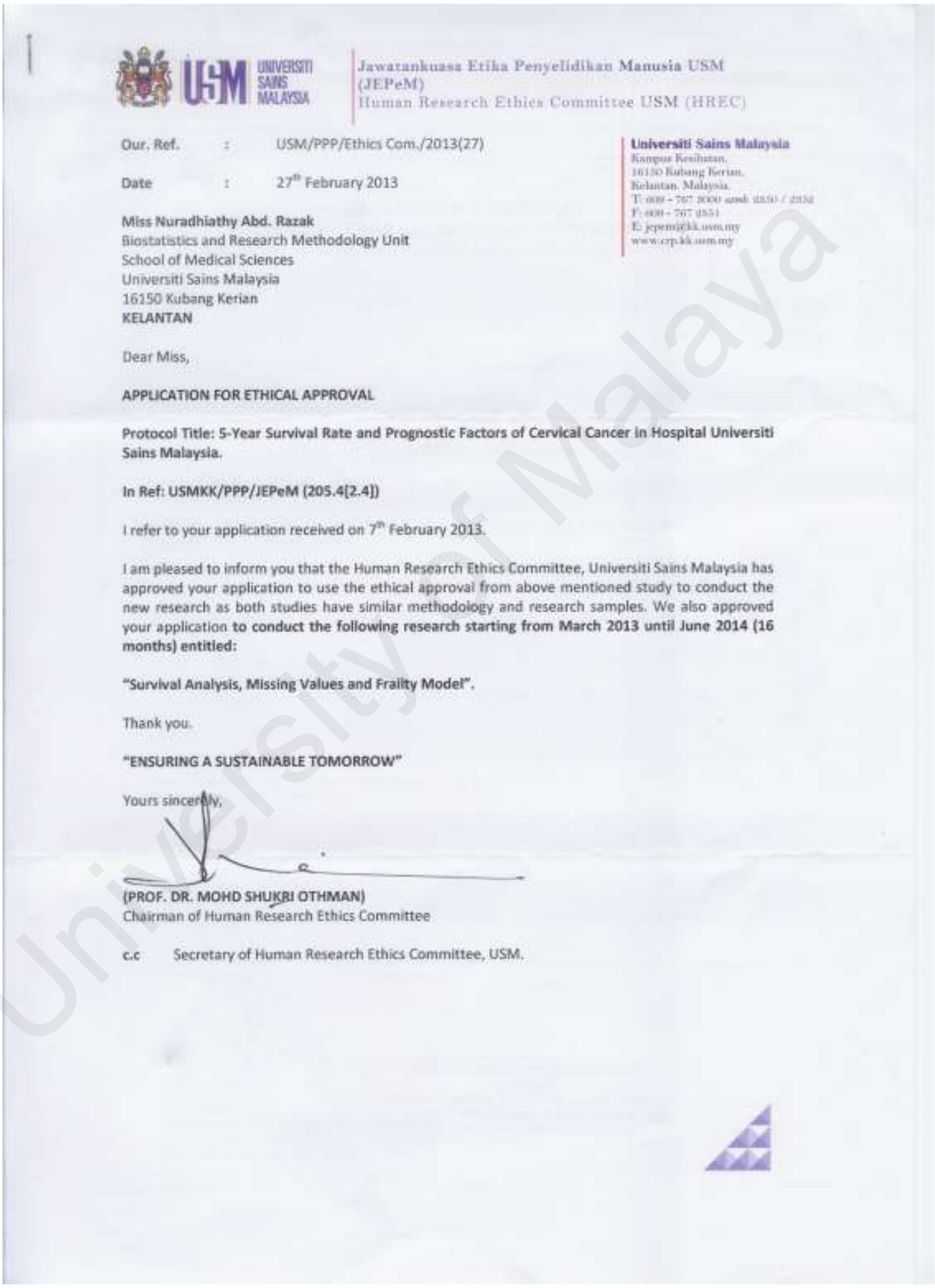
3. 1st ISM International Statistical Conference 2012, Persada Johor, Malaysia. (2012).
   Title: *Probability Distribution of PM_{10} Concentration with Missing Values*.

