Survival Analysis in Statistical Modelling

Timothy Kiprotich Lang'at (tklangat@aims.ac.za)

African Institute for Mathematical Sciences (AIMS)

Supervised by: Dr Viani A. Djeundje Biatat
University of Edinburgh, Scotland

18 May 2017

Submitted in partial fulfillment of a structured masters degree at AIMS South Africa

Abstract

Widely known and frequently used statistical tools for analysing survival data are reviewed in this essay. Due to censoring, standard data analysis techniques such as regression analysis cannot be employed. Inventions of powerful statistical software to analyse data have led to developments of non-parametric and semi-parametric techniques to analyse survival data. Non-parametric method employed include the Kaplan-Meier method for estimating the survivor function and the log-rank test to test the similarity in survival experiences in groups. The semi-parametric methods are the Cox proportional hazards model and the Cox regression model which is also known as the Cox extended model. Both models identify the covariates that have an effect on the hazard and estimate the hazard function. Whereas the former deals with time independent covariates, the latter deals with covariates that depend on time. Though these techniques can be applied to the analysis of survival data from different fields such as engineering, marketing, e.t.c, examples and the data set analysed in this essay are from medical experiments.

Declaration

I, the undersigned, hereby declare that the work contained in this research project is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.

Timothy Kiprotich Lang'at, 18 May 2017

Contents

1. Introduction

1.1 Background of Study

Survival analysis is a branch of statistics that deals with the study of distributions of times from a particular origin(marriage, starting work in a new job, treatment e.t.c) to the occurrence of event of interest(divorce, loosing a job, death, recovery from a disease, e.t.c). Observations in survival data are often incomplete. These type of observations are called censored. Censoring occurs when the exact time of occurrence of an event of interest is unknown. The types of censoring include interval , left and right censoring.

Right censoring occurs if the exact time of occurrence of the event of interest is unknown but is greater than the last known survival time. Under left censoring, the event of interest takes place before the study began. Interval censoring occurs when the exact time the event of interest takes place is unknown but is in between two known observed times.

In order to avoid bias, random and non-informative censoring is required.

1.2 Literature Review

According to Westergaard (1932) survival analysis traces its origin to the fields of demography and actuarial science. The techniques employed to construct life tables in actuarial science are not different from the ones used to estimate survivor functions for lives in a study. Makeham in (Makeham, 1860), Gompertz in (Gompertz, 1825) and De moive in (de Moivre, 1731) are credited to be the pioneers of the parametric methods used in survival analysis. The Gompertz Makeham law of mortality has been tested and found to predict the mortality rates for individuals in a population aged between ages 30 and 80 with accuracy. D Bernoulli in (Bernoulli, 1766) laid grounds for the theory of competing risks. The problems emanating from competing risks led to development of non-parametric methods of estimating survivor functions in continuous time, especially for studies where there are late entries into the study and right censoring. In his paper (Greenwood et al., 1926), Greenwood has worked in integrating survival analysis into theoretical statistics. This allowed usage of already developed statistical tools in the analysis of survival data. The development of survival analysis as statistical theory laid the grounds for D.R Cox in (Cox, 1972) to come up with the now famous Cox regression model for the hazard function. Cox regression model is classified as semi-parametric, that is, it is a hybrid of both parametric and non-parametric models. The hazard function is modelled to parametrically depend on the explanatory variables and to non-parametrically depend on time.

In this essay, we focus on the non-parametric and semi-parametric methods used to analyse survival data. In section 1.3, the functions used in survival analysis are defined. In Chapter 2 we cover the non-parametric methods. Both the Kaplan-Meier method of estimating the survivor function and the log-rank test for testing the difference in survival experiences among groups are discussed. In Chapter 3, the underlying theory of Cox proportional hazard's model is explained. The Cox regression model with time dependent explanatory variables is reviewed in Chapter 4. At the end of chapters 2,3 and 4 we have sections where the given survival data is analysed by fitting models and doing diagnosis so as to choose the best fitting model. Finally in chapter 5, we conclude by giving brief remarks and recommendations.

1.3 Preliminaries

1.3.1 Definition of key functions. Let T denote the random variable representing time taken for an event of interest to be observed. Examples of these events include; death, machine failure, recovery, e.t.c. T can take either continuous or discrete values. In this essay, death is the event and T will take continuous values.

Under survival analysis, functions of T that are of interest are cumulative distribution function, probability density function, survivor function and hazard function. These functions are denoted by $F(t)$, $f(t)$, $S(t)$ and $h(t)$ respectively.

The cumulative distribution function gives the probability of death occurring within a given time period. If this period is of length t, it is given by

$$F(t) = \text{prob}(T < t). \tag{1.3.1}$$

The probability density function gives the probability that death will occur at a given time. Mathematically, this time is expressed as an infinitesimal interval. For example the probability that an individual dies in the interval $(t-h, t+h)$ for infinitesimal h,is given by

$$f(t) = \lim_{h\to 0} \text{prob}\frac{(t \le T < t+h)}{h}. \tag{1.3.2}$$

The survivor function gives the probability that death will not be observed for at least a period of length of time. If this period is of length t, it is given by;

$$S(t) = \text{prob}(T \ge t) = 1 - F(t) = \int_t^\infty f(u)du. \tag{1.3.3}$$

The hazard function gives the probability of death occurring at a particular time given survival to that time. If the time is t then the hazard function is given by;

$$h(t) = \lim_{h\to 0} \text{prob}\frac{(t \le T < t+h | T > t)}{h}. \tag{1.3.4}$$

Hazard rate is also referred to as the force of mortality or the age-specific failure rate.

To obtain the cumulative hazard function, denoted by $H(t)$ the formula given below is used

$$H(t) = \int_0^t h(u)du. \tag{1.3.5}$$

1.3.2 Relationship between key functions. The survival function and the cumulative distribution function are complementary functions. This implies that

$$S(t) = 1 - F(t) \quad t \ge 0.$$

The relationship between the density function and the survivor function is given by

$$f(t) = \frac{d}{dt}F(t) = \frac{d}{dt}(1 - S(t)) = -\frac{dS(t)}{dt}. \tag{1.3.6}$$

The hazard function can be expressed in terms of the probability density function and the survivor function as follows

$$\begin{aligned} h(t) &= \lim_{h\to 0} \textbf{prob}\frac{(t \le T < t+h | T > t)}{h} \\ &= \lim_{h\to 0} \textbf{prob}\frac{(t \le T < t+h)}{h \cdot \text{prob}(T > t)} \\ &= \frac{1}{S(t)} \cdot \lim_{h\to 0} \textbf{prob}\frac{(t \le T < t+h)}{h} \end{aligned}$$

$$\implies h(t) = \frac{f(t)}{S(t)}, \quad t \ge 0 \tag{1.3.7}$$

From equations (1.3.6) and (1.3.7) it is easy to show that

$$\ln S(t) = \int_0^t \Big(- h(u) \Big) du, \tag{1.3.8}$$

and taking exponents of both sides of the above equation we get

$$S(t) = \exp\big(\int_0^t -h(u) du) \tag{1.3.9}$$

$$= \exp\big(- H(t)), \quad t \ge 0 \tag{1.3.10}$$

This implies that the cumulative hazard function expressed in terms of the survivor function is given by

$$H(t) = - \ln S(t), \quad t \ge 0 \tag{1.3.11}$$

1.4 Data for Analysis

In this section, we give a brief description of the data used for analysis. The dataset is mortality data from UK and was supplied by Dr. Viani, my supervisor. It consist of six columns. The columns are named as reference to the information they capture about individuals in the study. The name of the columns are; year of birth, gender, year of entry, year last observed, month last observed, and status(Alive or dead).

Some of the entries in the data lacked in consistency, for example, one individual had his year of entry being 2084 and year last observed being 1988. Out of the 1224 entries, 15 had this problem. Because 15 is a small number relative to the total entries, we decided to delete them and remained with 1209 entries in the final working sample.

All observations were made at the beginning of the month.

Indicator variables I_j, and δ_j are defined to capture the status and gender of an individual respectively. I_j takes the value 1 if the j^{th} life is observed to die and takes the value 0 if censoring occurs. δ_j takes the value 1 if the j^{th} life is male and 0 otherwise.

To obtain the age of individuals at the beginning of study we simply subtract the year of birth from the year of entry.Thus, our survival times are measured in years.

2. Theory and Results for Non-parametric Methods

Given survival data, two important functions to be estimated are the survivor and the hazard functions. However, since these two functions are related, once an estimate of one of them has been found, the other can be estimated by use of the relationships defined in subsection 1.3.2.

In the absence of censoring, the survivor function defined in equation (1.3.3) can be estimated using empirical survivor function abbreviated as $\hat{S}(t)$. $\hat{S}(t)$ is defined as the ratio number of lives alive and uncensored at time t to the total number of lives in the study, through the following formula

$$\hat{S}(t) = \frac{\text{Number of lives alive and uncensored at time t}}{\text{Total number of individuals in the study}}. \tag{2.0.1}$$

A plot of $\hat{S}(t)$ against time will give a step function. However, in reality censoring is present in survival data. Therefore, different ways to obtain the survivor function in the presence of censoring have to be employed.

In this chapter we will be looking at one of the non-parametric methods, the Kaplan-Meier method of estimating the survivor function.

2.1 Kaplan-Meier Estimate of The Survivor Function

This method was first developed in (Kaplan and Meier, 1958).

The model holds under the following assumptions.

(i) No distribution for the random variable of interest, T, is specified.

(ii) Censoring times are independent and random.

(iii) There exists uniformity within time intervals defined by death times. Hence we have the survival function being constant at each interval.

(iv) Deaths occur independently.

Definition of variables

We denote the number of lives being observed by n. This is the total sample size.

Let t_i for $i = 1, 2, \ldots, n$ be the survival time observed for life i and r be the number of times death is observed to occur, $r \leq n$.

$t_{(j)}$ for $j = 1, 2, \ldots, r$ is the observed j^{th} death event. These are the times the event of death occurs. Since death times are ordered, we have, $t_{(j)} < t_{(k)}$ if $j < k$.

n_j for $j = 1, 2, \ldots, r$ is the number of lives facing risk of death just before time $t_{(j)}$. It includes all those who will actually die at time $t_{(j)}$.

d_j for $j = 1, 2, \ldots, r$ is the number of deaths observed at time $t_{(j)}$.

