Approaches for Handling Time-Varying Covariates in Survival Models

Nwoko, Onyekachi Esther
NWKONY001

A minor dissertation submitted to the
Department of Statistical Sciences
University of Cape Town, South Africa

In partial fulfillment of the requirements of the degree of
Master of Science

Under the supervision of
Associate Professor Francesca Little

April 2, 2019.

Acknowledgements

The execution and completion of this research is a product of synergized contributions from numerous partners in the course of my academic pursuit. My admission into this programme was made possible through the University of Cape Town (UCT)/African Institute for Mathematical Sciences (AIMS) scholarship and the enabling platform provided by African Institute for Mathematical Sciences to encourage Africans like myself. I feel honoured, and remain grateful to these bodies for the privilege to have enjoyed such exposure.

I also appreciate the National Research Fund (NRF) Innovation Scholarship as well as the UCT International and Refugee Scholarship programmes. Their contribution was very instrumental to the completion of my programme here.

I will always remain grateful for the opportunity to interact and learn from very savvy intellectuals including my supervisor, Associate Professor Little, who made available, herself and all necessary intellectual resources to ensure I was comfortable enough to produce a dissertation. Senior colleagues at the department running their doctoral programme were similarly of immeasurable assistance in the course of my study. I appreciate all your contributions.

Some extracurricular activities kept my mind stable in the course of the Study. Top of them was my spiritual engagements with the Deeper Life Campus Fellowship and her loving members. I thank Associate Professor Falowo and his family for going the length to ensure my stay was well spent. I pray that your investment in me would yield results that would please you and God.

To my very loving parents and siblings who have never relented to dish out their support and prayers for me and through this programme, I deeply appreciate your love and look forward to stronger bonds.

To my ever supportive and cheering uncles, Mr Chinedum Nwoko and Pastor Boniface Azih, as well as their families, I am honoured and humbled to be a beneficiary of your consistent kindness and show of love. Thank you for your generosity and mentorship. I believe they made me a better person.

God bless you.

Nwoko, Onyekachi Esther
April 2, 2019.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Longitudinal studies are common in medical research particularly in follow-up studies where measurements on the same response variables of the subjects (animals, humans, plants, etc) are obtained at specific observation times (Tsiatis and Davidian, 2004). Longitudinal studies usually generates two kinds of outcome.

Firstly, longitudinal studies generate repeated measurements of the same variables at different time points. This results in each subject having a response profile. For example, in a cohort study of subjects infected with HIV, the CD4 count and viral load of each subject recruited into the study were collected repeatedly at each visit day. Since the measurements are taken from the same subjects, within subject correlation is inevitable. In order to obtain valid results and make valid inferences from the analysis, special statistical methods that take within subject correlation into consideration will be required. Repeated measurements are often analysed using mixed effect models. They model within subject correlation through the inclusion of subject-specific random effects.

Secondly, another type of outcome measured in longitudinal studies is time until the occurrence of a specific event. In this research, our specific event of interest is death. Time until the occurrence of a specific event is typically analysed using survival analysis methods which includes the estimation of the survival curve and models for the effect of covariates on the relative hazard of the event of interest. Standard methodologies used in performing survival analysis focus on baseline covariates whose association with the relative hazard does not change over time.

Figure 1.1 summarizes different methodological approaches for the two types of responses. This dissertation will in particular focus on the analysis of time-to-event responses.

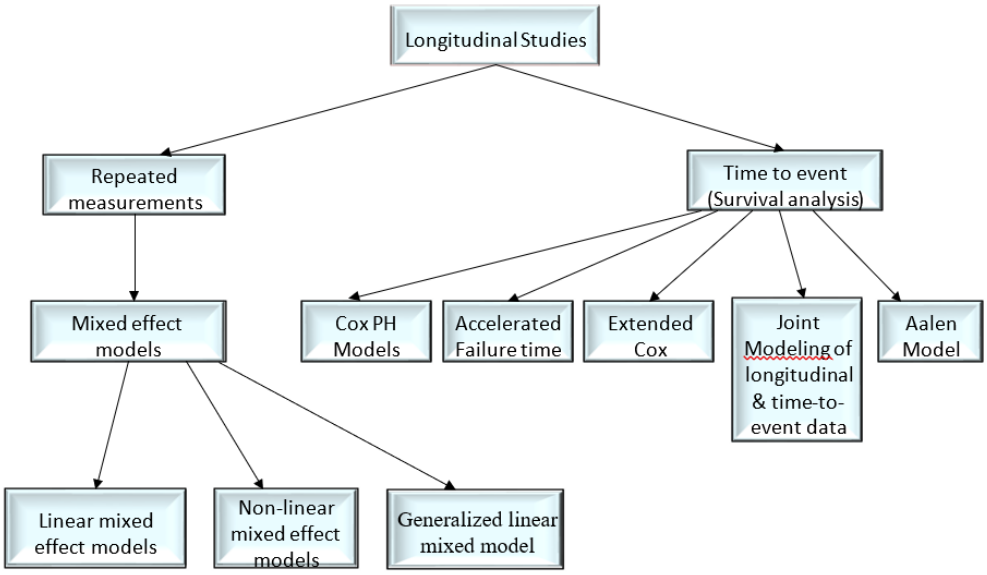


Figure 1.1: Illustration of the two response outcomes measured in longitudinal studies and different methodologies used for analysing them.

In particular, the focus of this dissertation is to (1) look at methods available to handle the association of baseline covariates with the relative hazard of the event of interest where this association vary with time, (2) model the effect of time-varying/ time-updated covariates. Therefore, the dissertation will discuss (1) the standard Cox proportional hazard model, (2) how to model time-varying effects of baseline covariates in the Cox model, (3) the extended Cox model for time-varying covariates, (4) the joint model for incorporating a time-varying effect in the Cox model and (5) the Aalen and semi-parametric Aalen model used to estimate time-varying effects of covariates on survival. These methods will be illustrated through the analysis of data from a cohort of HIV infected subjects on HAART. The clinical objective is to investigate the impact of treatment and CD4 on time to death.

The structure of this dissertation as shown in Figure 1.2 is as follows. In Chapter 2, I will describe the dataset, its features and variables of interest that will be used in this dissertation as well as perform some exploratory analysis. The concept of survival models and analysis of time to event data will be discussed in details in Chapter 3 while various methodologies for handling time-varying covariates effects will be presented in Chapter 4. Chapter 5 presents results and interpretation of results obtained after using each of the methodologies discussed in Chapter 4 on the dataset. By way of concluding, we will compare the results obtained from the different methodologies

used in analysis the dataset and suggestions for future research will be made in Chapter 6.

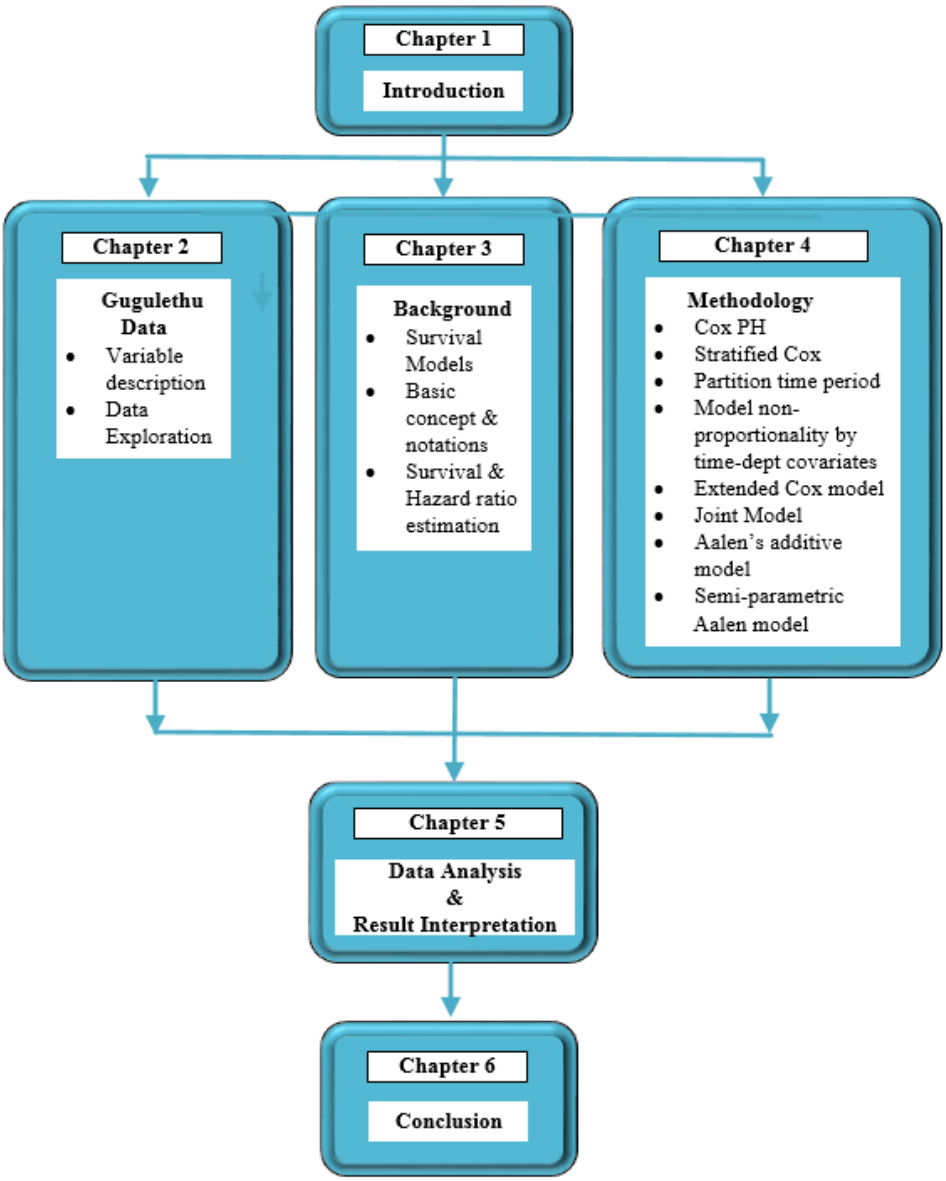


Figure 1.2: Dissertation Outline.

Chapter 2

Gugulethu Data

2.1 Introduction

This chapter gives brief description on how the data used for this research were collected, defines the variables measured and carries out some exploratory analysis on the variables of interest. The data collection procedure was approved by the University of Cape Town Health Sciences ethics committee.

In 2002, an Anti-Retroviral Treatment (ART) service was established in the Cape Town township of Gugulethu. Data from all patients who initiated ART within the programme between September 2002 and June 2007 were included. HIV/AIDS patients who were eligible to receive ART service were referred to the service from various primary health care clinics in the community. The eligibility criteria was based on the National ART guidelines to provide ART to those who have been diagnosed with AIDS or those whose CD4 count was less than 200 cells/μl. There is a median time of approximately 1 month between enrolment and the ART initiation by the patients. The first-line treatment comprises of three antiretroviral medications (Stavudine, Lamivudine and a reverse transcriptase inhibitor) together with prophylaxis and infection treatment before and during the ART (Lawn et al., 2009).

Blood CD4 cell count and plasma viral load measurements were obtained at baseline on entry into the study and thereafter 4 monthly, though exact visit intervals varied. Other variables such as the patient's age and gender were obtained. If patients did not visit the clinic for consultation or to collect their prescription for 12 weeks, they were considered as being lost to follow-up.

2.2 Variable Description

Variables of interest that will be used for the data analysis are described below.

1. PIDNo: Are unique identity numbers used in identifying the subjects that were involved in the study.
2. Age: represents the baseline age of the subjects at registration into the study in years.
3. Gender: represents the gender of the subject with 1 representing males and 0 representing females.
4. Stage: denotes the World Health Organisation (WHO) Acquired Immune Deficiency Syndrome (AIDS) stage with values ranging from 1 to 4. Stage 1 represents least severe and stage 4 represents full-blown AIDS. The WHO guidelines define Stage 4 as having CD4 cell count of less than 200 per micro-litre for adults and CD4 percentage of less than 20% for children less than 5 years old, (WHO et al., 2005)
5. Treatment: represents the treatment status of a subject. 1 represents the patient is on antiretroviral treatment and 0 that the patient is not.
6. CD4: represents the blood CD4 cell count values. They are usually measured in cells per micro-litre of blood. They are used in accessing how well one's immune system is functioning or compromised. The higher the CD4 cell count value of patients, the better.
7. Death: in this study, death is our event of interest. A patient that died was coded as 1 and coded 0 if otherwise.

2.3 Data Exploration

This section discusses the distributions and summary statistic for each of the variables in the data set. The complete data set contained 9152 subjects with 818 patients that experienced the event of interest. The percentage frequencies for the baseline categorical variables and the outcome in the data set are reported in Table 2.1.

Table 2.1: Frequency table for categorical variables in the Gugulethu data.

Variables	Category	Freq	Percent
Gender	Male (0)	2820	30.81
	Female (1)	6332	69.19
Stage	Stage1	3054	33.47
	Stage2	1479	16.21
	Stage3	3369	36.92
	Stage4	1250	13.70
Death	Yes (1)	818	8.94
	No (0)	8334	91.06

Table 2.1 presents a frequency table which provides some summary statistics for the categorical variables in the Gugulethu data. Out of 9,152 patients involved in the study, a total of 818 experienced the event of interest which is death. Of the 9,152 patients, 3054 of them are `Stage1` HIV patients, 1479 are `Stage2` HIV patients, 3369 are `Stage3` AIDS patients and 1250 are `Stage4` AIDS patients. The variable `Gender` in Table 2.1 shows that 69.19% of the subjects involved in the study are females.

Histograms were generated for the continuous variables, Age and CD4 and are shown in the figures below.

The distribution of `Age` is approximately normal with a mean score of 32.98, minimum and maximum values of 1.00 and 85.00 respectively.

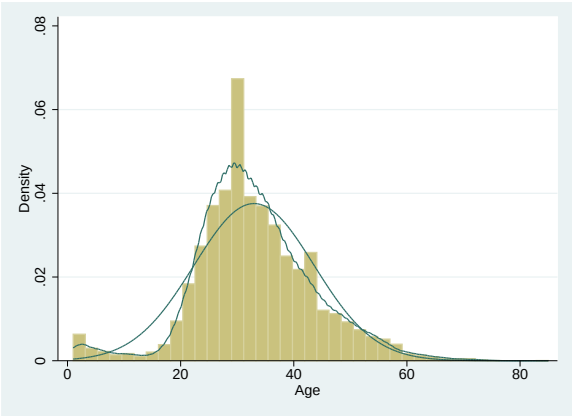


Figure 2.1: Histogram showing the distribution of Age.

Figure 2.2 below shows the distribution of CD4. CD4 has a right (positively) skewed distribution. Log and square root transformations succeed in bringing in the tail as shown in Figures 2.3 and 2.4 respectively. In this research, the $\log_{10}$(CD4) which will be represented as lCD4 will be used though it still shows some right skewness.

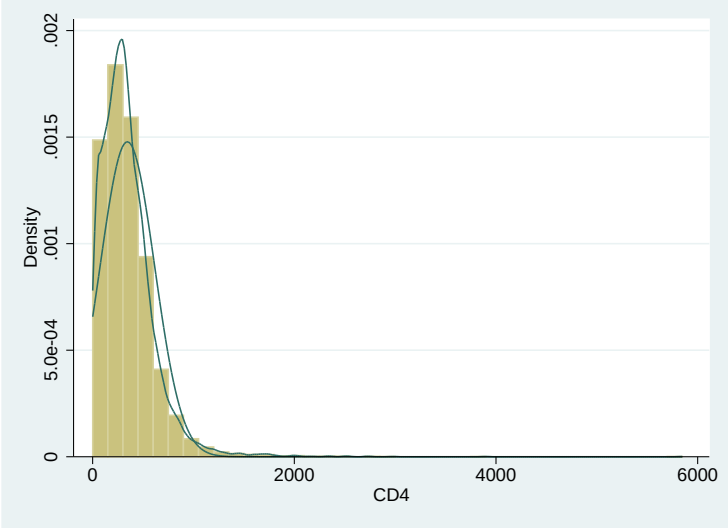


Figure 2.2: Histogram showing the distribution of CD4

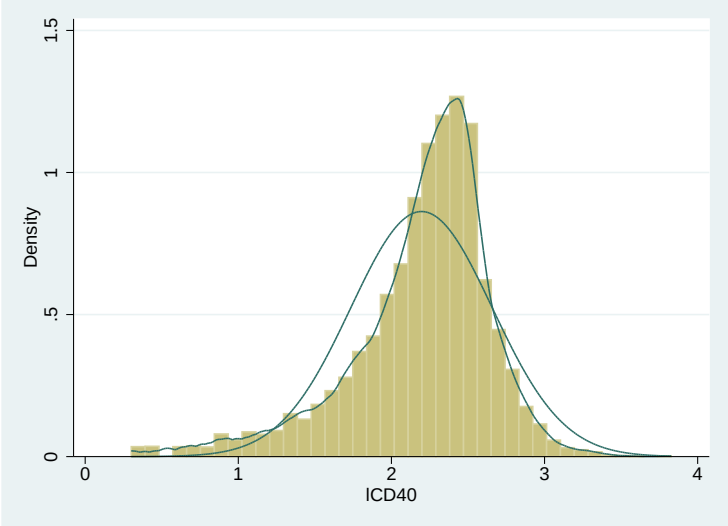


Figure 2.3: Histogram showing the distribution of lCD4

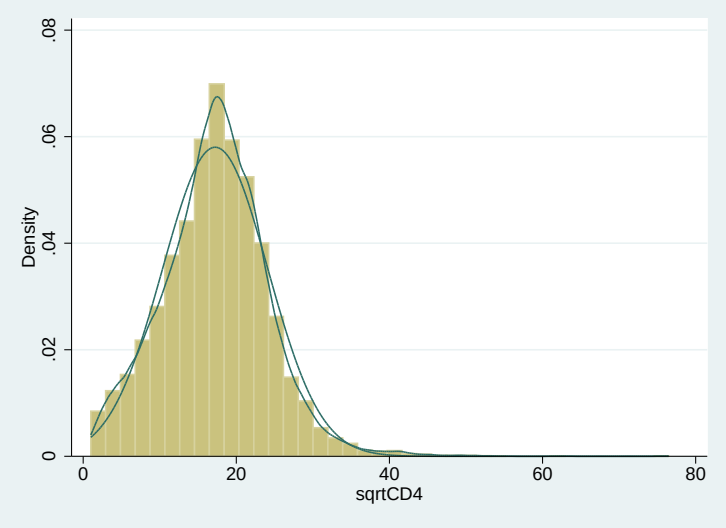


Figure 2.4: Histogram showing the distribution of square root of CD4

The Kaplan-Meier survival curve is used in estimating the survival time of individuals as well comparing the survival experience of two or more groups. Hence, Figures 2.5 and 2.6, show the survival function for time to treatment and the survival function of the whole HIV/AIDS cohort in the study respectively. Details of the Kaplan-Meier estimator can be found in section 3.5.1.

In Figure 2.5, at time $t = 0$ months, no patient was on treatment but at time $t = 68$ months, all the patients in the cohort under study had been placed on treatment. The median time to treatment is about 27 months.

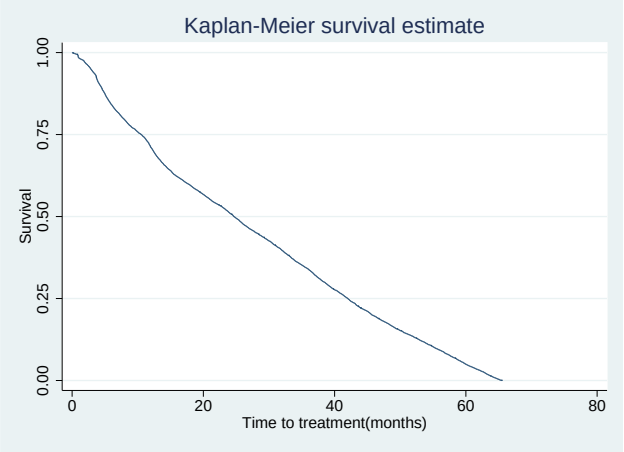


Figure 2.5: Kaplan- Meier Curve for time to treatment.

Figure 2.6 below is the Kaplan- Meier Curve showing the survival function of the HIV/AIDS patients. There is a sharp decline corresponding to increasing death rate at the start of the programme. This effect slows down but shows an increase again after 60 months. At 60 months, the survival rate is above 85% showing the success of HAART.

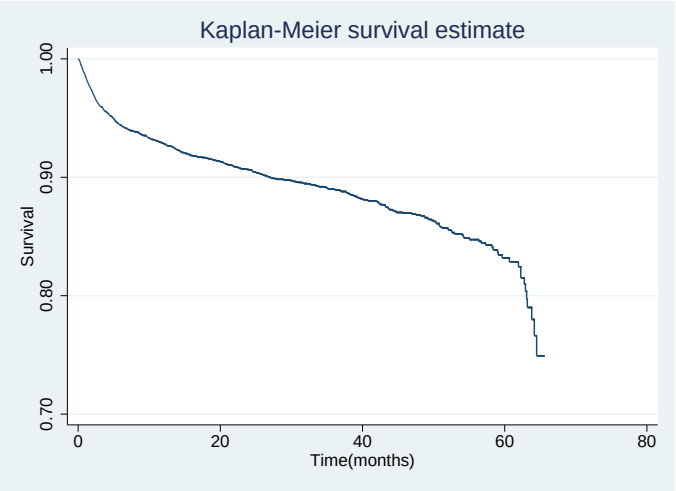


Figure 2.6: Kaplan- Meier Curve showing the survival function of the HIV/AIDS patients.

Figure 2.7 below, shows the survival plots for 12 strata representing the combination of three `lCD40` categories namely $< \log(200)$, $< \log(200 - 499)$ and $\geq \log(500)$ and four `Age` categories: `0-5`, `6-19`, `20-39` and `40-85`. Strata 1 are individuals in the 0-5 years age category and have a `lCD40` count of less than $\log(200)$. Strata 2 are individuals in the 0-5 years age category and have a `lCD40` count between $\log(200 - 499)$. Strata 3 are individuals in the 0-5 years age category and have a `lCD40` count greater than or equal to $\log(500)$.

Strata 4 are individuals in the `6-19` years age category and have a `lCD40` count of less than $\log(200)$. Strata 5 are individuals in the `6-19` years age category and have a `lCD40` count between $\log(200 - 499)$. Strata 6 are individuals in the `6-19` years age category and have a `lCD40` count greater than or equal to $\log(500)$.

Strata 7 are individuals in the `20-39` years age category and have a `lCD40` count of less than $\log(200)$. Strata 8 are individuals in the `20-39` years age category and have a `lCD40` count between $\log(200 - 499)$. Strata 9 are individuals in the `20-39` years age category and have a `lCD40` count greater than or equal to $\log(500)$.

Strata 10 are individuals in the `40-85` years age category and have a `lCD40` count of less than log(200). Strata 11 are individuals in the `40-85` years age category and have a `lCD40` count between log(200 – 499). Strata 12 are individuals in the `40-85` years age category and have a `lCD40` count greater than or equal to log(500).

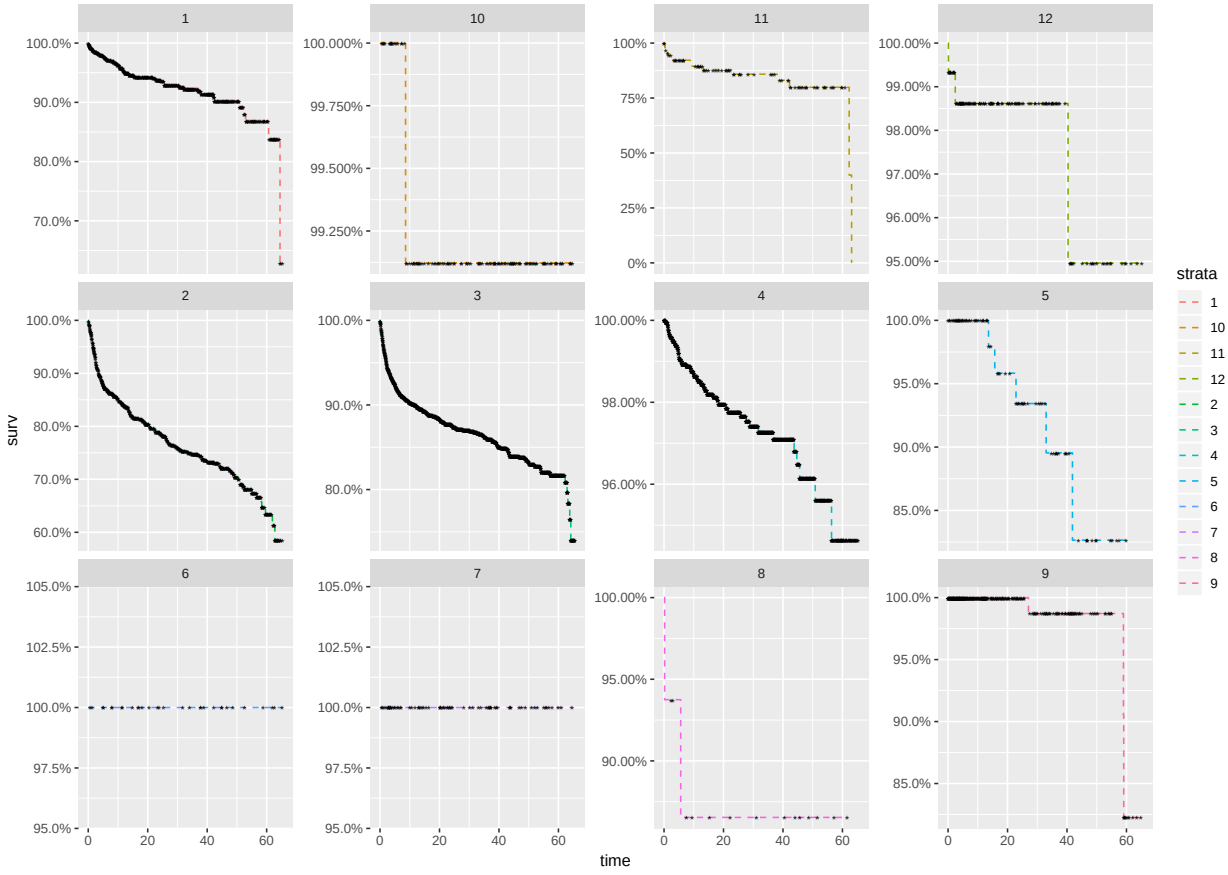


Figure 2.7: Kaplan-Meier curves showing the number of events in each stratum

Chapter 3

Background

3.1 Introduction

This chapter gives a brief background about the concept and features of survival models. Basic notation used in survival analysis will be introduced. In addition, different methods for estimating the survival function such as the Kaplan-Meier estimator and Nelson-Aalen estimator will be discussed. Regression models for modelling the effect of covariates on the hazard of an event will be presented. The concepts of time-varying coefficients and covariates will be introduced. Examples of each will be given and the differences between both will be discussed. This chapter will give some intuition into the main ideas underlying the methodology.

