UNIVERSITY OF LEEDS

Survival Analysis: Model Building and Frailty Models

MATH3001: Project in Mathematics

Faye Williams
201308646
Supervisor: Dr Leonid Bogachev

A report presented for the degree of
Mathematics, Bachelor of Science

Department of Mathematics
University of Leeds
United Kingdom
April 2022

Contents

1 Introduction

This report introduces survival analysis in a practical and convenient way, starting with the theory needed to secure ones knowledge of survival analysis in section I. We shall introduce the basic concepts governing survival analysis in chapter 2, leading us on to methods of non-parametric estimation in chapter 3. Next we will look at ways of assessing the fit of survival curves made using non-parametric estimators by the log-rank test in chapter 4. As well as estimating the variance, mean and median for non-parametric estimators in chapter 5. We will progress on to a description of the key parametric survival distributions in chapter 6, which leads us on to maximum likelihood estimation and its application in the presence of censoring in chapter 7. We shall finish the theory by introducing the two key models in survival analysis, namely the Cox model in chapter 8 and the Accelerated Failure Time model in chapter 9. Beginning section II is an introduction of model building, which takes the reader through the steps taken to produce a model and applying this to a data set in chapter 10 . This leads on to estimating the regression coefficients for our chosen model in chapter 11. We finish the report by introducing frailty models in section III, discussing what they are and the many versions of them in chapter 12. Lastly we will apply this knowledge to build a frailty model based on a particular data set in R in chapter 13.

1.1 Use of R

I shall be using the statistical software R to conduct my analysis of survival data. Within R there are many examples of survival data sets, under the ‘survival’ library. I have chosen to use the data set ‘ovarian’ throughout my report to demonstrate the theory underpinning survival analysis that we shall cover in section I. This records the time taken to death or censoring due to Ovarian cancer. After this, we shall continue with this data set and look at building a suitable model for this data in section II. In the final section I will also introduce another survival data set named ‘kidney’, which records the time taken for an infection to occur once a kidney catheter is inserted. We shall conduct further analysis on the ‘kidney’ data set by building an appropriate frailty model of this data. I will provide a glimpse of this data in the appendix A.2.

1.2 Ethics in my Report

In my research I have included data from medical studies, specifically those involving Ovarian cancer and kidney issues. The benefits of research in these sectors is immense, as it can help to determine which treatment is most effective in curing cancer, and it can provide us with insight into what factors may influence the onset of conditions and diseases. Other fields of research involved include the engineering and technology sectors, for which survival research allows us to see which factors cause technology to fail. This helps aid development of future technology to make it more effective and less likely to fail, as well as expanding our knowledge in this field. This benefits not only the engineering industry but society as a whole. I have taken sample data from the ‘ovarian’ and ‘kidney’ data sets, these data sets are reputable and protect the participant’s privacy and dignity by keeping them anonymous, meaning there is no breach in confidentiality.

Despite this, research in this field comes with risks, those of which include making false conclusions. For example, I have used the ‘ovarian’ data set in R throughout my report which measures the time taken until death or censoring occurs in patients of Ovarian cancer. The data set includes the treatment the subjects are receiving, which is given as one of two possible treatments, here we may falsely conclude that one treatment is more effective at curing Ovarian cancer than the other. The seemingly more effective treatment may then be taken as the recommended treatment for patients with ovarian cancer, which could potentially cause the mortality rate to increase, which could have been prevented if the participants had been given the alternative treatment. Therefore conducting survival research can come with many issues including false conclusions and misinterpretation of results.

I General Theory

2 What is Survival Analysis?

Survival analysis, also known as ‘time-to-event’ analysis is a branch of statistics used for analysing the expected duration of time until one specific event occurs, often denoted as failure and the time at which it occurs is termed failure time. Usually we aim to study the relation between this failure and other primary variables we have recorded, such as an individual’s age or time since a previous event occurred [9]. Despite these being of greatest interest in survival research, we may also choose to study the influence of explanatory variables observed for each subject, such as ones height, sex and blood type, on their likelihood of failure. Alternatively, we may also wish to compare failure times between two or more different groups, often by examining whether one group is more prone to failure than the other.[8]

Uses. Survival analysis is most often used in a medical context, for which we measure the time taken for death of infected patients to occur, or for the development of a disease to take place, or more optimistically, time taken to recover. Nonetheless, survival analysis has a wide application of uses outside medical science, each with their own terms for the practice, such as reliability theory in engineering, event history analysis (EHA) in sociology, and duration analysis in economics. For instance, we may be investigating the lifetime of an electronic or engineering device in reliability analysis, or we may look at the time until a company reaches a break-even point in duration analysis, or time taken to divorce in EHA.

2.1 Censoring

Censoring occurs when we have a lack of information about the survival time of a certain individual, in this case the survival time is censored. Note that, despite censored data being less informative than typical data where t is known, we still record our observation of censored individuals in our data set [8]. Censoring can be caused by many factors, which we can group into 3 categories – right, left and interval censoring.

Right censoring. This is the most common type of censoring we come across in survival research and occurs when an individuals observed survival time is smaller than the actual survival time, of which is unknown to us. This can be due to subjects leaving the study early, loss of contact with a subject, or the study ending while the subject is still alive, as the event has not occurred for this subject.

Left censoring. Here the event occurs prior to a certain time, how long before this event however is unknown. In this instance we record the event at the first recorded time since that event has taken place. An example of this taking place is in a study of age at cancer reoccurring, a subject in the study may be found to have this type of cancer but were previously unaware. Therefore, the event is recorded to happen when they first enter the study, despite not knowing the exact time the cancer became present in the patient.

Interval censoring. Alternatively, interval censoring occurs when the survival from a given exposure to a given event, T, lies within a boundary of times, namely T_1 and T_2. This happens when the study takes observations periodically, therefore we cannot be certain what time within this interval the event happened and the exact survival time is unknown but restricted to a boundary.

2.2 Uniqueness of Survival Research

You may ask why survival analysis is a necessary branch of research and how it offers something different to the usual procedures used to build statistical models, such as linear least squares or best linear unbiased estimator. This can be explained by many factors, including the influence of censoring as defined above. In addition, the use of distribution functions in survival research is not what we would expect compared to most statistical research, and the use of time-to-event data opens doors to a wider scope of comparison and data analysis in contrast to alternative methods of statistical research.

Influence of censoring. The answer is that it allows us to include the influence of censored data in our estimations in comparison to other models, where this data is completely ignored. Therefore we are potentially losing out on extremely useful and influential data. This is particularly important when researching such serious topics like the effectiveness of drugs on severely ill patients.

Use of distribution functions. In many branches of statistics, we assume data to have a normal distribution, however in survival research we do not usually see this type of distribution [13]. This is because survival distributions are usually positively skewed, meaning most readings are on the left-hand side of the graph. In addition to this, we know the survival time is always non-negative; therefore our distribution will not be symmetric as we usually see in normal distributions.

2.3 Functions Used in Survival Analysis

To define our functions in survival data, we firstly need rigorous definitions of time and event to fit this context. The event we consider can be defined differently depending on the context of the research, most often it is defined as the death of a subject. However outside of the medical field, the event may be recorded as the time until an insurance claim in actuarial science, or time until a relapse in addiction sufferers in event history analysis. We define time as the duration the research is carried out over, this is from the start of the observation at time t=0 and lasts until the event occurs or the study is called to an end; therefore the time variable is non-negative. Alternately the time might finish if a subject withdraws from the study or there is a loss of communication in the case of censoring.

Firstly we shall define the probability density and cumulative distribution functions in this context and using these we shall derive the three main functions in survival data, listed below.

The probability density function. By definition the probability density function, $f(t)$, is the instantaneous rate of failure at a given time t, given by

$$f(t) = F'(t) = \frac{dF(t)}{dt} \tag{1}$$

where F(t) denotes the cumulative distribution function and is given below.

The cumulative distribution function.

$$F(t) = Pr(T \leq t) \tag{2}$$

where T is a random variable that denotes the time between the start of the trial and the event occurring and t is a specific value for T. Therefore, in survival analysis $F(t)$ denotes the probability that our event will occur before a specified time t.

The survival function. The complement of the cumulative distribution function is the survival distribution function, $S(t)$, which gives the probability that our subject of interest will survive past a certain time t, by the equation

$$S(t) = 1 - F(t) = Pr(T{>}t). \tag{3}$$

The survival function has some specific properties, such as when $t < 0$, $S(t) = 1$, as we cannot have negative readings of time. Furthermore, when t tends to infinity, the survival function equals zero, as our subject cannot have an endless lifetime and must fail or die at some point. Also, as expected $S(t)$ is non-increasing in t, in the same way $F(t)$ is increasing in t, as the probability of failure or death of a subject typically increases as time progresses. If we combine (1) and (3) we find $f(t) = \frac{-dS(t)}{dt}$ i.e., that the probability density function of t is given by the negative derivative of the survival function.

The hazard function. Conceptually, we can define the hazard function as 'the instantaneous rate of failure at T=t conditional upon survival to time t' [14], which can be represented by the equation

$$h(t) = \lim_{\triangle t \to \infty} \frac{Pr(t{\leq}T < t + \triangle t | T{\geq}t)}{\triangle t} \tag{4}$$

where $h(t)$ quantifies the instantaneous risk of failure occurring within time $[t, \triangle t)$, given survival up to time t and $\triangle t$ defines a small interval of time. We can alternately write an equation for $h(t)$ that relates to the survivor function above. Building on from (3), we define $h(t)$ as follows

$$h(t) = \frac{f(t)}{S(t)} = \frac{-d \log S(t)}{dt}. \tag{5}$$

The cumulative hazard function. $H(t)$, defines the risk of failure occurring before time t. We can think of this as the sum of all the risks faced going from time 0 to time t. Therefore, to find $H(t)$ we need to integrate the hazard rate over this interval, as seen below,

$$H(t) = \int_0^t h(s)\, ds = -\log S(t) \tag{6}$$

by the definition of the hazard rate given above. Here we will introduce the hazard ratio, which is defined as the ratio of the risk of a hazard occurring at time t in the trial compared to another group. For instance, if we have the hazard rates for two groups, h_1 and h_2, the hazard ratio at time t_1 is given by $\frac{h_1(t_1)}{h_2(t_1)}$. We will implement the hazard ratio when examining the log-rank test and the Cox proportional hazard model, where we assume the ratio remains constant over time [11], i.e. where for all $t_i \in t$ we have $\frac{h_1(t_i)}{h_2(t_i)}$.