Since in the infinitesimal time interval $t_{(j)} - h$ to $t_{(j)}$ one death event is observed, the probability of a life in the study dying within this time interval is given by $\frac{d_j}{n_j}$. The corresponding probability of survival is then $(1 - \frac{d_j}{n_j})$ as shown in Collett (1994). In the case that censoring occurs at the same time as the event of death, we assume that censoring occurs after the death event, this implies that the censored lives will be included in the computation of $n_j, j = 1, 2, \ldots, r$.

In the construction of the intervals, we assumed that there is only one death time in each interval, this means that no death is observed in the time interval $t_{(j)}$ to $t_{(j+1)} - h$. On this regard, the probability of a life surviving in the time interval $t_{(j)}$ to $t_{(j+1)} - h$ is equal to 1. By the assumption of independence of intervals, the probability of surviving from time $t_{(j)} - h$ to time $t_{(j+1)} - h$ is easily shown to be the product of the probability of surviving from time $t_{(j)} - h$ to time $t_{(j)}$ and the probability of surviving from time $t_{(j)}$ to time $t_{(j+1)} - h$. Which is $(1 - \frac{d_j}{n_j}) \times 1 = (1 - \frac{d_j}{n_j}) = (\frac{n_j - d_j}{n_j})$. Taking the limit $h \to 0$, $(\frac{n_j - d_j}{n_j})$ gives the estimate of surviving from time $t_{(j)}$ to $t_{(j+1)}$.

Thus, the Kaplan-Meier estimate of the survivor function as shown in (Kaplan and Meier, 1958) is given by;

$$\hat{S}(t) = \prod_{j=1}^{k} (\frac{n_j - d_j}{d_j}), \tag{2.1.1}$$

for $t_{(k)} \leq t < t_{(k+1)}; k = 1, 2, \ldots, r$, $\hat{S}(t) = 1$ for $t < t_{(1)}$ and $t_{(r+1)} = \infty$. If the largest observed time is censored survival time , t^* say, $\hat{S}(t)$ is undefined for $t > t^*$. If the largest observation is uncensored then $n_r = d_r$ and $\hat{S}(t) = 0$ for $t \geq t_{(r)}$ (Collett, 1994)

Obviously, since Kaplan-Meier estimate, $\hat{S}(t)$ is constant between death times, a plot of $\hat{S}(t)$ against time will give a step function.

2.1.1 Standard error of the Kaplan-Meier estimate. Let p_j be the true probability of surviving from time $t_{(j)}$ to time $t_{(j+1)}$. Clearly, the estimate of p_j is given by $\hat{p_j} = \frac{n_j - d_j}{n_j}$. Suppose the number of survivors,$n_j - d_j$ in the interval $t_{(j)}$ to $t_{(j+1)}$ follows a binomial distribution with parameters n_j and p_j then the mean and the variance of the number of individuals who survive through the interval $t_{(j)}$ to $t_{(j+1)}$ is given by $n_j p_j$ and $n_j p_j (1 - p_j)$ respectively. Since the mean and the variance cannot be computed as defined because the value of p_j is unknown, we use the value of $\hat{p_j}$ to approximate the unknown parameter p_j.Therefore, the mean and variance becomes $n_j \hat{p_j}$ and $n_j \hat{p_j}(1 - \hat{p_j})$ respectively.

The following steps are followed to compute the standard error of $\hat{S}(t)$.

$$\hat{S}(t) = \prod_{j=1}^{k} \hat{p}_j. \tag{2.1.2}$$

Taking the natural logarithm of equation (2.1.2), we get,

$$\ln \hat{S}(t) = \sum_{j=1}^{k} \ln \hat{p}_j. \tag{2.1.3}$$

Note that we have used the $\ln$ to represent the natural logarithm. In this essay we will be using both $\ln$ and log interchangeably. After obtaining the natural logarithm, the variance of equation (2.1.3) is

computed so as to get

$$\text{var}\{\ln \hat{S}(t)\} = \sum_{j=1}^{k} \text{var}\{\ln \hat{p}_j\}. \tag{2.1.4}$$

Using Taylor series approximation to the variance of a function of a random variable, both the left hand side and the right hand side of equation (2.1.4) are approximated by

$$\text{var}\{\ln \hat{S}(t)\} \approx \left\{\frac{1}{\hat{S}(t)}\right\}^2 \text{var}[\hat{S}(t)]. \tag{2.1.5}$$

and

$$\begin{aligned}
\sum_{j=1}^{k} \text{var}\{\ln \hat{p}_j\} &\approx \sum_{j=1}^{k} \left\{\frac{1}{\hat{p}_j}\right\}^2 \text{var}(\hat{p}_j) \\
&\approx \sum_{j=1}^{k} \left\{\frac{1}{\hat{p}_j}\right\}^2 \text{var}\left(\frac{n_j - d_j}{n_j}\right) \\
&\approx \sum_{j=1}^{k} \left\{\frac{1}{\hat{p}_j}\right\}^2 \left(\frac{n_j \hat{p}_j (1-\hat{p}_j)}{n_j^2}\right) \\
&\approx \sum_{j=1}^{k} \frac{1-\hat{p}_j}{n_j \hat{p}_j} \\
&\approx \sum_{j=1}^{k} \frac{1-\left(\frac{n_j - d_j}{n_j}\right)}{n_j\left(\frac{n_j - d_j}{n_j}\right)}
\end{aligned}$$

respectively.

Therefore,

$$\sum_{j=1}^{k} \text{var}\{\ln \hat{p}_j\} \approx \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}. \tag{2.1.6}$$

Equating the righthand sides of equations (2.1.5) and (2.1.6) we get

$$\left\{\frac{1}{\hat{S}(t)}\right\}^2 \text{var}[\hat{S}(t)] \approx \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}.$$

Which can be re-written as shown in the equation (2.1.7) below

$$\text{var}[\hat{S}(t)] \approx [\hat{S}(t)]^2 \sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}. \tag{2.1.7}$$

The square root of equation (2.1.7) gives the standard error.

$$\text{se}\{\hat{S}(t)\} \approx \hat{S}(t)\left\{\sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}\right\}^{\frac{1}{2}}, \tag{2.1.8}$$

for $t_{(k)} \leq t < t_{(k+1)}, k = 1, 2, \ldots, r$. This formula was first proposed in (Greenwood et al., 1926).

2.1.2 Confidence intervals for values of the estimated survivor function. An interval constructed and a probability assigned to it indicating the chance of containing the true value of the survivor function is called a confidence interval. Confidence interval are constructed for each estimated value of the survivor function at time t, hence they are called point-wise confidence intervals (Collett, 1994).

For a survivor function, $S(t)$, to construct a $(1-\alpha)100\%$ confidence interval at time t we assume that the estimate of the survivor function $\hat{S}(t)$, has a Gaussian(normal) distribution whose mean is $S(t)$ and the estimated variance is shown in equation (2.1.7) .

α is known as the confidence level. It can take take a range of values from 1% to 15%. Most studies in medical statistics set $\alpha = 0.5$. The $(1-\alpha)100\%$ confidence interval for the true value of the survivor function at time t is given by $[\hat{S}(t) \pm Z_{\alpha/2}\text{se}\{\hat{S}(t)\}]$. Where $Z_{\alpha/2}$ is the value corresponding to the upper $\alpha/2$ percentage point of the standard normal distribution, $\hat{S}(t)$ is the estimate of the survivor function at time t and $\text{se}\{\hat{S}(t)\}$ is the standard error as defined in (2.1.8)

In the event that the upper limit of the confidence intervals for the survivor function at time t closer to 1 exceed 1, the limit is set to 1. On the other hand, the lower limits of the confidence intervals of the survivor function at time t that are closer to 0 are set to 0 if they are less than 0.

Due to positive skewness in the distribution of the survivor function, the median is the preferred location summary measure. The median survival time is a time t_{50} such that $S(t_{50}) = 0.5$. However, since the estimate of the survivor function is a step function, it is not easy to obtain a survival time that yields a survival probability value of exactly 0.5. To obtain its estimate, $\hat{t}_{50}$ we use of the following formulae;

$$\hat{t}_{50} = \min\{t_i | \hat{S}(t_i) < 0.5\} \quad t_i \geq 0, i = 1, 2, \ldots, n \tag{2.1.9}$$

t_i is the observed survival time for life i

Since the changes in the estimate of the survivor function only occurs at times when death event is observed, the formula given by equation (2.1.9) can be written as

$$\hat{t}_{50} = \min\{t_{(j)} | \hat{S}(t_{(j)}) < 0.5\} \quad t_{(j)} \geq 0, j = 1, 2, \ldots, r \tag{2.1.10}$$

$t_{(j)}$ is the j^{th} time that death is observed to occur.

If $\hat{S}(t_{(j)}) = 0.5$ then the median is approximated by $\frac{t_{(j)}+t_{(j+1)}}{2}$

Since the median is the fiftieth percentile to compute its variance we will use the general formulae of obtaining the p^{th} percentile. Using Taylor series approximation we get

$$\text{var}[\hat{S}(t_p)] \approx \left(\frac{d\hat{S}(\hat{t}_p)}{d\hat{t}_p}\right)^2 \text{var}\{\hat{t}_p\}, \tag{2.1.11}$$

where $\hat{t}_p$ is the p^{th} percentile of the distribution and $\hat{S}(t_p)$ is the Kaplan-Meier estimate of the survivor function at t_p. From equation (1.3.6) we have that

$$-\frac{d\hat{S}(\hat{t}_p)}{d\hat{t}_p} = \hat{f}(\hat{t}_p), \tag{2.1.12}$$

using (2.1.12), equation (2.1.11) becomes

$$\text{var}[\hat{S}(t_p)] \approx (\hat{f}(\hat{t}_p))^2 \text{var}\{\hat{t}_p\} \tag{2.1.13}$$

$$\implies \text{var}\{t_p\} \approx \left(\frac{1}{\hat{f}(\hat{t}_p)}\right)^2 \text{var}[\hat{S}(\hat{t}_p)] \tag{2.1.14}$$

$$\implies \text{se}\{t_p\} = \frac{1}{\hat{f}(\hat{t}_p)} \text{se}[\hat{S}(\hat{t}_p)]. \tag{2.1.15}$$

se$[\hat{S}(\hat{t}_p)]$ is calculated as shown in equation (2.1.8). $\hat{f}(\hat{t}_p)$ is estimated by use of the following formula;

$$\hat{f}(\hat{t}_p) = \frac{\hat{S}\{\hat{a}(p)\} - \hat{S}\{\hat{b}(p)\}}{\hat{b}(p) - \hat{a}(p)}.$$

Where

$$\hat{a}(p) = \text{maximum}\{t_{(j)} | \hat{S}(t_{(j)} \geq 1 - \frac{p}{100} + \epsilon\},$$

and

$$\hat{b}(p) = \text{minimum}\{t_{(j)} | \hat{S}(t_{(j)} \leq 1 - \frac{p}{100} - \epsilon\},$$

for $j = 1, 2, \ldots, r$. In our computations we will take $\epsilon = 0.05$. However, larger values of ϵ is required if $\hat{a}(p)$ and $\hat{b}(p)$ are equal.