3.2 Survival Models

Survival analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is the *time until an event occurs* (Kleinbaum and Klein, 2010). Time refers to the period from the beginning of follow-up of an individual until an event occurs or until censoring. These could be in years, months, weeks, or days. On the other hand, event means any designated experience of interest. The event of interest could be death, disease incidence, relapse from remission or recovery (Kleinbaum and Klein, 2010). The Gugulethu HIV/AIDS study followed HIV/AIDS patients over about 65 months until death or current end of follow up. The event of interest is death, and the outcome is time in months until a person dies.

3.3 Features of Survival Models

This section presents the features of survival models in terms of censoring and the structure of survival data.

3.3.1 Censoring

An important feature to be considered in survival analysis is censoring. An observation is censored when the event time for all the subjects in the study is not fully known. Standard statistical tools such as sample mean, standard deviation, t-test, linear and logistic regression cannot be used as they are based on the assumption that the information used is complete (Rizopoulos, 2012).

There are basically three forms of censored observations: `right`, `left` and `interval` censored observations. Right censoring which is the most common form occurs when the complete survival time has been cut off at the right side. That is, the event of interest was not observed due to study period expiration or withdrawal of individual or lost to follow up. Interval censoring is a special form of censoring as it incorporates both right-censoring and left-censoring. An observation is said to be interval censored if the exact time the event of interest occurred is unknown but it is known to have occurred within a particular interval. An observation is left censored if the event of interest has occurred before enrolment.

Conditional on the value of any covariates in a survival model and on an individual's survival to a particular time, censoring must be independent of the future value of the hazard for the individual. Censoring that meets this requirement is non-informative. A common instance of non-informative censoring occurs when a study terminates at a predetermined date. If this condition is not met, the estimates of the survival distribution can be seriously biased. For example, if individuals tend to drop out of a clinical trial shortly before they die, and therefore their deaths go unobserved, survival time will be over-estimated.

Figure 3.1 below shows the study time of ten subjects. "•" represents the start of follow up. The subjects were observed at times 0, 2, 4, 6, 8, 10 and 12 only. Subjects 1, 7 and 8 died (D) during the course of the study, subjects 2 and 9 were lost to follow-up (L), subjects 3 and 6 were withdrawn (W) from the study due to the negative effect the treatment had on them.

Finally, subjects 4, 5 and 10 were still alive (A) at the end of the study. For a study where the event of interest is death, subjects 4, 5 and 10 are said to be right censored because the event of interest was not observed during the course of the study due to administrative censoring in that the end of follow-up was reached. In addition, subjects 3, 6 and 9 are also right censored because the event of interest was not observed over the study period due to withdrawal and loss to follow-up (LTFU). On the other hand, the subjects 1, 7 and 8 could be regarded as a interval censored observation because the exact time the event of interest (death) occurred isn't known since their death would only be recorded at the scheduled visit following the actual death date. However, we are aware it occurred within a particular interval.

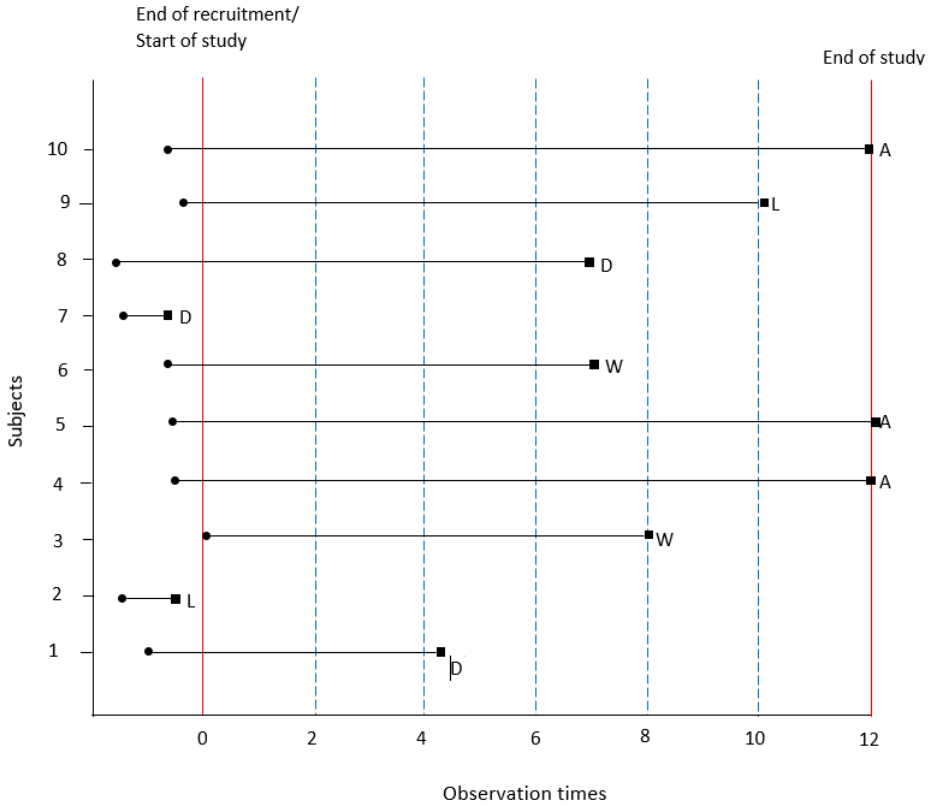


Figure 3.1: Visual illustration of the different types of censoring.

3.3.2 Structure of Longitudinal / Survival Data

Longitudinal data can be constructed or presented in two forms: (1) the wide form which is basically when repeated responses from each subject is recorded in a single row with each successive response recorded in a separate column and (2) the long form, where repeated measurements at each time point for each subject is recorded in a separate row. Variables whose value do not change over time remain the same in all the rows for a given subject. The choice of which form to present your data in is dependent on the type of analysis to be carried out. If one is interested in performing survival analysis for example, the required data format is the long form since multiple time points per subject will allow time-varying characteristics to be captured. In this section, we will describe the structure of survival data. The key components of survival data are:

1. a unique subject identifier
2. time variable: which could be date, time. This variable stores the time the event of interest was observed. It could be time to death, remission, exacerbation etc.
3. Failure or event variable: this is an indicator variable with 1 if the event of interest occurred and 0 if censored. This is applicable if we have only one event of interest. In the presence of competing risk, the indicator will have as many categories as the competing event categories.

The data structure when fitting the standard Cox proportional hazard (PH) model is such that only the baseline observations on the covariates for each subject is used. Hence, just one observation per subject including baseline covariate values, total follow up time and an event indicator is necessary. However, for models like the extended Cox used for time-dependent covariates, we create several time intervals for each subject using the (start, stop] notation. This means that all the changes that occurred in a covariate is recorded at the stop time. Here, the structure of the data is the long format with each subject's observation recorded in multiple rows for a specific time interval.

Figure 3.2 presents a subset of the HIV/AIDS dataset generated as a result of repeated observations taken on some explanatory variables of interest that will be used in this research. Subject 2 had a follow-up time from 0 to 26.46 months. The variable `log CD4, (lCD4)` is a time varying covariate and

was measured each time the subject visited the clinic. The start and stop variables denotes the time interval limits when the CD4 count was recorded. For subject 2, the `log CD4` equals 2.11 at baseline, 2.29 at 4.13 months, and 2.54 at 7.80 months and so on. The variable `DEATH` is an event indicator and equals 1 if the subject died at the end of the corresponding time interval. Variables like Age and Gender that do not change over time remains the same for each subject.

	PIDNo	Cumtime	DEATH	Age	Gender	Stage2	Stage3	Stage4	lCD4	Tx	START	STOP
1	1	0.3934426	1	42	Male	0	0	1	2.356026	0	0.000000	0.3934426
2	2	4.1311475	0	41	Female	1	0	0	2.107210	0	0.000000	4.1311475
3	2	7.8032787	0	41	Female	1	0	0	2.285557	1	4.131148	7.8032787
4	2	11.4754098	0	41	Female	1	0	0	2.536558	1	7.803279	11.4754098
5	2	11.5409836	0	41	Female	1	0	0	2.324282	1	11.475410	11.5409836
6	2	17.5081967	0	41	Female	1	0	0	2.324282	1	11.540984	17.5081967
7	2	20.0327869	0	41	Female	1	0	0	2.442480	1	17.508197	20.0327869
8	2	20.9836066	0	41	Female	1	0	0	2.389166	1	20.032787	20.9836066
9	2	23.7049180	0	41	Female	1	0	0	2.383815	1	20.983607	23.7049180
10	2	26.4590164	0	41	Female	1	0	0	2.459392	1	23.704918	26.4590164
11	3	1.1475410	0	30	Female	0	0	1	2.079181	0	0.000000	1.1475410
12	3	39.8688525	0	30	Female	0	0	1	2.096910	1	1.147541	39.8688525
13	3	41.9344262	0	30	Female	0	0	1	2.800029	1	39.868852	41.9344262
14	3	42.3934426	0	30	Female	0	0	1	2.800029	1	41.934426	42.3934426
15	3	44.2295082	0	30	Female	0	0	1	2.800029	1	42.393443	44.2295082
16	3	45.1475410	0	30	Female	0	0	1	2.779596	1	44.229508	45.1475410
17	3	47.9016393	0	30	Female	0	0	1	2.779596	1	45.147541	47.9016393
18	3	51.6393443	0	30	Female	0	0	1	2.831870	1	47.901639	51.6393443
19	3	53.4098361	0	30	Female	0	0	1	2.831230	1	51.639344	53.4098361
20	3	54.3278689	0	30	Female	0	0	1	2.831230	1	53.409836	54.3278689
21	3	55.2459016	0	30	Female	0	0	1	2.831230	1	54.327869	55.2459016
22	3	58.9180328	0	30	Female	0	0	1	2.984077	1	55.245902	58.9180328

Figure 3.2: Subset of the HIV/AIDS survival dataset generated as a result of repeated observations taken on some explanatory variables of interest.

3.4 Basic Concepts and Notation

This section presents a summary of the concepts and notations of survival models discussed in (Kleinbaum and Klein, 2010) and (Collett, 2015). Let T_i^* denote the true survival time for the ith individual, T_i be a non-negative random variable which denotes an individual's observed survival time, defined as $min(C_i, T_i^*)$. C_i the censoring time, $\delta_i = I(T_i \leq C_i)$ the event indicator

for the ith subject and t is a specific value of T. Let $y_i(t)$ be the observed value of a time-dependent covariate at time point t.

3.4.1 Probability Density Function f(t)

The probability density function, $f(t)$, the probability of death at time t, is given by

$$\begin{aligned} f(t) &= P(T = t). \\ &= \lim_{\Delta t \to 0} \frac{Pr(t \leq T \leq t + \Delta t)}{\Delta t}. \end{aligned} \tag{3.1}$$

3.4.2 Cumulative Distribution Function F(t)

The cumulative distribution function, $F(t)$, the probability that the event of interest has occurred on or before time t, is given by

$$\begin{aligned} F(t) &= Pr(T \leq t) \\ &= \int_0^t f(u)du. \end{aligned} \tag{3.2}$$

3.4.3 Survivor Function S(t)

The survivor function,$S(t)$, is the compliment of the cumulative distribution function. It gives the probability of an individual surviving beyond the specified time t. This is a decreasing function with $S(t) = 1$ when $t = 0$ to $S(t) = 0$ when $t = \infty$,

$$\begin{aligned} S(t) &= P(T > t) = 1 - F(t) \\ &= 1 - P(T \leq t) \\ &= 1 - \int_0^t f(u)du. \end{aligned} \tag{3.3}$$

3.4.4 Hazard Function h(t)

The hazard function, $h(t)$, gives the instantaneous risk at time t for the event to occur, given that the individual has survived up to time t,

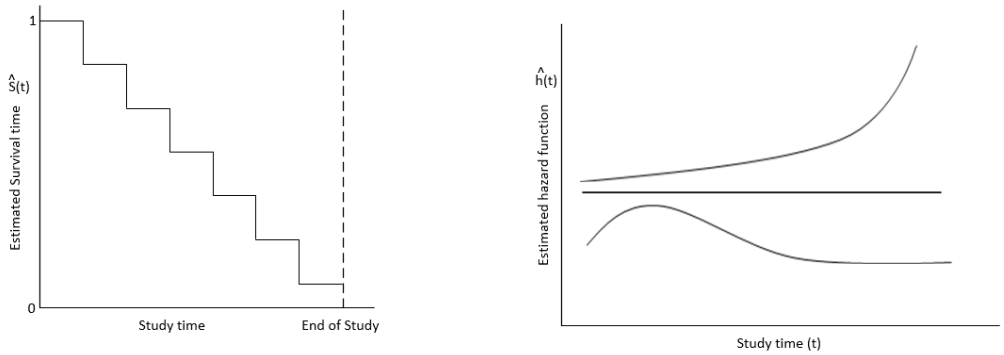
$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \tag{3.4}$$

The hazard function is related to the survivor function in the following way:

$$S(t) = \exp\Big[- \int_0^t h(u)du. \Big] \tag{3.5}$$

$$h(t) = -\Big[\frac{dS(t)}{dt}\Big]. \tag{3.6}$$

Equation (3.5) shows that the survival function equals the exponent of the negative of the integrated or cumulative hazard function whereas equation (3.6) specifies the hazard function as the negative derivative of the survival function.



(a) The survival function (b) The hazard function

Figure 3.3: Plots showing the survival and hazard functions.

As can be seen from Figure 3.3(a), the survival function is a decreasing function. It ranges from 1 to 0. At time $t = 0$, the probability of a subject surviving a particular condition is 1 (S(0) = 1), that is all the patients are still alive. As time progresses, the probability of survival reduces and tends towards 0 (S(∞) = 0), implying that all the patients have died. The survival function is a step function because events are observed at discrete times.

Unlike the survival function that has an upper bound of 1, the hazard function does not have an upper bound. Figure 3.3(b) presents three different

hazard functions. The hazard function can commence at any point and move in any direction over time. When the hazard function is a straight line, it is said to have a constant hazard. This implies that at any given time, the value of the hazard function remains the same. Hence, a survival model is said to have an exponential distribution if the hazard function is constant.

3.5 Survival and Hazard Ratio Estimation

In this section, I will discuss both parametric and non-parametric techniques for estimating the survival and cumulative hazard function.

3.5.1 Kaplan Meier (K-M) Estimator

The K-M estimator was said to have been proposed by Böhmer (1912) but was not used until in 1958 when a paper was released by Kaplan and Meier on the subject, (Andersen et al., 2012). The K-M estimator (Kleinbaum and Klein, 2010) also known as the product limit estimator (the survival probability is limited to product terms up to the survival time being specified) is a non-parametric (that is no underlying distribution is assumed) technique, popularly used in the estimation of survival time. This technique takes into account censored observations, especially the right censored observations. This is achieved by assuming that subjects censored at time t survive longer than the deaths at time t and no adjustment for the number at risk is required. At the next time point, the censored observations are removed from the number at risk. The K-M estimator is computed as the product of survival estimates at successive time points. Let $t_1 < t_2 < \cdots t_j$ be the time the event of interest is observed, d_j be the number of individuals that have experienced the event at t_j and r_j is the number of individuals who are yet to experience the event. It takes the form below:

$$\hat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{r_j}\right) \tag{3.7}$$

where $\frac{d_j}{r_j}$ estimates the proportion of the d_j that experience the event at time t_j of the r_j exposed or under observation at time t_j.

The variance is estimated using the Greenwood's formula

$$\hat{\sigma}^2(t) = \hat{S}(t)^2 \sum_{t_j \leq t} \frac{d_j}{r_j(r_j - d_j)}. \tag{3.8}$$

The K-M survival curve can be used to illustrate the survival experience of two or more groups. However, a statistical test is required to ascertain if the difference between the groups are statistically significant or otherwise. There are many options such as the log-rank test, the generalised Wilcoxon test, the Peto's generalised Wilcoxon test, the Tarone-Ware test and the Flemington-Harrington test to access this difference. The choice of which test to use depends on whether all failure times are treated with equal or differential importance. The "log-rank test" is the most commonly used method. It tests the null hypothesis that two or more survival functions are the same but does not give an estimate of the difference between groups. This approach gives equal weight to all deaths whereas other variants (generalised Wilcoxon, Peto's generalised Wilcoxon, Tarone-Ware and Flemington-Harrington test) apply different weights at either the early or later time.

The log rank test statistic used to compare the survival function of two groups is formed by using the squared difference between the summed observed and expected scores for one of the groups divided by the variance of the difference between the summed observed and expected scores. Under the null hypothesis that two survival functions or curves are the same, the log rank test statistic has an approximately chi-square distribution with one degree of freedom. The log rank test for two groups takes the form below:

$$Z = \frac{(O_i - E_i)^2}{\text{var}\,(O_i - E_i)}. \tag{3.9}$$

where,

$$O_i - E_i = \sum_{f=1}(m_{if} - e_{if}).$$

$$e_{1f} = \left(\frac{n_{1f}}{n_{1f} + n_{2f}}\right) \times (m_{1f} + m_{2f})$$

$$e_{2f} = \left(\frac{n_{2f}}{n_{1f} + n_{2f}}\right) \times (m_{1f} + m_{2f})$$

$$\text{var}\,(O_i - E_i) = \sum_j \frac{n_{1f}n_{2f}(m_{1f} + m_{2f})(n_{1f} + n_{2f} - m_{1f} - m_{2f})}{(n_{1f} + n_{2f})^2(n_{1f} + n_{2f} - 1)}.$$

where,
Z denotes the log rank test statistics, $i = 1,\ 2$, O_i & E_i are the summed

observed and expected scores for group i. e_{if} is the expected number at time f for the ith group. It is the proportion of individuals at risk multiplied by the number of failures in both groups. m_{if} and n_{if} are the number of failures and risks at time f for the ith group respectively.

3.5.1.1 Likelihood Estimation

According to Dobson and Barnett (2008), let y_j be the response observation recorded for the j^{th} subject, δ_j is event indicator with $\delta_j = 0$ for censored event times and $\delta_j = 1$ if the event of interest was observed. Let $y_1, \cdots, \cdots, y_r$ and $y_{r+1}, \cdots, y_n$ be the observations that are uncensored and censored respectively and let x_j denote the vector of explanatory variables.

The expression below is the likelihood function for the uncensored variables

$$\prod_{j=1}^{r} f(y_j).$$

The expression below is the likelihood function for the censored variables

$$\prod_{j=r+1}^{n} S(y_j).$$

The complete likelihood is

$$L = \prod_{j=1}^{n} f(y_j)^{\delta_j} S(y_i)^{1-\delta_j}. \tag{3.10}$$

The log-likelihood function is expressed as

$$\begin{aligned} l &= \sum_{j=1}^{n} \big[\delta_j \log f(y_j) + (1-\delta_j) \log S(y_j)\big]. \\ &= \sum_{j=1}^{n} \big[\delta_j \log f(y_j) - \delta_j \log S(y_j) + \log S(y_j)\big]. \end{aligned}$$

Recall that $\frac{f(y_j)}{S(y_j)} = h(y_j)$. Hence,

$$l = \sum_{j=1}^{n} \big[\delta_j \log h(y_j) + \log S(y_j)\big]. \tag{3.11}$$

Parameter estimation is done using numerical methods such as the Newton-Raphson approach to maximise the log-likelihood function.

3.5.2 Nelson-Aalen Estimator

Aalen (1978, 1975) introduced the Nelson-Aalen estimator which is the generalization of the empirical cumulative intensity estimator proposed by Nelson (1969, 1972) and Altshuler (1970). It is a non-parametric estimator used in estimating the cumulative hazard function for censored survival data (Andersen et al., 2012). Let $t_1 < t_2 < \cdots t_j$ be the time the event of interest is observed, d_j be the number of individuals that have experienced the event at t_j and r_j is the number of individuals who are yet to experience the event. The Nelson-Aalen estimator is given by:

$$\hat{A}(t) = \sum_{t_j \leq t} \frac{d_j}{r_j}. \tag{3.12}$$

The variance of $\hat{A}$ is estimated by

$$\hat{\sigma}^2(t) = \sum_{t_j \leq t} \frac{(r_j - d_j)d_j}{(r_j - 1)r_j^2}. \tag{3.13}$$

Thus, whereas the K-M estimator is multiplicative, the Aalen estimator is additive. It can be shown that $S(t) = \exp[-A(t)]$ or equivalently, $A(t) = -\ln S(t)$.

3.6 Regression Models

Whereas the above non-parametric estimators generate estimates of survival and hazard functions, regression models are needed to measure the association between several categorical and continuous covariates and the hazard or survival functions. The logistic or multiple regression models are not appropriate in this situation as they do not account for censored observations.

There are two broad categories of regression models used in survival analysis. They are the semi-parametric and parametric regression models. A model is said to be semi-parametric if no distributional assumption is made for the baseline hazard (this is the non-parametric bit) but the effect of predictors assumes a parametric form (this is the parametric bit). The Cox proportional hazard (PH) regression model is a common example of a semi-parametric model. On the other hand, parametric models are models in which the hazard function has an underlying distribution. The class of parametric models can

be divided into the proportional hazard (PH) models and accelerated failure-time (AFT) models.

In the PH regression model, the effect of covariates is obtained on the hazard function and covariates act multiplicatively on the hazard. Common parametric PH models include the exponential, Weibull or Gompertz models. The general form of a semi-parametric PH models is given below:

$$h_i(t) = h_0(t)\exp(\beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}) \tag{3.14}$$

where,
$h_0(t)$ is an arbitrary baseline hazard rate, reflecting the hazard rate when all X's = 0. $X_{i1}, \cdots, X_{ip}$ are the p-covariates or risk factors of interest and $\beta_1, \cdots, \beta_p$ are regression coefficients estimated, by maximizing the partial likelihood.

The semi-parametric Cox PH model is obtained when no underlying distribution is assumed for the baseline hazard function. However, when the baseline hazard function is assumed to follow a distribution, the parametric PH model is obtained.

We will focus on the non-parametric Cox proportional hazard model in this dissertation. It will be discussed in details in section 4.2.

Chapter 4

Methodology

4.1 Introduction

This chapter presents a review of the Cox proportional hazards (PH) model (which is popularly used in analysing time to event data) and the techniques used in estimating the parameters. In addition, we review various model diagnostics to ascertain the adequacy of the model fit. However, this model is based on the PH assumption. Violation of PH assumption often leads to biased results and inferences. Once non-proportionality is established, there is need to consider time-varying effects of the covariates.

Several models have been developed that relax the proportional hazard assumption making it possible to analyse data with time varying effects of both baseline and time-updated covariates. Various approaches for handling time varying covariates and time-varying effects in time to event models will be discussed. They include the stratified Cox model and the extended Cox model which handles exogenous time-dependent covariates using the counting process formulation introduced by Andersen and Gill (Rizopoulos, 2012).

Another is the Aalen model, an additive model which easily accounts for time-varying effects. However, there are situations where not all the covariates of interest have time-varying effects. Hence, the semi-parametric additive model is used. The Cox-Aalen model is an alternative model which combines Cox proportional hazards model for covariates with constant effects and the Aalen additive model for time-varying effects in a single model. Finally, we will consider joint models for longitudinal and time-to-event processes.

4.2 The Cox Proportional Hazards (PH) Model

Cox (1972) in a seminal paper introduced the Cox proportional hazards regression model. It is the most popularly used model in analysing survival data. This model is used in describing the hazard function of an event of interest as a function of multiple prognostic factors thereby adjusting for other covariates unlike in the K-M approach. In this section, we will discuss the form of the Cox PH model, what the PH assumption entails and estimation of the model parameters.

4.2.1 The Cox Proportional Hazards Model Form

The Cox proportional hazards model is of the form:

$$\begin{aligned} h(t, \mathbf{X}) &= h_0(t) \exp\big(X\beta\big). \\ &= h_0(t) \exp(\beta_1 X_{i1} + \cdots + \beta_p X_{ip}). \end{aligned} \tag{4.1}$$

where,
$h_0(t)$ is an arbitrary baseline hazard rate, reflecting the hazard rate when all X's = 0. $X_{i1}, \cdots, X_{ip}$ are the p-covariates or risk factors of interest and $\beta_1, \cdots, \beta_p$ are regression coefficients, estimated by maximizing the partial likelihood.

Equation 4.1 is basically the product of the arbitrary baseline rate and the exponential expression of a linear combination of the covariates. Hence, it is referred to as a multiplicative model. Taking the exponent of the estimated regression coefficient ($e(\hat{\beta})$) provides a hazard ratio estimate. For categorical variables, a hazard ratio greater than 1 ($\beta > 0$), implies there is an increased risk of the event of interest occurring for the subjects in one category compared to the subjects in the reference category, while a hazard ratio lower than 1 ($\beta < 0$) indicates a decreased risk. In the case of a continuous variable, a hazard ratio greater than 1 ($\beta > 0$) indicates an increased risk associated with a unit increase in the covariate while a hazard ratio less than 1 ($\beta < 0$) indicates an decreased risk for a unit increase in the covariate.

The following admirable properties account for the popularity of this approach. Firstly, it is a semi-parametric model, that is, no distribution is assumed for the baseline hazard function. Secondly, it is a "robust" model, results obtained from the Cox PH model are similar to the results from the correct parametric model. Thirdly, in the absence of covariates in the model,

the Cox PH reduces to the baseline hazard function. The baseline hazard function is the hazard function obtained when all the covariates in a model are zero. Fourthly, the set up of the model, specifically the exponential expression ensures that the estimated hazards are always positive. Fifthly, the regression coefficients in the model can be estimated irrespective of not assuming any distribution for the baseline hazard function. Without estimating the baseline hazard function, the measure of effect (hazard ratio) can be obtained.

Despite the admirable properties of the Cox PH model as listed above, it is based on the underlying proportional hazard assumptions.