We are able to graph the hazard function to give us greater understanding of the hazard in a sample over time. When the hazard rate is constant, i.e., $h(t) = h_0$, the graph simply shows a horizontal line, passing through h_0. In this case the survival model follows an exponential distribution, which we will explore further in chapter 6. Alternatively we can have an increasing hazard when $h(t_2){\geq}h(t_1)$ for $t_1{\leq}t_2$ and conversely, a decreasing hazard when $h(t_2){\leq}h(t_1)$. An example of a strictly increasing hazard function is an increasing Weibull model, and similarly for a strictly decreasing hazard function we have the decreasing Weibull model [17]. Additionally we may have functions containing both increasing and decreasing hazards, which we see this in the case of the log-normal survival model. Also the hazard function may be more difficult to understand than the survival function, nonetheless we can apply it to a wider range of techniques in survival analysis. To build on this, we will see applications of the hazard function being used in R in the following chapter, by applying the Nelson-Aalen estimator function. But firstly, let us introduce this notion of non-parametric estimation, which the Nelson-Aalen estimator falls under.

3 Nonparametric Estimation

We begin our research into the theory underpinning survival analysis by looking at non-parametric estimation, which is what builds the basis for most survival research. We introduce two non-parametric estimators, one an estimate of the survival function and another of the hazard function. Firstly let us define what we mean by non-parametric models. This is a model that does not rely on parameters or probability distributions to formulate it, instead using observations from data. This will be made up of independent, identically distributed data in this section and although we are assuming the true distribution of this data to be continuous, we estimate it using a discrete distribution.

As stated previously, standard distributions are not suitable for plotting survival data, as we require a positive survival time. This makes the use of non-parametric methods particularly beneficial in survival analysis, as they allow for the time variable to have these restrictions. In addition survival data often has features that are difficult to express using parametric models [13]. For instance in the case of plotting mortality rates (Appendix B.1), there is a decreasing hazard in the first roughly ten years of life, but this steadily increases after that. We would struggle to find a good fit for this model if we were limited to parametric models, so by assuming a non-parametric model we can overcome these hurdles. This leads us on to our two non-parametric estimators, the Kaplan-Meier (KM) and the Nelson-Aalen (NA) estimates.

3.1 The Kaplan Meier Estimator

The Kaplan Meier (1958) [15] estimator works by taking observations from a data set and generating an estimate of the survival function based on these, it calculates the probability that the survival time will be greater than time t. The estimate of survival time is given by the equation

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{7}$$

where t_i denotes the time passed to the next observation from the start of the study, d_i is given by the number of events that have occurred at the exact time t_i and n_i defines the number of subjects that have survived up until immediately before time t_i.

Assumptions. When using this estimator, we assume that all observations have the same survival probability, independent of whether they are censored or uncensored. We also assume the likelihood of the event taking place is the same for participants enrolled early or late and that the event occurs at the exact defined time. Lastly, we accept that the probability of censoring is the same for distinct groups, meaning censoring is independent of any extraneous factors.

Example. Here we shall be using the 'ovarian' data set. A glimpse of this data can be seen in figure 7 in appendix A.1 and we see there are 26 participants in total, with readings of their ages as well as values to quantify their treatment groups ('rx'), the regression of tumours ('resid.ds') and the patients performance ('ecog.ps'). The first column 'futime' denote the time the patients were tracked until death or censoring and whether the patients are censored or not is given by the second column, 'fustat', where 0 is censored and 1 non-censored. We use the function 'ggsurvplot' to display the curve, combined with the 'survfit' function to create the curve using the KM estimator [19]. We define the time and event variables for our data within the 'surv' function, which we relate to 1 to include the non-censored patients only. This code is given in figure 8 in A.1 and implementing this produces the graph below,

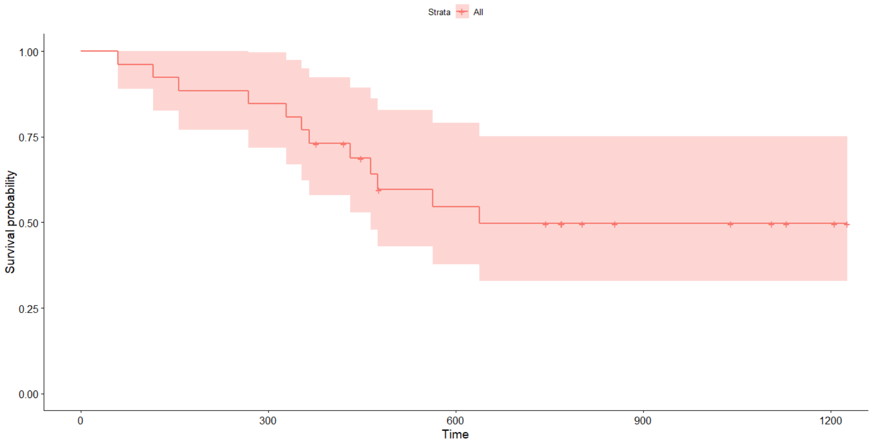


Figure 1: Kaplan Meier curve of 'ovarian' data set

As we can see, the plot shows a step function rather than a curve, this is due to our small data size as we only have 12 uncensored patients and therefore 12 steps on our plot. We see that as soon as one person dies, the survival rate decreases by $\frac{1}{26}$, due to there being a total of 26 participants in the study. We would expect to eventually reach a 0% survival rate, occurring once all patients have died, but as 14 of our patients are censored, the plot plateaus at 46.2%. The paler red area represents the confidence interval of the curve, which is at the default of 5%. Note that the drop in the step function increases in size over time, as we lose a greater percentage of the remaining people from death or censoring.

3.2 Nelson-Aalen Estimator

Originally this estimator was proposed by Altshuler [2], who looked at measuring competing risks in animal experiments. However Nelson went on to develop this research using counting processes, which made the estimator applicable to a greater scope of research. Here we will discuss the estimator, it's purpose and assumptions, as well as an example of its use in R.

The Nelson-Aalen (NA) estimator gives us an estimate for the cumulative hazard function. By definition, it calculates the instantaneous risk of failure occurring within a specific time interval, given the subject has survived up until that point based on some survival data. This is given by

$$\hat{H}(t) = \sum_{i:t_i \leq t} \frac{d_i}{n_i} \tag{8}$$

Here, d_i signifies the number of events occurring at time t_i and n_i represents the number of subjects at risk, i.e., subjects who have not yet failed or died, just prior to time t_i. Therefore, the sum of these for each recorded interval before time t gives the estimator.

Assumptions. Further, assumptions of this estimator include taking a constant hazard rate between successive event times and assuming censoring to be independent, i.e., the additional knowledge of censoring before any time t does not alter the risk of failure at time t. We require no assumptions on the distribution.

Example. Taking the code for the KM curve from A.1, we plot the 'ggsurvplot' using the function 'function(y)-log(y)' on the KM code, as this is how we convert the survival function into the cumulative hazard function, which we derived in equation (8). Thus, by the code in figure 9 A.1, the Nelson Aalen curve for the 'ovarian' data set is given as follows,

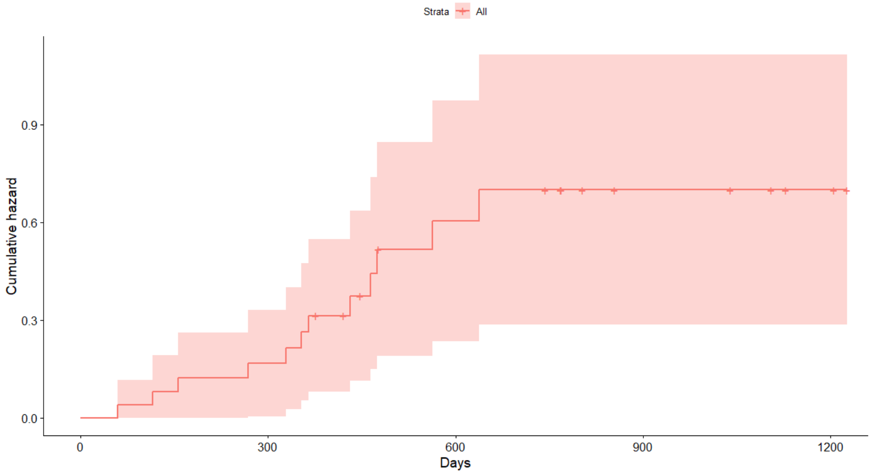


Figure 2: Nelson Aalen curve of 'ovarian' data set

Here the hazard steadily increases over the first roughly 650 days, which then levels off due to censoring, as indicated by the crosses along the line. Again, we see there is a large margin for the confidence intervals after 650 days, which is due to the high number of censored individuals making our estimate less accurate.

4 Comparison of Survival Curves

Often in survival research we seek to compare survival rates among different groups of participants, the Kaplan Meier estimator is useful for this. This has many applications in real life, for instance, we can take two groups of participants, one group trialling a new drug and the other group receiving a placebo drug and we can compare the differences in survival rates, here survival would be defined as the participant being cured of a specific illness. We compare two survival curves statistically by testing the null hypothesis, this being the notion that there is no difference in survival among the two groups. We shall investigate this using the Log Rank test, in two versions- the chi squared version and the normal version.

4.1 The Log-rank Test

Also known as the Mantel-Haenszel test (1959), the log-rank test is a non-parametric test, where we test for significance the null hypothesis against the alternative hypothesis. In this context, we often denote the null as there being no difference between the probability of death in two populations at any given point. If enough evidence is given against the null hypothesis, we can reject this and accept an alternative hypothesis, that there is a noticeable difference in survival between the two populations. Let us define $\hat{S}_1(t)$ and $\hat{S}_2(t)$ as the two survival functions, we have the null hypothesis and alternative hypothesis:

$$H_0 : \hat{S}_1(t) = \hat{S}_2(t) \quad \text{and} \quad H_1 : \hat{S}_1(t) \neq \hat{S}_2(t). \tag{9}$$

We continue with the same assumptions we have seen for the KM estimator, this includes the fact that censoring is independent, the likelihood of the events occurrence is the same for the participants enrolled early and late, and that the events happened at the times specified. We are most likely to find a significant difference between groups when the risk of an event is consistently greater for one group than for the other.

4.2 Chi-squared Version of the Log-rank Test

To carry out the log rank test, we start by calculating the test statistic, W_2, given as the weighted sum of squared deviations of the two groups,

$$W_2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \tag{10}$$

where E_1 and E_2 denote the expectation of the number of events in both populations, and O_1 and O_2 are the total number of events observed in each group. In this case we can use an approximation of the chi-squared distribution to give the test statistic, $W_2 \sim \chi_1^2$, and the P-value, $P(W_2 \geq w_2)$, and as previously mentioned we usually take the 95% confidence interval. Lastly, we need to determine the critical value, which we can find using the chi-squared table, which we will see more of in the example. If we find the test statistic value to be less than our p-value, we can reject the alternative hypothesis and instead accept there to be no significant difference between $\hat{S}_1(t)$ and $\hat{S}_2(t)$, with 95% confidence [8].