4. International Conference on Applied Analysis And Mathematical Modelling, Yildiz Technical University, Turkey. (2013).
   Title: *The applicability of Weibull based critical values for detecting frailty in Gompertz lifetime data*.

5. Bioinformatics and Biostatistics Applications in Cancer Genomics Research, Qatar University, Doha. (2015).
   Title:*Analysis of Missing Covariate Values in Parametric Survival Model And its Application in a Cancer Data*

APPENDIX

Appendix A: Ethical approval



Jawatankuasa Etika Penyelidikan Manusia USM (JEPeM)
Human Research Ethics Committee USM (HREC)

Universiti Sains Malaysia
Kampus Kesihatan,
16150 Kubang Kerian,
Kelantan. Malaysia.
T: 609 - 767 3000 samb. 2350 / 2352
F: 609 - 767 2351
E: jepem@kk.usm.my
www.crp.kk.usm.my

Our. Ref. : USM/PPP/Ethics Com./2013(27)

Date : 27th February 2013

Miss Nuradhiathy Abd. Razak
Biostatistics and Research Methodology Unit
School of Medical Sciences
Universiti Sains Malaysia
16150 Kubang Kerian
KELANTAN

Dear Miss,

APPLICATION FOR ETHICAL APPROVAL

Protocol Title: 5-Year Survival Rate and Prognostic Factors of Cervical Cancer in Hospital Universiti Sains Malaysia.

In Ref: USMKK/PPP/JEPeM (205.4[2.4])

I refer to your application received on 7th February 2013.

I am pleased to inform you that the Human Research Ethics Committee, Universiti Sains Malaysia has approved your application to use the ethical approval from above mentioned study to conduct the new research as both studies have similar methodology and research samples. We also approved your application **to conduct the following research starting from March 2013 until June 2014 (16 months) entitled:**

"Survival Analysis, Missing Values and Frailty Model".

Thank you.

"ENSURING A SUSTAINABLE TOMORROW"

Yours sincerely,

(PROF. DR. MOHD SHUKRI OTHMAN)
Chairman of Human Research Ethics Committee

c.c Secretary of Human Research Ethics Committee, USM.