4.2.2 The PH Assumption

A fundamental underlying assumption of the Cox PH model is that of proportional hazards. The PH assumption requires that the hazard ratio is constant over time, that is, the hazard ratio is time independent. Suppose that the event of interest is death, and we are interested in its association with p covariates, $X_1, X_2, \cdots, X_p$, then the hazard for a given set of values for these covariates is given by:

$$h(t) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p).$$

Assume that we are interested in a single covariate then the hazard is:

$$h(t) = h_0(t)\exp(\beta x).$$

The hazard ratio of two subjects with covariate values x_1 and x_2 is expressed as:

$$HR = \frac{h_{x2}(t)}{h_{x1}(t)} = \exp[\beta(x_2 - x_1)]. \tag{4.2}$$

Equation 4.2, shows that the hazard ratio (which compares two subjects who differ with respect to their value for X by $x_2 - x_1$) is constant. This implies that it is independent of time(t).

From Equation 4.1 we observed that the baseline hazard function is a function of time only whereas the exponential expression is a function that includes X_i and β_i alone. However, there could be situations whereby either the X_i's or/and the β_i's is a function of time.

The X's are referred to as time-independent variables because their values do not change over time. However, there could be situations whereby the X's are functions of time. These types of variables are referred to as time dependent variables. When the β_i's depend on time, the Cox proportional hazard assumption has been violated. Hence, using the Cox PH model will be inadequate. Other models such as the stratified and extended Cox models, amongst others could be used.

4.2.3 Model Estimation

Fitting the Cox proportional hazards model to a set of observed data requires the estimation of both the baseline hazard function and the unknown regression coefficients. However, both components can be estimated separately. This section, will only present how the regression coefficients are estimated.

Regression coefficients, β's, of the Cox proportional model are estimated using the maximum likelihood approach. A likelihood is basically the joint probability of the data that was observed and it's treated as a function of the unknown parameters in the model. The aim is to find the set of parameter values that maximises the likelihood function. It has been observed that it is computationally easier to maximise the logarithm of the likelihood function. The log-likelihood is maximized using the numerical methods such as the Newton-Raphson method.

Suppose n subjects have $t_1, t_2, \cdots, t_n$ as their event times with the ordered event times of r individuals as $t_1 < t_2 < \cdots < t_r$. Let $t_{(i)}$ be the i^{th} ordered event time. $R(t_{(i)})$ is said to be the number of subjects at risk at time $t_{(i)}$.

The partial likelihood function for the Cox proportional hazards is given by

$$L(\beta) = \prod_{i=1}^{r} \frac{\exp(\beta' x_{(i)})}{\sum_{l \in R(t_{(i)})} \exp(\beta' x_l)}. \tag{4.3}$$

where $x_{(i)}$ is a vector of covariates for subjects that died at the i^{th} ordered event time, $t_{(i)}$. Equation 4.3 considers uncensored observations only. That is, only subjects that have either observed the event of interest or those that are at risk. This is evident in equation 4.3 as the denominator is summed over the subjects yet to experience the event of interest. However, event times of censored subjects contribute to the risk set but this is just before

they are censored. In addition, the product is over the subjects whose event times are known.

Suppose n subjects have $t_1, t_2, \cdots, t_n$ as their observed event times and δ_i is the event indicator with its value being 0 if the ith event time is censored and 1 otherwise. δ_i is used as an exponent because it is a suitable way of getting all the observations included in the likelihood function without excluding the event times. The likelihood function in 4.3 can then be expressed as

$$= \prod_{i=1}^{n} \left\{ \frac{\exp(\beta' x_{(i)})}{\sum_{l \in R(t_{(i)})} \exp(\beta' x_l)} \right\}^{\delta_i}. \tag{4.4}$$

with $R(t_{(i)})$ being the risk set at time t_i. Since δ_i is 0 if the ith event time is censored and 1 otherwise. It implies that censored observations do not contribute to the likelihood function (recall the laws of indices: anything to the power of zero = 1).

Equation 4.5 below presents the log-likelihood function.

$$\log L(\beta) = \sum_{i=1}^{n} \delta_i \left\{ \beta' x_i - \log \sum_{l \in R(t_{(i)})} \exp(\beta' x_l) \right\}. \tag{4.5}$$

4.2.4 Model Diagnostics for the Cox Regression Model

Most models are based on assumptions. Hence, diagnostic techniques are used in determining whether the fitted model describes the data adequately. In addition, the validity of those assumptions are checked and ways in which they are being violated are identified. This implies studying and investigating certain aspects of the model fit such as the selection of explanatory variables to be included in the model, functional forms of the variables, outlying and influential observations and verifying that the PH assumption is satisfied.

Residuals are commonly used in assessing specific aspects of model adequacy. Some of the residuals defined for the Cox regression model are the scaled Schoenfeld residuals, used to verify the violation of the proportionality assumption. The Cox-snell residuals will be used in examining the overall fit of the model. The Martingale residuals are used to determine the functional form of a covariate. The dfbeta statistics are used in determining influential observations.

4.2.5 Cox-Snell residuals

According to Cleves (2008), Cox-Snell residuals are used in checking the overall fit of a model. A model fit is considered as being satisfactory if the estimated survival function for the kth individual at time t, $(\hat{S}_k(t_k))$, is similar to the observed/true survival function for the kth individual at time $(S_k(t_k))$. In addition, when an appropriate model has been fitted, Cox-Snell residuals are assumed to have an exponential distribution with unit mean. This means that the slope of the cumulative hazard should be 1 (i.e. 45° angle). The Cox-Snell residual for the kth individual, is given by

$$r_{Ck} = \exp(\hat{\beta}' \mathbf{x}_k)\ \hat{H}_0(t_k), \tag{4.6}$$

where $\hat{H}_0(t_i)$ is the estimated cumulative baseline hazard, this is also known as the Nelson-Aalen estimator. The major drawback of the Cox-Snell residuals are that they do not highlight the reason why the model does not fit the data.

4.2.6 Martingale residuals

The Martingale residuals are used in examining how necessary each of the covariates included in the model are. In addition, it is used to ascertain whether the functional forms of the covariates in the model are appropriate. If the functional form of the covariates are inappropriate, a transformation of covariates may be appropriate. The expression for the martingale residual is given below:

$$\hat{M}_k(t) = \delta_k - r_{Ck}, \tag{4.7}$$

where,
r_{Ck} is the Cox-Snell residual and δ_k is the observed number of deaths for the kth individual in the interval [0,t] with $\delta_k = 0$ if the survival time is censored.

From Equation 4.7, martingale residuals at each time t can be defined as the difference between the observed number of deaths for the kth subject in the interval [0,t] and its corresponding estimated number of deaths on the basis of the fitted model. Properties of this residuals under the correct model specifications include: (1) the residuals sum up to zero at any given time point i.e $\sum M_k(t) = 0$, (2) martingale residuals have an expected value of zero, $E(\hat{M}_k(t) = 0)$, (3) the residuals are uncorrelated with one another, $cov(\hat{M}_k, \hat{M}_j) = 0$, (4) martingale residuals takes values between $-\infty$ to 0.

4.2.7 Deviance residuals

Therneau et al. (1990) proposed the deviance residuals due to the limitations and disadvantages of the martingale residuals; its heavy skewness and interval values that makes the identification of outlying observations difficult. Deviance residuals are basically martingale residuals that are rescaled to give a symmetrical distribution about zero. They are used in identifying outlying observations. The deviance residual for the Cox regression model is defined below as

$$d_j \quad = \quad sgn(\hat{M}_k)\big[-2\big\{\hat{M}_k + \delta_k \log\big(\delta_k - \hat{M}_k\big)\big\}\big]^{\frac{1}{2}}, \tag{4.8}$$

where
$\hat{M}_k$ = the martingale residual of the kth subject.
sgn = the signed function.

Equation 4.8 is intuitive as the value of the deviance residual can be zero only if $\hat{M}_i = 0$. In addition, the value of the martingale residual is increased to be closer to one by the log function whereas the square root function tends to shrink large negative values (Therneau et al., 1990).

4.2.8 Dfbeta statistics

Dfbeta statistics are used for detecting influential observations. They do this by measuring the degree to which the regression coefficient changes or is affected if the i^{th} observation were deleted. Each subject i has one dfbeta value for each covariate in the model. This change is measured in terms of the standard deviation units. An observation is said to have a considerable amount of influence on the j^{th} regression coefficient if it has large dfbeta values. Influential positive (negative) dfbeta values increases (decreases) the regression coefficient (Braun, 2011). Equation 4.9 represents the dfbeta statistic .

$$\Delta_{\beta j,i} \quad = \quad \frac{\hat{\beta}_j - \hat{\beta_{j(i)}}}{S^2_{(i)} C_{jj}}, \tag{4.9}$$

where,
C_{jj} is the j^{th} diagonal element of $(X'X)^{-1}$, $\hat{\beta}_j$ is the j^{th} regression coefficient from the whole data set. $\hat{\beta_{j(i)}}$ is the j^{th} regression coefficient obtained when the i^{th} observation is removed and $S^2_{(i)}$ is the mean square error.

Note that

$$\Delta_{\beta j,i} > \begin{cases} 1 & \text{for small } n \\ \frac{2}{\sqrt{n}} & \text{for large } n \end{cases} \tag{4.10}$$

serves as a guideline for deciding whether observations are influential or not.

In situations where there are influential observations, it might be necessary to perform sensitivity analysis, that is present results with and without the influential observation(s). When it is obvious that error in data entry is the reason for the influential observation(s), they could be deleted permanently.

4.2.9 Schoenfeld residual

Schoenfeld (1982) proposed the Schoenfeld residuals also known as partial residuals which is used in investigating if the proportional hazards assumption holds. The Schoenfeld residual at time t is the difference between the observed value of the ith explanatory variable for the kth individual and its conditional expectation given the set of all individuals at risk at time t_k (Fitrianto and Jiin, 2013). This is given by

$$\begin{aligned} \hat{r}_{ki} &= X_{ki} - \hat{E}(X_{ki}|R_k), \\ \text{where} & \\ \hat{E}(X_{ki}|R_i) &= \frac{\sum_{l \in R(t_k)} X_{il} \exp(\beta' X_l)}{\sum_{l \in R(t_k)} \exp(\beta' X_l)}. \end{aligned} \tag{4.11}$$

The estimates of β are obtained by maximizing the partial likelihood function. Schoenfeld residuals are time independent. If the proportional hazard holds, a plot of the kth residual($\hat{r}_{ki}$) against t_k should be centered around zero and should have a random pattern.

(Grambsch and Therneau, 1994) suggested the scaled Schoenfeld residual are obtained by multiplying the covariance matrix of the residuals by the vector of the partial residuals. This approach has been proven to yield residuals with greater diagnostic power when compared to the unscaled Schoenfeld residual (Fitrianto and Jiin, 2013).

$$\hat{r}_k^* = [Var(\hat{r}_k)]^{-1} \hat{r}_k, \tag{4.12}$$

where,
$\hat{r}_k^*$ is the scaled Schoenfeld residual and $\hat{r}_k$ is the Schoenfeld residual proposed by Schoenfeld.

4.3 Time-varying effects (or coefficients)

The effect of covariates on the hazard of an event of interest occurring could vary over time. This is reflected in terms of regression coefficients that vary over time. In section 4.2.2, we discussed the PH assumption, which assumes that covariates regression coefficient remain the same over time, that is $\beta(t) = \beta$. For time-varying coefficients that violate the PH assumption, $\beta(t)$ is a function of time.

The structure of a Cox model with time-varying effects is presented in the equation below:

$$h(t) = h_0(t)\exp\{\beta_1(t)X_1 + \beta_2(t)X_2 + \cdots + \beta_p(t)X_p\}. \quad (4.13)$$

As reviewed in subsection 4.2.9, Schoenfeld residuals are commonly used in assessing the proportional hazard assumption.

(Grambsch and Therneau, 1994) proved that when β is the regression coefficient from an ordinary Cox model, then,

$$E(s_{kj}^*) + \beta_j \approx \beta_j(t_k),$$

where s_{kj}^* is the scaled Schoenfeld residual. Plotting $E(s_{kj}^*) + \beta_j$ against time or any function of time gives a graphical perspective as to whether PH assumption has been violated or not. The Schoenfeld residual test whether the effect of a covariate is constant over time by testing for a non-zero slope in a regression model of the residuals on time. If the test is rejected, it implies a non-zero slope and hence the hazard is non-proportional (Therneau and Grambsch, 2000). The shape of the plot gives us an idea of the appropriate function of time to use for a specific covariate. This is useful when we have to model non-proportionality by a time-dependent covariate.

In this section, we will discuss three different models that can be used in modelling covariates with time-varying effects. They are the stratified Cox model, partitioning the follow-up time or time period and modelling non-proportionality by a time-dependent covariate.

4.3.1 The Stratified Cox Model

The stratified Cox model is one of the extensions of the Cox PH model used for handling covariates that violate the Cox PH assumption. Covariates are incorporated into the model as stratification factors if the assumption is violated, whereas covariates that meet the assumption are included in the model as regressors. The concept behind the model where no interaction is assumed is that the baseline hazard for different strata are different but the impact of the covariates are the same across the strata. On the other hand, when interaction is assumed, the baseline hazard function and the regression coefficient is dependent on each of the strata.

4.3.1.1 The General Stratified Cox Model

Suppose there are k and p variables that violate the PH assumption and satisfy the PH assumption respectively. $Z_1, Z_2, \cdots Z_k$ denote the variables that do not satisfy the assumption whereas $X_1, X_2, \cdots X_p$ denote the variables that satisfy the assumption. Then the stratified model formulation is as follows.

$$h_g(t, \mathbf{X}) = h_{0g}(t) \exp\left[\beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_p\right], \tag{4.14}$$

where
g $= 1, 2, \cdots k^*$, strata defined from Z^*, k^* = the total number of strata in Z^*, β_j = the regression coefficient for the jth X-variables with $j = 1, \cdots, p$ and Z^* = a variable defined by first identifying the Z_i variables not satisfying the PH assumption. We then categorize each Z and form combinations of categories of each of the Z's. Each combination represents a different stratum making up the variable Z^*.

Despite the effectiveness of this approach in solving the problem of non-proportionality, there are still some drawbacks. They include:

1. Estimation efficiency is reduced as a result of not being able to estimate the effect of the stratified factors on other variables as they were not explicitly included in the model. This is the trade-off when using this approach. Hence, if one is interested in estimating the effect of a particular variable that happens to violate the PH assumption, this approach is not the best. An appropriate approach will be to model the changing effect of the predictor over time directly. This will be discussed in details later.

2. It becomes more complicated and messy with continuous and multiple stratification variables. This is because each covariate that violates the PH assumption is categorized (including continuous variables and this is difficult because one needs to have a reference category motivating the reason behind the way the variable was categorized) and then combinations will be made for all the categorization of each of the variables that violated the PH assumption.

4.3.1.2 Model Estimation

Parameters in this model are estimated by maximizing the partial likelihood function. The likelihood function in this model is obtained by multiplying the likelihood functions for each strata. Hence, the partial likelihood function for the g^{th} stratum is given by

$$l_g(\beta) = \prod_{i=1}^{n_g} \left\{ \frac{\exp(\beta' x_{(gj)})}{\sum_{l \in R(t_{(gj)})} \exp(\beta' x_{gl})} \right\}^{\delta_i}, \tag{4.15}$$

where,
n_g = number of observations in the g^{th} stratum, $x_{(gi)}$ = is a vector of covariates, $R(t_{(gi)})$ = subjects at risk in the g^{th} stratum at time $t_{(gi)}$, $t_{(gi)}$ = the time the i^{th} subject in the g^{th} stratum was observed and δ_i = is the event indicator.

The overall stratified Cox likelihood function is obtained by multiplying the likelihood from each stratum. It is given as the expression below:

$$l_G(\beta) = \prod_{g=1}^{G} l_g(\beta). \tag{4.16}$$

4.3.2 Partition the time period

According to Therneau and Grambsch (2000), partitioning the time period is another way of dealing with covariates that have time-varying effects. Since the PH assumption was violated over the entire time period, we fit different Cox models on shorter time periods. Except in cases where there are specific cut-off time values that may have some clinical implications, the median

event time is commonly used to partition the time period into sub-intervals. Everyone still at risk beyond the chosen cut-off time is censored in the first analysis whereas the individuals considered in the second part of the analysis are those still at risk after the chosen cut-off time. Results should thus be interpreted with caution as the interpretation of the models is conditional on the length of the survival time. It is a necessary to ensure that the PH assumption is not violated within the short time periods. The setback to implementing this approach is the reduction in power due to fewer event times within the intervals.

4.3.3 Model non-proportionality by time-dependent covariates

Finally, modelling non-proportionality by time-dependent covariates is another technique proposed by David et al. (1972) to use in handling time varying effects. A time-dependent variable is created by forming an interaction between the variable that violates the PH assumption and a function of time t as seen below.

Suppose the regression coefficient for a covariate is a function of time (non-proportional hazards). We can write

$$\begin{aligned} h(t) &= h_0(t)\exp\{\beta(t)X\} \\ &= h_0(t)\exp\{\beta X(t)\}, \end{aligned} \tag{4.17}$$

where
$X(t) = X \times g(t)$
and $g(t)$ is a function of time (t). For example, linear or quadratic.

Equation 4.17 shows that a covariate with a time-varying coefficient can be a expressed as a time-dependent covariate with a constant coefficient. Common functional forms of time include linear, quadratic and log functions of time (Bellera et al., 2010). The precision of the final model is dependent on correctly specifying the function of time. Information on the choice of the function of time may be governed by knowledge about the behaviour of the variable over time or the Schoenfeld residuals. The shape of the Schoenfeld plot of each covariate versus time gives the analyst an idea of the form or function of time to be assumed. This will be shown in section 5.3.3 where we allowed the shape of the Schoenfeld residual plot to infer the choice of our function of time on the Gugulethu data.

4.4 Time-varying covariates

In all the survival models we have discussed thus far, values of the explanatory variables incorporated into these models were baseline observations recorded at the beginning of the study. We have assumed that the values of all covariates were determined at the point when follow-up began on each subject (time zero) and that these values did not change over the period of observation. However, most survival data are generated as a result of repeated and multiple observations taken on some explanatory variables of interest. There may be situations where one or more of the covariates are measured during the period of follow up and their values change. In these settings, the case may be that the value of the hazard for the event depends more on the current values of those covariates than on their values at time zero. An example is the CD4 cell count and viral load of HIV/AIDS patient in a study being recorded at every hospital visit. The quality of this data gives a better and more appropriate indication of the impact of those variables on the hazard of the event of interest (Collett, 2015). Hence, covariates whose values change over time are commonly called **time-varying or time-dependent covariates**.

Time-varying covariates are variables whose values change over time. For example the CD4 count of an HIV/AIDS patient changes each time it is measured. However, a time-varying coefficient or effect is one in which the hazard ratio doesn't remain constant over time. An example of a time-varying coefficient or effect is the effect of a treatment which can be strong immediately after treatment but fades away with time.

4.4.1 Types of Time-varying covariates

Let $x_i(t)$ be the covariate vector at time t for the i^{th} subject and $X_k(t) = \{x_k(u); 0 \leq u < t\}$ denote the covariate history up to time t. It is necessary to consider the different types of time-varying covariates as different methods are used in handling them. There are two types of time-varying covariates, endogenous time-varying covariates and exogenous time-varying covariates. According to Kalbfleisch and Prentice (2011), a time-varying covariate is said to be `exogenous or external` if the following condition is satisfied,

$$P\{T \in [u, u+du) \,|T \geq u, X(u)\} \quad = \quad P\{T \in [u, u+du) \,|T \geq u, X(t)\}, \tag{4.18}$$

for all u, t such that $0 \leq u < t$. Equivalently, condition 4.18 can be presented as

$$P\big[X(t)|X(u), T \geq u\big] \;=\; P\big[X(t)|X(u), T = u\big]. \tag{4.19}$$

which describes the idea that though their value change over time, the values are not dependent on whether or when the event occurred. A time-varying covariate is said to be endogenous or internal if the condition in 4.19 is violated. Internal or endogenous covariates arise typically when the probability and timing of the event changes the subsequent values of the covariate.

1. Exogenous covariates: are those whose value at a particular time does not require subjects to be under direct observation. For example, a subject's age. If we follow subjects for a long enough period of time, their current age may have more of an effect on survival than their age when the study began. However, once we know a subject's birth date, age may be computed at any point in time, regardless of whether the subject is still under observation. The survival function given an exogenous covariate can be defined as

$$\begin{aligned} S_i(t|X_i(t)) &= Pr(T_i > t|X_i(t)) && (4.20)\\ &= \exp\Big\{ -\int_0^t h_i(s|X_i(s))ds \Big\}. && (4.21) \end{aligned}$$

2. Endogenous covariates: are those whose value is subject-specific, requires the subject to be under periodic observation and requires the survival of the subject for their existence. For example, a, HIV/ AIDS study where we are interested in investigating the survival rate of patients on treatment. Hence, the CD4 count and viral load of the patient needs to be taken. However, the endpoint of the study is the death of the patient. Suppose that we have a covariate measured at baseline, but whose value can change over time. It may be the case that the hazard depends on a more recent value than on the baseline value. For values of this covariate to be measured during the follow up period, the subjects in this study must be under direct observation. The survival function of an endogenous covariate is defined below as

$$S_i(t|X_i(t)) \;=\; Pr(T_i > t|X_i(t)) = 1. \tag{4.22}$$

4.5 The Extended Cox model

The Cox PH regression model and its variants that have been fitted thus far include only baseline covariate measurements. That is, data consist of only one observation per subject, the covariate measurements are obtained at study entry and the covariates are treated as time-invariant. The concept and types of time-varying covariates have been reviewed in sections 4.4 and 4.4.1 respectively. The Cox PH regression model will be an inappropriate model to use in modelling the survival relationship as it is not designed to handle time-varying covariates. Therefore, the extended Cox regression model, which is an extension of the Cox model to handle exogenous time-dependent covariates using the counting process formulation investigated by Andersen and Gill (1982), should be used instead.

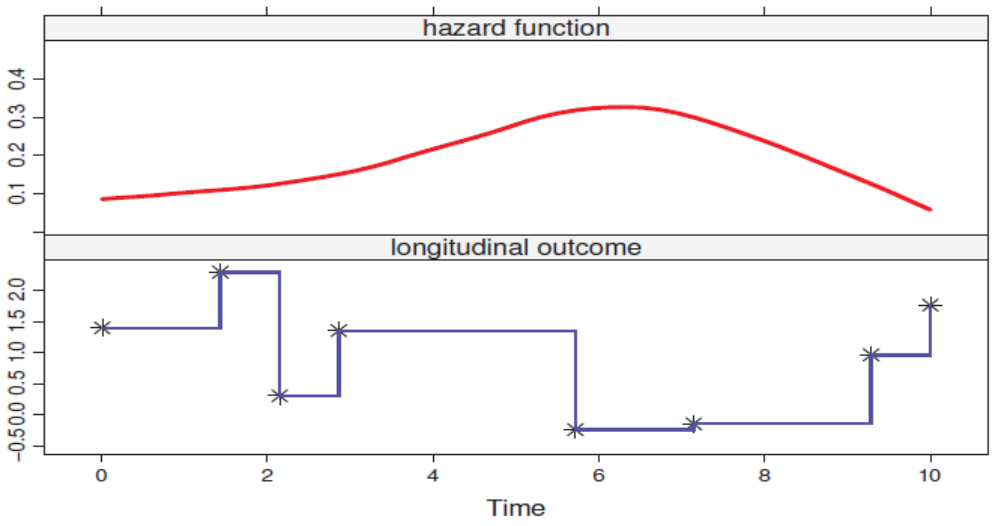


Figure 4.1: Intuitive representation of the extended Cox model, Rizopoulos (2012)

Figure 4.1 gives an intuitive idea on how time-varying covariates are handled in the extended Cox model framework. It assumes that the value of longitudinal marker remains constant between visits.

The extended Cox model (also known as the Andersen-Gill model) is written as

$$h(t, \mathbf{X}(t)) \quad = \quad h_0(t) \exp \left\{ \sum_{i=1}^{p1} \beta_i X_i + \sum_{j=1}^{p2} \alpha_j X_j(t) \right\}, \qquad (4.23)$$

where,
$h_0(t)$ = is the baseline hazard function, p_1 & p_2 = the number of baseline covariates and number of time-varying covariates, X_i = are time-independent covariates, $X_j(t)$ = are time-dependent covariates and $\beta_{\mathbf{i}}$ & $\alpha_{\mathbf{j}}$ = are the regression coefficient vectors of the two sets of covariates.s

In equation 4.23, some of the covariates are time-dependent hence the hazard ratio for any two individuals is also time-dependent and no longer constant. The regression coefficient vectors ($\beta_{\mathbf{i}}$ and $\alpha_{\mathbf{j}}$) have the same interpretation. Thus, the regression coefficient in an extended Cox model is interpreted as the log-hazard ratio of an event at time t that results from one unit increase in X_i(t) at the same time point.

Estimation of the regression coefficients is by maximizing the partial log likelihood function below:

$$l(\beta) = \sum_{i=1}^{n} \delta_i \left\{ \sum_{i=1}^{p1} \beta_i X_i + \sum_{j=1}^{p2} \alpha_j X_j(t) - \log \sum_{l \in R_i(t)} \exp\left(\sum_{i=1}^{p1} \beta_i X_i + \sum_{j=1}^{p2} \alpha_j X_j(t) \right) \right\},$$

where
$R_i(t)$ is the risk process with $R_i(t) = 1$ if individual i is at risk and 0, otherwise. δ_i is the event indicator with $\delta_i = 0$, if survival time of individual i is censored and 1 otherwise.

4.6 Joint Modeling for Longitudinal and Time-to-Event Data.

This section presents a summary of the discussion in Rizopoulos (2012) and Rizopoulos (2018). The extended Cox model is only appropriate for handling exogenous time-varying covariates. It does not handle endogenous time-varying covariates such as biomarkers appropriately because of the assumption that the values of these covariates only change at visits where measurements were taken which is unrealistic. Hence, joint modeling is a more appropriate technique used in assessing the association between an endogenous time-varying covariate and survival. The motivating idea behind the joint model is to couple the survival model, which is of primary interest with a suitable model for the repeated measurements of the endogenous covariate that accounts for its special features.

In this work, the joint model will be used as one of the approaches to handle time-varying covariates with the aim of studying the patient's survival. We will begin by specifying the submodels of the joint model, then discuss how the parameters and random effects are estimated using the maximum likelihood estimation technique and the Bayes theory respectively. Finally, extensions of the joint model such as parametrizations (association between the longitudinal outcome and the hazard of the event) and multivariate joint models will be considered.