Example. As we require two models to compare, continuing with 'ovarian', I have separated the subjects into two groups based on their age. If subjects are below or exactly 50 years they belong to the "young" group, whereas subjects over 50 are in the "old" group, as shown in figure 10 in A.1. We shall test the null hypothesis; there is no difference in survival between the two age groups, against the alternate hypothesis that there is a significant difference in survival between these two groups. We apply the 'survdiff' function to give us the required information to carry out the log-rank test [19]. From figure 11 in A.1, we have that the chi-squared test statistic is 2.7 on 1 degree of freedom with the corresponding p-value as 0.1. This p-value indicates that we do not have a significant difference in survival between the two groups at a 5% significance level. Therefore we cannot conclude that there is a statistically significant difference in survival survival between the "young" and "old" subjects. However, as this p-value is small, we may see this separation by age covariate make some impact on the model we choose for the data when investigating further in part II.

4.3 Normal Version of the Log-rank Test

For this version of the log-rank test, we require a different equation for the test statistic. In this instance, we have the test statistic

$$U = O_1 – E_1 \tag{11}$$

where O_1 and E_1 are defined as before in 4.2. denote the expectation of the number of events in the two populations, and O_1 and O_2 are the total number of events we observe in each group.

5 Estimation of the Variance

We require an estimation of the variance of our survival function to help us construct confidence intervals, which we require to carry out hypothesis tests. From chapter 3 we know the KM formula gives us an estimate for the survival function using observed data, but this data is not always error free and we know that there is some level of uncertainty involved in the estimator. To evaluate the uncertainty we use Greenwood's formula.

5.1 Greenwood's Formula

This formula was first proposed by Greenwood in 1926 as the asymptotic variance of a life table estimator, since then its application has become more extensive and can now be used to estimate the variance of the Kaplan Meier estimator. Given that the distribution of t is discrete with finite many data points [14], we find the variance of the KM estimate by the equation:

$$Var(\hat{S}(t)) = \hat{S}(t)^2 \sum_{i:t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}. \tag{12}$$

In general, the square root of the variance gives the error term for our estimator, $\sigma(\hat{S}(t))$, thus the margin of error for the KM estimator is given by $\sqrt{Var(\hat{S}(t))}$. Now we have a formula for the variance of $\hat{S}(t)$, we can calculate confidence intervals for $\hat{S}(t)$. Here, we obtain the pointwise confidence interval for the survival function at time t_0 via the normal approximation of $\hat{S}(t)$, i.e.

$$\hat{S}(t) \pm z_{\frac{1-\alpha}{2}} \sqrt{Var(\hat{S}(t))}. \tag{13}$$

Although these intervals are useful, they are restricted to being valid only at a single point t_0. We can construct better intervals that we are able to use over an interval by altering the distribution $S(t)$.

5.2 Mean

We apply the general definition for expectation of a continuous function to find the mean survival time, denoted μ. Using the probability density function and survival function defined in section 2.3, we have

$$E[T] = \int_0^\infty tf(t)\ dt = \int_0^\infty S(t)\ dt. \tag{14}$$

Note the interval is given by $t \in [0, \infty)$ as t is a variable of time, which is non-negative. Further, we can use this equation to find the mean of the Kaplan Meier estimator, as seen below

$$\hat{\mu}_\tau = \int_0^\tau \hat{S}(t)\ dt \tag{15}$$

for $t \in (0, \tau)$. Furthermore, combining Greenwood's formula and the mean of $\hat{S}(t)$, we have the variance of the sample mean of the KM estimator given by

$$Var(\hat{\mu}_\tau) = \sum_{i:t_i \le t} \left[\int_0^\tau \hat{S}(t)\ dt\right]^2 \frac{d_i}{n_i(n_i - d_i)}. \tag{16}$$

5.3 Median

Typically, the median represents the centre value according to size of any given data set. In survival analysis, we can find the median in any set of uncensored observations by locating the centre observation from a list of n survival time $t_1, \dots, t_n$, ordered by $t_1 \le t_2 \le \dots \le t_n$. In this case, the median, M, is given by

$$\begin{cases} t_{\frac{n+1}{2}} & \text{if n is odd} \\ \frac{1}{2}\left(t_{\frac{n}{2}} + t_{\frac{n}{2}+1}\right) & \text{if n is even.} \end{cases} \tag{17}$$

Nonetheless, when including censored observations we instead find an estimate for the median using the KM estimator. We find the value of M by taking the point in the estimator where the survival function is equal to $\frac{1}{2}$, i.e., $S(M) = 0.5$ [7]. this can be easily done using the KM curve and drawing a horizontal line from 0.5 on the y-axis to reach the curve.

Use of mean and median in distribution location. Finding the mean of observations can be beneficial in finding the location of our distribution when our data follows an approximately normal distribution [18]. As from the bell-shaped curve of a normal distribution, we know the mean lies in the centre. Despite this, the mean can be influenced by extreme values in a data set, as it takes account of every observation no matter how large or small and weights them evenly. We often see skewed distributions in survival analysis as we run into extreme values in our data. Therefore, when looking at survival data we may use the median to estimate our distribution's location, as the extreme values do not impact the value of the median.

5.4 Variance of the Nelson-Aalen Estimator

Similarly to Greenwood's formula, we can estimate the variance of the Nelson-Aalen estimator by the equation

$$Var(\hat{H}(t)) = \sum_{i:t_i \le t} \frac{(n_i - d_i)d_i}{(n_i - 1){n_i}^2}. \tag{18}$$

We find the estimator to be normally distributed. Thus, our $100(1-\alpha)\%$ confidence intervals are

$$\hat{H}(t)\ \pm\ z_{\frac{1-\alpha}{2}}\ \hat{\sigma}(t) \tag{19}$$

with $z_{\frac{1-\alpha}{2}}$ being the $\frac{1-\alpha}{2}$-fractile of the standard normal distribution. To improve this further we are able to apply a log transformation, this changes the confidence interval to $\hat{H}(t) \sim N(\hat{\mu}, \hat{\sigma}^2)$. This is useful when dealing with small sample sizes, as the original confidence intervals do not give a sufficiently good fit.

6 Parametric Survival Models

An alternative way of modelling data sets without using non-parametric methods is by applying a transformation. You may question the need for this when we have already outlined an effective way of modelling data, the answer is that one method often works better than the other depending on our data. For instance, we choose non-parametric models when we struggle to interpret results from a transformation, since we can only analyse and interpret the data referring to the transformed scale instead of the original data scale. But this is only the case for some data sets and the preference between the two will depend wholly on the data.

In this section we will introduce and discuss some important parametric survival distributions often seen in survival analysis. Examples of distributions that are commonly used include: the exponential, Weibull, Gompertz, Gamma, the log-normal and the log-logistic. In order to use parametric survival models, we must first parametrise the value of t. We start with the random variable, say W, with a standard distribution in $t \in (-\infty, \infty)$. From this we generate the distribution, which will be of the form

$$\log(T) = Y = \alpha + \sigma W \tag{20}$$

where T is our random variable of survival time, α our location parameter and σ the scale parameter.

6.1 Exponential Distribution

The probability density function for the exponential distribution of parameter λ is $f(t) = \lambda e^{-\lambda t}$ for $0 \leq t < \infty$ and $\lambda > 0$. Assuming that the hazard function is constant over time, i.e., h(t) = λ, for some $\lambda \in \mathbb{R}$ we obtain the survival function, $S(t) = e^{-\lambda t}$. We have the mean survival time as follows

$$\mu = E[T] = \int_0^\infty e^{-\lambda t} dt = \frac{1}{\lambda}. \tag{21}$$

Also, we have the variance, and at $S(t_{med}) = 0.5$, we have the median survival time,

$$E[T] = \frac{1}{\lambda^2} \quad \text{and} \quad t_{med} = \frac{\log(2)}{\lambda}. \tag{22}$$

Exponential case of the extreme value distribution. The exponential distribution is one of many forms of the extreme value distribution (EVD) that we will look at, specifically it is known as the standard extreme value distribution. The EVD is a limiting distribution which allows us to approximate the shape of a distribution of extreme values when considering large random samples. For our distribution, $T \sim Exp(\lambda)$, we parametrise T using the equation (26), where W denotes the random variable with the standard extreme value distribution. The probability density function of $Y = \log T$ is given by

$$exp(y - \alpha - e^{y-\alpha}) \tag{23}$$

for $-\infty < y < \infty$, where $\sigma = 1$ and $\alpha = -\log(\lambda)$. Thus, we have the distribution $Y = \alpha + W$ which has the density function $f(w) = e^{w-e^w}$ for $-\infty < w < \infty$.

Graphical assessment of exponential fit. We can check whether the exponential distribution is an appropriate model for a set of survival data by plotting the log of the survival function estimate against time t [14]. If we find this plot to be approximately a straight line, the exponential distribution is appropriate. We plot this by choosing the transformation $y = -\log(S(t))$ with $x = t$, and as $-\log(S(t)) = \lambda t$, we have $x = t$ and $y = \lambda t$, representing the linear graph with gradient λ and y-intercept at 0.

6.2 Weibull Distribution

The Weibull model involves the scale parameter, λ, and the shape parameter, γ, where both are non-negative. It is a generalisation of the exponential distribution meaning we can fit this distribution to a wider scope of data and we have the hazard function $h(t) = \lambda\gamma t^{\gamma-1}$. From this we see the hazard rate is constant at $\gamma = 1$ and is equivalent to the exponential distribution. Otherwise, we find the hazard function to be monotonically increasing for $\gamma > 1$ and monotonically decreasing for $\gamma < 1$. The probability density function $f(t)$ is given by $f(t) = \lambda\gamma(t)^{\gamma-1}e^{-\lambda t^\gamma}$ and the survival function is $S(t) = e^{-\lambda t^\gamma}$. We also have the mean

$$\int_0^\infty e^{-\lambda t^\gamma} dt \;=\; \frac{\Gamma(1 + \frac{1}{\gamma})}{\lambda^{\frac{1}{\gamma}}}. \tag{24}$$

This distribution is arguably the most widely used in survival research, due to its versatility in being able to change it's size and shape by altering the parameters. This allows us to model both the proportional hazards and accelerated failure time models using this distribution, meaning it can be either fully or semi-parametric, as we shall discuss further when fitting the Weibull PH and AFT models in chapters 8 and 9.