2.1.3 Hypothesis testing. Under hypothesis testing, there are two types of hypothesis, the null hypothesis and the alternative hypothesis.

The null hypothesis is a generally accepted statement about the process generating the given data where as the alternative hypothesis counters the null hypothesis.

To test the null hypothesis we need to compute a measure called the test statistic from which probability values commonly called $p - values$ are obtained.

The test statistic measures the extent the observed data departs from the null hypothesis. The larger the test statistic's value, the larger the departure.

For two sided tests, suppose the test statistic is represented by the random variable Y, then the $p-value$ is obtained by adding the probability that Y is less than the negative of the absolute value, $|y|$ i.e. $p(Y \leq -|y|)$ and the probability that Y is greater than the absolute value, $|y|$ i.e. $p(Y \geq |y|)$. If symmetry is assumed then $p - value = 2p(Y \geq |y|)$.

For one sided tests, departure in one direction is considered. If departure in the left direction is of interest the test is a left tailed test and the $p - value$ is given by $p(Y \leq y)$ on the other hand if departure in the right direction is of interest, the test is a right tailed test and the $p - value$ is given by $p(Y \geq y)$.

If the p-value is less than the level of significance, α we reject the null hypothesis and infer that the observed test statistic is significant at α-level of significance.

2.1.4 Results and Discussion. In this section we find the estimate of the survivor functions for males and females. The plot of the two Kaplan-Meier estimates of the survivor function split by gender is given in figure 2.1. These graphs enable us to compare the survivor experience of the two genders and also to estimate the survival probabilities. From the graph the following is observed;

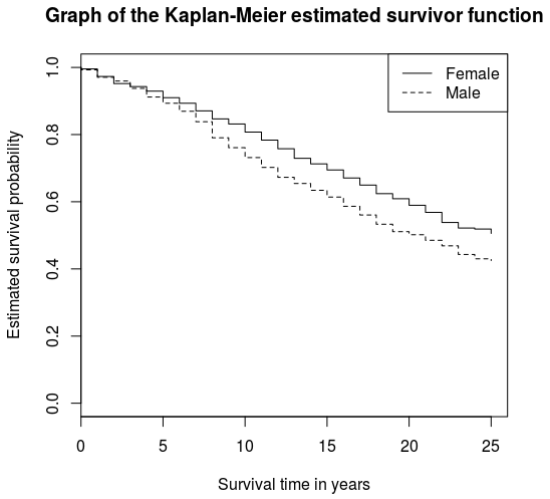

Figure 2.1: Kaplan-Meier estimate of the survivor function for both males and females from the data

(i) The Kaplan-Meier estimate of the survivor functions for both genders are observed not to cross each other. This is an evidence of the proportional hazard.

(ii) The plot of the Kaplan-Meier estimate of the survivor function of males is observed to lie in the same line with that of females for the first 4 years. But after 4 years it is observed to lie above that of females. This can be interpreted to imply that males and females have a similar survival experience during the first 4 years of the study. However, after 4 years males experience a higher mortality rate than their female counterparts.

(iii) After 20 years around 50% of males and around 60% of females will be alive.

2.2 The Log-rank Test

Part of survival analysis is the analysis of survival experience among groups. In medical studies for example, to analyse the effectiveness of two types of drugs, one group of patients may be given a dose of a new drug and the other may be given a dose of a standard drug. Several methods exist that can be employed to quantify the extent of these differences. Two examples of such methods are

(i) The Mantel-Haenszel log-rank test

(ii) The wilcoxon test

The log-rank test is based on the assumption that survival times are continuous or ordinal and the hazards are proportional. The null hypothesis to be tested is that there is no significant differences in the survival experiences among groups. To obtain a log-rank test statistic, a 2×2 contingency table is constructed each time death is observed, then contingent on the number of individuals at risk of death, the death rates between the two groups is compared. The tables obtained at each death time are then put together to give the overall test statistic (Mantel and Haenszel, 1959).

To compare two groups namely Group 0 and Group 1, we need to define the following: Let $t_{(1)}, t_{(2)}, \ldots, t_{(r)}$ be the r-ordered times death is observed to occur across the two groups. Suppose at death time $t_{(k)}$,

d_{0k} individuals out of the n_{0k} individuals at risk of death in Group 0, and d_{1k} individuals out of the n_{1k} individuals at risk of death in Group 1 die.This implies that at time $t_{(k)}$, a total of $d_k = d_{0k} + d_{1k}$ individuals die out of a total of $n_k = n_{0k} + n_{1k}$ individuals at risk of dying.

We summarize this discussion in the table below.

Table 2.1: Log-rank 2×2 contingency table

Group	Number who die at $t_{(k)}$	Number who live past $t_{(k)}$	Number in the risk set before$t_{(k)}$
0	d_{0k}	$n_{0k} - d_{0k}$	n_{0k}
1	d_{1k}	$n_{1k} - d_{1k}$	n_{1k}
Total	d_k	$n_k - d_k$	n_k

If the marginal totals are taken to be fixed, knowledge of d_{0k} is sufficient to know the values of the entries of the table. Assuming the null hypothesis holds, d_{0k} follows a $hypergeometric \quad distribution.$ That is

$$p(D_{0k} = d_{0k}) = \begin{cases} \frac{\binom{d_k}{d_{0k}}\binom{n_k - d_k}{n_{0k} - d_{0k}}}{\binom{n_k}{n_{0k}}}, & \text{for } d_{0k} = 0, 1, 2, \ldots, \min(d_k, n_{0k}) \\ 0, & \text{otherwise} \end{cases}.$$

Let m_{0k} and v_{0k} be the mean and variance of d_{0k} respectively. Therefore,

$$m_{0k} = \frac{n_{0k} d_k}{n_k} \quad \text{and}$$
$$v_{0k} = \frac{n_{0k} d_k (n_k - d_k) n_{0k} n_{1k}}{n_k^2 (n_k - 1)}.$$

m_{0k} is the expected number of individuals in Group 0 who will die at death time $t_{(k)}$. Of interest, is the deviation of the observed number of deaths from those expected on the assumption that the null hypothesis holds. The sum of these deviations is given by $L_D = \sum_{k=1}^{r} (d_{0k} - m_{0k})$. We then standardise the computed deviation by diving by the standard deviation which is obtained by finding the squareroot of the variance denoted by $L_V = \sum_{k=1}^{r} v_{0k}$. Therefore, the formula for computing the test statistic,L is given by

$$L = \frac{L_D}{\sqrt{L_V}}.$$

Since L has a standard normal distribution, L^2 will have a chi-square distribution with one degree of freedom.

The test statistic L^2 summarizes the extent of deviation of the observed survival times of the two groups from those expected under the null hypothesis. The larger the value of L^2 the larger the deviation and hence the more the evidence against the null hypothesis.

2.2.1 Results and Discussion of the Log-rank test. The log-rank test tests the null hypothesis that there is no difference in survival experience of males and females.

The summary of the results obtained from R software is given in the table (2.2) below.

Table 2.2: Log-rank results

Sex	Total number	Total observed (O)	Total expected (E)	$\frac{(O-E)^2}{E}$
Females	665	323	363	4.31
Males	544	311	271	5.75

The test statistic can be taken to be either 4.31 or 5.75. The corresponding p-values are 0.0379 and 0.0165 respectively. Under the assumptions of the null hypothesis, the test statistic is assumed to follow a chi-squared distribution with one degree of freedom.

Comparing the $p-values$ to the level of significance given by 0.05, we observe that $p-value < 0.05$. Therefore we reject the null hypothesis at $\alpha = 0.05$ level of significance and conclude that males and females have a different survival experience.

The major limitation of the non-parametric methods is that they do not incorporate covariates in the model, yet covariates have been found to have an effect on the hazard rates of individuals.

3. Cox Proportional Hazards Model

3.1 Model Description

In practice, during the data collection, additional quantities in respect to each individual being observed are recorded as well.These quantities are called covariates.

The covariates can be classified broadly into two categories. Variates and factors. Variates are quantitative variables that are obtained from direct measurements and take values from a continuous measuring scale e.g. age, height or weight. On the other hand, factors are qualitative variables; they take values which subdivide the data into segments called levels, e.g. treatment as a factor can have two levels , placebo treatment and new treatment.

The Cox proportional hazards model was first introduced by D.R Cox in (Cox, 1972). To explain the model let us look at the following example.

Suppose an experiment to compare the superiority between two drugs is conducted. To one group of patients drug A is administered and to the other group drug B . Let $h_A(t)$ and $h_B(t)$ denote the hazard functions at time t for patients in the two groups.If we assume that the hazard of death at time t for individuals using drug A is proportional to the hazard of death at time t for those on drug B. Then;

$$h_A(t) = \kappa h_B(t) \quad , t \geq 0. \tag{3.1.1}$$

κ, the constant of proportionality is called the relative hazard or hazard ratio. If $\kappa > 1$ then for each time t, the hazard of death for patients put on drug A is higher than that for patients put on drug B. Therefore, drug B is superior to drug A. On the other hand, if $\kappa < 1$ then at each time t, the force of mortality for those taking drug A is less than for those on drug B. Therefore, drug A performs better.