Appendix B: R code for the comparison of missing data methods

```
MISSING=function(n,S,lamdaC,shape,beta0,beta1,beta2,a02,a12,a11,a22,a3
3,a00,b00,c00){

# Generate data from Weibul distribution#
#----------------------------------------#

lambdaT = exp(beta0)                              # Baseline hazard
x1=rbinom(n,1,0.6)                                # Generate x1{0,1}
pi.x2=exp(a02+a12*x1)/(1+exp(a02+a12*x1))         # Prob(x2)
x2=rbinom(n,1,pi.x2)                              # Generate x2{0,1}
# True survival time
T = rweibull(n, shape=shape, scale=lambdaT*exp((beta1*x1+beta2*x2)))
C = rweibull(n, shape=shape, scale=lambdaC)       #censoring time
time = pmin(T,C)         # observed time is min of censored & true
event = time==T          # set to 1 if event is observed
sort(event)

A=cbind(time,event,x1,x2)
x=data.frame(A)                                   # Data

#-----------------------------------------------------------------------#

# generate missing at random covariate #
#---------------------------------------#

x.star=(T-mean(T))/sd(T)
miss.mech=function(x,a00,a11,a22,a33,x.star){

PR=exp(a00+a11*x.star+a22*x1+a33*x1*x.star)/      #Probability MAR
      (1+exp(a00+a11*x.star+a22*x1+a33*x1*x.star))
r.x2.mar=rbinom(n,1,PR)
x2.mar<-x2*(1-r.x2.mar)+r.x2.mar*99999
x2.mar[x2.mar==99999]=NA

z<-data.frame(cbind(A[,1:3],x2.mar))              #Data MAR

#-----------------------------------------------------------------------#

      # Complete case analysis method #
      #-------------------------------#

      #--- Survival analysis (beta estimation) ---#

      omit<-na.omit(z)                            # remove NA
      t=survreg(Surv(omit[,1], omit[,2]) ~ omit[,3]+omit[,4], omit,
      dist="weibull")
      coeff=cbind(t$coefficients)
      p=t$scale
      b0=coeff[1,]
      b1=coeff[2,]
      b2=coeff[3,]
      b0.omit=b0
      b1.omit=b1
      b2.omit=b2
      p.omit=p
      #--- Logistic regression (alpha estimation) ---#

      lrfit <- glm( omit[,4] ~ omit[,3], family = binomial)
      coef=cbind(lrfit$coefficients)
      a0=coef[1,]
      a1=coef[2,]

      alpha.omit=cbind(a0,a1)

# Note: Estimation values (.omit) are used for initial values for EM #

#--------------------------------------------------------------------#

      # EM by method of weight #
      #------------------------#

      EM.weight=EM(S,z,b0,b1,b2,p,a0,a1)
      b0.EM=EM.weight$b0.EM
      b1.EM=EM.weight$b1.EM
      b2.EM=EM.weight$b2.EM
      p.EM=EM.weight$p.EM

#--------------------------------------------------------------------#

      # Hot deck imputation method #
      #----------------------------#

      # Call library(rminer) #

      HD <- imputation("hotdeck",z)
      HDR=survreg(Surv(HD[,1],HD[,2])~HD[,3]+HD[,4],HD,dist="weibull")
      coeff.HD=cbind(HDR$coefficients)

      #--- Hot deck estimation ---#