Weibull case of the extreme value distribution. Another version of the extreme value distribution is the Weibull model, which takes the log transformation $Y = \alpha + \sigma W$, where W has the EVD. In this case we have $\alpha = -\log(\lambda)$ and $\sigma = \frac{1}{\gamma}$, thus the probability density function of Y is given by

$$\sigma^{-1} exp\left(\frac{y-\mu}{\sigma} - e^{\frac{y-\mu}{\sigma}}\right) \tag{25}$$

where $-\infty < y < \infty$. We have a fixed shape of the density for this distribution Y, as the only two parameters are γ and λ, which impact the location and scale of the distribution only.

Graphical assessment of the Weibull distribution using R. A graphical check of the Weibull distribution for any set of data is given by a plot of $\log(-\log(\hat{S}(t))$ against time [21], where $\hat{S}$ is the Kaplan Meier estimate of the survival function based on given data. This transformation gives us the equation

$$\log(-\log(\hat{S}(t)) = \gamma(\log t + \log\lambda). \tag{26}$$

For a good fit, we expect to see an approximate straight line, where the intercept gives an estimate of $-\log\lambda$ and the intercept an estimate of γ.

Example. Recall the 'ovarian' data set. To find the exponential and Weibull fits to our KM curve we use the 'flexsurvreg' function and specify the distributions as exponential and Weibull. We plot the exponential and Weibull lines to fit the model (figure 12 in A.1), along with the confidence intervals, as shown below,

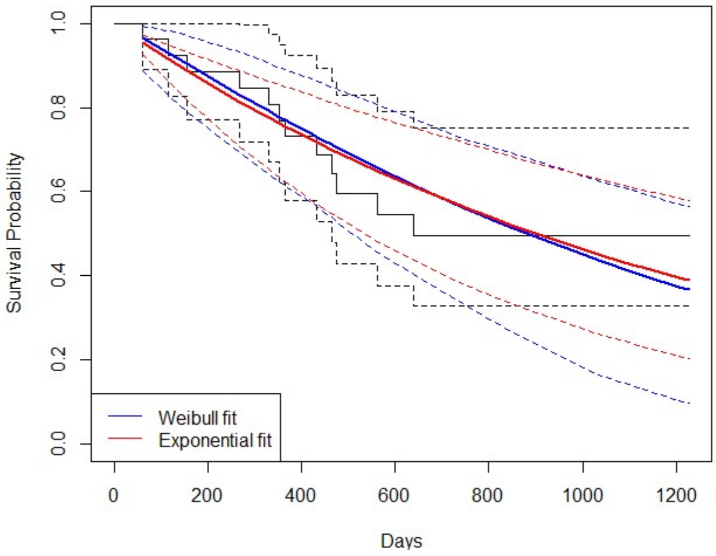


Figure 3: Exponential and Weibull fit on 'ovarian' data set

6.3 Gompertz Distribution

For certain parameters, the standard EVD has a very small value of $Pr(-\infty < x < 0)$. In this case we instead take the distribution for non-negative values of x only using the distribution introduced by Gompertz in 1825. We introduce the shape parameter, θ, which gives us the density and survival functions

$$f(t) = \lambda e^{\theta t} exp\left\{\frac{\lambda}{\theta}(1-e^{\theta t})\right\} \quad \text{and} \quad S(t) = exp\left\{\frac{\lambda}{\theta}(1-e^{\theta t})\right\} \tag{27}$$

respectively. From this we can derive the hazard function as $h(t) = \lambda e^{\theta t}$. Further, we can add a constant to the hazard rate, giving us the Gompertz-Makeham distribution with the hazard function $h(t) = \alpha + \lambda e^{\theta t}$.

6.4 Gamma Distribution

The density and survival functions for the gamma distribution, with parameters α and λ, are defined as

$$f(t) = \frac{\lambda^{\alpha} t^{\alpha-1} e^{-\lambda t}}{\Gamma(\alpha)} \quad \text{and} \quad S(t) = 1 - I_{\lambda t}(\alpha) \tag{28}$$

where $I_{\lambda t}(\alpha)$ denotes the incomplete gamma function, given by

$$I_{\lambda t}(\alpha) = \frac{1}{\Gamma(\alpha)} \int_0^{\lambda t} u^{\alpha-1} e^{-u} du. \tag{29}$$

The mean and variance are given as $E[T] = \frac{\alpha}{\lambda}$ and $Var(T) = \frac{\alpha}{\lambda^2}$. We find that the gamma function is related to the exponential distribution, which we see if we define $\alpha = 1$ in the density and survival functions.

Generalised gamma distribution function. We can generalise the gamma function above by introducing a scale parameter θ to the model, where $\theta > 0$. The new density function is given by

$$f(t) = \frac{\theta\lambda^{\alpha\theta}t^{\alpha\theta-1}exp(-\lambda t^{\alpha})}{\Gamma(\alpha)} \tag{30}$$

For $t \in [0, \infty)$. The new survival function is written as $S(t) = 1 - \Gamma_{(\lambda t)^{\theta}}(\alpha)$. Notice how this distribution is more versatile than the usual gamma distribution, as the we can define the new to variable model other distributions we have seen. For instance, taking $\alpha = 1$, the distribution becomes the Weibull distribution, and similarly for $\alpha \to \infty$, we have the log-normal distribution [7], as we shall see in the next section.

6.5 Log-normal Distribution

We apply the log transformation to our random variable T by defining $Y = \log(T)$, where Y is a normally distributed variable, $Y \sim N(\mu, \sigma^2)$. The probability density and survival functions of T are then given by

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma t}e^{-\frac{1}{2}(\frac{\log t - \mu}{\sigma})^2} \quad \text{and} \quad S(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right) \tag{31}$$

for non-negative t and $\sigma > 0$, and where $\Phi()$ denotes the standard normal distribution function, i.e.

$$\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} exp\left(\frac{-u^2}{2}\right) du. \tag{32}$$

We have the hazard function $h(t) = f(t)/S(t)$, and we find this to equal 0 when $t = 0$, after it increases to its maximum and then decreases, approaching zero as $t \to \infty$. This model is simple to apply when we have no censoring in our data, however as we often do we typically choose the log-logistic model instead.

6.6 Log-logistic Distribution

The log-logistic distribution is made up of two parameters, α and θ, where $\alpha \geq 0$. We often use this distribution over others due to the simplicity of the survival and hazard functions and along with the log-normal distribution, are the only two distributions included that allow for non-monotonic hazard function [19]. For a log-logistically distributed T, the probability density function is given by

$$f(t) = \frac{e^{\theta}\alpha t^{\alpha-1}}{(1 + e^{\theta}t^{\alpha})^2}. \tag{33}$$

Further, we have the survival function and hazard function as follows

$$S(t) = \frac{1}{1 + e^{\theta}t^{\alpha}} \quad \text{and} \quad h(t) = \frac{e^{\theta}\alpha t^{\alpha-1}}{1 + e^{\theta}t^{\alpha}}. \tag{34}$$

We find the hazard to be decreasing for $\alpha \leq 1$ and increasing for $\alpha > 1$, which resembles the lognormal hazard, as this is also found to increase from zero to a maximum, then decrease back to zero.

7 Maximum Likelihood Estimation in the Presence of Censoring

Survival data is a particularly unique branch of statistics, due to the inclusion of censoring and censored observations in our data. We accommodate for this by creating special models for the survival distribution that facilitate censoring in the data. In this chapter, we will look at another method of analysis that takes censoring into account, this being maximum likelihood estimation (MLE). To derive this we first need knowledge of how to construct the likelihood function without censoring, which is given in appendix B.2.

The maximum likelihood estimate of θ is the value at which the likelihood function, $L(\theta)$, attains a maximum. The derivation of the maximum likelihood estimate for uncensored data can be seen in appendix B.2.1 and derivation of it's asymptotic properties in B.2.2. This asymptotic property means we can construct confidence intervals for θ which allow us to conduct hypothesis tests. This will be investigated further in the section II where we carry out a likelihood ratio test. However, in this section we shall focus on finding a MLE in the case of right censored data and apply it to an example using the exponential distribution.

7.1 Maximum Likelihood Estimation with Right Censoring

The initial step to finding a maximum likelihood estimator for θ is finding the likelihood function for θ which we will firstly do. In the case of censoring, we shall assume that the event of censoring occurring for each participant is random and independent of one another, this will include the case of type 1 censoring which occurs when the censoring time of each subject is fixed in advance [14]. Let us assume that each participant will have a failure time T and a censoring time C, with T and C as independent continuous random variables and survival functions given by $S_{T_i}(t)$ and $S_{C_i}(t)$ respectively. We are assuming that given n observations $t_1, ..., t_n$, the censoring times C_i and failure times T_i are mutually independent. We find that this censoring time C_i is fixed for each individual, such that if $T_i \leq C_i$, censoring has not occurred and if $T_i > C_i$, censoring has occurred. Therefore, we have $\epsilon_i = min(T_i, C_i)$. We can denote whether the subject is censored using δ_i, where $\delta_i = 0$ if the i_{th} participant is censored, $T_i < C_i$, and $\delta_i = 1$ otherwise, $T_i \leq C_i$. Therefore, taking $f_{T_i}(t)$ and $f_{C_i}(t)$ to be the probability density functions of (t_i, δ_i) for T_i and C_i, we have

$$Pr(\epsilon_i = t, \delta_i = 0) = Pr(C_i = t, T_i > t) = f_{C_i}(t)S_{T_i}(t) \tag{35}$$

and

$$Pr(\epsilon_i = t, \delta_i = 1) = Pr(T_i = t, t \leq C_i) = f_{T_i}(t)S_{C_i}(t). \tag{36}$$

Putting these together in a single expression, we have

$$Pr(\epsilon_i = t, \delta_i) = \{f_{T_i}(t)S_{C_i}(t)\}^{\delta_i}\{f_{C_i}(t)S_{T_i}(t)\}^{1-\delta_i}. \tag{37}$$

Therefore to find the distribution, we simply take the product of the probability density functions for each subject $t_1, ..., t_n$. As we are assuming that censoring is non-informative, we can take out the parameters G(t) and g(t) as they do not involve any of the parameters in f(t), so have no influence on this parameter. Thus our likelihood function for right censored data in the case of independent random censoring is given by

$$L(\theta) = \prod_{i=1}^{n} f_{T_i}(t_i)^{\delta_i} S_{T_i}(t_i)^{1-\delta_i}. \tag{38}$$

Now we have obtained our likelihood function we apply the method described in B.2.2 to find the maximum likelihood estimator. This will be shown below, using the exponential distribution as an example.