4.6.1 Submodels specification

The joint model is made up of two submodels that are linked together: the longitudinal and survival submodels. The objective of using the joint model is to measure the association between the longitudinal marker level and the risk of an event. The joint model will be achieved in three (3) steps. They are:

1. The covariate history for each individual is reconstructed from the observed longitudinal response ($y_i(t)$).

$$\begin{aligned} y_i(t) &= m_i(t) + \epsilon_i(t) \\ &= x_i^T(t)\beta + z_i^T(t)b_i + \epsilon_i(t),\ \epsilon_i(t) \sim N(0, \sigma^2), \end{aligned} \tag{4.24}$$

   where,
   x_i & β = make up the fixed effect part of the model and z_i & b_i = make up the random effect part of the model, $b_i \sim N(0, D)$.

2. Assume that the term $m_i(t)$ that denotes the true and unobserved value of the longitudinal outcome/marker at time t is known. Hence, we can define the standard relative risk model as

$$h_i(t|\mathcal{M}_i(t), w_i) = h_0(t)\exp\{\gamma^T w_i + \alpha m_i(t)\}, t > 0, \tag{4.25}$$

   where,
   $\mathcal{M}_i(t) = m_i(s), o \leq s < t$ is the longitudinal history, α = measures the degree of relationship between the longitudinal marker and risk of an event, w_i = are the baseline covariates and γ = is the vector of the regression coefficients for the baseline covariates.

3. In the third step, a joint distribution is described for the two processes discussed in steps 1 & 2

$$P(y_i, T_i, \delta_i) = \int P(y_i|b_i)\ \{h(T_i|b_i)^{\delta_i}\ S(T_i|b_i)\ p(b_i)db_i\}, \quad (4.26)$$

where,
b_i = denotes a vector of random effects that explains the ith subject deviation from the population, $p(\cdot)$ = denotes the density function and $S(\cdot)$ represents the survival function.

Based on the key assumption of the joint model (conditional independence), the longitudinal outcome is independent of the survival outcome and the repeated measurements in the longitudinal outcome are independent of each other,

$$\begin{aligned} p(y_i, T_i, \delta_i|b_i) &= p(y_i|b_i)\ p(T_i, \delta_i|b_i) \\ p(y_i|b_i) &= \prod_j p(y_{ij}|b_i). \end{aligned} \quad (4.27)$$

The survival function depends on the entire longitudinal marker history.

$$S_i(t|b_i) = \exp\left(-\int_0^t h_0(s) \exp\{\gamma^T w_i + \alpha m_i(s)\} ds \right). \quad (4.28)$$

In the Cox PH model, it is usual to leave the baseline hazard function unspecified to avoid specifying a wrong distribution for the survival times. In the joint modeling framework, the standard errors of the parameter that were estimated are underestimated if the baseline hazard function is left unspecified. This limitation is handled by explicitly assuming an underlying distribution for the baseline hazard function. Distributions such as the Weibull, log-normal and the Gamma are commonly used. On the other hand, several approaches that involve non-parametric models with flexible specifications of the baseline hazard function such as the B-splines, restricted cubic splines, piecewise-constant model and regression splines have been proposed.

The baseline hazard function for the piecewise-constant model is of the form:

$$h_0(t) = \sum_{q=1}^{Q} \xi_q I(v_q - 1 < t \leq v_q), \quad (4.29)$$

where $0 = v_0 < v_1 < \cdots < v_Q$ represents the cuts of the time scale, v_Q has a value that is larger than the largest observed time, the value of the hazard in the interval $(v_{q-1}, v_q]$ is represented by ξ_q.

In the regression model, the logarithm of the baseline hazard function is expanded into B-spline basis functions for the cubic spline. This is as shown in the equation below:

$$\log h_0(t) \;=\; \kappa_0 + \sum_{q=1}^{m} \kappa_d B_d(t, q), \tag{4.30}$$

where $\kappa^T = (\kappa_0, \kappa_1, \cdots, \kappa_m)$ represent the coefficients of the spline, q is the degree of the B-splines basis functions $B(\cdot)$ and $m = \ddot{m} + q - 1$, where $\ddot{m}$ is the number of interior knots.

In both the piecewise-constant and regression model approach, the more the number of knots, the better the probability of estimating the baseline hazard function. However, to avoid over fitting, variability and bias must be brought into perspective.

4.6.2 Model Estimation

The log-likelihood function of the ith subject is as follows,

$$\begin{aligned} \log p(T_i, \delta_i, y_i; \theta) &= \log \int p(T_i, \delta_i, y_i, b_i; \theta) db_i \\ &= \log \int p(T_i, \delta_i | b_i; \theta_t, \beta) \big[\prod_j p\{y_i(t_{ij}) | b_i; \theta_y\} \big] p(b_i; \theta_b) db_i. \end{aligned} \tag{4.31}$$

Equation 4.31 illustrates the dependence of the log likelihood on a time to event process and a longitudinal process. The density function for the time-to-event process is as follows,

$$\begin{aligned} p(T_i, \delta_i | b_i; \theta_t, \beta) &= h_i(T_i | \mathcal{M}_i(T_i); \theta_t, \beta)^{\delta_i} \mathcal{S}_i(T_i |\, \mathcal{M}_i(T_i); \theta_t, \beta) \\ &= \big[h_0(T_i) \exp\{\gamma^T w_i + \alpha m_i(T_i)\} \big]^{\delta_i} \\ &\times \exp\Big(- \int_0^{T_i} h_0(s) \exp\{\gamma^T w_i + \alpha m_i(s)\} ds \Big). \end{aligned} \tag{4.32}$$

where the baseline hazard function $h_0(\cdot)$ could assume underlying distribution of time as discussed above. To estimate the parameter θ, the log-likelihood

function in 4.31 is maximized using some numerical optimization algorithms such as the Expectation - Maximization (EM) algorithm or the Newton-Raphson algorithm amongst others.

The joint density function for longitudinal responses with random effects is as shown below:

$$\begin{aligned} p(y_i|b_i;\theta)p(b_i;\theta) &= \prod_j p\{y_i(t_{ij})|b_i;\theta_y\}p(b_i;\theta_b) \\ &= (2\pi\sigma^2)^{n_i/2}\exp\{||y_i - \mathbf{X}_i\beta - Z_ib_i||^2/2\sigma^2\} \\ &\times (2\pi)^{qb/2}\det(D)^{-1/2}\exp(-b_i^TD^{-1}b_i/2). \end{aligned} \tag{4.33}$$

Estimation of the random effects as proposed by Rizopoulos (2012) is carried out using the Bayesian methods. Hence, estimates for the random effects are based on the posterior distribution. With a prior distribution in of the form $p(b_i;\theta)$ and a conditional likelihood function $p(b_i|Y_i,\delta_i,y_i,\theta)p(y_i|b_i;\theta)$,.

$$\begin{aligned} p(b_i|Y_i,\delta_i,y_i,\theta) &= \frac{p(Y_i,\delta_i|b_i,\theta)p(y_i|b_i;\theta)p(b_i;\theta)}{p(Y_i,\delta_i,y_i;\theta)} \\ &\propto p(Y_i,\delta_i|b_i,\theta)p(y_i|b_i;\theta)p(b_i;\theta). \end{aligned} \tag{4.34}$$

Since equation 4.34 does not have closed solution, it is solved numerically and the measures of location used are the mean or the mode with their forms presented below:

$$\begin{cases} \bar{b}_i & \int b_i p(b_i|Y_i,\delta_i,y_i;\theta)\text{and} \\ \hat{b}_i & \arg\max_b\{\log p(b_i|Y_i,\delta_i,y_i;\theta)\}. \end{cases}$$

4.6.3 Extensions of the Joint Model

In this section, we will consider extensions of the standard joint model we have considered before. These extensions include the parametrization of the association structure between the longitudinal and event outcomes and fitting joint models for multiple longitudinal responses.

4.6.4 Parametrization

In the standard joint model, we assumed that the risk of an event of interest at time t, depends on the true level of the longitudinal marker at that time. The parameter α measures the degree of association or dependency. This assumption may be violated. In a situation whereby the assumption is invalid, the results obtained will be inappropriate and lead to incorrect medical conclusions. In this section, different assumptions are made on the association structure between the risk of an event and the longitudinal outcome. They include the lagged effect parametrization, the interaction effect parametrization, the cumulative effect parametrization, the weighted cumulative effect parametrization and the time-dependent slopes parametrization.

The risk of an event at a particular time in a lagged effect parametrization, is assumed to be associated with the level of the longitudinal outcome at a preceding time point.

$$h_i(t) = h_0(t)\exp[\gamma^T w_i + \alpha m_i\{\max(t-c,0)\}] \tag{4.35}$$

where c is the time lag one is interested in.

The interaction effect parametrization allows the effect of the risk of an event on the longitudinal outcome to differ in all subgroups of the intended population. This is done by introducing an interaction term as this will accommodate the different measures of associations in the subgroups. The formulation of this type of parametrization is as shown below:

$$h_i(t) = h_0(t)\exp[\gamma^T w_{i1} + \alpha^T\{w_{i2} \times m_i(t)\}] \tag{4.36}$$

where,
w_{i1} = denotes a vector that includes the direct effects of baseline covariate to the risk for an event, w_{i2} = contains covariates that are allowed to interact with $m_i(t)$, thereby allowing different associations for different subgroups of the data.

The time-dependent slopes parametrization allows the risk of an event to depend on other features of the longitudinal trajectory such as the slope. Hence, the formulation below is when we assume that the risk of an event depends on the current value and slope of the longitudinal trajectory.

$$h_i(t) = h_0(t)\exp\{\gamma^T w_i + \alpha_1 m_i(t) + \alpha_1 m_i^{'}(t)\} \tag{4.37}$$

where

$$m_i^{'}(t) = \frac{d}{dt}m_i(t) = \frac{d}{dt}\{x_i^T(t)\beta + z_i^T(t)b_i\}$$

α_1 measures the degree of the association between the risk of an event at time t and the current longitudinal outcome at the same time. On the other hand, α_2 measures the degree of the association between the risk of an event at time t and the slope of the longitudinal outcome at the same time provided $mi(t)$ is constant.

The cumulative effect parametrization is more or less an extension of the time-dependent slopes parametrization. In the time-dependent parametrization, the risk of an event at time t is assumed to depend on other features of the longitudinal outcome either at the same time or lagged time. In addition, this is assumed for a single time point. However, in the cumulative effect parametrization, the risk of an event at time t is assumed to depend on the cumulative trajectory of the longitudinal outcome.

$$h_i(t) = h_0(t)\exp\{\gamma^T w_i + \alpha\int_0^t m_i(s)ds\} \tag{4.38}$$

The strength of the association between the risk of an event at time point t and the cumulative longitudinal trajectory up to the same time t is measured by α.

The cumulative effect parametrization assumes the same weights for all the longitudinal values whether they were observed currently or in the past. However, in the weighted cumulative effect parametrization, weight functions that places different weights at different time points are chosen. The formulation is as shown below:

$$h_i(t) = h_0(t)\exp\{\gamma^T w_i + \alpha\int_0^t \varpi(t-s)m_i(s)ds\} \tag{4.39}$$

where $\varpi(\cdot)$ denotes the weight function. A desirable property of $\varpi(\cdot)$ would be to place smaller weights in points further in the past.

4.6.5 Joint models for multiple longitudinal responses

All along we have fitted joint models that have only one longitudinal outcome. However, there are situations were multiple longitudinal outcomes/biomarkers are useful in predicting the hazard of an event. For example, in the Gugulethu HIV/AIDS data, `lCD4` and `Tx` are longitudinal outcomes essential in predicting the risk of death.

The multivariate generalized linear mixed model is used to accommodate the different longitudinal outcomes in a joint model framework with the conditional distribution of kth outcome given a vector of random effects b_{ki} being a member of the exponential family, with linear predictors given by

$$g_k[E\{y_{ki}(t)|b_{ki}\}] = \eta_{ki}(t) = \mathbf{x}_{ki}^T\beta_k + \mathbf{z}_{ki}^T\mathbf{b}_{ki}, \;\; \mathbf{b}_{ki}^T \sim MVN(0, D) \quad (4.40)$$

where,
g_k = represents the one-on-one monotone link function, y_{ki} = is the kth longitudinal outcome for the ith individual at time t, $\mathbf{x}_{ki}^T$ & $\mathbf{z}_{ki}^T$ = the vector of the fixed and random effects respectively, D = the variance/ covariance matrix of the random effects.

The hazard model for the multiple longitudinal outcome is given as

$$\begin{aligned} h_i(t) &= h_0(t)\exp[\gamma^T\mathbf{w}_i + \sum_{k=1}^{K}\sum_{l=1}^{L_k} f_{kl}\{\mathcal{H}_{ki}(t), w_i(t), b_{ki}, \alpha kl] \\ &= h_0(t)\exp\{\gamma^T\mathbf{w}_i + \sum_{k=1}^{K}\alpha_k\eta_{ki}(t)\} \end{aligned} \quad (4.41)$$

where
$\mathcal{H}_{ki}(t)$ = $\{\eta_{ki}(s), 0 \leq s < t\}$ is the history of the longitudinal process up to t,
$h_0(t)$ = is the baseline hazard function modeled using the B-splines,
$\mathbf{w}_i$ = denotes the vector of covariates with ts regression coefficient γ.

Just like in the standard joint model where different parametrization structure for the association parameter were assumed, the same is applicable in joint models for multiple longitudinal responses.

$$f\{\mathcal{H}_i(t), w_i(t), b_i, \alpha\} = \alpha\eta_i(t), \quad (4.42)$$

$$f\{\mathcal{H}_i(t), w_i(t), b_i, \alpha\} = \alpha_1\eta_i(t) + \alpha_2\eta_i^{'}(t) \;\text{ with }\; \eta_i^{'}(t) = \frac{d\eta_i(t)}{dt}, \quad (4.43)$$

$$f\{\mathcal{H}_i(t), w_i(t), b_i, \alpha\} = \alpha\int_0^t \eta_i(s)ds \quad (4.44)$$

Equation 4.42 assumes that the risk of an event at a particular time t is associated with the longitudinal outcome at the same time point. Equations 4.43 and 4.44 assume that the risk of an event at a particular time t is associated with the slope of the longitudinal outcome at the same time point or the cumulative trajectory up to time t respectively.

4.6.5.1 Model Estimation

The parameters in the multivariate joint model are estimated by maximising the log-likelihood function in equation 4.45 below using the Bayesian approach.

The posterior distribution of the model parameters given the observed data is derived under the assumptions that given the random effects, the longitudinal outcomes are independent from the event times, the multiple longitudinal outcomes are independent of each other, and the longitudinal responses of each subject in each outcome are independent. Under these assumptions the posterior distribution is analogous to:

$$p(\theta, b) \propto \prod_{i=1}^{n} \prod_{k=1}^{K} \prod_{j=1}^{n_{kl}} p(y_{kij}|b_{ki}, \theta) p(T_i, T_i^U, \delta_i|b_{ki}, \theta) p(b_{ki}|\theta) p(\theta) \quad (4.45)$$

where,
θ = denotes the full parameter vector, $p(\theta, b)$ = the posterior distribution of the model given the random effects, $p(y_{kij}|b_{ki}, \theta)$ = the density for the longitudinal part and $p(T_i, T_i^U, \delta_i|b_{ki})$ = the density distribution of the survival part.

The longitudinal part of the model is

$$p(y_{kij}|b_{ki}, \theta) = \exp\{[y_{kij}\psi_{kij}(b_{ki}) - c_k\{\psi_{kij}(b_{ki})\}]/a_k(\varphi) - d_k(y_{kij}\varphi)\} \quad (4.46)$$

with $\psi_{kij}(b_{ki})$ and φ denoting the natural and dispersion parameters in the exponential family respectively, $c_k(\cdot), a_k(\cdot)$ and $d_k(\cdot)$ are known functions specifying the member of the exponential family.

For the survival part accordingly we have

$$\begin{aligned} p(T_i, T_i^U, \delta_i|b_{ki}, \theta) &= \{h_i(T_i|\mathcal{H}_i(T_i), w_i(T_i))\}^{I(\delta_i=1)} \exp\{-\int_0^{T_t} h_i(s|\mathcal{H}_i(s), w_i(s))ds\} \\ &\times \{1 - \exp\{-\int_0^{T_t} h_i(s|\mathcal{H}_i(s), w_i(s))ds\}\}^{I(\delta_i=2)} \\ &\times \{\exp\{-\int_0^{T_t} h_i(s|\mathcal{H}_i(s), w_i(s))ds\} \\ &- \exp\{-\int_0^{T_t} h_i(s|\mathcal{H}_i(s), w_i(s))ds\}\}^{I(\delta_i=3)} \end{aligned} \quad (4.47)$$

where $I(\cdot)$ is the indicator function, T_i^U, $\delta_i = 2$ and $\delta_i = 3$ are used when the data is interval censored.

4.7 Aalen's additive hazard regression models

The following section comprises of the Aalen's additive hazard regression models and the semi-parametric additive hazards model are summaries from Martinussen and Scheike (2007). The multiplicative regression models such as the Cox PH model have been our focus until this point. In the multiplicative regression models specifically the Cox PH model, the hazard function is associated with the multiplicative effect of the covariates on the baseline hazard (, 2008), and these effects are expected to remain constant over time. However, there are situations whereby the additive effect of the covariates is what is of interest. These are referred to as additive models. It is based on the assumption that the effect of covariates on a subject accumulates over time in its action to either cause or prevent the event. Unlike in multiplicative regression models were instantaneous effect of the event of interest is obtained, in additive regression models, the effects are allowed to be realised some time after the change in the covariate has occurred.

In 1980, Aalen introduced the Aalen's additive hazard regression model. This model was developed further in 1989 and 1993 (Hosmer et al., 2011). It is a flexible non-parametric (that is no distributional assumption is made for the hazard rate) model as it accommodates the inclusion of time-varying effects/covariates.

In Aalen's additive model with p-covariates, $\mathbf{x}(t) = (x_1(t), \cdots, x_p(t))^T$, the conditional hazard function at time t is of the form

$$\begin{aligned} h(t, \mathbf{x}, \beta(t)) &= Y(t)\mathbf{x}^T(t)\beta(t) \\ &= Y(t)(x_1(t)\beta_1(t) + \cdots + x_p(t)\beta_p(t)), \end{aligned} \tag{4.48}$$

where,
$\beta(t) = (\beta_1(t), \cdots, \beta_p(t))^T$ is the vector of regression coefficients, $Y(t) =$ is the risk indicator and $\mathbf{x}^T(t) = (x_1(t), \cdots, x_p(t))^T$ is the vector of time-varying covariates.

The regression coefficient in 4.48 is a function of time. Hence, it is time dependent which implies that the effect of the covariates may change with time.

The cumulative hazard function obtained from the hazard function in 4.48 is

$$\begin{aligned} H(t, \mathbf{x}, \mathbf{B}(t)) &= \int_0^t h(u, \mathbf{x}(u), \boldsymbol{\beta}(u)) du \\ &= \sum_{k=0}^{p} x_k(u) \int_0^t \beta_k(u) du \\ &= \sum_{k=0}^{p} x_k(t) B_k(t) \end{aligned} \tag{4.49}$$

The k^{th} covariate has the cumulative regression coefficient, $B_k(t)$. Aalen noted that the cumulative regression coefficients are easier to estimate compared to the regression coefficient. The former is known to converge at a faster rate compared to the latter.

The cumulative regression functions are plotted against time for each covariate. This plot describes how covariates influence survival over time. If the estimated cumulative regression coefficient is constant over time, the plot looks like a straight line through the origin with slope equal to the coefficient's value. This implies that the effect of the covariate does not vary with time. However, if there are deviations from the straight line at any time interval shows an evidence of the covariate having a time-varying effect (Hosmer et al., 2011). In addition, the slope of plots are intuitive as this indicates whether a specific covariate has time-varying effect or not. For plots with positive slopes, we conclude that increasing the covariate increases hazard whereas negative slopes occur when increasing the covariate decreases hazard during periods (BAŞAR, 2017). Examples of these plots will be shown in section 5.8.

4.7.1 Model Estimation

The cumulative regression coefficient for Aalen's additive model can be estimated using several techniques such as the non-parametric approach with a uniform asymptotic and using the ordinary least-squares. Aalen (1980) introduced estimating the cumulative regression coefficient using the ordinary least-squares (Martinussen and Scheike, 2007).

The Aalen additive model has the intensity for the counting process $N(t)$ conditioned on the covariates as given in equation 4.48. $\Lambda(t) = \int_0^t \lambda(s) ds$ is the n-dimensional cumulative intensity with $M(t) = N(t) - \Lambda(t)$ as the

n-dimensional martingale.

$$\begin{aligned} dN_i(t) &= h(t)dt + dM(t) \\ &= \mathbf{X}(t)\beta(t)dt + dM(t) \end{aligned} \tag{4.50}$$

Equation 4.50 has the form of the ordinary linear regression model where $dN_i(t)$ is the response, $\mathbf{X}(t)$ the covariates , $\beta(t)$ is the parameter vector to be estimated and $dM(t)$ denotes the random error term.

$\beta(t)$ can be rewritten as $dB(t)$ and can be estimated using the multiple linear regression method. The ordinary least squares estimator gives

$$d\hat{B}(t) = (X(t)^T X(t))^{-1} X(t)^T dN(t). \tag{4.51}$$

When $X(t)$ is a full rank, $J(t)$, a Jacobian is introduced as an indicator with the value of 1 in the existence of the inverse and 0 otherwise. The least squares generalised inverse is given as

$$X^-(t) = (X(t)^T X(t))^{-1} X(t)^T, \tag{4.52}$$

which satisfies the relation $X^-(t)X(t) = J(t)I_p$ where I_p is an identity matrix. Substituting the values of $X^-(t)$ in equation 4.52 into 4.51 we have an estimator

$$d\hat{B}(t) = X^-(t)dN(t) \tag{4.53}$$

Equation 4.53 can be re-written in the integral form as

$$\hat{B}(t) = \int_0^t X^-(s)dN(s) \tag{4.54}$$

4.7.2 Inference for Aalen's additive hazard regression models

This section presents approaches for conducting inference for the additive hazards model. The hypothesis can be stated in terms of either the regression coefficients (β) or the cumulative regression coefficients (B). Two hypothesis will be considered:

$$H_{01} : B_p(t) \equiv 0 \quad \text{vs} \quad H_{A1} : B_p(t) \not\equiv 0 \tag{4.55}$$

$$H_{02} : B_p(t) \equiv \gamma t \quad \text{vs} \quad H_{A2} : B_p(t) \not\equiv \gamma t \tag{4.56}$$

Hypothesis 4.55 tests whether one of the components differs significantly from zero and hypothesis 4.56 tests whether at least one of the components has constant effect with time. Hypothesis 4.55 can be evaluated using either the Hall - Wellner confidence band approach obtained by the resampling technique or the simulation based confidence band and observing whether or not the zero function is contained within the band. The effect of a variable is significant if the zero function of the confidence band is contained outside the bands. However, this approach can't evaluate hypothesis 4.56 because the uncertainity of not knowing γ is not reflected by looking at the Hall-Wellner confidence band. Hypothesis 4.56 can be tested using either the Kolmogorov-Smirnov test or the Cramér-von Mises test (Martinussen and Scheike, 2007).

Several test statistics can be used in analysing the hypothesis in equation 4.55. Some of them include the Hall - Wellner confidence band approach presented in equation 4.61 below, the maximal deviation test statistics

$$\tilde{T}_{1S} = \sup_{t\in[0,\tau]} |\hat{B}_p(t)|, \tag{4.57}$$

or revised version of the maximal deviation test statistics that considers the variability of $\hat{B}_p(t)$

$$\sup_{s,t\in[0,\tau]} |\hat{B}_p(s) - \hat{B}_p(t)|. \tag{4.58}$$

The Hall-Wellner confidence band is based on a Gaussian Martingale, $U_p(t)$, with covariance function $\Phi_{pp}(t)$,

$$U_p(t)\frac{\Phi_{pp}(\tau)^{\frac{1}{2}}}{\Phi_{pp}(\tau)+\Phi_{pp}(t)}, \tag{4.59}$$

which has the same distribution as

$$B^0\frac{\Phi_{pp}(t)^{\frac{1}{2}}}{\Phi_{pp}(\tau)+\Phi_{pp}(t)}. \tag{4.60}$$

where B^0 is the Brownian bridge. Thus, the Hall-Wellner confidence band is given by

$$\hat{B}_p(t) \pm n^{\frac{1}{n}} d_\alpha \Phi_{pp}(t)^{\frac{1}{2}}\left(1+\frac{\Phi_{pp}(t)}{\Phi_{pp}(\tau)}\right), \quad t\in[0,\tau] \tag{4.61}$$

where the pth diagonal element of Φ is Φ_{pp} and d_α is the upper α- quantile of $\sup_{t\in[0,1/2]} |B^0(t)|$.

After the confidence band has been calculated, examining if the zero function is within the band or not is used as a criteria in testing the hypothesis presented in equation 4.55. This approach cannot be used in evaluating the hypothesis that the regression coefficients are time-invariant because the uncertainties and variabilities surrounding γ is not known.

Testing the hypothesis presented in equation 4.56, the following test statistics are considered,

$$T_{2S} = n^{\frac{1}{2}} \sup_{t\in[0,\tau]} |\hat{B}_p(t) - \hat{B}_p(\tau)\frac{t}{\tau}|, \tag{4.62}$$

and

$$T_{21} = n\int_0^\tau \left(\hat{B}_p(t) - \hat{B}_p(\tau)\frac{t}{\tau}\right)^2. \tag{4.63}$$

Based on the results obtained from from equations 4.62 and 4.63, some test such as the Kolmogorov-Smirnov test or the Cramér-von Mises test are constructed. They are used in testing the hypothesis presented in equation 4.56.

The Kolmogorov-Smirnov test rejects at level α if

$$\sup_{t\leq\tau_1} |\hat{V}_p^*(t,\tau_2)/(n\hat{\Phi}_{pp})^{\frac{1}{2}}| \geq f_\alpha \tag{4.64}$$

where f_α = the $(1-\alpha)$-quantile in the distribution of $\sup_{0\leq x\leq 1} |B(x)|$ with B the standard Brownian motion and $\Phi_{pp} = \Phi_{pp}(\tau_1)$. $V_p^*(t) = n^{\frac{1}{2}}\hat{B}_p(t) - \hat{B}_p(\tau)\frac{t}{\tau}$. f_α is assigned specific values.

The Cramér-von Mises test rejects at level α if

$$\int_0^{\tau_1} \left(\frac{\hat{V}_p^*(t,\tau_2)/(n\hat{\Phi}_{pp})^{1/2}}{1+\hat{\Gamma}(t)}\right)^2 d\left(\frac{\hat{\Gamma}(t)}{1+\hat{\Gamma}(t)}\right) \geq e_\alpha \tag{4.65}$$

where e_α = the $(1-\alpha)$-quantile in the distribution of $\int_0^{1/2} B^0(u)^2 du$ with B^0 the standard Brownian bridge and $\hat{\Gamma}(t) = \Phi_{pp}(\hat{t})/\Phi_{pp}$.

f_α and e_α in the Kolmogorov-Smirnov and Cramér-von Mises tests are assigned specific values.