      p.HD=HDR$scale
      b0.HD=coeff.HD[1,]
      b1.HD=coeff.HD[2,]
      b2.HD=coeff.HD[3,]

#--------------------------------------------------------------------#

      # multiple Imputation (MICE-PMM)method #
      #--------------------------------------#

      # Call library(mice) #


      #--- Imputation (m=10) ---#

      imp <- mice(z, m=10)
      com <- complete(imp, "long")
      data1<-com[1:n,]
      data2<-com[(n+1):(2*n),]
      data3<-com[(2*n+1):(3*n),]
      data4<-com[(3*n+1):(4*n),]
      data5<-com[(4*n+1):(5*n),]
      data6<-com[(5*n+1):(6*n),]
      data7<-com[(6*n+1):(7*n),]
      data8<-com[(7*n+1):(8*n),]
      data9<-com[(8*n+1):(9*n),]
      data10<-com[(9*n+1):(10*n),]



      #--- Analysis ---#

      M1<-survreg(Surv(time,event) ~ x1+x2.mar, data1, dist="weibull")
      coeff.M1=cbind(M1$coefficients)
```

```
p.M1=M1$scale
b0.M1=coeff.M1[1,]
b1.M1=coeff.M1[2,]
b2.M1=coeff.M1[3,]

M2<-survreg(Surv(time,event) ~ x1+x2.mar, data2, dist="weibull")
coeff.M2=cbind(M2$coefficients)
p.M2=M2$scale
b0.M2=coeff.M2[1,]
b1.M2=coeff.M2[2,]
b2.M2=coeff.M2[3,]

M3<-survreg(Surv(time,event) ~ x1+x2.mar, data3, dist="weibull")
coeff.M3=cbind(M3$coefficients)
p.M3=M3$scale
b0.M3=coeff.M3[1,]
b1.M3=coeff.M3[2,]
b2.M3=coeff.M3[3,]

M4<-survreg(Surv(time,event) ~ x1+x2.mar, data4, dist="weibull")
coeff.M4=cbind(M4$coefficients)
p.M4=M4$scale
b0.M4=coeff.M4[1,]
b1.M4=coeff.M4[2,]
b2.M4=coeff.M4[3,]

M5<-survreg(Surv(time,event) ~ x1+x2.mar, data5, dist="weibull")
coeff.M5=cbind(M5$coefficients)
p.M5=M5$scale
b0.M5=coeff.M5[1,]
b1.M5=coeff.M5[2,]
b2.M5=coeff.M5[3,]

M6<-survreg(Surv(time,event) ~ x1+x2.mar, data6, dist="weibull")
coeff.M6=cbind(M6$coefficients)
p.M6=M6$scale
b0.M6=coeff.M6[1,]
b1.M6=coeff.M6[2,]
b2.M6=coeff.M6[3,]

M7<-survreg(Surv(time,event) ~ x1+x2.mar, data7, dist="weibull")
coeff.M7=cbind(M7$coefficients)
p.M7=M7$scale
b0.M7=coeff.M7[1,]
b1.M7=coeff.M7[2,]
b2.M7=coeff.M7[3,]

M8<-survreg(Surv(time,event) ~ x1+x2.mar, data8, dist="weibull")
coeff.M8=cbind(M8$coefficients)
p.M8=M8$scale
b0.M8=coeff.M8[1,]
b1.M8=coeff.M8[2,]
b2.M8=coeff.M8[3,]

M9<-survreg(Surv(time,event) ~ x1+x2.mar, data9, dist="weibull")
coeff.M9=cbind(M9$coefficients)
p.M9=M9$scale
b0.M9=coeff.M9[1,]
b1.M9=coeff.M9[2,]
b2.M9=coeff.M9[3,]

M10<-survreg(Surv(time,event)~x1+x2.mar, data10, dist="weibull")
coeff.M10=cbind(M10$coefficients)
p.M10=M10$scale
```

```
     b0.M10=coeff.M10[1,]
     b1.M10=coeff.M10[2,]
     b2.M10=coeff.M10[3,]

     #--- Combine ---#

     p.all=rbind(p.M1,p.M2,p.M3,p.M4,p.M5,p.M6,p.M7,p.M8,p.M9,p.M10)
     b0.all=rbind(b0.M1,b0.M2,b0.M3,b0.M4,b0.M5,b0.M6,b0.M7,b0.M8,b0.
     M9,b0.M10)
     b1.all=rbind(b1.M1,b1.M2,b1.M3,b1.M4,b1.M5,b1.M6,b1.M7,b1.M8,b1.
     M9,b1.M10)
     b2.all=rbind(b2.M1,b2.M2,b2.M3,b2.M4,b2.M5,b2.M6,b2.M7,b2.M8,b2.
     M9,b2.M10)