Example. We shall derive the MLE in the case of right censored data with exponentially distributed survival. Suppose we have T_i independent failure times which follow an exponential distribution, we find the likelihood function in this case is given by

$$L(\lambda) = \prod_{i=1}^{n} (\lambda e^{-\lambda t_i})^{\delta_i} (e^{-\lambda t})^{1-\delta_i}. \tag{39}$$

Taking the log of L, we obtain the log-likelihood, $l(\lambda) = d\log(\lambda) - \lambda \sum_{i=1}^{n} t_i$ where $d = \sum_{i=1}^{n}(\delta_i)$, i.e. the total number of failures that have occurred. Further, taking the derivative we find the score vector is $U(\lambda) = \frac{d}{\lambda} - \sum_{i=1}^{n} t_i$. Thus, we find the maximum likelihood estimator to be the solution to $U(\lambda)$, given as

$$\hat{\lambda} = \frac{d}{\sum_{i=1}^{n} t_i} \tag{40}$$

i.e. the total number of failures divided by the total time we are at risk. We can find the asymptotic variance by taking the second derivative of the log likelihood with respect to λ. In conclusion, for large n we have

$$\hat{\lambda} \sim N\left(\lambda, \frac{d}{(\sum_{i=1}^{n} t_i)^2}\right) = N(\lambda, \frac{\hat{\lambda}^2}{d}) \tag{41}$$

which we can use to construct confidence intervals for λ, i.e. a $(1-\alpha)$ confidence interval for λ is given by

$$\hat{\lambda} \pm z_{\frac{\alpha}{2}} \frac{\hat{\lambda}}{\sqrt{d}}. \tag{42}$$

8 The Cox Proportional Hazards Model

Previously we have investigated differences between populations using the log-rank test and although this is beneficial in finding if there is a significant difference between data or not, we fail to quantify how important this difference is to us. This explains the popularity of the Cox (1972) proportional hazards model, as it allows us to quantify how meaningful the difference is between data. We most often consider one covariate in our data, i.e. whether the subject is in the control group or not, for instance in medical studies testing the efficacy of new drugs. The comparison of these can be expressed by the proportional hazards model, which we will demonstrate with an example in the appendix B.3, focusing more on the case of multiple covariates in this section.

Despite chapter 6 being focused on survival models, we did not include this model as this model is not fully parametric, but semi-parametric [14]. This is because despite our regression parameters, β, being known and modelled parametrically, the baseline hazard remains non-parametric. However, we can build a fully-parametric proportional hazards model by assuming we can parametrise the baseline hazard also depending on a given distribution, as we shall demonstrate below using the Gompertz and Weibull distributions.

Over the next few subsections, we will demonstrate how to use this model and will also fit the proportional hazards model with the assistance of survival distributions and likelihood theory that we have covered previously. But firstly, we must state some key assumptions we make on the survival and hazard rates.

Assumptions. There are a few crucial assumptions we require for this model to work, firstly we must assume that our survival times are continuously distributed [8]. More importantly, we require the effect of a risk factor between subjects to remain constant over time, i.e. the hazard for any individual, i, is a fixed proportion of the hazard for another individual, j, as suggested by the name of the model. Cox found that this proportionality assumption enable us to estimate the influence of our parameters on survival without any consideration of the hazard function.

8.1 Multiple Proportional Hazards Model

We can expand on this further by considering how we can compare survival between multiple groups rather than just two, which we do using multiple covariates in the generalised proportional hazards model. Note that in this instance the covariates denote the group the subjects belong to, but they can also signify any other factors in the data that we suspect may influence our outcome. Firstly, we have n individuals each with their own hazard function, $h_i(t)$, for $i = 1, \ldots, n$ and when dealing with multiple covariates, we shall define $\mathbf{x} = (x_1, \ldots, x_p)^T$ to be the 1xp vector of explanatory variables and we denote the hazard rate as $h(t; \mathbf{x})$. In addition, for multiple regression coefficients we introduce the vector of regression coefficients, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, which represent the relationship between the covariates and the time taken for the event to occur. The hazard function for the i_{th} individual is given by

$$h_i(t; \boldsymbol{x}) = \psi(\boldsymbol{x}_i; \boldsymbol{\beta}) h_0(t) \tag{43}$$

where $\mathbf{x}$ remains constant over time for each subject. There are three versions of the hazard ratio we can choose from - log-linear, linear and logistic. We most often select the log-linear model, in particular when examining data with an exponential, Gompertz or Weibull distribution. For the log-linear model, we have the hazard ratio ($\psi(\boldsymbol{x}; \boldsymbol{\beta}) = exp(\boldsymbol{\beta}^T \boldsymbol{x})$) and therefore the hazard function as follows:

$$h(t; \mathbf{x}) = h_0(t) exp(x_1\beta_1 + ... + x_p\beta_p). \tag{44}$$

We also have the survival function and density function as follows

$$S(t; \mathbf{x}) = exp\left[-e^{\mathbf{x}\boldsymbol{\beta}} \int_0^t h_0(u)du\right] \quad \text{and} \quad f(t; \mathbf{x}) = h_0(t)e^{\mathbf{x}\boldsymbol{\beta}} exp\left[-e^{\mathbf{x}\boldsymbol{\beta}} \int_0^t h_0(u)du\right]. \tag{45}$$

Alternatively, the linear and logistic models have the hazard ratios $\psi(\boldsymbol{x}; \boldsymbol{\beta}) = 1 + \boldsymbol{\beta}^T \boldsymbol{x}$ and $\psi(\boldsymbol{x}; \boldsymbol{\beta}) = \log(1 + exp(\boldsymbol{\beta}^T \boldsymbol{x}))$ respectively.

Example. When examining data with a Gompertz distribution and covariates $\boldsymbol{x}$, we know the hazard rate is given by $h(t) = \lambda e^(\theta t)$. As this is an example of the log-linear case, we find the hazard rate to be

$$h(t; \mathbf{x}) = \lambda e^{\theta t} exp\left\{\sum_{u=1}^{p} \beta_u x_u\right\} = e^{\theta t} exp\left\{\sum_{u=0}^{p} \beta_u x_u\right\} \tag{46}$$

with $\beta_0 = \log\lambda$ and $x_0 = 1$, which we write more succinctly as $\log h(t; \mathbf{x}) = \theta t + \sum_{u=0}^{p} \beta_u x_u$.

8.2 Fitting the Model

The main task of this model is finding appropriate estimates to the parameters $\boldsymbol{\beta}$, we do this using a partial maximum likelihood method [18]. In addition to this, we may also need an estimation of the baseline hazard function, which we construct using the estimates of $\boldsymbol{\beta}$ [7]. Using this we also find the standard errors, a statistical test with a specific p-value and our confidence interval. Therefore, using the partial likelihood to estimate $\boldsymbol{\beta}$ is particularly important, as it enables us to make inferences about the influence of our p explanatory variables $\boldsymbol{z}$ on the hazard ratio, without needing an estimate for the baseline hazard. By Cox's suggestion, we treat the new partial likelihood function as a regular likelihood function and find the maximum likelihood estimate for $\boldsymbol{\beta}$ using the same method that we used in chapter 7. This means we maximise the partial likelihood to give an estimate of $\boldsymbol{\beta}$ and find the fisher information matrix by taking the second derivative of the log partial likelihood with respect to $\boldsymbol{\beta}$.

Finding partial likelihood for $\boldsymbol{\beta}$. In the case of censored data, suppose we have n observed survival times, given by $t_1, \dots, t_n$ and the number of subjects at risk at time t_i will be given by $R(t_i)$. This is the number of individuals who are alive and uncensored just before time t_i. We shall reintroduce the indicator variable to determine whether the i_{th} survival time is censored or not. We then have the likelihood function for $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ as follows

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{exp(\boldsymbol{\beta}^T \mathbf{x}_i)}{\sum_{l \in R(t_i)} exp(\boldsymbol{\beta}^T \boldsymbol{x}_l)} \right\}^{\delta_i}. \tag{47}$$

The next step is to take the logs of this function to give the log partial likelihood function of $\boldsymbol{\beta}$ as

$$l(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left\{ \boldsymbol{\beta}^T \mathbf{x}_i - \log \sum_{l \in R(t_i)} exp(\boldsymbol{\beta}^T \boldsymbol{x}_l) \right\}. \tag{48}$$

Next, we shall define the mean of $\boldsymbol{x}$ to be

$$\bar{\mathbf{x}}(t, \boldsymbol{\beta}) = \frac{\sum_{l \in R(t)} \mathbf{x}_l exp(\boldsymbol{\beta}^T \mathbf{x}_l)}{\sum_{l \in R(t)} exp(\boldsymbol{\beta}^T \boldsymbol{x}_l)}. \tag{49}$$

Using this to write the score vector more succinctly, we have the score vector and fisher information matrix for $\boldsymbol{\beta}$ as follows

$$U(\boldsymbol{\beta}) = \left(\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_1}, ..., \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_p} \right)^T = \sum_{i=1}^{n} \delta_i \left[\mathbf{x}_i - \bar{\boldsymbol{x}}(t_i, \boldsymbol{\beta}) \right] \tag{50}$$

$$I(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\frac{\sum_{l \in R(t_i)} exp(\boldsymbol{\beta}^T \mathbf{x}_i) [\mathbf{x}_l - \bar{\mathbf{x}}(t_i, \boldsymbol{\beta})][\mathbf{x}_l - \bar{\mathbf{x}}(t_i, \boldsymbol{\beta})]^T}{\sum_{l \in R(t_i)} exp(\boldsymbol{\beta}^T \mathbf{x}_i)} \right]. \tag{51}$$

Fitting a Weibull model. As we know the exponential distribution has the proportional hazards property in the log-linear case, we find that the Weibull distribution also has this property as it is simply a generalisation of the exponential distribution. This means that all hazards given by $h_i(t)$ will be Weibull too. The 'partial' likelihood is found depending on the model we choose, in the instance of the Weibull model where the hazard rate is given by equation $h_0(t) = \lambda \gamma t^{\gamma - 1}$, we have the Cox model $h_i(t) = \lambda \gamma t^{\gamma-1} exp(\boldsymbol{\beta}^T \mathbf{x})$.