Suppose $h_j(t), j = 1, 2, \ldots, n$ and $h_0(t)$ are the hazard functions for individuals put on drugs A and B respectively then the hazard function for an individual on drug A will be $\kappa h_0(t)$. Since the hazard ratio, κ cannot be negative, it is expressed as $\exp(\beta)$ which implies $h_j(t) = \exp(\beta)h_0(t), t \geq 0, j = 1, 2, \ldots, n$. β the natural logarithm of κ takes values in the range $(-\infty, \infty)$

To generalize the model in equation (3.1.1), let Z be an indicator variable such that ;

$$Z = \begin{cases} 1, & \text{if a patient is on drug A} \\ 0, & \text{if a patient is on drug B} \end{cases}$$

If z_j is the realisation for the random variable Z for the j^{th} patient, $j = 1, 2, \ldots, n$, then, $h_j(t)$ is expressed as

$$h_j(t) = \exp(\beta z_j)h_0(t) \quad t \geq 0, j = 1, 2, \ldots, n \tag{3.1.2}$$

To obtain the Cox proportional hazards model, we generalise the model given in equation (3.1.2)

Suppose for each individual in the study, additional q measurements denoted by random variables $Z_1, Z_2, \ldots, Z_q$ are taken whose realisations are represented by vector $\mathbf{z} = (z_1, z_2, \ldots, z_q)^T$. The T in the exponent denotes transpose. Then the function of the hazard rate of the j^{th} individual can be expressed as a function of both the covariates and time as shown below;

$$h_j(\mathbf{z_j}, t) = \kappa(\mathbf{z_j})h_0(t), \quad t \geq 0, \quad j = 1, 2, \ldots, n.$$

where $h_0(t)$ the baseline hazard function is the hazard function for individuals whose values of the explanatory variables take the value zero. The function $\kappa(\mathbf{z_j})$ at any time t is the ratio of the hazard function for a life whose set of explanatory variables is given by vector $\mathbf{z_j}$ to the hazard function for a life whose set of explanatory variables take the value zero.

To avoid negative values, we set $\kappa(\mathbf{z_j}) = \exp(\psi_j)$, where $\psi_j = \beta_1 z_{1j} + \beta_2 z_{2j} + \ldots + \beta_q z_{qj} = \sum_{i=1}^{q} \beta_i z_{ij} = \boldsymbol{\beta}^T \mathbf{z_j}$. Vector $\boldsymbol{\beta}$ is a vector of coefficients of the covariates. i.e $\boldsymbol{\beta}^T = (\beta_1, \beta_2, \ldots, \beta_q)$. ψ_j is referred to as the linear component, risk score or the prognostic index (Collett, 1994).

The general Cox proportional hazard's model is thus given by

$$h(t; \mathbf{z_j}) = h_0(t) \exp(\beta^T \mathbf{z_j}). \tag{3.1.3}$$

The model can as well be expressed linearly as the natural logarithm of the relative hazards as shown below

$$\ln\left(\frac{h(t; \mathbf{z_j})}{h_0(t)}\right) = \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_q z_q. \quad t \geq 0 \tag{3.1.4}$$

3.2 Linear Component of The Model

3.2.1 Including variates. To include a variate, each variate is assigned a coefficient, for example in a study where two variates Z_1 and Z_2 are measured, the hazard function is given by

$$h_j(t) = h_0(t) \exp(\beta_1 z_{1j} + \beta_2 z_{2j}), \quad t \geq 0, j = 1, 2, \ldots, n.$$

3.2.2 Including a factor. To include a factor B with b levels into the model the following terms are defined $\gamma_1, \gamma_2, \ldots, \gamma_b$. They denote the contributions of each level of factor B. The hazard function at time $t \geq 0$ for an individual at level i of factor B is $h_0(t) \exp(\gamma_i)$. For consistency in the definition of the baseline hazard function, one of the parameters representing the effects of the levels of the factor is set to zero. Suppose γ_1 is set to zero, then the baseline hazard is the hazard for an individual at the first level.

Factors can be expressed as linear combination of covariates. A set of indicator variables corresponding to each factor is defined . Since $\gamma_1 = 0$, for the remaining main effects we define $b-1$ indicator variables $Z_2, Z_3, \ldots, Z_b$ such that

$$z_k = \begin{cases} 1, & \text{if a life is at level k} \\ 0, & \text{otherwise} \end{cases}$$

γ_i in the model is then replaced by $\gamma_2 z_2 + \gamma_3 z_3 + \ldots + \gamma_b z_b$, where z_k is the realisation of Z_k for $k = 2, 3, \ldots, b$, the levels of B. The model thus becomes;

$$h_j(t) = h_0(t) \exp(\gamma_2 z_{2j} + \gamma_3 z_{3j} + \ldots + \gamma_b z_{bj}) \quad \text{for} \quad j = 1, 2, \ldots, n.$$

The $b-1$ parameters defined to capture the main effects of factor B implies B has $b-1$ degrees of freedom.

3.2.3 Interaction. If two or more factors are incorporated into the model, sometimes there will be interaction between the factors. Interaction occurs when the function of the hazard rate at time t depends on the combination of factors in the model.

If G and H are two factors with g and h levels respectively, let $\alpha_i, i = 1, 2, \ldots, g$ represent the effect due to the levels of factor G and let $\gamma_l, l = 1, 2, \ldots, h$ represent the effect due to the levels of factors H. The interaction effect is denoted by $(\alpha\gamma)_{il}$ for $i = 1, 2, \ldots, g$ and $l = 1, 2, \ldots, h$.

3.2.4 Example. Suppose that G and H have 3 and 2 factors respectively. If we use indicator variables W_2, W_3 and V_2, they will be expressed as shown in the table below:

Table 3.1: Table showing levels of G

Levels of G	W_2	W_3
1	0	0
2	1	0
3	0	1

Table 3.2: Table showing levels of H

Levels of H	V_2
1	0
2	1

If w_2, w_3 and v_2 are the realisation of W_2, W_3 and V_2 respectively, then the effect due to interaction, $(\alpha\gamma)_{il}$ is fitted by including the variates formed from the products of W_i and V_l for $i = 2, 3$ and $l = 2$. The interaction term will thus be given by;

$$(\alpha\gamma)_{22}w_2v_2 + (\alpha\gamma)_{32}w_3v_2$$

If G and H have g and h levels respectively, the two factor interaction GH has $(g-1)(h-1)$ parameters. This implies that the number of degree of freedom is given by $(g-1)(h-1)$.

3.2.5 Mixed terms. If the coefficient of a variate take different values at different levels then it implies that the levels of the factor have an effect on the coefficient of the variate, this effect has to be captured in the model as well.

3.2.6 Example. Suppose in a study of 6 individuals, values of a variate M and levels of a factor B were measured. Suppose B has 3 levels this implies that we have to define 2 indicator variables V_2 and V_3 as shown in Table 3.3 below

Table 3.3: Table showing mixed terms

Individual	levels of B	M	V_2	V_3	V_2M	V_3M
1	1	m_1	0	0	0	0
2	1	m_2	0	0	0	0
3	2	m_3	1	0	m_3	0
4	2	m_4	1	0	m_4	0
5	3	m_5	0	1	0	m_5
6	3	m_6	0	1	0	m_6

In addition to variate Y, the explanatory variables V_2M and V_3M are included into the linear component of the model as well. If γ_2', γ_3' and β are the coefficients of V_2M, V_3M and M respectively, the linear component of the model will then be given by $\beta m + \gamma_2'(v_{2m}) + \gamma_3'(v_{3m})$.

This implies that when $v_2 = v_3 = 0$ as is the case at level 1, coefficients of m is β. When $v_2 = 1$ and $v_3 = 0$ the coefficient of m is $\beta + \gamma_2'$ and when $v_2 = 0$ and $v_3 = 1$ the coefficient of m is $\beta + \gamma_3'$. This

shows that had we not included the term βm, the model would not have captured the effect of the variate M on the hazard function for individuals in the first level of factor B.

3.3 Fitting of The Proportional Hazards Model

In most applications of proportional hazard models, the main interest is to estimate the coefficients, i.e. the beta's. Since the baseline hazard function,$h_0(t)$ accounts for the shape of the hazard function for individuals while the linear component accounts for the differences between individuals, it is the linear component that is responsible with quantifying the effects of the covariates. Therefore, the $\beta's$,are estimated first then their values, in conjunction with the data, are used to construct the baseline hazard function. This technique is known as the semi-parametric approach.

To estimate the $\beta's$ the method of maximum likelihood is often used. Under this method, we first construct the likelihood function of the unknown parameter $\beta's$ and the survival times. After obtaining the likelihood, we solve for parameters that maximize the likelihood. Since, the parameters that maximize the likelihood are the same parameters that will maximize the natural logarithm of the likelihood function, we maximize the latter because it is computationally easier. Numerical techniques are used to maximize the log-likelihood, in our case we use the Newton-Raphson numerical method.

The following assumptions will guide us :

(i) Only one life is observed to die at each death time. The case of ties will be considered later.

(ii) If censoring and death are observed to occurs simultaneously, then censoring takes place immediately after the death event has taken place.

Let there be n lives whose observed survival time are given by $t_1, t_2, \ldots, t_n$. Suppose out of these n individuals r die. Let the death times be ordered such that $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$. Let $R(t_{(j)})$ denote the set of lives in the study who have not experienced death or censoring at a time just before death time $t_{(j)}, j = 1, 2, \ldots, r$. This set is referred to as the risk set.

According to Cox (1972) the partial likelihood for the model given in equation (3.1.3) is

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp(\beta^{\mathbf{T}}\mathbf{z_j})}{\sum_{k \in R(t_{(j)})} \exp(\beta^{\mathbf{T}}\mathbf{z_k})}, \tag{3.3.1}$$

Where $\mathbf{z_j}$ is the vector representing the set of the explanatory variables associated with the life whose death is observed at death time $t_{(j)}, j = 1, 2, \ldots, r$.

The likelihood can be expressed in terms of the indicator variable by defining an indicator variable ρ_i;

$$\rho_i = \begin{cases} 1, & \text{if the } i^{\text{th}} \text{ life is observed to die} \quad i = 1, 2, \ldots, n \\ 0, & \text{if the } i^{\text{th}} \text{ life is censored} \end{cases}.$$

The likelihood then becomes

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\beta^{\mathbf{T}}\mathbf{z_i})}{\sum_{k \in R(t_i)} \exp(\beta^{\mathbf{T}}\mathbf{z_k})} \right)^{\rho_i}, \tag{3.3.2}$$

where $R(t_i)$ is the risk set at time t_i . The log likelihood function calculated from (3.3.2) is given by

$$\log L\beta = \sum_{i=1}^{n} \rho_i \left(\beta^{\mathbf{T}}\mathbf{z_i} - \log \left(\sum_{k \in R(t_i)} \exp(\beta^{\mathbf{T}}\mathbf{z_k}) \right) \right) \tag{3.3.3}$$

3.3.1 Derivation of the likelihood. To derive the likelihood function, it is often assumed that since there are no observed deaths between successive death times the hazard rate is assumed to take the value zero. Therefore, we consider the probability of the i^{th} life whose set of covariates is represented by vector $\mathbf{z_{(j)}}$ dying at a time $t_{(j)}$ conditional on this time being among the times death is observed to occur, $t_{(v)}$ for $v = 1, 2, \ldots, r$.