4.8 The Semi-parametric additive hazards model

The semi-parametric additive hazard model is an extension of the additive Aalen model. Several researchers have proposed different forms of the semi-parametric additive hazard model. Lin and Ying (1994) proposed a model whereby the hazard function is positive and the regression coefficients are time-invariant. Unlike the additive Aalen model where all its regression coefficients are time-varying, the semi-parametric additive hazards model structure proposed by McKeague and Sasieni (1994) combines the time-varying coefficients of the Aalen model with time-invariant coefficients. We will make use of the semi-parametric structure proposed by McKeague and Sasieni.

The semi-parametric additive hazards model is applicable when one is interested in investigating the nature of covariates effect, as to whether all the regression coefficients are time dependent. In addition, this approach can be used in reducing the degrees of freedom in the case of limited amount of data, thus ensuring precise results. This model verifies if the regression coefficients with time-varying effects are in fact significantly varying with time. The effect of each covariates should be carefully studied to investigate if they are time-varying before reducing to the semi-parametric model. An effective and efficient approach to handle this is to start with a model that allows the effect of all variables to be time-varying. Further investigation should be made to ascertain if some of the effects are well described by constants and then successively simplifying the model as appropriate. The structure of the semi-parametric additive hazard model is presented in equation 4.66 below.

$$\lambda(t) \quad = \quad Y(t)(\mathbf{x}^T(t)\beta(t) + \mathbf{z}^T(t)\gamma), \tag{4.66}$$

where,
$\lambda(t)$ is the hazard rate, $\mathbf{x}(t)$ & $\mathbf{z}(t)$ are time-varying covariates, $Y(t)$ is the at risk indicator, that is the variable that signifies if an individual is at risk or not, $\beta(t)$ is a vector of time-varying regression coefficient and γ is a vector of time-invariant coefficient.

4.8.1 Model estimation

Let $X(t) = (X_1(t), \cdots, X_n(t))^T$ denote the covariates with time-varying coefficients and $Z(t) = (Z_1(t), \cdots, Z_n(t))^T$ denote the covariates whose effects are time-invariant. The martingale decomposition of the counting process is given as

$$\begin{aligned} dN_i(t) &= \lambda(t)dt + dM(t), \\ &= X(t)\beta(t)dt + Z(t)\gamma dt + dM(t). \end{aligned} \tag{4.67}$$

$dB(t)$ and γ will be estimated using the following least squares equations.

$$X^T(t)(dN(t) - \lambda(t)dt) = 0, \tag{4.68}$$

$$\int Z^T(t)(dN(t) - \lambda(t)dt) = 0, \tag{4.69}$$

Solving equation 4.69 for the regression coefficient of the covariates with time-invariant effects, we have

$$d\hat{B}(t) = X^-(t)\{dN(t) - Z(t)\gamma dt\}. \tag{4.70}$$

Substituting the results from equation 4.70 into equation 4.69 and solving for γ as well as integrating gives

$$\hat{\gamma} = \left\{ \int_0^T Z^T(t)H(t)Z(t)dt \right\}^{-1} \int_0^T Z^T(t)H(t)dN(t) \tag{4.71}$$

where $H(t) = (I - X(t)X^-(t))$. Using equation 4.70 with $\hat{\gamma}$ gives

$$\hat{B}(t) = \int_0^t X^-(s)(dN(s) - Z(s)\hat{\gamma}ds). \tag{4.72}$$

Estimation of the variance of $\hat{\beta(t)}$ in the Aalen additive model and $\hat{\beta(t)}$ and γ in the semi-parametric model is based on the regularity conditions being satisfied, the root-n difference between the estimator and true cumulative regression function converges in distribution to a Gaussian martingale. The results obtained follows some functional forms of the strong laws of large numbers. The variances can be estimated using various approaches such as the optimal variation processes and the independent and identically distributed (iid) martingale decomposition. Details on the estimation of the variance terms for $\hat{\beta(t)}$ in the Aalen additive model and $\hat{\beta(t)}$ and γ in the semi-parametric model is found in Martinussen and Scheike (2007).

4.8.2 Inference for the semi-parametric additive hazard model

In this section, approaches for conducting inference in the semi-parametric additive hazards model will be presented. Two hypothesis will be considered:

$$H_{01} : B_p(t) \equiv 0 \quad \text{vs} \quad H_{A1} : B_p(t) \not\equiv 0 \tag{4.73}$$

$$H_{02} : B_p(t) \equiv \gamma_{q+1} t \quad \text{vs} \quad H_{A2} : B_p(t) \not\equiv \gamma_{q+1} t \tag{4.74}$$

It should be noted that the effects of the **z**-covariates in 4.66 are time invariant. Just like in the inference for the Aalen additive hazard model, hypothesis 4.73 can be evaluated using the Hall- Wellner confidence band approach obtained by either the resampling technique or the Khmaladze transformation.

The procedure that has been constructed to test hypothesis 4.74 is the observed test-process and this shows the nature of deviation from the null hypothesis and where it occurs or the modified observed test-process (this takes the variance into account).

Chapter 5

Data Analysis and result interpretation

5.1 Introduction

Chapter 4 covered the methodology for the analysis of time-to-event data, with specific focus on methods for dealing with time-varying effects and for incorporating time-varying covariates. This chapter represent an application and assessment of these methods using the data from the HIV cohort.

5.2 The Cox Regression model

The analysis is based on data for 9,152 individuals followed up for up to 65 months. 818 of them experienced the event of interest, which is death. The Cox model is used to determine the association between the relative hazard of the death and baseline covariates. The structure of the model fitted is shown below:

$$\begin{aligned} \log\left\{\frac{h_i(t)}{h_0(t)}\right\} &= \beta_1 \text{ Age}_i + \beta_2 \text{ Male}_i + \beta_3 \text{ Stage2}_i + \beta_4 \text{ Stage3}_i + \beta_5 \text{ Stage4}_i \\ &+ \beta_6 \text{ lCD40}_i \end{aligned} \tag{5.1}$$

The output from the Cox proportional model is presented below. All the covariates (`age, gender, Stage2, Stage3, Stage4`) have statistically significant associations with the relative risk of death.

Table 5.1: Results of the Cox regression model.

Covariate	Coef.	HR	SE	lower .95	upper .95	p-value
Age	0.03	1.03	0.00	1.022	1.036	<0.001
Male	0.18	1.19	0.07	1.036	1.375	0.014
Stage2	0.38	1.46	0.19	1.013	2.113	0.042
Stage3	1.29	3.62	0.15	2.718	4.826	<0.001
Stage4	1.75	5.75	0.15	4.263	7.751	<0.001
ICD40	-1.18	0.31	0.06	0.275	0.346	<0.001

Table 5.1 presents the result of the Cox regression model. The estimated regression coefficients are given in the second column of Table 5.1 and the estimated relative hazards which is the exponentiated coefficients or hazard ratio is in the third column. It is interpreted as the multiplicative effects on the hazard. The covariates `Stage2`, `Stage3`, `Stage4` are binary with the value 1 for an individual whose HIV/AIDS status is in any of the respective Stages. `Gender`, is also a binary variable with the value 1 when the subject is male and 0 when female.

Holding other covariates constant, the estimated relative hazard in patients whose HIV/AIDS status is Stage4 is 5.75 times that of patients in Stage1. In addition, keeping the value of the other covariates fixed, `Stage3` has 3.62 times an estimated relative hazard compared to `Stage1`. Furthermore, `Stage2` has 1.46 times an estimated relative hazard compared to `Stage1` when every other covariate is kept constant. From the results, we observed a progressive increase in the relative hazard as a patients status changes from `Stage1` to other Stages. This implies that the higher the stage in which an HIV/AIDS patient is in, the higher the relative hazard.

`CD4` is fitted using a log transformation due to its skewness. It has a negative estimated coefficient. CD4 molecules are used to measure normal immunity and HIV infection. Each 1 unit increase in the log of the baseline CD4 is associated with about 69% decrease in a patient's relative hazard of death. Finally, holding other covariates constant, an additional year in age increases the relative hazard of death by a factor of 1.03 on average, that is by, 3%.

5.2.1 Model Checking

In section 4.2.4 we reviewed several model diagnostics that have been proposed to be used in verifying the adequacy of fitted Cox regression models. In

this section, we will be applying the following model diagnostics: Schoenfeld residuals, martingale residuals, Cox-Snell residuals, deviance residuals and dfbeta residuals to investigate the model fitted to the Gugulethu HIV/AIDS dataset.

5.2.1.1 Cox-Snell Residuals

The Cox-Snell residual is used in examining the overall fit of the Cox model. A straight line through the origin with a slope of 1 is expected if the Cox model provides a good fit of the data. Figure 5.1 below presents a plot of the Cox-Snell residuals. The line shows some deviation from the 45-degree reference line thus indicating that the Cox model may not provide a good fit for the model.

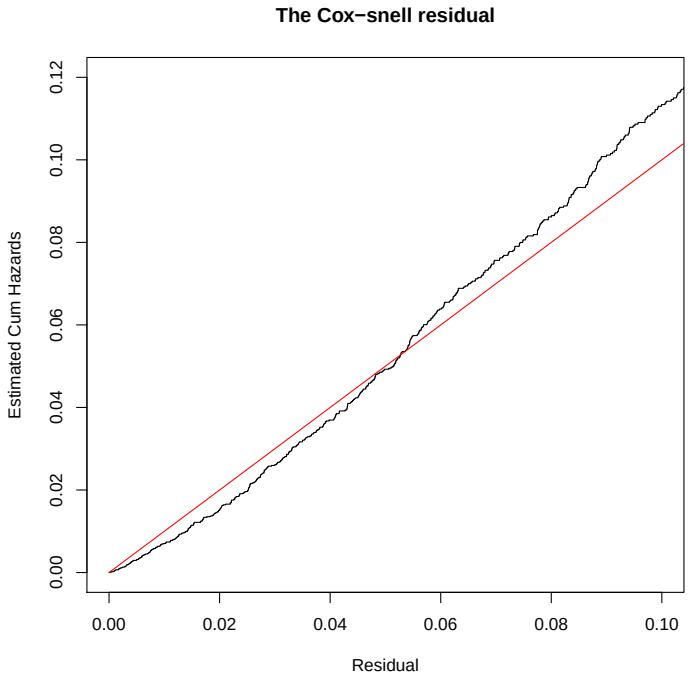


Figure 5.1: Cox-Snell residual plot

The limitation of the Cox-Snell residual is that it is not specific on which part of the model is not fitting properly. Hence, we will consider other model diagnostics to evaluate the goodness of fit for specific parts of the model.

5.2.1.2 Martingale Residuals

The martingale residual as reviewed in section 4.2.6 is used in examining the importance of certain variables in the model and in verifying if the functional forms of the covariates are appropriate after other covariates have been included in the model or if there is a need for transformation. Hence, we plot the martingale residuals against the continuous covariates, if there are patterns in the plots, it suggests that the functional form of the variable is not appropriate.

There are two continuous variables amongst the variables of interest in the Gugulethu HIV/AIDS dataset. They are `Age` and `lCD40`. To investigate if those variables are important in the model, we fit a model excluding each variable one at a time. Figure 5.2 below shows the plot of the residuals of these model versus the excluded variables.

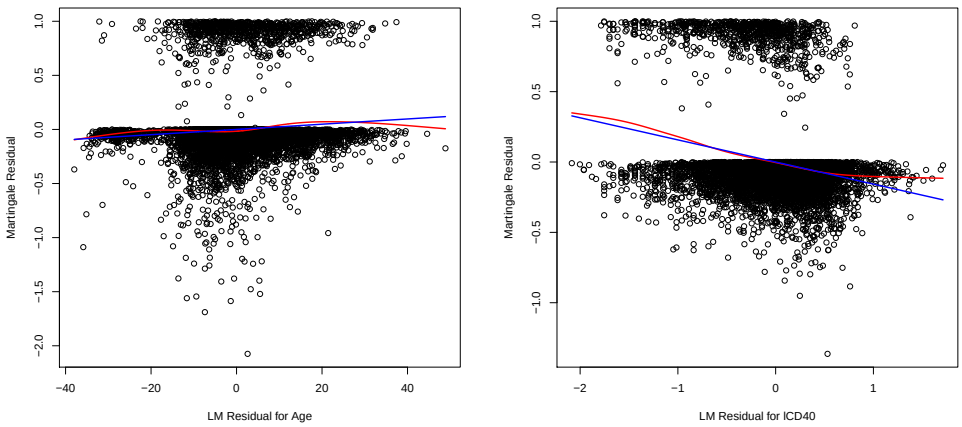


Figure 5.2: Martingale residual plots for Age and lCD40 when not included in the model

The plot of residual versus age shows a positively linear relationship whereas the plot of residual versus the logged baseline CD4 shows a linearly decreasing relationship.

Furthermore, both variables were included into the model to observe their impact and ascertain that the functional form of the variables are suitable. This is shown in Figure 5.3 below.

Both plots in Figure 5.3 provide a good fit. `Age` being brought into the model in a linear form provides a good fit as the loess line wiggles around zero and

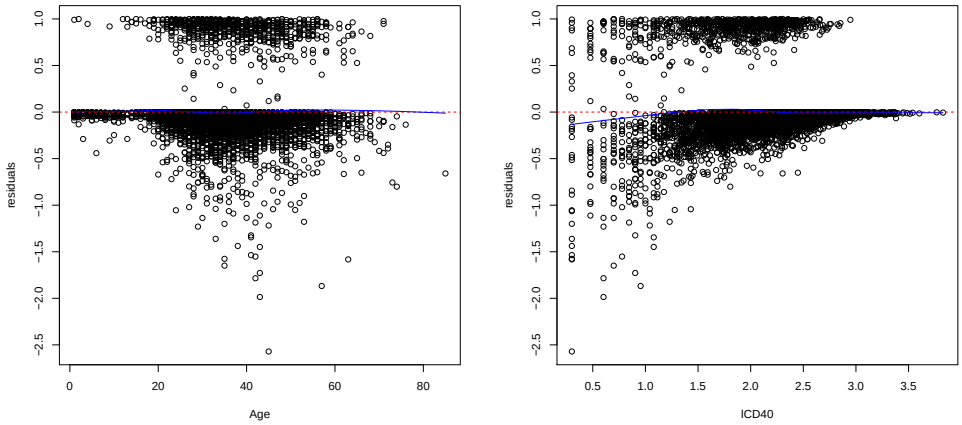


Figure 5.3: Martingale residual plots for Age and lCD40 when they are included in the model.

there seems to be no trend. The plot of residuals versus the logged baseline `CD4` shows some deviation from the horizontal line indicating a possible non-linear association.

5.2.1.3 Deviance Residuals

Recall from section 4.2.7 that the deviance residual is a scaled martingale residual that has a considerably symmetrical distribution about zero. Due to this, outlying observations are considered to values beyond of the range of ± 2, (n.d.) and Braun (2011).

The aim is to determine if there are outlying observations in the Cox model with all the covariates in it. The linear predictor estimates and the deviance residuals are obtained from the model. A plot of deviance residuals versus the linear predictors for the event of interest (death) is presented in Figure 5.4 below. 5.4 shows the characteristic of two cluster of residuals for subjects who did experience the event and those who were censored (residuals < 0). A large number of residual were greater than $+2$ among the subjects who experienced the event of interest, indicating that the Cox PH model did not fit the data well.

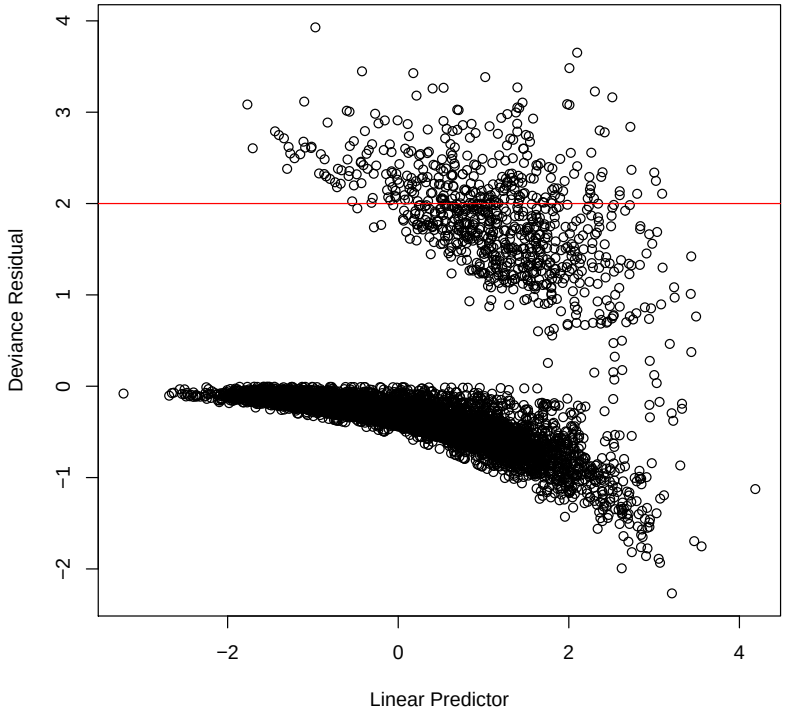
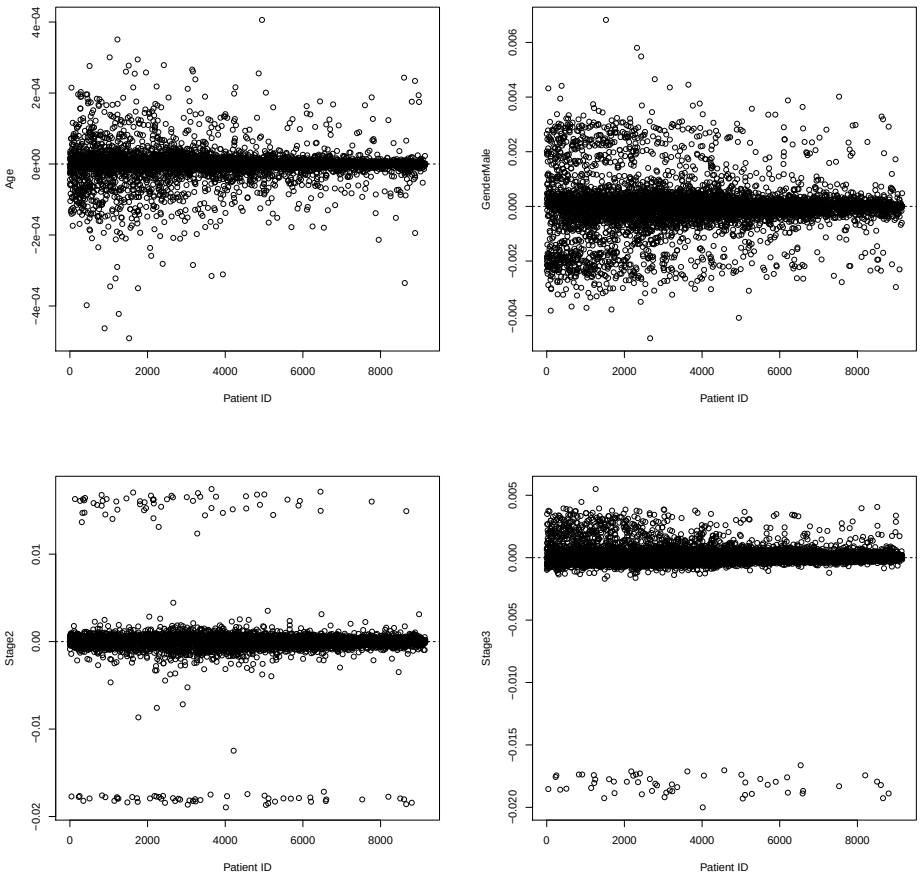


Figure 5.4: Deviance residual plot for the Cox PH model

5.2.1.4 DfBeta Residuals

Just as reviewed in section 4.2.8, Figure 5.5 represents the index plot of the dfbeta for the Cox PH regression model. Comparing the magnitudes of estimated regression coefficients upon deleting each observation in turn with that of the estimated regression coefficients from the whole dataset and dividing them by their standard errors suggests that none of the observations is influential since none of the dfbeta values exceeded ± 0.02. This value was based on the guideline in 4.10 where `n = 9152`. However, clusters or isolated points of potential influence are visible in figure 5.5, especially for the Stage dummy variables. They are in fact the subjects who were in each specific stage.


Age
Patient ID
GenderMale
Patient ID
Stage2
Patient ID
Stage3
Patient ID

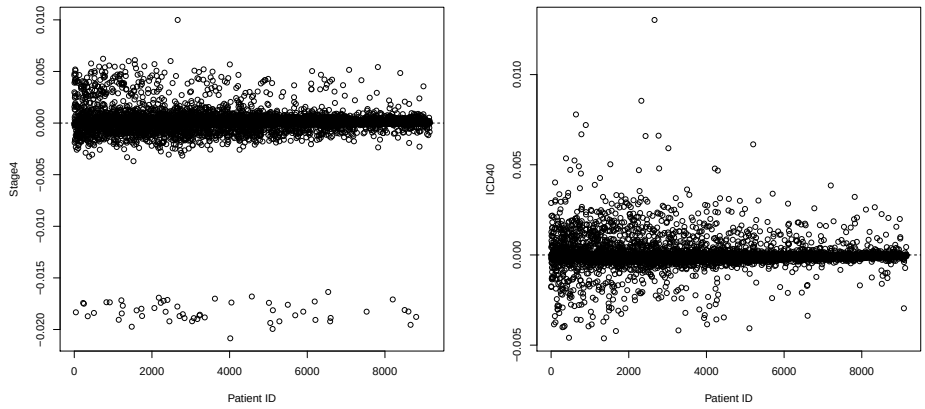


Figure 5.5: Index plots of dfbeta for the Cox PH regression model

5.2.1.5 Schoenfeld Residuals

The Cox proportional hazards model assumes that the effect of covariates on the relative hazard of the event stays constant over time. The Schoenfeld residual is used to test the proportional hazard assumption in Cox model.

For each covariate, scaled Schoenfeld residuals were plotted over time, and tests for a zero slope were performed. Table 5.2 below reports the corresponding p-value for each covariate, as well as the p-value associated with a global test of non-proportionality. The global test suggested strong evidence of non-proportionality ($p < 0.001$). The covariates `Age` and `lCD40` with (p-values = `0.033` and <`0.001`) respectively contributed to the overall non-proportionality of the model. These numerical findings suggest a non constant hazard ratio for these variables. On the other hand, the covariates `Male`, `Stage2`, `Stage3` and `Stage4` were found to follow the proportional hazards assumption with p-values of `0.712, 0.389, 0.972 and 0.480` respectively.

Table 5.2: Test for non-proportionality based on the scaled Schoenfeld residuals for the Cox model.

Covariate	rho	chisq	p-value
Age	0.074	4.556	0.033
Male	0.013	0.136	0.712
Stage2	0.030	0.743	0.389
Stage3	0.001	0.001	0.972
Stage4	-0.024	0.498	0.480
lCD40	0.238	37.498	<0.001
GLOBAL	NA	55.764	<0.001

Residuals help in visualizing the log hazard ratio β over time for each covariate see Figure 5.6 below. The PH assumption is violated if the plot of Schoenfeld residuals against time shows a non-random pattern and if the plotted curves are not roughly constant with time. The graphical approach confirms the results in Table 5.2. The plot of the log hazard ratio for `Gender`, `Stage2`, `Stage3` and `Stage4` confirms that the effect on the hazard of death is constant throughout the follow-up time. The residuals of the covariates that violate the PH assumption (`Age` and `lCD40`) seem to follow some pattern and the curves aren't approximately constant over time. With respect to the baseline `Age` variable, the plot suggests a strong effect over the first four months which tends to diminish over time.

At the early to mid stages, from zero to 30 months, the effect of the logged baseline CD4 count is very negative on the hazard of death, that is, `lCD40` is highly protective. This effect decreases with time and by 60 months it seems that `lCD40` stops being protective from death (zero is the point of equivalence in the log hazard scale). The estimate of `lCD40` having a 69% decreased effect on the relative hazard of death suggested at the end of section 5.2 is misleading because if one regarded this as true, they would under-estimate the strong protective effect of treatment at the early stages and over-estimate the effect at the later stages. Without the plot of the effects, one would not only be missing out on valuable information of the variation of the effect over time, but may also reach incorrect conclusions or recommendations.

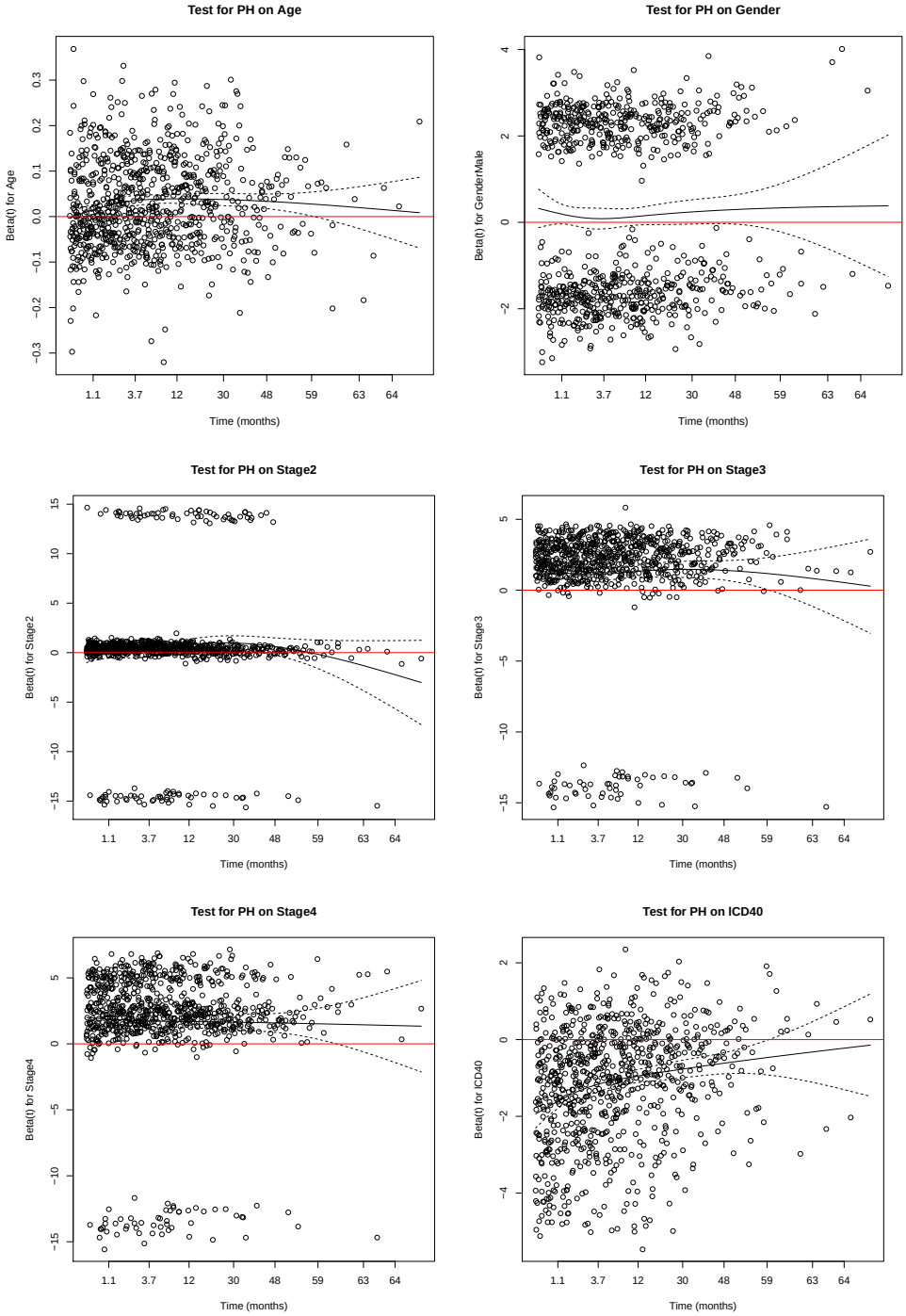


Figure 5.6: Schoenfeld residual plots with 95% pointwise confidence intervals for all the covariates

5.3 Time Varying Effects/ Coefficients

In 4.4, distinctions were made between time-varying coefficients and time-varying covariates. Subsequent sections will illustrate various approaches that have been proposed on how to handle time-varying coefficients since the basic assumption on which the Cox proportional hazard model is based on was violated in the in the HIV/AIDS data.