We can find the likelihood using the method for calculating maximum likelihood estimates in the presence of censoring, given in section 7.1. Initially, we define the survival and cumulative hazard functions in this context, which are given for the i_{th} individual as follows

$$H_i(t) = exp(\beta_i x_{i1} + ... + \beta x_{ip}) \lambda t^{\gamma} \quad \text{and} \quad S_i(t) = e^{-\lambda t^{\gamma}} exp\Big\{ -e^{(\beta_i x_{i1} + ... + \beta x_{ip})}. \Big\} \tag{52}$$

From equation (38), we have the likelihood function and log-likelihood for $\boldsymbol{\beta}$ with respect to γ and λ as follows,

$$L(\lambda, \gamma, \boldsymbol{\beta}) = \prod_{i=1}^{n} \{h_i(t_i)\}^{\delta_i} S_i(t_i), \tag{53}$$

$$\begin{aligned} l(\lambda, \gamma, \boldsymbol{\beta}) &= \sum_{i=1}^{n} \{ \delta_i \log h_i(t_i) + \log S_i(t_i) \} \\ &= \sum_{i=1}^{n} \left\{ \delta_i \left(\boldsymbol{\beta}^T \mathbf{x}_i + log(\lambda \gamma) + (\gamma - 1) \log(t_i) \right) - \lambda {t_i}^{\gamma} exp(\boldsymbol{\beta}^T \mathbf{x}_i) - \delta_i \log(t_i) \right\}. \end{aligned} \tag{54}$$

Therefore, to maximise $\boldsymbol{\beta}$ we must solve for each β_j, $j = (1, ..., p)$ the equation

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \delta_i x_{ij} - \lambda \sum_{i=1}^{n} \delta_i t_i^{\gamma} x_{ij} exp(\boldsymbol{\beta}^T \mathbf{x}_i) = 0. \tag{55}$$

We will most often do this using a statistical software, like R, as these equations may be very complicated or time consuming to solve.

9 The Accelerated Failure Time Model

The accelerated failure time model (AFT) is an alternative to the Cox proportional hazards model as it provides another way of explaining the effects of covariates in survival data. In Cox's model, the effect of covariates is multiplicative on the hazard scale, given by some constant, whereas the AFT model assumes the covariate has a multiplicative effect on the time scale, which accelerates or decelerates the survival time by some constant. We can illustrate this assumption using an example, take the survival times of a group of smokers and another for non-smokers, denoted $S_S(t)$ and $S_N(t)$ respectively. Assuming that the survival rates of these groups are proportional to one another, by the AFT model, for any time t we have

$$S_N(t) = S_S(\phi t) \tag{56}$$

where ψ denotes the acceleration factor [17], which is a positive constant. This signifies the stretching ($\phi > 1$) or shrinking ($\phi < 1$) of survival functions when comparing survival between two groups. In practice, this factor can provide us with valuable information about our data, take for instance in the example above. If our acceleration factor is $\phi = 2$, we have $S_N(t) = S_S(2t)$. Here we find the median survival time of the smokers to be half the median survival of the non-smokers, as this is the value of t at $S(t) = 0.5$. Further, as $\phi > 1$ in this case, we can say that the non-smokers are likely to live longer than the smokers, on average.

Another assumption of AFT models is that the ratio of survival times between groups remains constant for any value of $S(t)$, that is ϕ remains the same for all values of t. The value of ϕ we choose as our acceleration factor often depends on the explanatory variables, specific to the data we are analysing.

The parametric AFT model is beneficial as it provides us with a concise analysis of survival data involving censoring that is an alternative to Cox's model, which is particularly useful as we only have one initial distribution where the AFT model and Cox's proportional hazards model coincide [8]. This is the Weibull distribution and as the exponential distribution is a variation of this, we can apply this property to the exponential distribution also. Thus, we shall discuss the accelerated failure time form of these distributions in sections 9.2 and 9.3 and how they relate to Cox's model. Firstly, we shall consider the accelerated failure time model in the case of two generic groups, we shall find the associated hazard and density functions that relate these two groups in the next section.

9.1 Comparison of Two Groups

In general, when comparing any two groups with the accelerated failure time assumption, specifically groups 1 and 2 each with survival functions $S_1(t)$ and $S_2(t)$ respectively, we have

$$S_2(t) = S_1(\phi t) \tag{57}$$

which we can generalise to density and hazard functions using our definitions of these given in the introduction. We have the density and hazard functions for subjects in group 2, $f_2(t) = \phi f_1(\phi t)$ and $h_2(t) = \phi h_1(\phi t)$. In addition, the cumulative hazard function for group 2 against 1 is expressed as $H_2(t) = H_1(\phi t)$. As we require a non-negative acceleration factor, we often define it as $\phi = e^{\alpha}$, where α denotes the any real constant. In this case, we have the hazard function for an individual in group 2 as follows

$$h_2(t) = e^{\alpha} h_1(e^{\alpha} t). \tag{58}$$

Therefore, if we want the acceleration factor to be larger than 1, i.e., the survival time for subjects in group 2 being larger than survival time for subjects in group 1, we require α to be positive.

9.2 AFT Model Using the Exponential Distribution

As we know, the survival function for variable T with an exponential distribution is $S(t) = e^{-\lambda t}$. To ensure the survival function corresponds with the assumptions of the AFT model, we rescale our parameter using the acceleration factor as $\tilde{\lambda} = \lambda\psi$, which produces the new survival function, $\tilde{S}(t) = e^{-\lambda\phi t}$ [17].

9.3 AFT Model Using the Weibull Distribution

Furthermore, we apply a similar method as above to the Weibull distribution, with survival function $S(t) = e^{-\lambda t^{\gamma}}$, where we again rescale λ as $\hat{\lambda} = \lambda\psi^{\gamma}$ but keep γ the same. We have the survival function $\tilde{S}(t) = e^{-\lambda\phi^{\gamma}t^{\gamma}}$. As we have seen in the previous chapter, the Weibull distribution and therefore exponential distribution have the proportional hazards property and given we can find the hazard function for our rescaled survival function as $-\log(\tilde{S}(t)) = \lambda\phi^{\gamma}t^{\gamma}$. By the definition of cumulative hazard, the cumulative hazard function for a Weibull distribution is given by $H(t) = \lambda t^{\gamma}$. Using this and the proportional hazards property, we have that for any $\phi > 0$, our cumulative hazard can be written as follows

$$\psi H(t) = \psi(\lambda t^{\gamma}) = (\psi\lambda)t^{\gamma}. \tag{59}$$

Therefore, if we have two groups with a Weibull distribution satisfying the PH property, with cumulative hazard functions $H_1(t)$ and $H_2(t)$ respectively, we find that they also satisfy the AFT model assumption when we make the substitution of $\psi = \phi^{\gamma}$.

Example. We shall continue with the example as seen in 4.2, where we have separated the 'ovarian' patients into two groups based on their age. We can fit a Weibull and an exponential model to this variation of the data set using the code in figure 13 in A.1, and using 'summary' we find the acceleration factors of the two model variations. In figure 14, we find the Weibull model has an acceleration factor of $e^{1.502} = 4.49$, to three s.f.. Therefore, being in the 'old' group increases the hazard of death by a factor of 4.49. Alternatively when fitting an exponential model (figure 15 A.1) we find the acceleration factor to be larger, at $e^{1.746} = 5.73$ to 3 s.f., suggesting that being over 50 years increases a subjects hazard of death by a factor of 5.75.

II Model Building in Survival Analysis

Here we shall discuss the steps needed to find an appropriate survival model for a given set of data and will be applying this knowledge to build an adequate model for the 'ovarian' data set. Firstly, we must be certain of what we mean by an appropriate model, specifically in the context of survival analysis. We aim for our model to show an adequate representation of the patterns and trends we see in the data set, thus in our context we wish for the survival trends we see in the data to be reflected within the model which may vary depending on our covariates. Finding an appropriate model for survival data is extremely important, as the construction of an inappropriate model may lead to false conclusions being made. Therefore, taking necessary steps to build the best model are important and we shall cover many of these steps in our investigation to fit our model. Despite this, we will not be able to cover every possible step, due to the restriction of time and length of this report, nonetheless we shall present some of the main approaches we use to build an appropriate model and apply these to find a model for our data.

So far in this report, we have described statistical methods we use to analyse survival data, such as the KM and NA curves, as well as using statistical tests like the log-rank test, to decide whether explanatory variables have a significant impact on survival. Furthermore, we proposed the concept of proportional hazards and further the Cox model using this assumption, as well as introducing the accelerated failure time model. In this section, we will utilise this knowledge we have developed to build a Cox regression model for the 'ovarian' data set, again with the assistance of R. When building our Cox model, we must test whether this is appropriate by testing the main assumption of the model holds, that is, proportionality of risk factors. There are several approaches we can use, some being statistical tests and other being graphical assessments, for which we use our own judgement to assess the models adequacy.

Throughout this investigation, we shall develop further on this theory, as well as introducing new methods of analysis, which include comparing models using the likelihood ratio test, log-logistic plots of the model and generating confidence intervals for β based on the Cox model. Before this however, let us introduce the key steps we take in model building.

Preparing data for analysis in R We shall aim to develop a suitable model based on the 'ovarian' data set that we have used throughout the report so far. Before we were able to plot the figures seen so far in the project, we had to retrieve the data and prepare it for analysis in R, which was implemented using the code in figure 16 in appendix A.1. Firstly, we installed the relevant packages, which we then retrieved along with some pre-installed packages using the 'library' function. Within the 'survival' package is where we find the 'ovarian' data set, which we retrieved using the 'attach' command. Next, we defined each of the covariates 'rx', 'resid.ds' and 'ecog.ps' as factors with 2 levels and labelled these levels, using the 'as.factor' function [6].We have also split the numeric age category into two factors divided at the 50 years age boundary, this allows for comparison between the 'old' and 'young' patients of ovarian cancer. We shall continue with the use of the dichotomised version of the 'age' variable, given by 'age group', which allows us to use age as a predictive variable, so is better suited for model building.

10 Steps in Building a Cox Regression Model

We shall build a Cox regression model based on the 'ovarian' data set. Taking the hazard function for the multiple PH model and applying it here, we have the hazard for the i_{th} individual as follows,

$$\begin{aligned} h_i(t;\mathbf{x}) &= h_0(t)exp(\boldsymbol{\beta}^T\mathbf{x}_i) \\ &= h_0(t)exp(\beta_1 agegroup_i + \beta_2 treatment_i + \beta_3 residual_i + \beta_4 performance_i). \end{aligned} \tag{60}$$

The area we shall first focus on is the linear model of explanatory variables, given by

$$\boldsymbol{\beta}^T\mathbf{x}_i = \beta_1 agegroup_i + \beta_2 treatment_i + \beta_3 residual_i + \beta_4 performance_i. \tag{61}$$

Firstly, we shall decide which covariates are needed in our model, which will be determined using the log rank test. The next step we take is to decide which distribution will fit our data best, whether it be a Weibull distribution, exponential, log-logistic or lognormal, as decided by the comparison of the log-likelihood ratio test of our model fit to each of the four distributions. Further, we shall investigate the goodness of fit of our model, this will involve seeing whether a more complex model is needed to provide a better fit to the data, or whether this will simply over-complicate the model, as well as introducing residuals. Lastly, we shall discuss the appropriateness of the Cox model for our specified data set and test it's adequacy by plotting the logarithm of the cumulative hazard function and assessing its fit.