     #--- Multiple imputation estimation ---#

     p.MI=matrix(colMeans(p.all))
     b0.MI=matrix(colMeans(b0.all))
     b1.MI=matrix(colMeans(b1.all))
     b2.MI=matrix(colMeans(b2.all))


  list(b0.omit=b0.omit,b1.omit=b1.omit,b2.omit=b2.omit,p.omit=p.omit,
  b0.EM=b0.EM,b1.EM=b1.EM,b2.EM=b2.EM,p.EM=p.EM,b0.HD=b0.HD,
  b1.HD=b1.HD,b2.HD=b2.HD,p.HD=p.HD,b0.MI=b0.MI,b1.MI=b1.MI,
  b2.MI=b2.MI,p.MI=p.MI)

  }

 # Run for different percentage of missing values #

 NA.10=miss.mech(x,a00,a11,a22,a33,x.star)     # 10%  NA
 NA.30=miss.mech(x,b00,a11,a22,a33,x.star)     # 30%  NA
 NA.50=miss.mech(x,c00,a11,a22,a33,x.star)     # 50%  NA

#------- Estimation based on complete case analysis

 b0.omit.10=NA.10$b0.omit;b1.omit.10=NA.10$b1.omit;
 b2.omit.10=NA.10$b2.omit;p.omit.10=NA.10$p.omit

 b0.omit.30=NA.30$b0.omit;b1.omit.30=NA.30$b1.omit;
 b2.omit.30=NA.30$b2.omit;p.omit.30=NA.30$p.omit

 b0.omit.50=NA.50$b0.omit;b1.omit.50=NA.50$b1.omit;
 b2.omit.50=NA.50$b2.omit;p.omit.50=NA.50$p.omit


 #------- Estimation based on EM

 b0.EM.10=NA.10$b0.EM;b1.EM.10=NA.10$b1.EM;b2.EM.10=NA.10$b2.EM;
 p.EM.10=NA.10$p.EM

 b0.EM.30=NA.30$b0.EM;b1.EM.30=NA.30$b1.EM;b2.EM.30=NA.30$b2.EM;
 p.EM.30=NA.30$p.EM

 b0.EM.50=NA.50$b0.EM;b1.EM.50=NA.50$b1.EM;b2.EM.50=NA.50$b2.EM;
 p.EM.50=NA.50$p.EM
```

```
  #------- Estimation based on hot deck imputation

  b0.HD.10=NA.10$b0.HD;b1.HD.10=NA.10$b1.HD;b2.HD.10=NA.10$b2.HD;
  p.HD.10=NA.10$p.HD

  b0.HD.30=NA.30$b0.HD;b1.HD.30=NA.30$b1.HD;b2.HD.30=NA.30$b2.HD;
  p.HD.30=NA.30$p.HD

  b0.HD.50=NA.50$b0.HD;b1.HD.50=NA.50$b1.HD;b2.HD.50=NA.50$b2.HD;
  p.HD.50=NA.50$p.HD


  #------- Estimation based on multiple imputation

  b0.MI.10=NA.10$b0.MI;b1.MI.10=NA.10$b1.MI;b2.MI.10=NA.10$b2.MI;
  p.MI.10=NA.10$p.MI

  b0.MI.30=NA.30$b0.MI;b1.MI.30=NA.30$b1.MI;b2.MI.30=NA.30$b2.MI;
  p.MI.30=NA.30$p.MI

  b0.MI.50=NA.50$b0.MI;b1.MI.50=NA.50$b1.MI;b2.MI.50=NA.50$b2.MI;
  p.MI.50=NA.50$p.MI

list(b0.omit.10=b0.omit.10,b1.omit.10=b1.omit.10,
b2.omit.10=b2.omit.10,p.omit.10=p.omit.10,b0.omit.30=b0.omit.30,
b1.omit.30=b1.omit.30,b2.omit.30=b2.omit.30,p.omit.30=p.omit.30,
b0.omit.50=b0.omit.50,b1.omit.50=b1.omit.50,b2.omit.50=b2.omit.50,
p.omit.50=p.omit.50,b0.EM.10=b0.EM.10,b1.EM.10=b1.EM.10,
b2.EM.10=b2.EM.10,p.EM.10=p.EM.10,b0.EM.30=b0.EM.30,b1.EM.30=b1.EM.30,
b2.EM.30=b2.EM.30,p.EM.30=p.EM.30,b0.EM.50=b0.EM.50,b1.EM.50=b1.EM.50,
b2.EM.50=b2.EM.50,p.EM.50=p.EM.50,b0.HD.10=b0.HD.10,b1.HD.10=b1.HD.10,
b2.HD.10=b2.HD.10,p.HD.10=p.HD.10,b0.HD.30=b0.HD.30,b1.HD.30=b1.HD.30,
b2.HD.30=b2.HD.30,p.HD.30=p.HD.30,b0.HD.50=b0.HD.50,b1.HD.50=b1.HD.50,
b2.HD.50=b2.HD.50,p.HD.50=p.HD.50,b0.MI.10=b0.MI.10,b1.MI.10=b1.MI.10,
b2.MI.10=b2.MI.10,p.MI.10=p.MI.10,b0.MI.30=b0.MI.30,b1.MI.30=b1.MI.30,
b2.MI.30=b2.MI.30,p.MI.30=p.MI.30,b0.MI.50=b0.MI.50,b1.MI.50=b1.MI.50,
b2.MI.50=b2.MI.50,p.MI.50=p.MI.50)

}
```