This probability is given by;

$$\textbf{Prob}(i^{\text{th}} \text{ life with covariates } \mathbf{z_{(j)}} \text{ dies at time } t_{(j)} |\ t_{(j)} \text{ is among the observed times of death}),$$

and can be written as

$$\textbf{Prob}(i^{\text{th}} \text{ life with covariates } \mathbf{z_{(j)}} \text{ dies at time } t_{(j)} | \text{ there is a single death at time of death } t_{(j)}\ ). \tag{3.3.4}$$

From results of conditional probability;

$$\textbf{Prob}(\Xi|\Phi) = \frac{\textbf{Prob}(\Xi \text{and} \Phi)}{\textbf{Prob}(\Phi)},$$

expression (3.3.4) becomes

$$\frac{\textbf{Prob}(i^{\text{th}} \text{ life with covariates } \mathbf{z_{(j)}} \text{ dies at time } t_{(j)})}{\textbf{Prob}(\text{ there is a single death at time of death } t_{(j)}\ )}.$$

On the assumption of independence of survival times the expression becomes

$$\frac{\textbf{Prob}(i^{\text{th}} \text{ life with covariates } \mathbf{z_{(j)}} \text{ dies at time } t_{(j)})}{\sum_{k \in R(t_{(j)})} \textbf{Prob}(\text{life k dies at time } t_{(j)})}.$$

Substituting death time $t_{(j)}$ by the infinitesimal interval $(t_{(j)}, t_{(j)} + h)$ we get,

$$\frac{\textbf{Prob}(i^{\text{th}} \text{ life with covariates } \mathbf{z_{(j)}} \text{ dies in the time interval } (t_{(j)}, t_{(j)} + h))}{\sum_{k \in R(t_{(j)})} \textbf{Prob}(\text{life k dies in the time interval } (t_{(j)}, t_{(j)} + h))}$$

Dividing both the numerator and the denominator by h and taking limits as $h \to 0$, by equation (1.3.4) we get

$$\frac{\text{Hazard of death for the } i^{\text{th}} \text{ life with covariates } \mathbf{z_j} \text{ at time } t_{(j)}}{\sum_{k \in R(t_{(j)})} \text{Hazard of death for the } k^{\text{th}} \text{ life at time } t_{(j)}}. \tag{3.3.5}$$

Expression (3.3.5) is simplified to

$$\begin{aligned} \frac{h_i(t_{(j)})}{\sum_{k \in R(t_{(j)})} h_k(t_{(j)})} &= \frac{h_0(t_{(j)}) \exp(\beta^{\mathbf{T}} \mathbf{z_{(j)}})}{\sum_{k \in R(t_{(j)})} h_0(t_{(j)}) \exp(\beta^{\mathbf{T}} \mathbf{z_k})} \\ &= \frac{\exp(\beta^{\mathbf{T}} \mathbf{z_{(j)}})}{\sum_{k \in R(t_{(j)})} \exp(\beta^{\mathbf{T}} \mathbf{z_k})}. \end{aligned}$$

Taking the product over the r death times, we get the likelihood given in (3.3.1).

Because the obtained likelihood function does not directly incorporate the censored and uncensored times, it is referred to as a partial likelihood function (Cox, 1975)

3.3.2 Treatment of ties. In practice, tied survival times occur because survival times are usually rounded to the nearest unit of measurement for example year, month, week or day. In the presence of ties, the likelihood function needs some modification. The extended likelihood function that takes into account ties can be found in Kalbfleisch and Prentice (2002). However, due to its complication, a simplified version proposed in Breslow (1974) is widely used.

Suppose d_j lives are observed to die at time $t_{(j)}, j = 1, 2, \ldots, r$. We defined a new vector $\mathbf{s_j}$ whose elements are the sum of each corresponding explanatory variables for the lives observed to die at time $t_{(j)}$.

Thus, the Breslow formula approximating the likelihood function when there is existence of ties is given by;

$$L(\beta) = \prod_{j=1}^{r} \frac{\exp(\beta^{\mathbf{T}}\mathbf{s_j})}{\left(\sum_{k\in R(t_{(j)})} \exp(\beta^{\mathbf{T}}\mathbf{z_k})\right)^{d_j}}. \tag{3.3.6}$$

3.3.3 Newton-Raphson numerical method for estimating β's. Newton-Raphson method is among the most widely used methods in numerical analysis. It is used to estimate the roots of functions. The method is explained below.

Let $l(g)$ be a function whose root we want to estimate. Suppose g_r is the true root that is unknown and g_0 is the first initial estimate. Expanding $l(g)$ about g_0 by taylor series we get;

$$l(g) = l(g_0) + (g - g_0)l'(g_0) + \frac{1}{2!}(g - g_0)^2 l''(g_0) + \cdots$$

Equating the first two terms to zero and solvin for g we get

$$l(g) \approx l(g_0) + (g - g_0)l'(g_0) = 0 \tag{3.3.7}$$

$$\implies g = g_1 = g_0 - \frac{l(g_0)}{l'(g_0)} \tag{3.3.8}$$

g_1 is then used in place of g_0 to compute g_2 the new approximation of g_r. The process is repeated over and over until the newly approximated value is not significantly different from the previous approximation. Generalizing, we get the following iteration formula;

$$g_j = g_{j-1} - \frac{l(g_{j-1})}{l'(g_{j-1})} \quad , j = 1, 2, \ldots \tag{3.3.9}$$

The iteration is said to converge if

$$\lim_{j\to\infty} g_j \to g_r.$$

The Newton-Raphson iteration function given in (3.3.9) is generalized to solve a system of linear equations with multiple variables by using matrices and vectors. The equation to solve this type of system of equations is given by

$$\mathbf{w^{(j)}} = \mathbf{w^{(j-1)}} - J^{-1}(\mathbf{w^{(j-1)}})G(\mathbf{w^{(j-1)}}) \quad j = 1, 2, \ldots \tag{3.3.10}$$

where ;

$\mathbf{w}$ is a $n \times 1$ vector given by $\mathbf{w}^T = (w_1, w_2, \ldots, w_n)$, $w_i \in \mathbb{R}, i = 1, 2, \ldots, n$, G is a vector function such that $G : \mathbb{R}^n \to \mathbb{R}^n$, that is

$$G(w_1, w_2, \ldots, w_n) = \begin{bmatrix} g_1(w_1, w_2, \ldots, w_n) \\ g_2(w_1, w_2, \ldots, w_n) \\ \vdots \\ g_n(w_1, w_2, \ldots, w_n) \end{bmatrix}$$

where g_i maps $\mathbb{R}^n$ to $\mathbb{R}$ for $i = 1, 2, \ldots, n$.

$J(\mathbf{w})$ is a Jacobian matrix. The inverse on the Jacobian matrix is the one in use in (3.3.10). This inverse is given by

$$J^{-1}(\mathbf{w}) = \begin{bmatrix} \frac{\partial g_1(\mathbf{w})}{\partial w_1} & \frac{\partial g_1(\mathbf{w})}{\partial w_2} & \cdots & \frac{\partial g_1(\mathbf{w})}{\partial w_n} \\ \frac{\partial g_2(\mathbf{w})}{\partial w_1} & \frac{\partial g_2(v)}{\partial w_2} & \cdots & \frac{\partial g_2(\mathbf{w})}{\partial w_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_n(\mathbf{w})}{\partial w_1} & \frac{\partial g_n(\mathbf{w})}{\partial w_2} & \cdots & \frac{\partial g_n(\mathbf{w})}{\partial w_n} \end{bmatrix}^{-1}$$

To use the Newton-Raphson method, we follow the steps outlined below

Step 1: Obtain the initial vector $\mathbf{w}^{(0)}$ given by $\mathbf{w}^{(0)T} = (w_1^{(0)}, w_2^{(0)}, \ldots, w_n^{(0)})$.

Step 2: Calculate the Jacobian matrix $J(\mathbf{w})$ and the vector function $G(\mathbf{w})$ when $\mathbf{w} = \mathbf{w}^{(0)}$ to get $J(\mathbf{w}^{(0)})$ and $G(\mathbf{w}^{(0)})$ respectively.

Step 3: By Gaussian elimination solve for $\mathbf{e^{(0)}}$ in the linear system $J(\mathbf{w}^{(0)})\mathbf{e}^{(\mathbf{0})} = -G(\mathbf{w}^{(0)})$ to get $\mathbf{e}^{(\mathbf{0})} = -J^{-1}(\mathbf{w}^{(0)})G(\mathbf{w}^{(0)})$.

step 4: solve for $\mathbf{w}^{(\mathbf{1})} = \mathbf{w}^{(0)} + \mathbf{e}^{(\mathbf{0})} = \mathbf{w}^{(0)} - J^{-1}(\mathbf{w}^{(0)})G(\mathbf{w}^{(0)})$.

step 5: $\mathbf{w}^{(1)}$ is then used to compute $\mathbf{w}^{(2)}$ which is then used to compute $\mathbf{w}^{(3)}$ and so on. The iteration is stopped after say j iterations if the difference between $\mathbf{w}^{(\mathbf{j-1})}$ and $\mathbf{w}^{(\mathbf{j})}$ is negligible. $\mathbf{w}^{(\mathbf{j})}$ is the estimate of the roots of the system of equations.

To apply the Newton-Raphson method in the estimation of the coefficients of the explanatory variables we need to define the following terms appropriately to fit equation (3.3.10).

Let $G(\boldsymbol{\beta})$ be a $q \times 1$ vector function of the first derivatives of the log-likelihood. It is called the efficient scores function and is given by;

$$G(\boldsymbol{\beta}) = \left(\frac{\partial logL(\beta)}{\partial \beta_1}, \ldots, \frac{\partial logL(\beta)}{\partial \beta_q} \right).$$

The maximum likelihood estimate, $\hat{\boldsymbol{\beta}}$ is obtained by solving $G(\hat{\boldsymbol{\beta}}) = 0$.

The Jacobian matrix $J(\beta)$ is a $q \times q$ matrix of negative second order derivatives of the natural logarithm of the likelihood, such that the entry in the k^{th} row and l^{th} column is given by

$$J(\boldsymbol{\beta})_{kl} = -\frac{\partial^2 logL(\beta)}{\partial \beta_k \partial \beta_l}.$$

$J(\boldsymbol{\beta})$ is called the observed information matrix. The modification of equation (3.3.10) to be used to estimate the β's is thus given by;

$$\boldsymbol{\beta}^{(j)} = \boldsymbol{\beta}^{(j-1)} - J^{-1}(\boldsymbol{\beta}^{(j-1)})G(\boldsymbol{\beta}^{(j-1)})$$

Suppose that after k iterations it is observed that there is no significant change in the log-likelihood function, this will mean that the iteration has converged and the estimates of the β's is given by the entries of vector $\boldsymbol{\beta}^{(k)}$.