10.1 Selecting Significant Explanatory Variables

Often in real research settings, researchers are unaware of which covariates have a greater impact on survival than others. Therefore tests like the log-rank test can be carried our on data to investigate whether the impact of each covariate is statistically significant in altering survival times.

Log-rank test. Referring to the log-rank test defined in chapter 4, we can decide which covariates are of significance in our data. We have presented the results from the log rank test for all four of our covariates together using the 'ggforest' function, as shown in figure 17 A.1. From the output in figure 18, we find that the age group, treatment and residual disease covariates have a significant impact on survival, with p values of 0.047, 0.032 and 0.047 respectively. Therefore based on this test alone we would choose to include these three covariates only in our model, i.e. $\boldsymbol{\beta}^T\mathbf{x}_i = \beta_1 agegroup_i + \beta_2 treatment_i + \beta_3 residual_i$.

10.2 Assessing Goodness of Fit

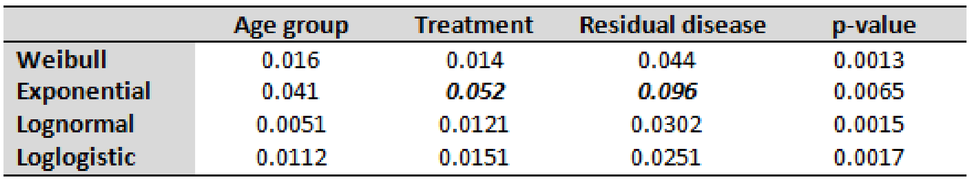
Another approach we take is using the likelihood ratio test to determine whether a simpler Cox regression model involving less covariates is sufficient, or whether a more complex model is necessary. We do this by calculating the deviance of the two models, given as minus twice the log of the likelihood ratio for each, fitted by the maximum likelihood. We test this difference using the chi-squared distribution with our specified degrees of freedom.

Likelihood ratio test. We shall use the likelihood ratio test to decide whether continue with the reduced model for 'ovarian', where we do not consider the influence of performance status on survival, in addition to the age group, treatment and residual disease covariates. We define the two models and fit them using the 'coxph' function and shall use ties handled by Breslow's approximation [5], as shown in the r code (figure 19, appendix A.1). The 'anova' function provides a summary of statistics based on the deviance of the two models. By the output (figure 20, appendix A.1), we see that the chi-squared test statistic has the value 0.8828. As the p-value is rather large at 0.3474, we still lack sufficient evidence to suggest a statistically significant difference between the simpler model and the advanced model at a 5% significance level. Therefore, we have further evidence to suggest the performance status covariate is not necessary in our model and we continue with the model defined from the log rank test above.

10.3 Choice of Distribution

As mentioned in the Weibull distribution section of chapter 6, we find the Weibull distribution to be a suitable fit when our hazard is monotonically increasing over time. As that is the case for our data set, this suggests that Weibull will be the most appropriate choice of distribution. Also in the case of the exponential distribution, the hazard is modelled as constant over time [3] which we would not expect from our data, meaning we would be less inclined to choose this distribution. Albeit, this is based on intuition only, to provide more rigid evidence for our claim, we can carry out the Log-likelihood ratio test on our so far decided model. We shall specify the distribution as Weibull, then as exponential, then log-normal and log-logistic and we shall compare the p-values from the chi-squared distribution.

Log-likelihood ratio tests. We formulate a table of the p-values by taking the results from the 'summary' function for each of the 'survreg' models with differing distributions. By the R code in figure 21 in A.1, we find the p-values for each of our chosen covariates and for the model overall in the table below,



	Age group	Treatment	Residual disease	p-value
Weibull	0.016	0.014	0.044	0.0013
Exponential	0.041	***0.052***	***0.096***	0.0065
Lognormal	0.0051	0.0121	0.0302	0.0015
Loglogistic	0.0112	0.0151	0.0251	0.0017

Figure 4: Table of p values for covariates and model with varying distributions

We see that the p-value remains below 0.05 for all four distributions, therefore suggesting all distributions show a sufficient fit for the model. Despite this, we see in bold and italic the p-values above 0.05 are present for the treatment and residual disease covariates in the exponentially distributed model. This suggests exponential would not be the most sufficient choice, as under this distribution the residual disease status and treatment one patient receives has no significant impact on their survival, when we have evidence to suggest otherwise. Therefore we choose to reject the notion that the exponential distribution would be the best fit for our model. Analysing the other three distributions, we find that the lowest p-value for the age group and treatment covariates is the log-normal distribution, thus suggesting that this may be the most appropriate distribution. However, as the overall p-value for the model is lowest when we take the default Weibull distribution, and as stated above that the increasing hazard can be well represented by a Weibull distribution, we shall choose this distribution overall.

10.4 Testing a Models Adequacy

We have chosen to use a Cox regression model on our data set and although this appears sufficient so far, we must answer the question of whether it would be better to use an AFT model. For instance, if the hazard for each covariate increases as time progresses, an AFT model would be better suited. We can test this statistically by introducing the predictor by time interaction effects to our model and testing whether these are statistically significant, known as a goodness-of-fit test. This method was originally proposed by Schoenfeld [20] where we test the correlation between the Schoenfeld residuals and survival time. However, we shall use a variation of this test proposed by Harrell [12], for which we utilise the 'cox.zph' function in R to test our covariates. If we find these values to be significant, there is evidence to suggest a time dependency and the proportionality function is violated. A further method of assessment is a graphical assessment of fit, that we take by plotting the $\log(-\log(\hat{S}(t))$ Kaplan Meier curves of the model against $\log(t)$ for each covariate separately. If we find that the lines on one plot are parallel to each other, we can say that these covariates are proportional to one another and therefore a Cox model is appropriate [21]. However if we find the covariates do not show this pattern an AFT model may be better suited.

Residuals test. We shall use the 'cox.zph' function on the regression model, including all the covariates we have in the data set to run the Schoenfeld residuals test. Further, we can represent the residuals graphically using the 'ggcoxzph' function in the 'survminer' package which for each covariate produces graphs of the scaled Schoenfeld residuals against time. This shall tell us whether a particular coefficient from a covariate is time-dependent. [19] We plot the graphs of the Schoenfeld residuals for each covariate against time, as shown by the code in figure 22. In each plot, the solid line represents a smoothing spline fit to the plot, with the confidence intervals given by the dashed lines. A horizontal line centered approximately around time zero indicates that the covariate meets the proportionality assumption. The output from our code is given in the figure on the next page.

Firstly for the statistical test, we see on the plot the p-value for each covariate according to the Schoenfeld residuals test, as well as the global p-value for the model at the top. We see the p-value for 'ecog.ps' to be rather low compared to the other p-values, which aligns with our findings so far as we lacked evidence of this covariate being significant and here we have evidence to suggest a non-proportional hazard. Despite this, the other p-values are shown to be relatively high, meaning there is sufficient evidence to suggest proportionality for the other covariates. Further we see all four graphs show somewhat of a horizontal line in the case of the residuals, age group and sex covariates, further supporting this claim. Alternately the performance status covariate shows a slightly increasing slope, representing a trend of the effect of performance status increasing over time. This means that there is a tendency for the residuals to increase with time.

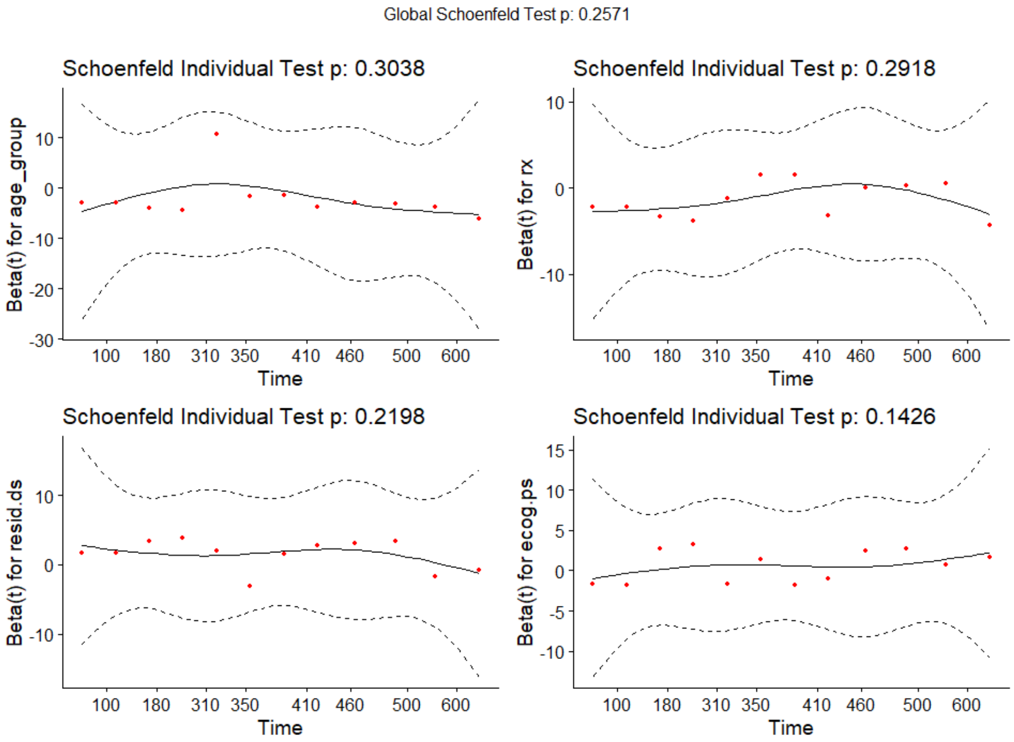


Figure 5: Plots of Schoenfeld residuals for each covariate

Graphical assessment of fit. Again, we shall utilise the 'survfit' function but instead change our plot to be a log-logistic plot of the survival function, along with adjusting the time scale to be log(t) by defining the 'logt' variable and replacing 'futime' with this. After applying this code to the Cox model of each covariate singularly as shown in figure 23 in A.1, we have the four plots as follows,