5.3.1 The stratified Cox model

As mentioned previously in section 4.3.1.1, the stratified Cox model is one of the methods used in dealing with covariates that have time-varying effects. In the Gugulethu HIV/AIDS data we analysed in section 5.2, two covariates were found not to have constant hazard ratios, `Age` and `lCD40`. However, both variables are continuous. To fit the stratified Cox model, the variables that violate the PH assumption ought to be categorical. Therefore, we split `Age` into 4 categories ($0-5$ `years`, $6-19$ `years`, $20-39$ `years and` $40-85$ `years`) and `lCD40` was split into 3 categories ($<\log(200)$, $\log(200-499)$, $\geq \log(500)$). The criteria on which the covariates were split can be found in WHO et al. (2005). After both variables have been categorised, we formed ($4 \times 3 = 12$) `Age` group-by- `lCD40` status combinations. This is presented in the table below.

Table 5.3: `Age` group-by-`lCD40` status combination.

		Age			
		0-5	6-19	20-39	40-85
lCD40	$<\log(200)$	1 (16)	2 (96)	3 (3426)	4 (1238)
	$\log(200-499)$	5 (31)	6 (153)	7 (2618)	8 (92)
	$\geq \log(500)$	9 (137)	10 (78)	11 (459)	12 (808)

These 12 combinations represent the different categories of a single new variable that we stratify on in the stratified Cox model. The values in brackets are the number of subjects in each stratum. In this model we assumed no strata by covariate interaction. Below is the structure of the model that was fitted:

$$h_{gi}(t) = h_{0gi} \exp\Big\{\beta_1 \text{ Male}_i + \beta_2 \text{ Stage2}_i + \beta_3 \text{ Stage3}_i + \beta_4 \text{ Stage4}_i\Big\}, \quad (5.2)$$

where $g = 1, \cdots 12$.

Table 5.4 below presents the results of the stratified Cox regression model.

Table 5.4: Results of the stratified Cox regression model.

Covariate	Coef.	HR	SE	p-value
Male	0.241	1.272	0.071	0.001
Stage2	0.393	1.481	0.187	0.036
Stage3	1.369	3.932	0.145	<0.001
Stage4	1.960	7.096	0.149	<0.001

Despite the many advantages of stratification; simple to implement and handles the issue of non-proportionality, the major drawbacks of stratification of not being able to obtain estimates for the effects of variables used in stratification, diminished precision and power of the estimated regression makes is less attractive. Hence, if we are interested in knowing the effect of `Age` and `lCD40` on the event of interest (death), this approach will not be appropriate as we cannot say anything about them because they weren't obtained. On the other hand, if we are not interested in the variables that were used as the stratifying variables, we will interpret the results in Table 5.4 as having adjusted for other covariates including the stratifying covariates, the hazard ratio for the effect of `Males` is given by `1.27`. This implies that `Males` have `1.27` times the risk of death compared to females. In addition, `Stage2`, `Stage3` and `Stage4` have `1.48`, `3.93` and `7.10` times the hazard of death respectively compared to `Stage1`. We see that while the magnitude of the effects of these variables have changed compared to the Cox PH model presentation in Table 5.1, the general associations/trends are the same.

Another limitation of the model fitted with no strata by covariate interaction assumption is that it is probably not the case for the HIV/AIDS data set as `Age` and `lCD40` would most likely affect other variables.

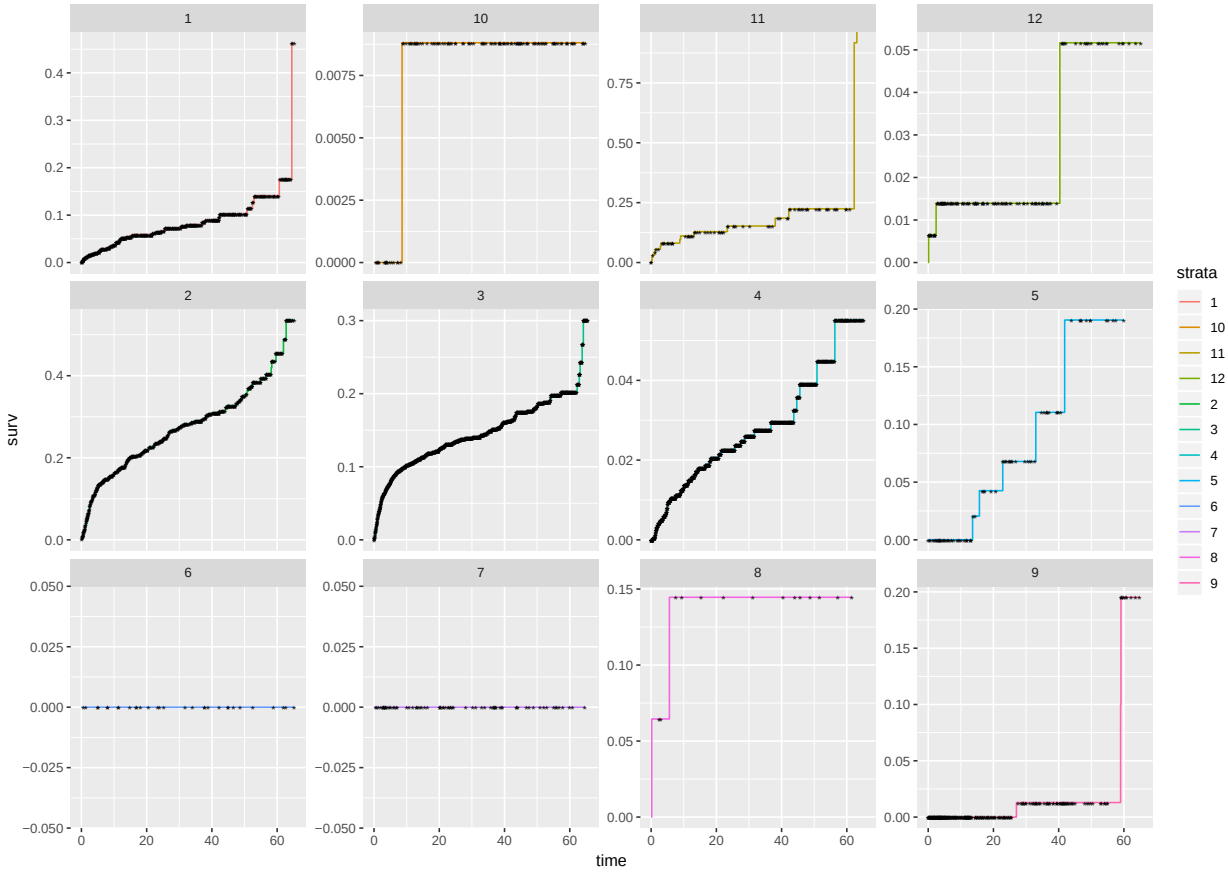


Figure 5.7: Cumulative baseline hazard function for the stratified model.

Figure 5.7 below presents the cumulative baseline hazards plots for the 12 strata representing the combination of three `lCD40` categories namely $< \log(200)$, $< \log(200 - 499)$ and $\geq \log(500)$ and four `Age` categories: `0-5`, `6-19`, `20-39` and `40-85`. Strata 1 are individuals in the 0-5 years age category and have a `lCD40` count of less than log(200). Strata 2 are individuals in the 0-5 years age category and have a `lCD40` count between $\log(200 - 499)$. Strata 3 are individuals in the 0-5 years age category and have a `lCD40` count greater than or equal to log(500).

Strata 4 are individuals in the `6-19` years age category and have a `lCD40` count of less than log(200). Strata 5 are individuals in the `6-19` years age category and have a `lCD40` count between $\log(200 - 499)$. Strata 6 are individuals in the `6-19` years age category and have a `lCD40` count greater than or equal to log(500).

Strata 7 are individuals in the `20-39` years age category and have a `lCD40` count of less than log(200). Strata 8 are individuals in the `20-39` years age category and have a `lCD40` count between $\log(200 - 499)$. Strata 9 are individuals in the `20-39` years age category and have a `lCD40` count greater than or equal to log(500).

Strata 10 are individuals in the `40-85` years age category and have a `lCD40` count of less than log(200). Strata 11 are individuals in the `40-85` years age category and have a `lCD40` count between $\log(200 - 499)$. Strata 12 are individuals in the `40-85` years age category and have a `lCD40` count greater than or equal to log(500).

The plot illustrates different hazards for `lCD40` categories. Strata with lower `lCD40` showed faster acceleration.

5.3.2 Partition the time axis

The partitioning of the follow-up time into sub-intervals is also an alternative approach used in modeling covariates with time-varying effects. Since some covariates violated the PH assumption over the entire follow-up period, we decide the partition the follow-up period and fit the Cox model for each partition. In this model, we partitioned the follow-up time into `0-2 months`, `2-6 months 6-18 months`, `18-40 months` and `40-65 months`. Results obtained after fitting the Cox regression model for each partitioned follow- up time is presented in the table below. The table contains the value of the

hazard ratio, p-values, value of the Schoenfeld residual test and the value of the global test.

Table 5.5: Results obtained from partitioning the time period.

Covariates	Parameters	0-2 months	2-6 months	6-18 months	18-40 months	40-65 months
Age	Hazard ratio	0.99	0.99	0.99	1.00	1.00
	P-value	0.92	0.14	0.77	0.80	0.22
	Schoenfeld test	0.05	0.16	0.16	0.96	0.40
Male	Hazard ratio	1.05	0.84	0.91	1.10	1.00
	P-value	0.53	0.00	0.05	0.02	0.94
	Schoenfeld test	0.65	0.12	0.39	0.34	0.58
Stage2	Hazard ratio	0.75	0.64	0.61	0.76	0.93
	P-value	0.01	< 0.001	< 0.001	< 0.001	0.34
	Schoenfeld test	0.75	0.70	0.21	0.71	0.16
Stage3	Hazard ratio	0.80	0.59	0.66	0.74	0.90
	P-value	0.01	< 0.001	< 0.001	< 0.001	0.09
	Schoenfeld test	0.43	0.02	0.05	0.80	0.93
Stage4	Hazard ratio	1.12	0.71	0.64	0.66	0.80
	P-value	0.25	< 0.001	< 0.001	< 0.001	0.004
	Schoenfeld test	0.32	< 0.001	0.99	0.83	0.20
lCD40	Hazard ratio	0.73	0.91	1.09	1.09	1.17
	P-value	< 0.001	0.11	0.11	0.06	0.004
	Schoenfeld test	0.47	0.44	0.97	0.03	0.14
	Global	0.32	0.001	0.14	0.32	0.14

From Table 5.5 above, the Schoenfeld residual p-value for all the covariates in the various partitioned follow-up time are insignificant excluding the following covariates; `Stage3` and `Stage4` in the `0-2 months` segment and `lCD40` in the `18-40 months` segment. This shows that partitioning the follow-up time increases the probability of covariates in the Cox-model to satisfy the proportional hazards assumption.

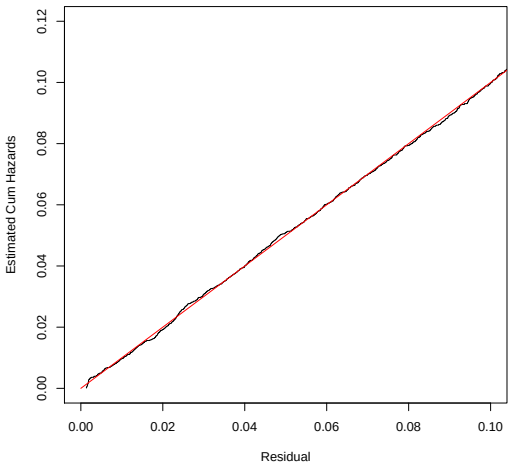
The effect of `Age` on death is constant for at the 5 intervals but Schoenfeld residual test indicates non-PH in the first interval. Thus the non-constant association is very early on. Intervals seem to induce bigger differences in `gender` and `Stage` effects, variables that were deemed to have a constant effect in the Cox model. The estimates for the effect of `lCD4` across the intervals successively illustrate the decreasing association of baseline logged CD4,(`lCD40`) on the relative hazard of death.

Another thing that was observed from the results presented in Table 5.5

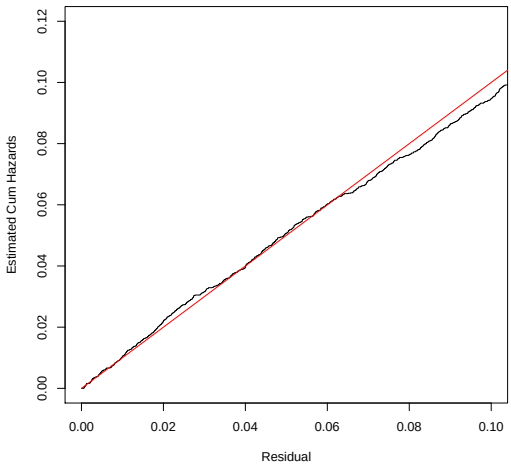
above is that the global test for each partitioned follow-up time met the proportional assumptions except that of `2-6 months` segment.

As discussed in section 4.2.5, the Cox-Snell residual is used in assessing the overall fit of a model. Hence, we show the Cox-Snell residual for models fitted on each one of the segments and compare them to the Cox-Snell residuals for the model fitted on the whole data set where some of the variables violated the PH assumption. A plot close to the 45^o line indicates that the model is good fit.

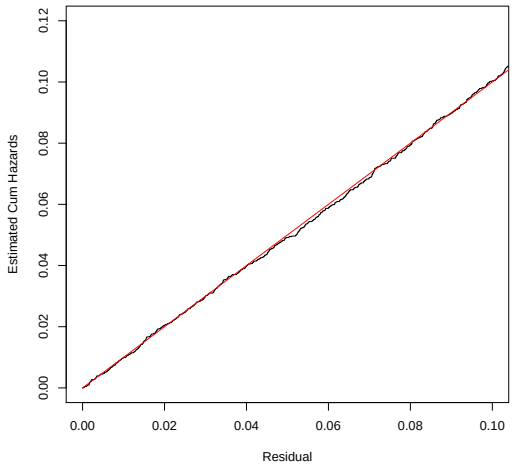
Comparing the plot in Figure 5.1 to those in Figure 5.8 below, we see that the model has a better fit when the time axis is divided into shorter segments compared to when the model is fitted on the whole time interval of study. We make this conclusion because each plots in the different time segments are very close to the 45^o line. This proves that if there are covariates that violate the PH assumption over the entire time interval, splitting the time axis into smaller segments or partitions and fitting the Cox model in each partition is a way to handle time-varying coefficients.



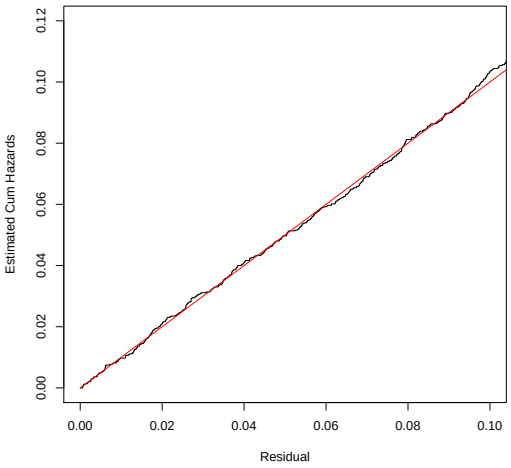
(a) Cox-Snell residual for follow-up time 0-2 months



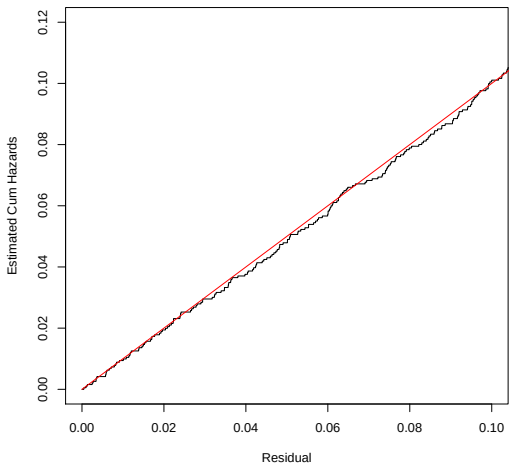
(b) Cox-Snell residual for follow-up time 2-6 months



(c) Cox-Snell residual for follow-up time 6-18 months



(d) Cox-Snell residual for follow-up time 18-40 months



(c) Cox-Snell residual for follow-up time 40-65 months

Figure 5.8: Cox-Snell residuals for the partitioned follow- up times.

5.3.3 Model non proportionality by time-dependent covariates

From section 5.2.1.5 where the PH assumption was tested and the covariates `Age` and `lCD40` violated the assumption. Model non proportionality by time-dependent covariates is another method for handling covariates that violated the PH assumption. In this section, various time functions were used in modelling the covariates, `Age` and `lCD40` since they were found not to remain the same over time. However, `Age` was not modelled with any function of time because its appropriate form was an interaction between the linear form of time and `Age`. This essentially means we are modelling time. Hence, the focus was on `lCD40` only.

Table 5.6 below presents the results obtained from modelling the covariate `lCD40` which violated the PH assumption as an interaction between `lCD40` and a quadratic form of time.

Table 5.6: Results obtained from modelling non- proportionality by time-dependent covariates.

Covariate	Coef.	HR	SE	p-value
Age	0.029	1.029	0.003	<0.001
Male	0.169	1.184	0.072	0.019
Stage2	0.371	1.449	0.188	0.048
Stage3	1.276	3.581	0.146	<0.001
Stage4	1.719	5.578	0.153	<0.001
lCD40	-1.315	0.269	0.064	<0.001
tt(lCD40)	4.5e-04	1.00045	9.3e-05	<0.001

The row in Table 5.6 titled `tt(lCD40)` is the measure of the time-varying effect of the covariate. As mentioned earlier, this was modelled as an interaction term between `lCD40` and a quadratic form of time. Comparing the results from Table 5.6 to that of the Cox regression model in Table 5.1, even though their coefficients have changed a bit, both models have the same association trends.

The results in Table 5.6 are interpreted similarly to the results in Table 5.1 except that for the variable `lCD40`. The `lCD40` effect is measured as $-1.315\ \texttt{lCD40} + 0.00045\ \texttt{lCD40}^2$.

5.4 The extended Cox model

As stated in section 4.5, the extended Cox model is an extension of the Cox proportional hazard regression model used in fitting exogenous time-varying covariates. Hence, while fitting this model, we made use of the time-updated data whereas in the Cox PH model, only baseline observations of covariates are used. This model combines baseline covariates as well as time-varying covariate. For the Gugulethu data the following variables are baseline covariates `Age`, `Male`, `Stage2`, `Stage3` and `Stage4` whereas `Tx` and `lCD4` are time-varying covariates. The model structure of the model fitted is as shown below:

$$\begin{aligned} h_i(t) &= h_0(t)\exp\Big\{\gamma_1 \text{Age}_i + \gamma_2 \text{Male}_i + \gamma_3 \text{Stage2}_i + \gamma_4 \text{Stage3}_i + \gamma_5 \text{Stage4}_i \\ &\quad + \alpha_1 \text{Tx}(t) + \alpha_2 \text{lCD4}(t)\Big\} \end{aligned} \tag{5.3}$$

The results obtained from fitting the model in equation 5.3 is shown in the table below.

Table 5.7: Results of the extended Cox model.

Covariate	Coef.	HR	SE	lower .95	upper .95	p-value
Age	0.03	1.03	0.00	1.027	1.041	<0.000
Male	0.16	1.17	0.07	1.011	1.342	0.031
Stage2	0.24	1.27	0.19	0.881	1.834	0.202
Stage3	0.99	2.70	0.15	2.084	3.689	<0.000
Stage4	1.37	3.92	0.15	3.024	5.488	<0.000
Tx	-0.62	0.54	0.09	0.350	0.485	<0.000
lCD4	-1.63	0.20	0.06	0.184	0.232	<0.000

All the covariates in Table 5.7 except for `Stage2` have a significant effect on the risk of death. In addition, the association trends in the extended Cox are similar to those in the Cox PH model. However, there is a distinction while interpreting the time-varying covariates such as `lCD4` and `Tx` in the extended Cox. At any specific time point `t`, a unit increase in `lCD4` results in about 80% decrease in the risk of death at the same point. For a patient on treatment at a given time `t`, the relative hazard of death decreases by about 46%.

The results below refer to models that were fitted including each of the time-varying covariates one at a time to allow for comparison with Cox PH model and future joint models.

Table 5.8: Results of the extended Cox model.

Covariate	Coef.	HR	SE	lower .95	upper .95	p-value
Age	0.03	1.03	0.00	1.027	1.041	<0.000
Male	0.16	1.17	0.07	1.016	1.348	0.030
Stage2	0.23	1.26	0.19	0.871	1.814	0.222
Stage3	0.99	2.70	0.15	2.029	3.592	<0.000
Stage4	1.35	3.86	0.15	2.867	5.201	<0.000
lCD4	-1.67	0.19	0.06	0.168	0.211	<0.000

The covariate `lCD4` is strongly associated with the risk of death with a unit increase in the logged CD4 resulting to about 81% decrease in the risk of death. From Table 5.8 below, the impact of time-updated `lCD4` is stronger than that of baseline `lCD40` as modelled using the Cox PH model see table 5.1. In the Cox PH model, a unit increase in the logged value of the baseline CD4 resulted in about 69% decrease in the risk of death. This is because, in the Cox PH model only one observation is used in modelling the lCD40, that is the baseline lCD4. However, in the extended Cox model, the lCD4 measurements are updated each time follow up time and this provides more and better information thus leading to more reliable results.

Table 5.9: Results of the extended Cox model.

Covariate	Coef.	HR	SE	lower .95	upper .95	p-value
Age	0.03	1.03	0.00	1.025	1.038	<0.000
Male	0.29	134	0.07	1.160	1.537	<0.000
Stage2	0.44	1.55	0.19	1.073	2.233	0.020
Stage3	0.39	4.00	0.14	3.017	5.313	<0.000
Stage4	1.98	7.27	0.15	5.430	9.728	<0.000
Tx	-0.85	0.43	0.09	0.359	0.508	<0.000

The results from Table 5.9 shows that there is a significant decrease in the risk of death at a given time `t` by about 57% for a patient on treatment at the same time compared to a patient who is not on treatment.

5.5 The Joint model

Unlike the extended Cox model that is suitable for analysing exogenous time-varying covariates. The joint model is appropriate for modelling the association between endogenous time-dependent covariates and time-to-event as discussed in section 4.6. The dataset used in this research has two longitudinal outcome variables: `lCD4` and `Tx` that are endogenous to the process being analysed. Hence, we fit univariate joint models for each of the longitudinal outcome variables.

The results obtained from fitting a univariate joint model for the effect of lCD4 is presented in Table 5.10.

Table 5.10: A univariate joint model for the effect of lCD4.

Survival Outcome				
	PostMean	**95% CI**	**exp(postmean)**	**p-value**
Age	0.0122	0.0012, 0.0231	1.0123	0.022
Male	0.2605	0.0556, 0.4593	1.2976	0.016
Stage2	0.3290	-0.4002, 1.1219	1.3896	0.348
Stage3	1.5876	1.0582, 2.1716	4.8920	<0.000
Stage4	2.0787	1.5555, 2.6665	7.9941	<0.001
lCD4value	-2.0157	-2.205, -1.8098	0.1332	<0.001
Longitudinal Outcome				
	PostMean	**95% CI**	**p-value**	
Intercept	2.1759	2.1685, 2.1830	<0.001	
Time	0.0238	0.0234, 0.0241	<0.001	
sqTime	-0.0003	-0.0003, -0.0003	<0.001	
sigma	0.1776	0.1768, 0.1785	<0.001	

From the results for the survival outcome, presented Table 5.10, the parameter `lCD4value` is the parameter that measures the degree of relationship and association between `lCD4` and the risk of death. Hence, from the results there is a strong relationship between `lCD4` and the risk of death with a unit increase in `lCD4` resulting to about 87% decrease in the relative risk of death.

Comparing the results obtained from the extended Cox which included only one longitudinal outcome `lCD4` that was presented in Table 5.8 above, it was observed that the association between the `lCD4` and the risk of death is stronger in the joint model compared to the extended Cox model.

Table 5.11 presents the results obtained from fitting a univariate joint model for the effect of treatment.

Table 5.11: A univariate joint model for the effect of treatment.