Appendix C: The non-null mean and variance for $T^{*}_{(2)}$ and $T^{**}_{(2)}$

The density function of s_i under alternative hypothesis $H_1 : 0 < v < 1$ is given in (6.46) as the following:

$$f_1\left(s_i\right) = \exp\left(-s_i^{v}\right)\left\{v^2 s_i^{2v-1} - v\left(v-1\right)s_i^{v-1}\right\}.$$

Some expected values (Sarker, 2002) those are obtained from (6.46) are

$$E\left(s_i^{-1}\right) = \int_0^{\infty} s_i^{-1} f_1\left(s_i\right) ds_i = -\infty,$$

$$E\left(s_i^{-2}\right) = \int_0^{\infty} s_i^{-2} f_1\left(s_i\right) ds_i = \infty,$$

$$E\left(\ln s_i\right) = \int_0^{\infty} \ln s_i \, f_1\left(s_i\right) ds_i = 1 - \frac{\gamma}{v},$$

$$E\left(s_i \ln s_i\right) = \int_0^{\infty} s_i \ln s_i \, f_1\left(s_i\right) ds_i = \frac{1}{v^2}\left[\left\{2\Psi\left(\frac{1}{v}\right) + 3v\right\}\Gamma\left(\frac{1}{v}\right)\right],$$

$$E\left(\ln^2 s_i\right) = \int_0^{\infty} \ln^2 s_i \, f_1\left(s_i\right) ds_i = -\frac{1}{6v^2}\left[12v\gamma - \pi^2 - 6\gamma^2\right],$$

$$E\left(s_i \ln^2 s_i\right) = \frac{2}{v^3}\left\{\Psi\left(1, \frac{1}{v}\right) + v^2 + \Psi\left(\frac{1}{v}\right)^2 + 3\Psi\left(\frac{1}{v}\right)v\right\}\Gamma\left(\frac{1}{v}\right),$$

$$\begin{aligned}
E\left(s_i^2 \ln^2 s_i\right) = \frac{1}{2v^3}\Bigg[& 3\Psi\left(\frac{1}{2v}\left(2+v\right)\right)^2 + 3\Psi\left(\frac{1}{v}\right)^2 + 3\Psi\left(1, \frac{1}{2v}\left(2+v\right)\right) \\
& + 3\Psi\left(1, \frac{1}{v}\right) + 12\ln\left(2\right)^2 + 10v\Psi\left(\frac{1}{2v}\left(2+v\right)\right) \\
& + 10v\Psi\left(\frac{1}{v}\right) + 6\Psi\left(\frac{1}{v}\right)\Psi\left(\frac{1}{2v}\left(2+v\right)\right) \\
& + 12\ln\left(2\right)\Psi\left(\frac{1}{v}\right) + 12\ln\left(2\right)\Psi\left(\frac{1}{2v}\left(2+v\right)\right) \\
& + 20v\ln\left(2\right) + 4v^2\Bigg]\Gamma\left(\frac{2}{v}\right),
\end{aligned}$$

where γ is Euler's constant, $\Psi\left(\cdot\right) = \Gamma'\left(\cdot\right)/\Gamma\left(\cdot\right)$ is the di-gamma function, and $\Psi\left(n,.\right)$ is the n^{th} derivative of the di-gamma function.

From Sarker (2002), the non-null mean of $T_{(2)}^{*}$ may be written as

$$E_v\left(T_{(2)}^{*}\right)=n\left(4-\frac{1}{v^2}\left\{2\Psi\left(\frac{1}{v}\right)+3v\right\}\Gamma\left(\frac{1}{v}\right)-2\gamma\left(\frac{1}{v}\right)\right)=n\mu_v^{*},$$

whilst, the non-null variance of $T_{(2)}^{*}$ may be obtained as follows:

$$\begin{aligned}Var_v\left(T_{(2)}^{*}\right)=\frac{n}{6v^4}\Big\{&-24\Gamma\left(\frac{1}{v}\right)^2\Psi\left(\frac{1}{v}\right)^2-72\Gamma\left(\frac{1}{v}\right)^2\Psi\left(\frac{1}{v}\right)v-54\Gamma\left(\frac{1}{v}\right)^2v^2\\&-96\Gamma\left(\frac{1}{v}\right)v^2\Psi\left(\frac{1}{v}\right)+24\Gamma\left(\frac{1}{v}\right)v^3-48\Gamma\left(\frac{1}{v}\right)v\Psi\left(1,\frac{1}{v}\right)-48\Gamma\left(\frac{1}{v}\right)v\Psi\left(\frac{1}{v}\right)^2\\&-48\Gamma\left(\frac{1}{v}\right)v\gamma\Psi\left(\frac{1}{v}\right)-72\Gamma\left(\frac{1}{v}\right)v^2\gamma+12\Gamma\left(2\frac{1}{v}\right)v^3+9\Gamma\left(2\frac{1}{v}\right)v\Psi\left(\frac{1}{v}\right)^2\\&+36\Gamma\left(2\frac{1}{v}\right)v\ln\left(2\right)^2+9\Gamma\left(2\frac{1}{v}\right)v\Psi\left(1,\frac{1}{2}\frac{2+v}{v}\right)+9\Gamma\left(2\frac{1}{v}\right)v\Psi\left(\frac{1}{2}\frac{2+v}{v}\right)^2\\&+9\Gamma\left(2\frac{1}{v}\right)v\Psi\left(1,\frac{1}{v}\right)+30\Gamma\left(2\frac{1}{v}\right)v^2\Psi\left(\frac{1}{v}\right)+60\Gamma\left(2\frac{1}{v}\right)v^2\ln\left(2\right)\\&+30\Gamma\left(2\frac{1}{v}\right)v^2\Psi\left(\frac{1}{2}\frac{2+v}{v}\right)+36\Gamma\left(2\frac{1}{v}\right)v\ln\left(2\right)\Psi\left(\frac{1}{v}\right)\\&+36\Gamma\left(2\frac{1}{v}\right)v\ln\left(2\right)\Psi\left(\frac{1}{2}\frac{2+v}{v}\right)+18\Gamma\left(2\frac{1}{v}\right)v\Psi\left(\frac{1}{v}\right)\Psi\left(\frac{1}{2}\frac{2+v}{v}\right)\\&+4v^2\pi^2-24v^4\Big]\\&=n\sigma_v^{*2}.\end{aligned}$$

Meanwhile, the non-null mean and variance of $T_{(2)}^{**}$ are

$$E_v\left(T_{(2)}^{**}\right)=\sum_{i=1}^{n}E\left(\ln s_i\right)=n\left(1-\frac{\gamma}{v}\right)=n\mu_v^{**},$$

and

$$Var_v\left(T_{(2)}^{**}\right)=\sum_{i=1}^{n}E\left(\ln^2 s_i\right)-\sum_{i=1}^{n}E^2\left(\ln s_i\right)=n\left(\frac{\pi^2}{6v^2-1}\right)=n\sigma_v^{**2},$$

respectively.

Appendix D: R code for evaluating the convergence rate for uncensored case without nuisance parameters

```
nonu=function(n,shape,rate,cycle){

      a0=matrix(0,ncol=1)
      b0=matrix(0,ncol=1)
      c0=matrix(0,ncol=1)

      for (i in 1:cycle){
            y1=rgompertz(n,shape,rate)      # generate y1 & y2 by
            y2=rgompertz(n,shape,rate)      # calling library(flexsurv)


            #--- Find s ---#

            s=(exp(y1)-1)+(exp(y2)-1)       # compute s (shape=scale=1)


            #--- Calculate Tn ---#

            T2=sum(2+2*log(s)-s*log(s)-(1/s))
            T.2= sum(2+2*log(s)-s*log(s))
            T..2=sum(log(s))


            #--- Normalisation Sn ---#

            S2= T2/sqrt(1/2*n*log(n))
            S.2= (T.2-n)/sqrt(n*3.492929993)
            S..2=(T..2-(n*0.4227843351))/sqrt(n*0.6449340675)

            a0[i]=S2
            b0[i]=S.2
            c0[i]=S..2

            }

list(a0=a0,b0=b0,c0=c0)

}
```