To calculate the variance covariance matrix C we compute the negative inverse of the observed information matrix at $\hat{\boldsymbol{\beta}}$, the estimate of $\boldsymbol{\beta}$ obtained by the maximum likelihood method. This is given by

$$C = -J^{-1}(\bar{\boldsymbol{\beta}}) = -\begin{bmatrix} \frac{\partial^2 logL(\beta)}{\partial\beta_1^2}\Big|_{\beta=\hat{\beta}} & \frac{\partial^2 logL(\beta)}{\partial\beta_1\partial\beta_2}\Big|_{\beta=\hat{\beta}} & \cdots & \frac{\partial^2 logL(\beta)}{\partial\beta_1\partial\beta_q}\Big|_{\beta=\hat{\beta}} \\ \frac{\partial^2 logL(\beta)}{\partial\beta_2\partial\beta_1}\Big|_{\beta=\hat{\beta}} & \frac{\partial^2 logL(\beta)}{\partial\beta_2^2}\Big|_{\beta=\hat{\beta}} & \cdots & \frac{\partial^2 logL(\beta)}{\partial\beta_2\partial\beta_q}\Big|_{\beta=\hat{\beta}} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial^2 logL(\beta)}{\partial\beta_q\partial\beta_1}\Big|_{\beta=\hat{\beta}} & \frac{\partial^2 logL(\beta)}{\partial\beta_q\partial\beta_2}\Big|_{\beta=\hat{\beta}} & \cdots & \frac{\partial^2 logL(\beta)}{\partial\beta_q^2}\Big|_{\beta=\hat{\beta}} \end{bmatrix}^{-1}$$

The main diagonal entries of this matrix gives the variance of the maximum likelihood estimates $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_q$ whose square-root is the standard error of the corresponding estimates.

The estimates represented by vector $\hat{\boldsymbol{\beta}}$ are asymptotically unbiased. The $(1-\alpha)100\%$ confidence interval of the β's is given by the formula $[\hat{\beta}_j \pm z_{\alpha/2} se(\hat{\beta}_j)]$ for $j = 1, 2, \ldots, q$. Where $se(\beta_j)$ is the standard error of β_j for $j = 1, 2, \ldots, q$ and $z_{\alpha/2}$ gives the upper $\alpha/2$-point of the standard normal distribution. In the event that the computed confidence interval of a particular parameter β contains zero, then there is a chance that the true value of parameter β takes the value zero in the presence of the other parameters. In this regard, it will be prudent to conduct a hypothesis test to test the significance of the parameter to the model in the presence of the other parameters.

To test the null hypothesis that $\beta_j = 0, j = 1, 2, \ldots, q$ in the presence of all the other terms , we use the test statistic $W^2 = \left(\frac{\hat{\beta}_j}{se(\beta_j)}\right)^2$. This statistic is then compared to the percentage points of a chi-squared distribution on one degree of freedom. If the test statistic is big we will reject the null hypothesis and concluded that the covariate whose coefficient is β_j is not significant in the presence of the other covariates in the model. Its effect will have be analysed by fitting it alone. This is explained in model fitting subsection below.

3.4 Model Fitting

Practically, during studies values of several explanatory variables are recorded. As part of the modelling process, we need to determine which explanatory variables should be included and which should be left out. To do this, we need to come up with a way to asses the contribution of an explanatory variable to the model.

The statistic $-2logL(\hat{\beta})$ measure the fitness of a model to the sample data. The value of the statistic is obtained by computing the likelihood by replacing the parameter β's by the maximum likelihood

estimates, then taking the natural logarithm and then multiplying by -2. The smaller the value of the test statistic $-2logL(\hat{\beta})$ for a particular model signifies a better fit.

3.4.1 Assessing the effects of the covariates. Suppose the effects of adding additional explanatory variables into a model is to be assessed. Let two models , model A and B be fitted to the given data set. Suppose we fit u covariates $Z_1, Z_1, \ldots, Z_u$ in model A and $u+v$ covariates $Z_1, Z_1, \ldots, Z_u, Z_{u+1}, \ldots, Z_{u+v}$ in model B. Then the hazard functions will be given by $\exp\left\{\sum_{j=1}^{u}\beta_j z_j\right\}h_0(t)$ and $\exp\left\{\sum_{j=1}^{u+v}\beta_j z_j\right\}h_0(t)$ for models A and B respectively.

To determine whether the v covariates that have been added improves the model or not, we need to compute the log-likelihood statistic.

3.4.2 Log-likelihood statistic. Let $-2log\hat{L}(modelA)$ and $-2log\hat{L}(modelB)$ be the likelihood statistics associated with model A and B respectively. The log-likelihood ratio statistic(LRS) is given by

$$\begin{aligned}\text{LRS} &= -2log\hat{L}(modelA) - (-2log\hat{L}(modelB)) \\ &= -2[log\hat{L}(modelA) - log\hat{L}(modelB)]\end{aligned}$$

The log-likelihood ratio statistic is used for testing the null hypothesis that the coefficients of the additional explanatory variables, $\beta_{u+1}, \beta_{u+2}, \ldots, \beta_{u+v}$ take the value zero.

Asymptotically, the log-likelihood ratio statistic follows a chi-squared distribution whose degree of freedom is computed by taking the difference between the number of explanatory variables in the models being fitted. In our examples it will be v degrees of freedom.

In the event that the value of the test statistic is significantly small, the null hypothesis is not rejected. In this scenario both models are considered to be equally suitable. Thus the model with less covariates is preferred. However, if the test statistic's value is significantly large, the null hypothesis is rejected and the model with more terms is deemed a better fit hence it is preferred.

3.5 Model Selection Strategy

The preliminary stage in model selection process involves identifying a set of covariates that have a high chance of being included in the linear component of the model. This set of covariates will include the interactions as well. It is from this set that we will obtain the combination of covariates to be added in the model. If interaction is to be included in the model by the hierarchic principle (Nelder, 1977) the main effects have to be included as well.

If the number of covariates is not too large we can fit all possible models and compare their respective $-2log\hat{L}$ statistic and then pick the model(s) that give the least value of these statistics. However in most cases the number of explanatory variables is huge thus fitting all the models is computationally costly.

Several procedures have been developed to assist in selecting variables to be included in the model. They include forward selection procedure, backward selection procedure and step wise procedure.

Under forward selection procedure, the null model is fitted first, then variables are added one by one. At each step, variables that lead to the largest decrease in the value of $-2log\hat{L}$ on their addition are

the ones included in the model. The process stops when the next variable to be added does not reduce the value of $-2log\hat{L}$ by more than a predetermined amount called the stopping rule (Collett, 1994).

Under backward selection, all variables are first fitted then the variables are eliminated one by one. At each step, variables that lead to the smallest increase in $-2log\hat{L}$ are the ones to be removed. The procedure goes on and on until the next variable to be eliminated leads to an increase in $-2log\hat{L}$ that exceeds the amount determined by the stopping rule.

The step wise procedure is a combination of both the forward and the backward procedures. In this procedure, the variables already included will still be subjected to removal test at later steps. So after a variable has been added, the procedure checks if any of the variables already in the model can be omitted.

The disadvantage of these procedures is that they produce one set of variables that have the potential to be included in the model and different procedures lead to different results.

The following steps give a better way to select a model.

Step 1: Determine which variables significantly reduce the value of $-2log\hat{L}$ statistic. This is achieved by fitting models that contain the explanatory variables on their own then comparing them with the null model.

Step 2: All the variables that were viewed as significant in step 1 are fitted together then the variables are omitted one at a time just as is the case in the backward procedure. The ones that do not significantly decrease the value of $-2log\hat{L}$ are omitted.

Step 3: The variables that were left out in step 1 are added in this step one by one as was the case in the forward procedure. The ones that significantly reduce the size of $-2log\hat{L}$ are retained.

Step 4: The variables that finally make to the model in step 3 are tested once more to ensure that each will significantly increase the value of $-2log\hat{L}$ if they are removed. Also the variables left out are checked to ensure that none can significantly reduce the value of $-2log\hat{L}$ if included in the model.

When undertaking this procedure we allow flexibility in the choice of the significant levels.

3.6 Estimating The Hazard and Survivor Functions

Let vector $\mathbf{z_j}^T = (z_1, z_2, \ldots, z_q)$ be a vector representing the q covariates. Let $\boldsymbol{\beta}^T = (\beta_1, \beta_2, \ldots, \beta_q)$ be another vector representing the estimated coefficients of the explanatory variables. Then the hazard function of the j^{th} life in a study of n lives will be estimated by

$$\hat{h}_j(t) = \exp(\hat{\beta}^T \mathbf{z_j})\hat{h}_0(t)$$

where $\hat{h}_0(t)$ is the estimate of the baseline hazard function. Knowledge of $\hat{h}_0(t)$ will enable us to estimate the value $h_j(t)$, the hazard function for the j^{th} life.

To estimate the baseline hazard function Kalbfleisch and Prentice (1973) used the method of maximum likelihood. According to them, suppose there are d_j deaths and n_j lives at risk of death at each time of death $t_{(j)}, j = 1, 2, \ldots, r$, then the estimate of the baseline hazard function is given by

$$\hat{h}_0(t_{(j)}) = 1 - \hat{s}_j.$$

$\hat{s}_j$ is obtained by solving the following equation

$$\sum_{k\in D(t_{(j)})} \frac{\exp(\hat{\boldsymbol{\beta}}^T \mathbf{z_k})}{1-\hat{s}_j^{\exp(\hat{\boldsymbol{\beta}}^T \mathbf{z_k})}} = \sum_{k\in R(t_{(j)})} \exp(\hat{\boldsymbol{\beta}}^T \mathbf{z_k}) \quad j=1,2,\ldots,r \tag{3.6.1}$$

$D(t_{(j)})$ is the set of all d_j lives observed to die at the j^{th} death time, $t_{(j)}$ and $R(t_{(j)})$ is the risk set.

To derive equation (3.6.1) is tasking, therefore we will not look into here. However, its derivation can be found in (Kalbfleisch and Prentice, 1973).