Survival Outcome			
	PostMean	**95% CI**	**exp(postmean)**	**p-value**
Age	0.0126	0.0028, 0.0220	1.0127	0.012
Male	0.4494	0.2263, 0.6687	1.5674	<0.000
Stage2	0.3848	-0.3434, 1.1097	1.4693	0.310
Stage3	1.9432	1.4387, 2.5782	6.9811	<0.000
Stage4	2.8486	2.3584, 3.4401	17.2636	<0.001
Txvalue	-0.2636	-0.3037, -0.2212	0.7683	<0.001
Longitudinal Outcome			
	PostMean	**95% CI**	**p-value**
Intercept	-0.2330	-0.2919, -0.1755	<0.001
Time	0.2815	0.2718, 0.2911	<0.001

From the results presented in Table 5.11, there is a strong relationship between `Tx` and the risk of death with about 77% decrease in the risk of death for those on treatment when compared to those who are not on treatment. Comparing the results obtained in Table 5.11 to that obtained in Table 5.9, the degree of relationship between `Tx` and the risk of death is suppressed in the latter model compared to the former. This shows that the extended Cox model does not handle endogenous time-varying covariates appropriately.

5.6 Extensions of the Joint model

In the standard joint model, we assumed that the an association parameter measures the degree of association between the longitudinal outcome at a specific time point `t` and the risk of event at the same time point. However, this assumption may not be feasible all the time. Hence, in this section, the degree of association between the longitudinal outcome and the risk of event is modelled using different parametrization structure. In addition, a bivariate joint model that measures the degree of association between two longitudinal outcome (`lCD4` and `Tx`) and the risk of death is fitted.

The interaction effect parametrization is used to measure the degree of association between a longitudinal outcome and the risk of event in subgroups.

Hence, Table 5.12 below shows the results from models fitted to measure the degree of association between the longitudinal outcomes lCD4 and the risk of event in the gender subgroup.

Table 5.12: Results obtained from the univariate joint model on the interaction effect in the gender subgroup for lCD4.

Survival Outcome			
	PostMean	**95% CI**	**exp(PostMean)**	**p-value**
Age	0.0132	0.0006, 0.0249	1.0133	0.046
Male	-0.8824	-1.6582, -0.1295	0.04138	0.030
Stage2	0.4331	-0.3458, 1.2065	1.5420	0.274
Stage3	1.6247	1.0288, 2.3008	5.0769	<0.001
Stage4	2.1465	1.5213, 2.8320	8.5549	<0.001
lCD4value	-2.2893	-2.5984, -1.9711	0.1013	<0.001
lCD4value:Male	0.6785	0.2565, 1.1121	1.9709	0.006
Longitudinal Outcome			
	PostMean	**95% CI**	**p-value**	
Intercept	2.1759	2.1684, 2.1833	<0.001	
Time	0.0238	0.0234, 0.0241	<0.001	
sqTime	-0.0003	-0.0003, -0.0003	<0.001	
sigma	0.1776	0.1767, 0.1785	<0.001	

`lCD4` is strongly related with the risk of death. Each unit increase in the current value of `lCD4` is associated with about 90% decrease in the risk of death for those who are females and about 80% decrease in the risk of death for those who are males. From the results, the degree of association between `lCD4` and the risk of death for the gender subgroup is different.

In the time-dependent slopes parametrization structure, the risk of an event depends on the current value and slope of the longitudinal trajectory. Table 5.13 present results from the time-dependent slopes parametrization on the `lCD4` longitudinal trajectory.

Table 5.13: Results obtained from the univariate joint model on the time dependent slopes for lCD4

Survival Outcome				
	PostMean	**95% CI**	**exp(PostMean)**	**p-value**
Age	0.0113	0.0000, 0.0241	1.0114	0.050
Male	0.2613	0.0424, 0.4906	1.2986	0.016
Stage2	0.3307	-0.4352, 1.0841	1.3919	0.390
Stage3	1.5983	1.0537, 2.2689	4.9446	<0.001
Stage4	2.0926	1.5655, 2.7692	8.1060	<0.001
lCD4value	-2.7613	-5.3843, 0.4528	0.0632	0.074
lCD4slope	0.7857	-2.4739, 3.5114	2.1939	0.562
Longitudinal Outcome				
	PostMean	**95% CI**	**p-value**	
Intercept	2.1759	2.1684, 2.1833	<0.001	
Time	0.0238	0.0234, 0.0241	<0.001	
sqTime	-0.0003	-0.0003, -0.0003	<0.001	
sigma	0.1776	0.1767, 0.1785	<0.001	

From Table 5.13, it can be seen observed that the slope of the trajectory is not associated with the risk of death, (p=0.562). However, the log hazard ratio for a unit increase in the current slope of the `lCD4` trajectory is 0.7857 for patients with the same `lCD4` value. This is not significant.

The cumulative effect parametrization is based on the assumption that the risk of an event depends on the cumulative trajectory of the longitudinal outcome. Hence, we are interested in knowing the degree of relationship between the risk of event and the cumulative longitudinal trajectory. Table 5.14 presents the results of modelling the degree of relationship between the risk of death and the cumulative longitudinal trajectory of `lCD4`.

Table 5.14: Results obtained from the univariate joint model on the cumulative effect for lCD4

Survival Outcome					
	PostMean	**95% CI**		**exp(PostMean)**	**p-value**
Age	0.0126	0.0024	0.0234	1.0268	0.018
GenderMale	0.3797	0.1432	0.6003	1.4618	0.004
Stage2	0.5088	-0.2844	1.2437	1.6633	0.194
Stage3	1.9479	1.3926	2.5006	7.0139	<0.001
Stage4	2.6348	2.1023	3.2332	13.9405	<0.001
lCD4Area	-0.3698	-0.4296	-0.3067	0.6909	<0.001
Longitudinal Outcome					
	PostMean	**95% CI**		**p-value**	
Intercept	2.1759	2.1684	2.1833	<0.001	
Time	0.0238	0.0234	0.0241	<0.001	
sqTime	-0.0003	-0.0003	-0.0003	<0.001	
sigma	0.1776	0.1767	0.1785	<0.001	

From Table 5.14 above, there is a strong relationship between `lCD4` and the risk of death with a unit increase in the area under the log CD4 (`lCD4`) longitudinal profile resulting to a decrease in the risk of death.

5.7 Joint models for multiple longitudinal responses.

Joint models for multiple longitudinal responses was introduced in 4.6.5. This model accommodates multivariate longitudinal responses of different distribution types in a unified framework. In the Gugulethu HIV/AIDS data used in this analysis, two longitudinal responses of interest in predicting the risk of death are `lCD4` and `Tx`. The joint models for multiple longitudinal responses was used in fitting the longitudinal responses of interest. `lCD4` and `Tx` were modelled using the gaussian and binomial distributions respectively. The result for the fitted model is presented in Table 5.15 below.

Table 5.15: A bivariate joint model for treatment and CD4.

Survival Outcome				
	PostMean	**95% CI**	**exp(postmean)**	**p-value**
Age	0.0219	0.0071, 0.0370	1.0221	0.010
Male	0.1733	-1.227, 0.4734	1.1892	0.298
Stage2	0.0605	-1.1334, 1.1151	1.0624	0.886
Stage3	1.7041	1.0367, 2.4353	5.4964	<0.000
Stage4	2.4392	1.7548, 3.1901	11.4639	<0.001
lCD4value	-1.9710	-2.27887, -1.6688	0.1393	<0.001
Txvalue	-0.4437	-0.5268, -0.3662	0.6417	<0.001
Longitudinal Outcome				
lCD4	**PostMean**		**95% CI**	**p-value**
Intercept	2.1752	2.1659, 2.1829	<0.001	
Time	0.0237	0.0233, 0.0247	<0.001	
sqTime	-0.0003	-0.0003, -0.0003	<0.001	
sigma	0.1776	0.1767, 0.1786	<0.001	
Longitudinal Outcome				
Tx	**PostMean**	**95% CI**	**p-value**	
Intercept	-0.0227	-0.0657, 0.0185	0.302	
Time	0.2385	0.2334, 0.2436	<0.001	

From Table 5.15, `lCD4value` and `Txvalue` have significant effect on the risk of death with a unit increase in `lCD4` resulting to in about 86% decrease in the risk of death. In addition, there is about 36% decrease in the risk of death for patients on treatment compared to parents who are not on treatment.

Comparing the results in Table 5.7 which is an extended Cox model that estimates the effect of `lCD4` and `Tx` to the risk of death to results presented in Table 5.15 which is a bivariate joint model for `lCD4` and `Tx`. In the latter model, there is a higher percentage in the decrease in the risk of death compared to the former for a unit increase in `lCD4`. On the other hand, the percentage decrease in the risk of death for patients on treatment compared to patients that are not on treatment is higher in the extended Cox model compared to the bivariate joint model. The extended Cox model assumes step-wise updating of values at the observed time points while the joint model allows for a smoothed underlying change. This is more appropriate for the `lCD4` profiles than for the `Tx` profile, since the latter is essentially a discrete process where a patient is either off or on the treatment. Under the assumption that treatment initiation happened between visits, the smoothed `Tx` will be appropriate.

5.8 The Aalen model

In section 4.7, we reviewed the Aalen's additive hazard model as one of the approaches used in handling time-varying covariate coefficients. In this section, we will examine the effect the following covariates of interest; `Age`, `Male`, `Stage2`, `Stage3`, `Stage4`, `treatment` and `log(CD4)` have on survival using the Aalen's additive hazard model. This will be applied to the baseline and time-updated Gugulethu dataset. The model will be fitted using the `timereg` package in R.

5.8.1 Aalen's additive hazard regression model applied to the baseline covariates of the Gugulethu dataset.

In this section, we fit the additive hazards model while applying it to the baseline covariates of the Gugulethu HIV/AIDS dataset. In this framework we assume that all the covariates inclded in the model have non-parametric time-varying effects. The model is fitted with Age, Male, Stage2, Stage3, Stage4 and the logged baseline CD4 count (log(CD40)).

The fitted model is as shown below:

$$\begin{aligned} h(t) &= \beta_0(t) + \beta_1(t)\,\text{Age} + \beta_2(t)\,\text{Male} + \beta_3(t)\,\text{Stage2} + \beta_4(t)\text{Stage3} \\ &+ \beta_5(t)\,\text{Stage4} + \beta_7(t)\,\text{lCD40} \end{aligned} \tag{5.4}$$

Table 5.16: Tests associated with the Aalen's additive hazard regression model on the HIV baseline data.

	Test for non-significant effect		Test for time-invariant effects	
Covariates	**Statistic**	**p-value**	**Statistic**	**p-value**
Intercept	11.20	<0.001	0.209	<0.001
Age	7.40	<0.001	<0.001	0.580
Male	4.02	0.003	0.014	0.275
Stage2	4.71	<0.001	0.012	0.229
Stage3	7.95	<0.001	0.014	0.238
Stage4	9.50	<0.001	0.058	<0.001
lCD40	13.60	<0.001	0.088	<0.001

The output in Table 5.16 presents a two sets of summary statistics namely, the supremum test of significance used to verify if covariates have significant effect on the event of interest and the Kolmogorov-Smirnov test of constant effect used in checking if covariates have time-varying coefficients. All the covariates had significant effects on the risk of death. Two covariates indicated evidence of time-varying effects. They are `Stage4` and `lCD40` whereas the other covariates (`Age`, `Male`, `Stage2`, `Stage3`) appeared to have constant effect with time.

The plots in Figure 5.9 shows the estimated cumulative regression coefficients against time with 95% pointwise confidence intervals. This figure confirms the statistics presented in Table 5.16 as all the variables with constant time effect had their estimated cumulative function being approximately straight lines showing a constant slope, whereas those with time varying effects change slopes, often fast initially and slows down later. This is seen from figure 5.9 since all the confidence bands for all the variables other than `lCD40` and `Stage4` cross zero. `lCD40` has a negative slope indicating that an increase in these covariates results in a decrease in the relative hazard of death. On the other hand, `Stage4` has a positive slope indicating that subjects with Stage4 HIV/AIDS have an increased risk of death compared to subjects with Stage1 HIV/AIDS.

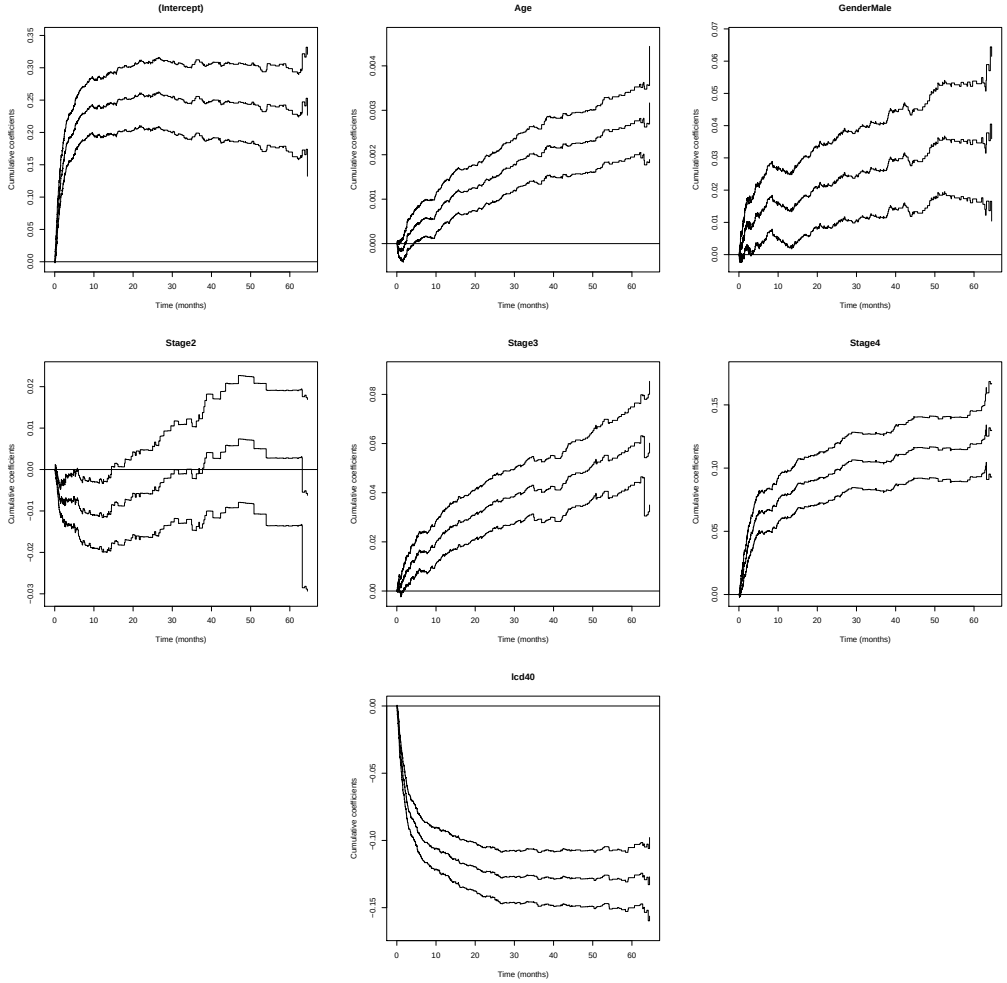


Figure 5.9: Estimated cumulative regression functions with 95% pointwise confidence intervals based on Aalen's additive model - baseline HIV data.

5.8.2 Inference for additive hazard models

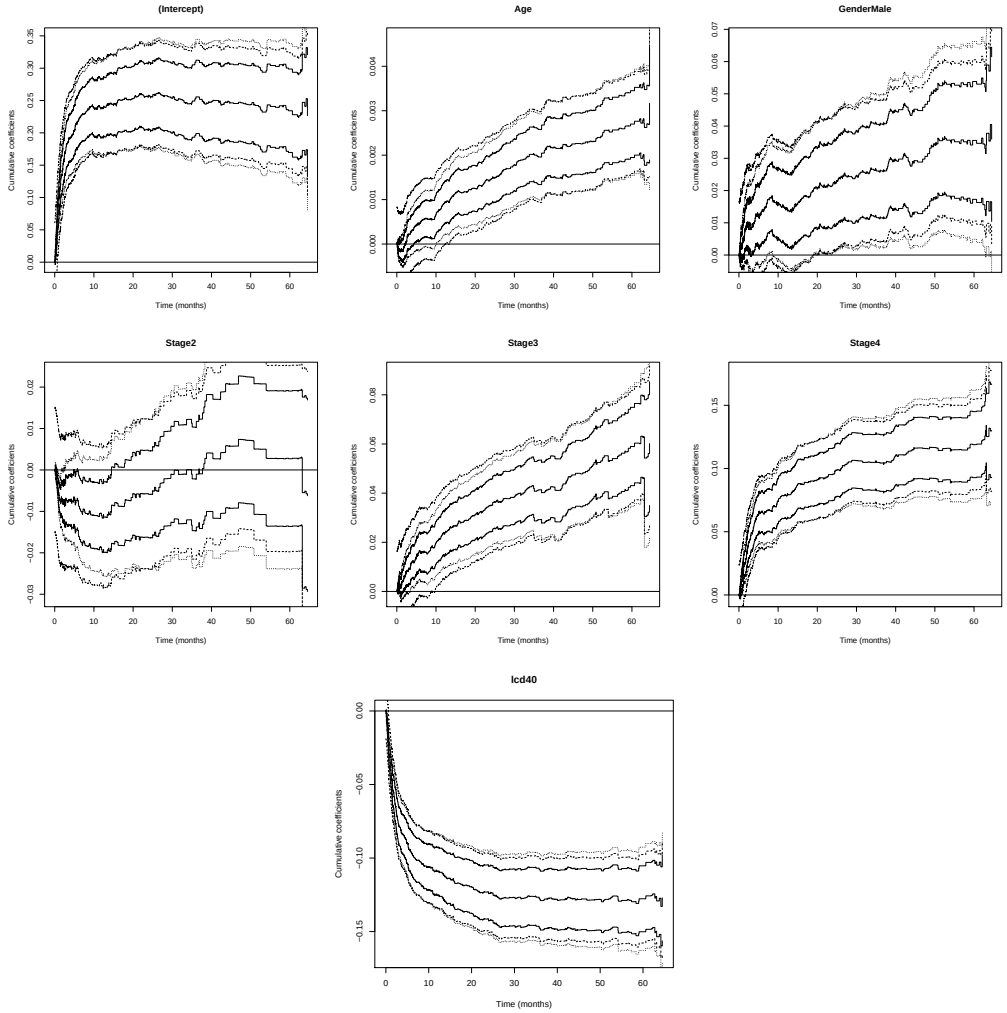


Figure 5.10: Estimated cumulative regression functions with 95% pointwise confidence intervals(solid lines), Hall-Wellner bands (broken lines) and simulation based bands(dotted lines) - baseline Gugulethu HIV/AIDS data.

The pointwise confidence intervals in figure 5.10 will be used for comparison. The zero-function is observed to be outside the bands for both the Hall-Wellner band and the simulation based band showing that the time-varying effects of `Stage4` and `lCD40` are significant.

Considering the estimates depicted in Figure 5.9 it is not clear based on the pointwise confidence interval shown there which of the components that

have time-varying effects. Figure 5.10 shows three sets of confidence bands and more clearly illustrates that only the confidence bands for `Stage4` and `lCD40` excludes zero. However, the bands do not reflect the uncertainty of the estimate of the constant effect. It is seen that the intercept, Stage4 and lCD40 do have effects that vary with time. Figure 5.11 below are plots of the processes with 50 random realizations under the null of constant effects for all the variables.

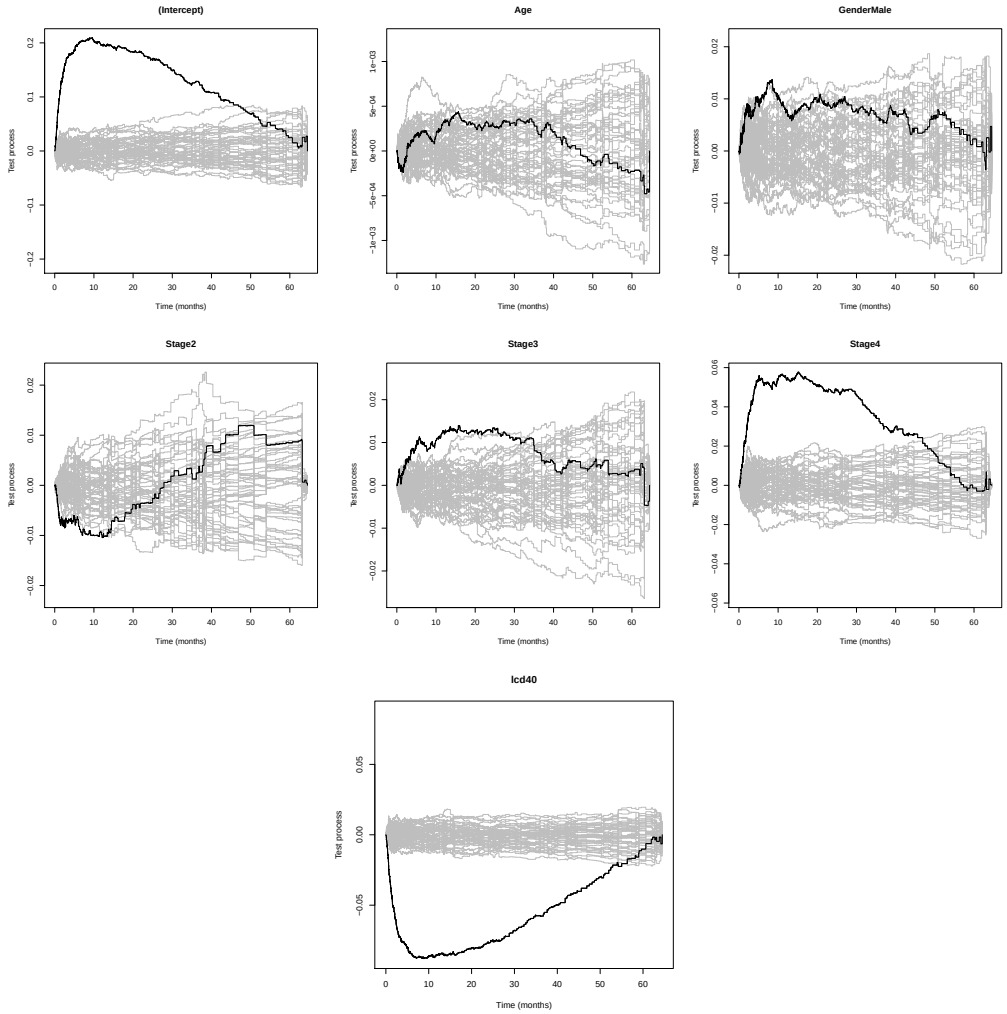


Figure 5.11: Test processes for testing constant effects with 50 simulated processes under the null - baseline HIV data.

In Figure 5.11, the grey lines are the 50 simulated processes under the null hypothesis used in testing for constant effects of the covariates. When the black solid line is outside the grey lines, that covariate is said to have a time-

varying effect. Hence, the `intercept`, `Stage4` and `lcd40` have time-varying effect while the other covariates have performances consistent with the null of time-invariant effect. As seen from the simulation based results in Table 5.16, the p-value was < 0.001 for the `intercept`, 0.580 for `Age`, 0.275 for `Male`, 0.229 for `Stage2`, 0.238 for `Stage3`, < 0.001 for `Stage4`, < 0.001 for `lCD40`. Thus, we reject the hypothesis of constant effect for `intercept`, `Stage4` and `lCD40`.

The results presented in Table 5.16 was obtained after fitting model specified in equation 5.4 and it showed that not all the covariates had time-varying effects. Hence, a number of successive test will be performed in order to achieve a parsimonous and best fitted model. First, we note that the `Male` does not seem to have a time-varying effect (p=0.580, using the test of time-invariant effects), which, as mentioned before, is consistent with the cumulative estimate being approximately a straight line in Figure 5.9. Fitting the model with the effect of `Male` being constant (output can be found in the Appendix A) we found that the effects of `Age`, `Stage2` and `Stage3` were also constant (p=0.564, 0.174, 0.143) respectively. Secondly, the model with both `Male` and `Age` having constant effects showed that the effect of `Stage2` and `Stage3` were constant also (p=0.197 and 0.191). Thirdly, a model with `Male`, `Age` and `Stage2` having constant effects was fitted. Finally, a reduced semi-parametric model was fitted with the covariates `Male`, `Age`, `Stage2` and `Stage3` having constant effects. The result is presented below.

Table 5.17: Test associated with the semi-parametric hazard regression model on the HIV baseline data.

	Tests for non-significant effect		**Test for time-invariant effects**	
Covariates	**Statistic**	**p-value**	**Statistic**	**p-value**
Intercept	14.00	<0.001	0.227	<0.001
Stage4	9.43	<0.001	0.053	<0.001
lcd40	14.20	<0.001	0.092	<0.001
Parametric terms	**Coeff**	**SE**	**95% CI**	**p-value**
const(Age)	5.13e-05	7.93e-06	3.58e-05, 6.68e-05	<0.001
const(Male)	7.63e-04	1.91e-04	3.89e-04, 1.14e-03	<0.001
const(Stage2)	-2.44e-05	1.67e-04	-3.52e-04, 3.03e-04	0.886
const(Stage3)	1.19e-03	1.68e-04	8.61e-04, 1.52e-03	<0.001

The fit of the semi-parametric model shows that `Stage4` and `lCD40` have

effects that are significantly time-varying (p<0.001 for all of them) using the test of time invariant effects. The effect of the other covariates is described by their constant effects. What is interesting to observe is how small the effects of the constant terms are. They are of course now additive effects as oppose to multiplicative effects. The time-varying effects are illustrated in figure 5.12 and are as estimated in previous models.

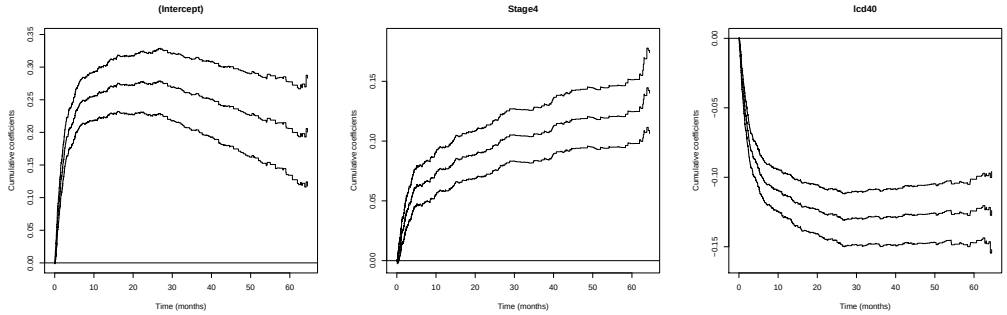


Figure 5.12: Estimated cumulative regression functions with 95% pointwise confidence intervals based on Aalen's semi-parametric additive model - baseline HIV data.