Simulation program

```
Simunonu<-function(n,shape,rate,cycle,simu){

S2<-matrix(0,nrow=cycle,ncol=simu)
S.2<-matrix(0,nrow=cycle,ncol=simu)
S..2<-matrix(0,nrow=cycle,ncol=simu)

      for(j in 1:simu){
            nonus=nonu(n,shape,rate,cycle)
            S2[,j]<-nonus$a0
            S.2[,j]<-nonus$b0
            S..2[,j]<-nonus$c0
            }

      a1=quantile(S2[,1], c(.10, .05, 0.025,0.01))
      a2=quantile(S2[,2], c(.10, .05, 0.025,0.01))
      a3=quantile(S2[,3], c(.10, .05, 0.025,0.01))
      a4=quantile(S2[,4], c(.10, .05, 0.025,0.01))
      a5=quantile(S2[,5], c(.10, .05, 0.025,0.01))
      a6=quantile(S2[,6], c(.10, .05, 0.025,0.01))
      a7=quantile(S2[,7], c(.10, .05, 0.025,0.01))
      a8=quantile(S2[,8], c(.10, .05, 0.025,0.01))
      a9=quantile(S2[,9], c(.10, .05, 0.025,0.01))
      a10=quantile(S2[,10], c(.10, .05, 0.025,0.01))

      b1=quantile(S.2[,1], c(.10, .05, 0.025,0.01))
      b2=quantile(S.2[,2], c(.10, .05, 0.025,0.01))
      b3=quantile(S.2[,3], c(.10, .05, 0.025,0.01))
      b4=quantile(S.2[,4], c(.10, .05, 0.025,0.01))
      b5=quantile(S.2[,5], c(.10, .05, 0.025,0.01))
      b6=quantile(S.2[,6], c(.10, .05, 0.025,0.01))
      b7=quantile(S.2[,7], c(.10, .05, 0.025,0.01))
      b8=quantile(S.2[,8], c(.10, .05, 0.025,0.01))
      b9=quantile(S.2[,9], c(.10, .05, 0.025,0.01))
      b10=quantile(S.2[,10], c(.10, .05, 0.025,0.01))

      c1=quantile(S..2[,1], c(.10, .05, 0.025,0.01))
      c2=quantile(S..2[,2], c(.10, .05, 0.025,0.01))
      c3=quantile(S..2[,3], c(.10, .05, 0.025,0.01))
      c4=quantile(S..2[,4], c(.10, .05, 0.025,0.01))
      c5=quantile(S..2[,5], c(.10, .05, 0.025,0.01))
      c6=quantile(S..2[,6], c(.10, .05, 0.025,0.01))
      c7=quantile(S..2[,7], c(.10, .05, 0.025,0.01))
      c8=quantile(S..2[,8], c(.10, .05, 0.025,0.01))
      c9=quantile(S..2[,9], c(.10, .05, 0.025,0.01))
      c10=quantile(S..2[,10], c(.10, .05, 0.025,0.01))

      a=rbind(a1,a2,a3,a4,a5,a6,a7,a8,a9,a10)
      b=rbind(b1,b2,b3,b4,b5,b6,b7,b8,b9,b10)
      c=rbind(c1,c2,c3,c4,c5,c6,c7,c8,c9,c10)

      a.mean=colMeans(a)
      b.mean=colMeans(b)
      c.mean=colMeans(c)

      a.sd=apply(a,2,sd)
      b.sd=apply(b,2,sd)
      c.sd=apply(c,2,sd)

list(S2=S2,S.2=S.2,S..2=S..2,a=a,b=b,c=c,a.mean=a.mean,b.mean=b.mean,
c.mean=c.mean,a.sd=a.sd,b.sd=b.sd,c.sd=c.sd)
}
```