Assuming a constant hazard rate between neighbouring times of death then the function of the baseline hazard will be a step function. To find the instantaneous baseline hazard function we divide the hazard function by the time interval as shown below.

$$\hat{h}_0(t) = \frac{1-\hat{s}(t)}{t_{(j+1)}-t_{(j)}}.$$

Since $\hat{s}(t)$ approximates the probability of surviving in the interval $t_{(j)}$ to $t_{(j+1)}$, the estimate of the baseline survivor function will be given by

$$\hat{S}_0(t) = \prod_{j=1}^{k} \hat{s}(t) \quad t_{(k)} \le t < t_{(k+1)}, k=1,2,\ldots,r-1$$

$\hat{S}_0(t)=1$ for $t<t_{(1)}$ and $\hat{S}_0(t)=0$ for $t \ge t_{(r)}$ if t^* the last censored survival time is less than $t_{(r)}$. If $t^* > t_{(r)}$ then $\hat{S}_0(t) = \hat{S}_0(t_{(r)})$ until $t=t^*$ but is undefined for $t>t^*$.

The baseline cumulative hazard function is given by;

$$\hat{H}_0(t) = -log\hat{S}_0(t) = -log\prod_{j=1}^{k}\hat{s}_t = -\sum_{j=1}^{k} log\hat{s}_t.$$

for $t_{(k)} \le t < t_{(k+1)}, k=1,2,\ldots,r-1$, $\hat{H}_0(t)=0$ for $t<t_{(1)}$.

The estimate of the cumulative hazard for the j^{th} life in the study whose explanatory variables are given by the vector $\mathbf{z_j}$, is thus given by;

$$\hat{H}_j(t) = \int_0^t \hat{h}_j(u)du = \exp(\hat{\boldsymbol{\beta}}^T \mathbf{z_j}) \int_0^t \hat{h}_0(u)du = \exp(\hat{\boldsymbol{\beta}}^T \mathbf{z_j})\hat{H}_0(t).$$

and the corresponding estimate of the survivor function is given by

$$\hat{S}_j(t) = \exp(-\hat{H}_j(t)) = \left(\hat{S}_0(t)\right)^{\exp(\hat{\boldsymbol{\beta}}^T \mathbf{z_j})} \quad , t_{(k)} \le t < t_{(k+1)}, k=1,2,\ldots,r-1. \tag{3.6.2}$$

In the absence of covariates, equation (3.6.1) becomes $\frac{d_j}{1-\hat{s}_j} = n_j$, which implies that $\hat{s}_t = \frac{n_j-d_j}{n_j}$ and $\hat{h}_0(t) = \frac{d_j}{n_j}$. Therefore, $\hat{S}_0(t) = \prod_{j=1}^{k} \frac{n_j-d_j}{n_j}$, which is the Kaplan-Meier estimate of the survivor function as given in equation (2.1.1). Thus, equation (3.6.1) is the generalisation of the kaplan-Meier to incorporate situations when the hazard function depends on the explanatory variables.

Table 3.4: Estimates of the coefficients of covariates obtained after fitting the cox proportional hazards to the data

Variable	$(\hat{\beta})$	$(\text{se}\hat{\beta})$	$\exp(coef)$	p-value
Age only				
Age	0.1128	0.0045	1.119	$< 2 * 10^{-16}$
Gender only				
$Gender$	0.2552	0.0795	1.291	0.00132
Month last observed only				
Mlo	0.4846	1.6235	0.0488	$< 2 * 10^{-16}$
Age + Gender				
Age	0.1136	0.0045	1.120	$< 2 * 10^{-16}$
$Gender$	0.3683	0.0797	1.445	$3.82 * 10^{-6}$
Age + MLO				
Age	0.1104	1.1167	0.0045	$< 2 * 10^{-16}$
MLO	0.2368	1.2672	0.0427	$2.94 * 10^{-8}$
Gender + MLO				
$Gender$	0.2176	1.2431	0.0797	0.00632
MLO	0.4746	1.6073	0.0487	$< 2 * 10^{-16}$
Age + Gender + MLO				
Age	0.1112	1.1177	0.0045	$< 2 * 10^{-16}$
$Gender$	0.3411	1.4065	0.0799	$1.96 * 10^{-5}$
MLO	0.2226	1.2493	0.0424	$1.54 * 10^{-7}$

Table 3.5: Values of $-2log\hat{L}$ for various models fitted to the given data

Variables in the model	$-2log\hat{L}$
$none$	8578.6
Age	7841.6
$Gender$	8568.4
MLO	8490.6
$Age + Gender$	7820.4
$Age + MLO$	7811.6
$Gender + MLO$	8483.2
$Age + Gender + MLO$	7793.4

3.7 Results and Discussion

The tables, Table 3.4 and Table 3.5 below give a summary of the results obtained after using R software.

The first step was to fit the null model. After this we fitted Age, $Gender$ and MLO(Month last observed) in models of their own alone. These covariates led to a reduction of 737, 10.2 and 88 in the value of $-2log\hat{L}$ respectively. Comparing these values with chi-square percentage points, the reductions are significant at 1% level of significance. For this reason, we fitted a cox model whose linear component contains the three covariates. when all the three covariates are fited, $-2log\hat{L}$ reduces to 7793.4 which is a significant reduction at 1% level of significance. The next step is all about assessing the effects

of omitting the covariates already fitted. We first began by removing the month last observed(MLO) explanatory variable. When MLO is omitted the value of $-2log\hat{L}$ increases by 27. This increase is significant at 1% level of significance. When $Gender$ is omitted, the new value of $-2log\hat{L}$ is given by 7811.6 which is a significant increase. The same thing happens when Age is omitted, $-2log\hat{L}$ increases by 689.8 which is significant at 1% level of significance.

Therefore, since we cannot remove the covariates from a model that contains all the three of them without causing a significant increase in the value of $-2log\hat{L}$, the most satisfactory model will be the model that contains Age, $Gender$ and Month last observed(MLO) in its linear component. That is,

$$h_j(t) = \exp(0.1112Age_j + 0.3411Gender_j + 0.2226MLO_j) \quad j = 1, 2, \ldots, n \tag{3.7.1}$$

3.8 Residuals for the Cox Proportional Hazard's Model

Once a model has been fitted, its adequacy has to be assessed. Model diagnosis in the Cox regression model is different and slightly complicated than those employed in linear regression because of the presence of censored data. Model diagnosis is based on residuals. Residuals are quantities that can be computed for each life in the sample. These residuals are such that their behaviour is known at least roughly when the fitted model is adequate.

There are a number of residuals that are used to check various aspects of model adequacy. In this section, we review the Cox-Snell residual

3.8.1 Cox-Snell Residuals. Cox-Snell residual was first proposed by Cox and Snell in their paper (Cox and Snell, 1968). For any individual $j, j = 1, 2, \ldots, n$ the Cox-Snell residual is given by

$$rcs_j = \exp(\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\mathbf{z_j})\hat{H}_0(t_j) = \hat{H}_j(t) = -log\hat{S}_j(t), \tag{3.8.1}$$

where, $\hat{H}_0(t_j)$, $\hat{H}_j(t)$ and $-log\hat{S}_j(t)$ are the estimates of the cumulative baseline hazard function, cumulative hazard function and the survival function for the j^{th} individual at time t_j.

The derivation of the residual is obtained from the result in mathematical statistics that if T is a random variable representing the survival times and $S(t)$ is the survival function, the random variable $Y = -logS(t)$ has an exponential distribution with a mean of one. Therefore, if an adequate model has been fitted, the estimated values for the j^{th} individual, $\hat{S}_j(t_j)$ at time t_j should be close to the true values $S_j(t_j)$. They should be having similar properties. That is, $-log\hat{S}_j(t_j)$ should approximately have an exponential distribution with a mean of one. Therefore, testing whether the residuals have an exponential distribution with unit mean is the test of model adequacy.

Once the residuals have been obtained, we plot a graph of the cumulative hazard function of the Cox-Snell residuals against the Cox-Snell residuals. A plot whose gradient is one and the intercept is zero implies an adequate model.

3.8.2 Results and Discussion. In this section we check model adequacy for the model given in equation (3.7.1) in the previous section. The plot of the cumulative hazard of Cox-Snell residual against the Cox-Snell residual is given in Figure 3.1. From the Figure we observe that the plot is a straight line with unit gradient and a zero intercept. This implies that our model is a better fit.

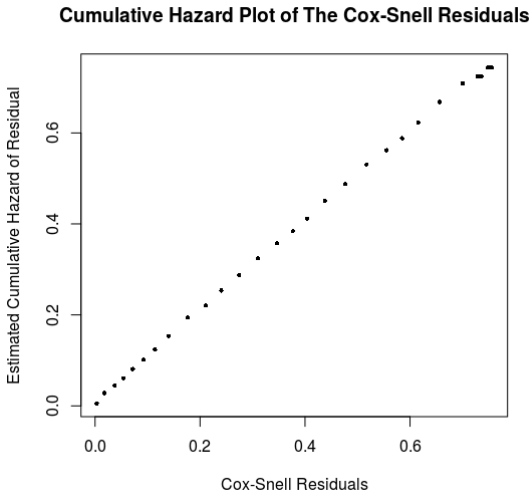

Figure 3.1: Cumulative hazard plot of the cox-snell residuals

4. Cox Model With Time-Dependent Variables

4.1 Time-Dependent Model

In this section, we investigate hazard functions that depend on covariates that vary with time.

Time dependent covariates are broadly classified into internal variables and external variables.

Internal variables are dependent on the life of an individual in the study. Their measurements can only be taken as long as the subject of interest is alive. In medical studies variables such as systolic blood pressure, size of tumor, the level a patient is in with regard to a particular disease are all categorized as internal variables.

External variables on the other hand, are variables that depend on time and can be measured even if the subject of interest in the study has left the study either by death or censoring. External variables evolve in a way that we can tell at the outset what values they will take in subsequent times. A good example is the age of an individual in the study.

When we have a term whereby the coefficients are time dependent and the variables are time independent, we can apply some transformation and have the term analysed as a time dependent variable with constant coefficient as shown in the following example.

Suppose Z is an explanatory variable whose coefficient is a function of time, βt say. The term βtz is expressed as $\beta z(t)$ where $z(t) = zt$ is a variable that depends on time. For this reason, models whose linear components contain some of its explanatory variables being time dependent cannot be called proportional models. Instead, they are called cox regression model.

The time dependent cox regression model of the hazard function at time $t \geq 0$ for the j^{th} life in a study of n lives is thus given by;

$$h_j(t) = \exp\left(\sum_{i=1}^{q} \beta_i z_{ij}(t)\right) h_0(t) \quad t \geq 0, \tag{4.1.1}$$

where $h_0(t)$ is the baseline hazard function which is the hazard function at time t for a life whose value of the explanatory variable are constant through time and take the value zero. $z_{ij}(t)$ is the value of the i^{th} time dependent covariate for the j^{th} life at time t.