5.8.3 Aalen's additive regression model applied to the time-updated Gugulethu data

Similar procedures applied to the Gugulethu HIV/AIDS baseline dataset will be carried out here except that they will be applied to different datasets. Here, the HIV time updated data will be used. The aim is to examine the effect of each covariate on survival. I commenced with the additive hazards model framework where all components of the model have nonparametric time-varying effects.

The fitted model is as shown below:

$$h(t) = \beta_0(t) + \beta_1(t)\,\text{Age} + \beta_2(t)\,\text{Male} + \beta_3(t)\,\text{Stage2} + \beta_4(t)\text{Stage3} + \beta_5(t)\,\text{Stage4} + \beta_6(t)\,\text{Tx(t)} + \beta_7(t)\,\text{lCD4(t)} \quad (5.5)$$

Results from the tests for non-significant effects and time-invariant effect are presented in Table 5.18 on the time updated HIV data. All the covariates excluding the covariate `Male` had significant effects on the risk of death. On

Table 5.18: Tests associated with the Aalen's additive hazard regression model on the HIV time updated data.

	Test for non-significant effects		Test for time-invariant effects	
Covariates	**Statistic**	**p-value**	**Statistic**	**p-value**
Intercept	12.30	<0.001	0.179	0.062
Age	3.37	0.019	0.001	0.620
Male	2.89	0.084	0.013	0.265
Stage2	4.83	<0.001	0.012	0.239
Stage3	7.82	<0.001	0.013	0.280
Stage4	9.70	<0.001	0.057	<0.000
Tx	5.75	<0.001	0.066	0.339
lCD4	13.60	<0.001	0.074	<0.000

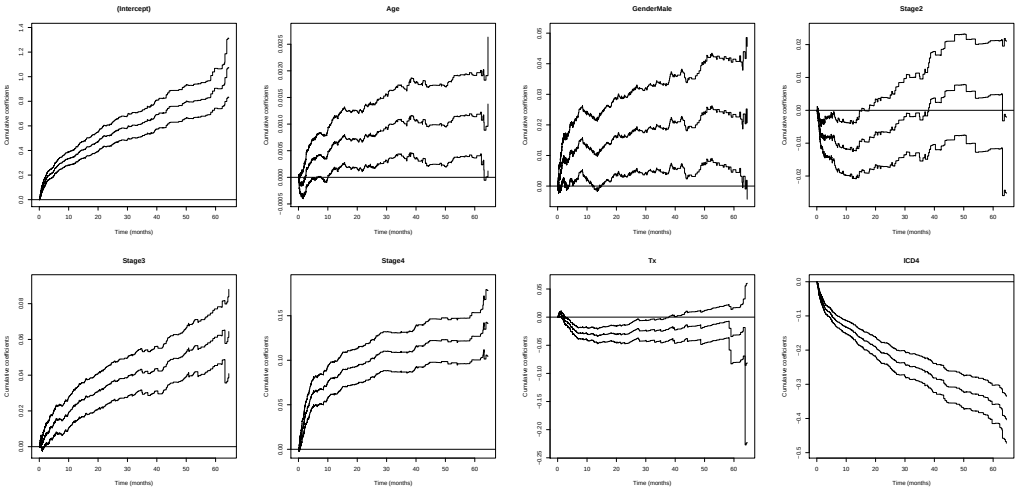


Figure 5.13: Estimated cumulative regression functions with 95% pointwise confidence intervals based on Aalen's additive model - time-updated HIV data.

the other hand, the covariates `Stage4` and `lCD4` indicate evidence of time-varying effects while other covariates (`Age`, `GenderMale`, `Stage2`, `Stage3`, `Tx`) appeared to have effects that is constant with time.

Figure 5.13 above is a plot showing the estimated cumulative regression coefficients with 95% pointwise confidence intervals. A careful look at this figure shows some estimated cumulative functions with approximately straight lines and others are not. They covariates with their estimated cumulative functions being approximately straight lines are `Age`, `Male`, `Stage2`, `Stage3`, `Tx`.

These covariates are said to have constant time effects while the other covariates `Stage4` and `lCD4` are said to have time varying effects since their estimated cumulative functions are not straight lines. This conforms with the statistics presented in Table 5.18.

5.8.4 Inference for additive hazard models

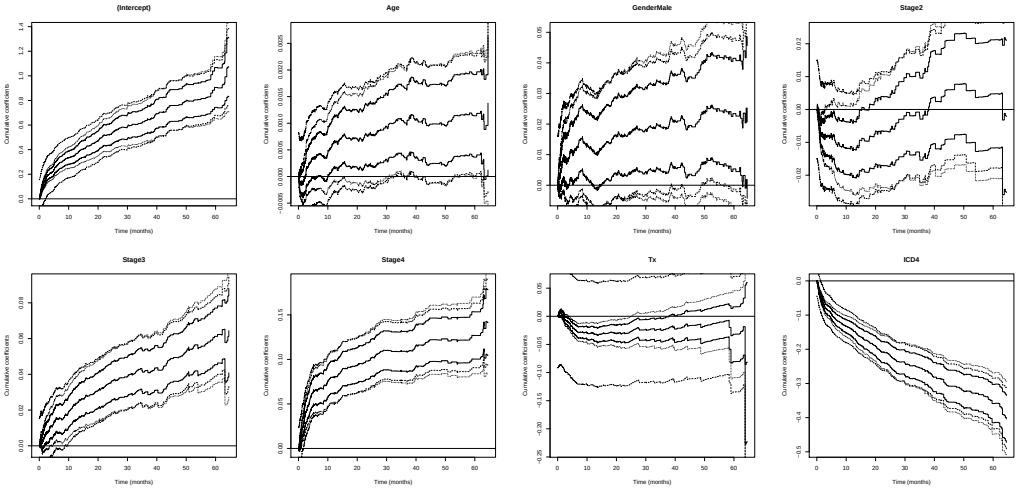


Figure 5.14: Estimated cumulative regression functions with 95% pointwise confidence intervals(solid lines), Hall-Wellner bands (broken lines) and simulation based bands(dotted lines) - time-updated HIV data.

Figure 5.14 above will be used to draw inference on the Aalen additive hazard model fitted to the updated HIV data. The pointwise confidence interval will be used in making comparison. `Stage4` and `lCD4` had their zero-functions outside the bands for both the Hall-Wellner band and the simulation based band showing that the time-varying effects are significant. Contrariwise, `Age`, `GenderMale`, `Stage2`, `Stage3` and `Tx` zero-functions inside the bands for both the Hall-Wellner band and the simulation based band depeting their time-invariant effects.

From Figure 5.15 we can reject the hypothesis of constant effect for `Stage4` and `lCD4` while the other covariates have performances consistent with the null of time-invariant effect.

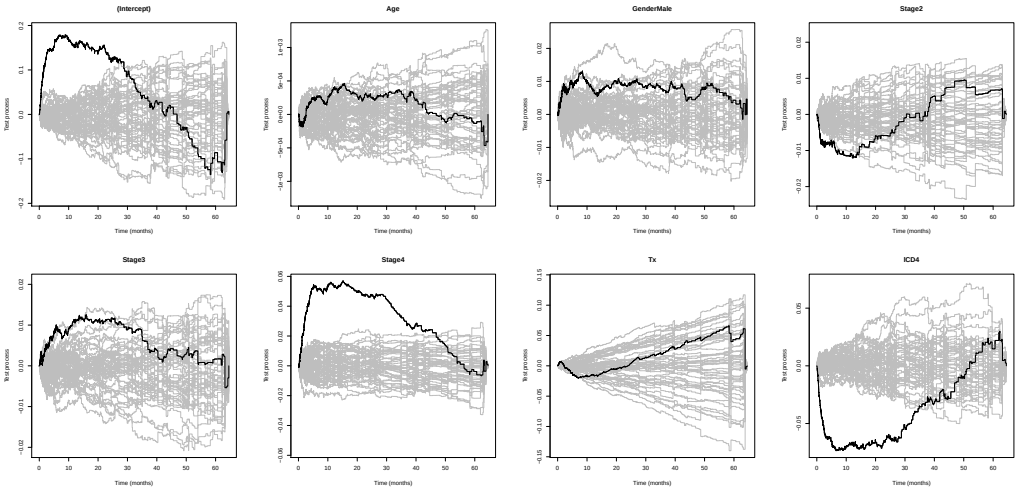


Figure 5.15: Test processes for testing constant effects with 50 simulated processes under the null - time updated HIV data.

At this point, it is obvious that some of the variables are time-invariant while others have time-varying effects. Hence, we will move ahead to simplify the model as shown below. First, we note that the `Age` does not seem to have a time-varying effect (p=0.620, using the test of time-invariant effects), which, as mentioned before, is consistent with the cumulative estimate being approximately a straight line in Figure 5.13. Fitting the model with the effect of `Age` being constant (output can be found in the Appendix) we found that the effects of `Male`, `Stage2`, `Stage3`and `Tx` were also constant (p=0.217, 0.229, 0.308, 0.377) respectively. Secondly, the model where both `Age` and `Tx` have constant effects showed that the effect of `Male`, `Stage2` and `Stage3` were constant also (p=0.172,0.190 and 0.263). Thirdly, a model with `Age`, `Tx` and `Stage3` having constant effects was fitted and `Male` and `Stage2` was found to have constant effects (p= 0.121 and < 0.001) respectively. Fourthly, a model with `Age`, `Tx`, `Stage3` and `Male` having constant effects was fitted and `Stage2` was found to have a time-varying effect (p= < 0.001). Finally, a reduced semi-parametric model was fitted with the covariates `Male`, `Tx`, `Age`, `Stage2` and `Stage3` having constant effects. The result is presented below.

Table 5.19: Test associated with the semi-parametric hazard regression model on the HIV time updated data.

	Tests for non-significant effect		Test for time-invariant effects	
Covariates	**Statistic**	**p-value**	**Statistic**	**p-value**
Intercept	14.30	<0.001	0.189	0.003
Stage4	9.92	<0.001	0.054	<0.001
lCD4	14.10	<0.001	0.079	0.001
Parametric terms				
	Coeff	**SE**	**95% CI**	**p-value**
const(Age)	2.66e-05	8.30e-06	1.03e-05, 4.29e-05	0.001
const(Male)	5.77e-04	1.91e-04	2.03e-04, 9.51e-04	0.004
const(Stage2)	3.61e-06	1.67e-04	-3.24e-04, 3.31e-04	0.983
const(Stage3)	1.25e-03	1.66e-04	9.25e-04, 1.58e-03	<0.001
const(Tx)	-1.85e-03	3.33e-04	-2.50e-03, -1.20e-03	<0.001

The fit of the semi-parametric model shows that `Stage4` and `lCD4` have effects that are significantly time-varying (p <0.001 for both covariates) using the test of time invariant effects. Increasing `Age` by a year, results in a significant increase in the risk of death by 2.66e-05 (se=6.57e-06). In addition, `Males` have significant increased intensity of 5.77e-04 (se=1.91e-04) when compared to females. Patients with HIV/AIDS status in Stage2 and `Stage3` have an estimated increased risk of death of 3.61e-06 and 1.25e-03 respectively compared to patients with HIV/AIDS status in `Stage1`. Furthermore, `Tx` has a reduced estimated risk of death of -1.85e-03 (se=3.33e-04).

The effect of the time-updated `lCD4` values was estimated to be time-varying too. Hence, even when using the time-updated current `lCD4`, the impact is larger during early follow-up period compared to later follow-up period.

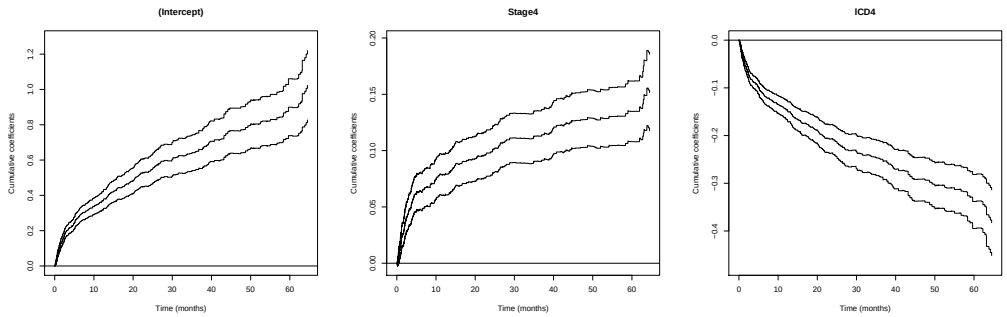


Figure 5.16: Estimated cumulative regression functions with 95% pointwise confidence intervals based on Aalen's semi-parametric additive model-time updated HIV data.

Chapter 6

Conclusions

In this dissertation, various approaches for handling time-varying coefficients and covariates in longitudinal and survival models were investigated. Each model discussed in Chapter 4 was applied to the Gugulethu HIV/AIDS data.

This research commenced by fitting the popular model used in analysing survival data, the Cox PH model and checking the PH assumption. Covariates that violate the PH assumption are said to have time-varying effects/coefficients. This led to fitting models that handle time-varying coefficients. They include stratified Cox model, partitioning the time period and modelling non-proportionality by time-dependent covariates.

The limitation of the stratified Cox model is that estimates are not obtained for the stratifying variable. Hence, the magnitude of their effect is unknown for the stratified variable. Partitioning of the time period seem to give more reliable results as different Cox models are fitted on smaller segments of the follow-up period. However, this approach demands a motivation for which a particular time interval is chosen. Another approach is to model non-proportionality by time-dependent covariates but the limitation of this method is that the validity of the results obtained from the model depends on the appropriateness of the selected function of time.

Furthermore, models were fitted to incorporate covariates that were time-varying. These include the extended Cox model, the joint model and its extensions, the Aalen model and the semi-parametric additive model. The extended Cox model handled exogenous time-varying covariates adequately but performed poorly in modelling endogenous time-varying covariates. Hence, the joint model and its extensions were used instead as they gave more accurate results for the estimate of effects of endogenous time-varying variables. The results from the extensions of the joint model were more precise compared to the standard joint model. This is because, the assumptions for each

of those extensions handled specific situations that were not incorporated in the standard joint model. In contrast to the models that have been fitted, the Aalen model is an additive model that analysis survival data with time-varying covariates as well as time-varying coefficients.

The semi-parametric additive model gives more precise results when compared to the Aalen model because the former assumes that all the covariates have time-varying effects which is not always the case while the former takes into cognisance the fact that some variables could have constant effect. The beauty of the Aalen model and the semi-parametric additive model is that the effect of variables with both time varying effects and time varying covariates can be modelled and measured.

Other models that could be explored include the Cox-Aalen model (Martinussen and Scheike, 2007) which is an alternative model that combines Cox proportional hazards model for covariates with constant effects and the Aalen additive model for time-varying effects in a single model. The dynamic path model (Gamborg et al., 2011) which combines path analysis and survival analysis in order to estimate the effect of covariates on the risk of an event.

Bibliography

(2008). Springer Science & Business Media.

(n.d.).

Aalen, O. (1975). *Statistical inference for a family of counting process*, PhD thesis, Ph. D. Thesis, University of California, Berkeley.

Aalen, O. (1978). Nonparametric inference for a family of counting processes, *The Annals of Statistics* pp. 701–726.

Altshuler, B. (1970). Theory for the measurement of competing risks in animal experiments, *Mathematical Biosciences* **6**: 1–11.

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (2012). *Statistical models based on counting processes*, Springer Science & Business Media.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study, *The annals of statistics* pp. 1100–1120.

BAŞAR, E. (2017). Aalens additive, cox proportional hazards and the cox-aalen model: Application to kidney transplant data, *Sains Malaysiana* **46**(3): 469–476.

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**(1): 20.

Böhmer, P. (1912). Theorie der unabhängigen wahrscheinlichkeiten, *Rapports Memoires et Proces verbaux de Septieme Congres International dActuaires Amsterdam*, Vol. 2, pp. 327–343.

Braun, W. J. (2011). Ch. 6 leverage and influence diagnostics.
URL: *http://www.stats.uwo.ca/faculty/braun/ss3859/notes/Chapter6old/ch6notes.pdf*

Cleves, M. (2008). *An introduction to survival analysis using Stata*, Stata Press.

Collett, D. (2015). *Modelling survival data in medical research*, CRC press.

Cox, D. (1972). Regression models and life tables (with discussion) j roy stat soc. 1972, *B34* pp. 187–220.

David, C. R. et al. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society* **34**: 187–220.

Dobson, A. J. and Barnett, A. (2008). *An introduction to generalized linear models*, CRC press.

Fitrianto, A. and Jiin, R. L. T. (2013). Several types of residuals in cox regression model: an empirical study, *Int J Math Anal* **7**: 2645–54.

Gamborg, M., Jensen, G. B., Sørensen, T. I. and Andersen, P. K. (2011). Dynamic path analysis in life-course epidemiology, *American journal of epidemiology* **173**(10): 1131–1139.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals, *Biometrika* **81**(3): 515–526.

Hosmer, D. W., Lemeshow, S. and May, S. (2011). *Applied survival analysis*, Wiley Blackwell.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*, Vol. 360, John Wiley & Sons.

Kleinbaum, D. G. and Klein, M. (2010). *Survival analysis*, Vol. 3, Springer.

Lawn, S. D., Little, F., Bekker, L.-G., Kaplan, R., Campbel, E., Orrell, C. and Wood, R. (2009). Changing mortality risk associated with cd4 cell response to antiretroviral therapy in south africa, *AIDS (London, England)* **23**(3): 335.

Lin, D. and Ying, Z. (1994). Semiparametric analysis of the additive risk model, *Biometrika* **81**(1): 61–71.

Martinussen, T. and Scheike, T. H. (2007). *Dynamic regression models for survival data*, Springer Science & Business Media.

McKeague, I. W. and Sasieni, P. D. (1994). A partly parametric additive risk model, *Biometrika* **81**(3): 501–514.

Nelson, W. (1969). Hazard plotting for incomplete failure data, *Journal of Quality Technology* **1**(1): 27–52.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data, *Technometrics* **14**(4): 945–966.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*, CRC Press.

Rizopoulos, D. (2018). Multivariate joint models.
URL: *http://www.drizopoulos.com/vignettes/multivariate%20joint%20models*

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model, *Biometrika* **69**(1): 239–241.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling survival data: Extending the Cox model*, Springer.

Therneau, T. M., Grambsch, P. M. and Fleming, T. R. (1990). Martingale-based residuals for survival models, *Biometrika* **77**(1): 147–160.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview, *Statistica Sinica* pp. 809–834.

WHO et al. (2005). Interim who: clinical staging of hiv/aids and hiv/aids case definitions for surveillance: African region, *Interim WHO: clinical staging of HIV/Aids and HIV/Aids case definitions for surveillance: African region*.

Appendix A

Additional Tables and Plots

This chapter contains additional results from the models fitted in this research presented in tables as well as plots that were generated.

Table A.1: Test associated with the semi-parametric hazard regression model on the HIV baseline data with the effect of the covariate `male` fitted to be time-invariant.

	Tests for non-significant effect		Test for time-invariant effects	
Covariates	**Statistic**	**p-value**	**Statistic**	**p-value**
Intercept	11.30	<0.001	0.210	<0.001
Age	7.37	<0.001	0.001	0.564
Stage2	4.50	<0.001	0.013	0.174
Stage3	7.93	<0.001	0.015	0.143
Stage4	9.63	<0.001	0.059	<0.001
lcd40	13.60	<0.001	0.088	<0.001
Parametric terms	**Coeff**	**SE**	**95% CI**	**p-value**
const(Male)	7.41e-04	1.91e-04	3.67e-04, 1.12e-03	<0.001

Table A.2: Test associated with the semi-parametric hazard regression model on the HIV baseline data with the effect of the covariates `male` and `Age` fitted to be time-invariant.

	Tests for non-significant effect		Test for time-invariant effects	
Covariates	**Statistic**	**p-value**	**Statistic**	**p-value**
Intercept	13.30	<0.001	0.218	<0.001
Stage2	4.62	<0.001	0.012	0.197
Stage3	7.95	<0.001	0.015	0.191
Stage4	9.64	<0.001	0.059	<0.001
lcd40	13.90	<0.001	0.089	<0.001
Parametric terms	**Coeff**	**SE**	**95% CI**	**p-value**
const(Age)	5.08e-05	7.92e-06	3.53e-05, 6.63e-05	<0.001
const(Male)	7.52e-04	1.91e-04	3.78e-04, 1.13e-03	<0.001

Table A.3: Test associated with the semi-parametric hazard regression model on the HIV baseline data with the effect of the covariates `male`, `Age` and `Stage2` fitted to be time-invariant.

	Tests for non-significant effect		Test for time-invariant effects	
Covariates	**Statistic**	**p-value**	**Statistic**	**p-value**
Intercept	13.30	<0.001	0.214	<0.001
Stage3	7.20	<0.001	0.018	0.010
Stage4	9.52	<0.001	0.062	<0.001
lcd40	13.90	<0.001	0.089	<0.001
Parametric terms	**Coeff**	**SE**	**95% CI**	**p-value**
const(Age)	5.06e-05	7.92e-06	3.51e-05, 6.63e-05	<0.001
const(Male)	7.47e-04	1.91e-04	3.73e-04, 1.12e-03	<0.001
const(Stage2)	-1.15e-04	1.67e-04	-4.42e-04, 2.12e-04	<0.001

Table A.4: Test associated with the semi-parametric hazard regression model on the HIV time updated data with the effect of the covariate `Age` fitted to be time-invariant.

	Tests for non-significant effect		Test for time-invariant effects	
Covariates	**Statistic**	**p-value**	**Statistic**	**p-value**
Intercept	13.90	<0.001	0.192	0.045
Male	3.00	0.071	0.014	0.217
Stage2	4.84	<0.001	0.012	0.229
Stage3	7.80	<0.001	0.013	0.308
Stage4	9.71	<0.001	0.057	<0.001
Tx	5.73	<0.001	0.0540	0.377
lCD4	14.00	<0.001	0.076	0.003
Parametric terms				
	Coeff	**SE**	**95% CI**	**p-value**
const(Age)	2.38e-05	8.27e-06	7.59e-06, 4e-05	3.18e-03

Table A.5: Test associated with the semi-parametric hazard regression model on the HIV time updated data with the effect of the covariates `Age` and `Tx` fitted to be time-invariant.

	Tests for non-significant effect		Test for time-invariant effects	
Covariates	**Statistic**	**p-value**	**Statistic**	**p-value**
Intercept	13.70	<0.001	0.177	0.007
Male	3.14	0.036	0.014	0.172
Stage2	4.92	<0.001	0.012	0.190
Stage3	8.21	<0.001	0.013	0.263
Stage4	9.89	<0.001	0.057	<0.001
lCD4	13.80	<0.001	0.077	<0.001
Parametric terms				
	Coeff	**SE**	**95% CI**	**p-value**
const(Age)	2.56e-05	8.28e-06	9.37e-06, 4.18e-05	1.54e-03
const(Tx)	-1.79e-03	3.32e-04	-2.44e-03, -1.14e-03	<0.001

Table A.6: Test associated with the semi-parametric hazard regression model on the HIV time updated data with the effect of the covariates `Age`, `Stage3` and `Tx` fitted to be time-invariant.

	Tests for non-significant effect		Test for time-invariant effects	
Covariates	**Statistic**	**p-value**	**Statistic**	**p-value**
Intercept	14.00	<0.001	0.186	0.003
Male	3.34	0.033	0.015	0.121
Stage2	5.77	<0.001	0.020	<0.001
Stage4	10.30	<0.001	0.049	<0.001
lCD4	13.90	<0.001	0.078	<0.001
Parametric terms				
	Coeff	**SE**	**95% CI**	**p-value**
const(Age)	2.57e-05	8.28e-06	9.47e-06 4.19e-05	1.51e-03
const(Stage3)	1.28e-03	1.70e-04	9.47e-04 1.61e-03	<0.001
const(Tx)	-1.81e-03	3.33e-04	-2.46e-03 -1.16e-03	<0.001

Table A.7: Test associated with the semi-parametric hazard regression model on the HIV time updated data with the effect of the covariates `Age`, `Gender`, `Stage3` and `Tx` fitted to be time-invariant.

	Tests for non-significant effect		Test for time-invariant effects	
Covariates	**Statistic**	**p-value**	**Statistic**	**p-value**
Intercept	14.30	<0.001	0.191	0.004
Stage2	5.76	<0.001	0.020	<0.001
Stage4	10.40	<0.001	0.050	<0.001
lCD4	14.10	<0.001	0.079	<0.001
Parametric terms				
	Coeff	**SE**	**95% CI**	**p-value**
const(Age)	2.67e-05	8.31e-06	1.04e-05 4.30e-05	1.03e-03
const(Male)	5.79e-04	1.91e-04	2.05e-04 9.53e-04	4.24e-03
const(Stage3)	1.30e-03	1.69e-04	9.69e-04 1.63e-03	<0.001
const(Tx)	-1.86e-03	3.33e-04	-2.51e-03 -1.21e-03	<0.001

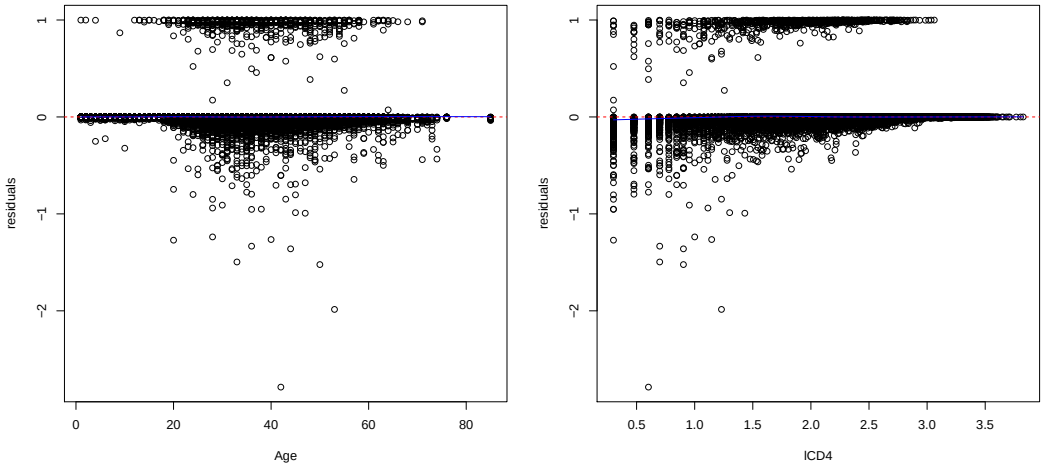


Figure A.1: Martingale residuals for the continuous variables in the extended Cox model.