Since the log hazard ratio $\frac{h_j(t)}{h_0(t)}$ depends on time, it begs the question, what is the interpretation of the β-parameters for such a model?

The interpretation of the β's can best be explained by considering the following example

Let $h_k(t)$ and $h_l(t)$ be the hazard functions of individuals k and l respectively. Then the ratio of their hazard function is given by

$$\frac{h_k(t)}{h_l(t)} = \exp\left(\sum_{i=1}^{q} \beta_i [z_{ki}(t) - z_{li}(t)]\right). \tag{4.1.2}$$

The coefficient $\beta_j, j = 1, 2, \ldots, q$ can now be interpreted to be the natural logarithm of the relative hazard for two lives whose values of the covariates for the two lives at time t are similar except for the j^{th} covariates where their values differ by one unit.

To fit the model, the log likelihood function given in equation (3.3.3) is generalised as shown below.

$$\sum_{j=1}^{n} \rho_j \Big(\sum_{i=1}^{q} \beta_i z_{ij}(t_j) - log \sum_{k \in R(t_j)} \exp\big(\sum_{i=1}^{q} \beta_i z_{ik}(t_j)\big) \Big) \tag{4.1.3}$$

From the model, $R(t_j)$ is defined as previously. It is the set of lives facing the risk of death at time $t_j, j = 1, 2, \ldots, n$ and ρ_j is an indicator variable given by;

$$\rho_j = \begin{cases} 1, & \text{if life j is observed to die} \\ 0, & \text{if life j is censored} \end{cases}$$

Expression (4.1.3) is maximised by numerical methods techniques such as the Newton-Raphson method to give the estimates of the β's

For equation (4.1.1) to be used, the values of the explanatory variables for each individual in the risk set has to be known. However, this is challenging for internal variables because the values of the explanatory variables have to be estimated for all individuals in the risk set at each death time.

This is best illustrated by the example given below.

Suppose three individuals u, v and w are in the study. Suppose that only one time dependent explanatory variable has been recorded. Let $Z_j(t)$ be the random variable for time depended explanatory variable for individual j where $j = u, v, w$ at time t.

Then the hazard function for the j^{th} life $h_j(t)$ is modelled by the formula;

$$h_j(t) = \exp(\beta z_j(t)) h_0(t).$$

Suppose that individual u dies at death time t_u, individual v dies at death time t_v and individual w has his survival time being right censored such that the last observation was recorded at time t_w. Suppose $t_u < t_v < t_w$. If $Z(t)$ is recorded at regular time intervals called patient times , then the survival times of these patients can be represented graphically as shown in Figure 4.1 below.

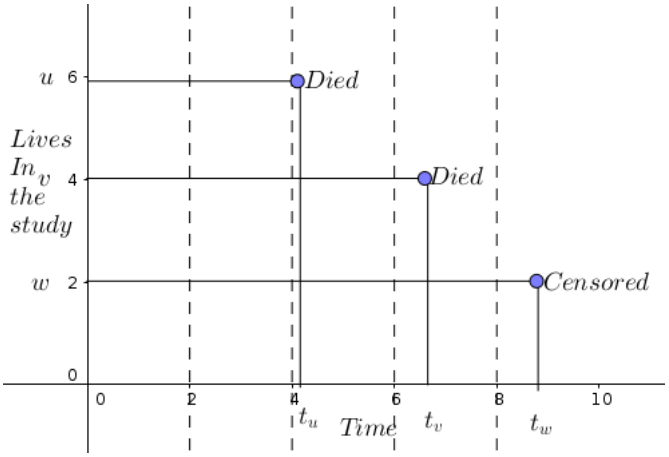

Figure 4.1: Survival experience of u, v and w in patient time

At death time t_u the contribution of individual u to the log likelihood function given in (4.1.3) is given by

$$\beta z_u(t_u) - log \sum_{k} \exp(\beta z_k(t_k)). \tag{4.1.4}$$

for $k = u, v, w$ and $z_u(t_u)$ is the realisations of the explanatory variable for life u at death time t_u. Expanding expression (4.1.4) we get

$$\beta z_u(t_u) - log\Big(\exp(\beta z_u(t_u)) + \exp(\beta z_v(t_v)) + \exp(\beta z_w(t_w))\Big).$$

This means that the values of the time dependent variables for individuals in the risk set at death times t_u and t_v are required. That is at death time t_u; z_u, z_v and z_w are required and at death time t_v; z_v and z_w are required.

Suppose the value of these time dependent explanatory variables are required at intermediate times as shown in figure 4.1 . We can either estimate these variables by either using the last recorded value, the closest value or linear interpolation if the explanatory variable is a quantitative variable.

4.2 Estimation of The Baseline Hazard

The baseline hazard function is estimated using the techniques discussed in section 3.6

Having obtained the hazard function, using the relationships given in section 1.3.2 it is easier to compute the values of the baseline cumulative hazard function and the baseline survivor function. Since the formula to estimate the survivor function that is given in equation (3.6.2) cannot be used because it no longer holds for the case of time dependent explanatory variables, we need to use a different type of relationship.

To obtain the estimate of the survivor function for the j^{th} life in the study, we use the relationship

$$S_j(t) = \exp\Big(-\int_0^t h_j(u)du\Big)$$
$$\implies S_j(t) = \exp\Big(-\int_0^t \exp\big(\sum_{i=1}^{q}\beta_i z_{ij}(u)\big)h_0(u)du\Big).$$

4.3 Results and Discussion

Our objective is to asses whether age as a time dependent variable has a significant effect on the hazard function compared to when it is fitted as a time independent variable. We denote the time dependent age variable by $age(t)$. The first step is to fit $age(t)$ alone. The output from R is given in the table below. From table 4.1 we observe that the $p-value$ is very small. This is an evidence that that the time

Table 4.1: Estimates of the coefficient of the time-dependent age variable obtained after fitting the cox regression model to the data

Variable	$(\hat{\beta})$	$(\text{se}\hat{\beta})$	$\exp(coef)$	p-value
age(t) only				
$age(t)$	0.1125	0.0045	1.119	$< 2 * 10^{-16}$

dependent age variable has a significant effect on the hazard rate. To asses whether a model containing

$age(t)$, MLO and $Gender$ is a better fit than the one containing a time independent age variable we fit the two models and compare there $-2log\hat{L}$ statistics. since the latter model had already been fitted in chapter 3, in this chapter we fit the model containing the time dependent covariate. The table below gives the summary of the model and the test statistic. From Table 4.2 , the value of $-2log\hat{L}$ statistic

Table 4.2: Values of $-2log\hat{L}$ for the model fitted to the given data

Variables in the model	$-2log\hat{L}$
$Age(t) + Gender + MLO$	8479.4

is 8479.4. In comparison to the value obtained in table 3.5 chapter 3 which is 7793.4, we see that $7793.4 < 8479.4$. Since the smaller the value of the test statistic implies a better model, we conclude that the time independent model fitted in chapter 3 gives a better fit.

5. Conclusion

We have seen that despite the incomplete nature of survival data, various techniques have been developed to analyse this special type of data. The Kaplan-Meier estimate of the survivor function is easy to compute and can be plotted in a graph to give a quick visual impression on the behaviour of the survival functions of individuals in a study. After getting a rough picture on the survival experience among groups, the log-rank test can be employed to test whether differences observed are any significant.

Despite all the advantages the non-parametric methods offer, the effects of covariates can not be captured in these models. For this reason, Cox proportional model under its main assumption that the hazard rates between groups is proportional to each other at each time $t \geq 0$ comes in handy. Cox proportional hazard model enables covariates with significant effect on the hazard function be incorporated into the model. Moreover, to capture the dynamism of real life scenarios where mortality depends on the values of covariates at that particular point in time the Cox proportional model is extended to accommodate time dependent covariates. The techniques discuss herein are just but a tiny drop in an ocean, several efficient and advanced methods to analyse survival data have been developed. Machine learning algorithms such as random forest are currently being employed to analyse survival data. Further research can be conducted on machine learning techniques and the results obtained can be compared to those generated by traditional methods.

Acknowledgements

I would like to first thank Almighty God for His blessings, providence and care. Secondly, I would like to thank my supervisor Dr Viani Biatat and tutor Carine Umulisa for continuous advice and guidance. Lastly I would like to thank my family and friends for their emotional and spiritual support throughout the essay phase.

References

D. Bernoulli. Essai d'une nouvelle analyse de la mortalité cause par la petite vérole et des avantages de l'inoculation pour la prévenir. histoire de l'académie royale des sciences avec les mémoires de mathématique et de physique tirés des registres de cette académie. paris 1766 (année 1760). *History of Actuarial Science*, 8, 1766.

N. Breslow. Covariance analysis of censored survival data. *Biometrics*, pages 89–99, 1974.

D. Collett. Modelling survival data. In *Modelling Survival Data in Medical Research*, pages 53–106. Springer, 1994.

D. R. Cox. Models and life-tables regression. *JR Stat. Soc. Ser. B*, 34:187–220, 1972.

D. R. Cox. Partial likelihood. *Biometrika*, pages 269–276, 1975.

D. R. Cox and E. J. Snell. A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 248–275, 1968.

A. de Moivre. *Annuities Upon Lives: Or, the Valuation of Annuities Upon Any Number of Lives; as Also, of Reversions. To which is Added, an Appendix Concerning the Expectations of Life, and Probabilities of Survivorship. By A. de Moivre. FRS*. S. Fuller, 1731.

B. Gompertz. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical transactions of the Royal Society of London*, 115:513–583, 1825.

M. Greenwood et al. A report on the natural duration of cancer. *A Report on the Natural Duration of Cancer.*, (33), 1926.

J. D. Kalbfleisch and R. L. Prentice. Marginal likelihoods based on cox's regression and life model. *Biometrika*, pages 267–278, 1973.

J. D. Kalbfleisch and R. L. Prentice. Relative risk (cox) regression models. *The Statistical Analysis of Failure Time Data, Second Edition*, pages 95–147, 2002.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

W. M. Makeham. On the law of mortality and construction of annuity tables. *The Assurance Magazine and Journal of the Institute of Actuaries*, 8(06):301–310, 1860.

N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. 1959.

J. Nelder. A reformulation of linear models. *Journal of the Royal Statistical Society. Series A (General)*, pages 48–77, 1977.

T. M. Therneau and P. M. Grambsch. *Modeling survival data: extending the Cox model*. Springer Science & Business Media, 2013.

H. Westergaard. Contributions to the history of statistics. 1932.