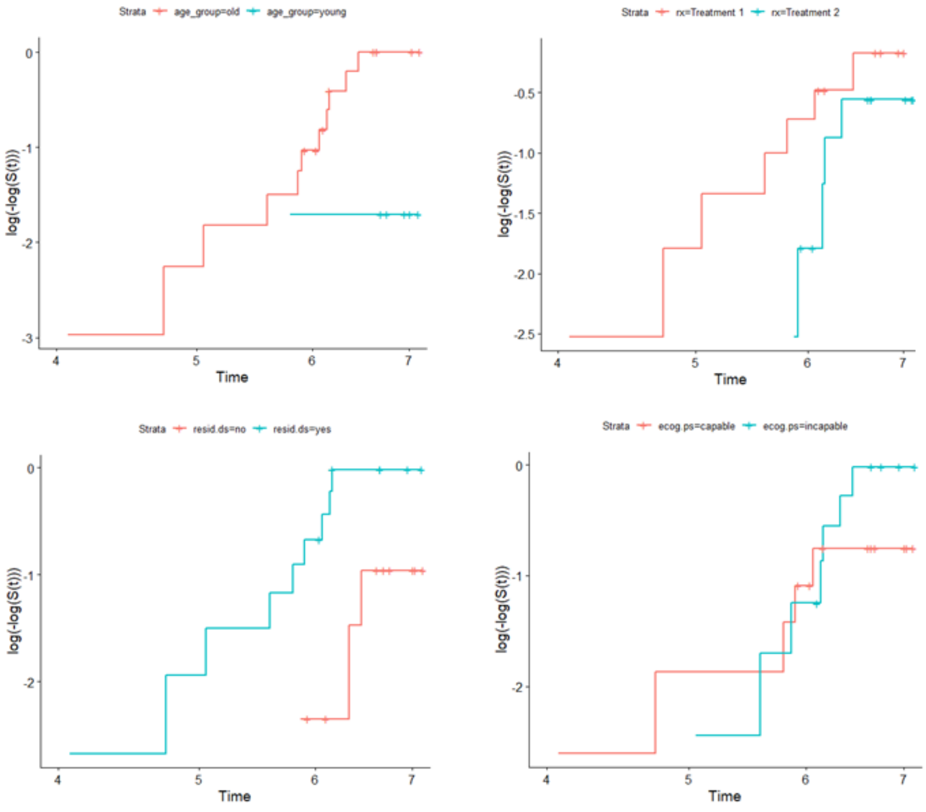


Figure 6: Plots of Cox model for each covariate (top left - age, top right - treatment, bottom left - residual disease and bottom right - performance status).

It is clear that the first plot fails to show two parallel lines when comparing the two age groups, suggesting that this covariate is not entirely proportional. On the other hand, one may argue that we lack data on the 'young' age group, as we can see by the graph there is only one non-censored patient that falls into this category. Therefore, we struggle to see proportionality here due to the small size of the 'young' data set. Additionally, the performance status covariate does not show a clear set of parallel lines, however they are shown to be rather similar and therefore somewhat parallel. This may be due to the similarity between the two groups, which we would expect as so far we have not found this covariate to be significant in altering survival rates. Despite these, on the treatment and residual disease plots we can clearly see the two curves to be closely parallel. This infers that these two variables are indeed proportional to one another, and as these variables are two out of three of the significant covariates in our model, we have evidence to suggest that the Cox model is a suitable fit for our data over the AFT model. Therefore we shall continue with this model and next we shall investigate the values given to our regression parameters, $\beta_1, ..., \beta_3$, and how we can use the proportional hazards model to obtain confidence intervals for these.

11 Estimating Regression Coefficients

To fit a proportional hazards model, we require a statistical distribution as a base and as each of these come equipped with their own variances, we can find the standard error of each unknown parameter β in each case. From this we are able to construct confidence intervals for β and therefore carry out hypothesis tests, where we investigate whether different groups have the same survival distribution. To formulate the confidence intervals for the hazard ratio ψ, we simply take the exponential of the confidence intervals for β [6], which depend on the survival data we are given.

As we have so far decided our Cox regression model has a Weibull distribution and is made up of the three parameter '*age_group*', '*rx*' and '*resid.ds*', our last step is to find values for the regression parameters which will give us our final model. In addition to this we will also construct the confidence intervals of these and therefore determine the multiplicative effect each of the three covariates has on our hazard.

Finding confidence intervals for $\beta_0, ..., \beta_3$. Using R we can estimate the value of β_1, β_2 and β_3 for 'ovarian' which gives us an estimate for the hazard ratios for each. In addition, when assuming asymptotic normality we can construct confidence intervals for $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$ at a 5% significance level. In this case, we have the hazard function for the i_{th} individual as follows,

$$h_i(t) = h_0(t)exp(\beta_1 agegroup_i + \beta_2 treatment_i + \beta_3 residual_i). \tag{62}$$

We fit our desired model with the 'coxph' function, specifying the relation between survival and all three covariates by '$s \sim agegroup + rx + resid.ds$' (figure 24 in A.1). We find that the p-value for all three tests shown at the bottom of figure 25 (likelihood, Wald, and score test) are significant, thus providing evidence that our chosen covariates are indeed significant and reason to suggest our $\boldsymbol{\beta}$ values are non-zero. We find the coefficient estimate for β_1 is -2.11, with a standard deviation of 1.09, while for β_2 we have an estimate of -1.28 with standard deviation 0.623. Lastly, our estimate for β_3 is 1.25 with standard deviation 0.692. Therefore the hazard ratios are estimated by $exp(\beta)$ as 0.121, 0.278 and 3.50 respectively. We find being 'young' reduces the hazard by a factor of 0.12, or equivalently 88%, and receiving treatment 1 also reduces the hazard by approximately 72%. Despite this, the hazard for those with residual disease present is much higher than those without, at almost a 350% increase in risk. Assuming asymptotic normality, we can use the standard deviations to construct our confidence intervals for $\hat{\beta}$ and $\hat{\psi}$, which are given as follows

$$\beta_1 \in [-4.25, 0.024] \quad \text{and} \quad \psi_1 \in [0.014, 1.02], \tag{63}$$

$$\beta_2 \in [-2.50, -0.062] \quad \text{and} \quad \psi_2 \in [0.082, 0.940], \tag{64}$$

$$\beta_3 \in [-0.105, 2.608] \quad \text{and} \quad \psi_3 \in [0.900, 13.6]. \tag{65}$$

We find in this test the covariates for age and residual disease fail to be significant, with p values of 0.0526 and 0.0705 respectively. Despite this, as the confidence intervals for ψ_1 and ψ_3 include 1, which suggests age and residual disease have a smaller impact on hazard rates in comparison to the treatment group covariate. Despite this, we shall still include these covariates in our model, as they have shown to be significant in previous methods, and their p-values remain low here, both being significant at the 10% significance level.

Final model. In summary, we shall choose the Cox regression model for 'ovarian' to be as follows,

$$h_i(t) = h_0(t)exp(-2.11 agegroup_i + -1.28 treatment_i + 1.25 residual_i). \tag{66}$$

However we often come across data in survival analysis where the Cox model is not appropriate. This often being due to a violation of the proportional hazard assumptions, which may be solved by using an AFT model instead when we have a nonlinear effect between covariates. Alternately, often in research we have the case where hazards are still proportional, but are not independent of each other, which we usually assume to be the case. A model which still incorporates the proportionality of the Cox model, but also involves the correlation of hazard between covariates is called a frailty model, which we shall discuss further in the final section III below.

III Frailty Models

12 What are Frailty Models?

Previously we have mostly assumed the survival times of individuals are independent of each other, and while this assumption has worked well so far enabling us to make concise models from our data, we may want to consider how we can increase the validity of our research, by taking the approach that survival times of individuals may depend on one another. Survival research has been focusing more on this question in recent years, which has given rise to the concept of frailty models. In particular with the first notable contribution by Aalen [1], where he provided motivation for the use of frailty models, explaining the importance of the impact of heterogeneity on analysis.

Correlational data occurs when we see individuals belonging to the same subgroup, which we call a 'cluster', all having a similar survival rate. Examples of clusters include members of the same family or perhaps individuals staying at the same hospital. Another example is measuring the time taken for a range of diseases to occur in one individual, as the onset of one disease may increase the likelihood of another disease occurring. We introduce these random variables and incorporate these into our survival models by the means of a frailty model. We use the word 'frailty' as in this model we assume each individual has their own disposition to failure and more 'frail' individuals have a higher mortality rate. In other words, we have unobserved heterogeneity in the hazard due to the unique risk factors specific to each individuals.

Frailty models are beneficial in survival research as we no longer need to assume a homogeneous population, meaning individuals in the study all have the same risk of the event occurring. Instead we can take a mixture of different hazard rates for the individuals, which are altered by the influence of unmeasured covariates. This is particularly useful when building survival models, as we can incorporate the heterogeneity into our model. Whereas if we were to ignore unobserved heterogeneity like the usual survival models do, we find the magnitude of regression coefficients is underestimated, which lessens the accuracy of our model.

We will most often see the frailty model being used as an extension of the Cox Proportional Hazards model, with the addition of a single random intercept in the frailty model that takes account for random effects in the univariate case. The multivariate frailty model is a further extension of this, which introduces a vector to account for multiple random effects. There are fives classes of frailty models we shall investigate, listed as univariate, multivariate (shared), nested, joint, and additive frailty.

12.1 Univariate Frailty Models

Vaupel et al. [21] first introduced the notion of frailty and applied it to survival data. In this case, we quantify ones frailty by the unobservable, age-dependent random variable Z, acting multiplicatively on the baseline hazard, i.e.

$$h(t; Z) = Zh_0(t). \tag{67}$$

We absorb a universal scale factor into the baseline hazard function, so that the random variable Z has a mean of one. This variable is assumed to be drawn independently from each individual, varying across the population with a variance parameter σ^2. We find that when σ^2 is small our values of Z are close to 1, whereas for large variance, the Z values are more greatly dispersed around 1. This differs from Cox model as we no longer have a proportional hazard between subjects, instead we find the hazards to be converging to one another when the distribution of the frailty function Z has a finite variance. This is similar to the behaviour we see in the accelerated failure time model, as this has a multiplicative effect on hazards. Therefore, we find that individuals with a higher frailty value are at greater risk of failure than those with lower frailty, so we expect the individuals with a greater frailty to experience failure the earliest [22]. We introduce known covariates to the model to give the univariate frailty model as follows

$$h(t; Z, \mathbf{x}) = Zh_0(t)exp(\boldsymbol{\beta}^T\mathbf{x}) \tag{68}$$

where $h_0(t)$ again denotes the baseline hazard, $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_p)$ is the vector of covariates and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ the vector of regression coefficients. Therefore, by equation (6) we find the survival function of an individual subject to frailty is given as follows

$$S(t; Z) = exp\left(-\int_0^t h(s, Z)ds\right) = exp\left(-Z\int_0^t h(s)ds\right) = e^{-ZH_0(t)} \tag{69}$$

where $H_0(t)$ denotes the baseline cumulative hazard function. Next we shall derive the survival and density functions for Z using the Laplace transform.

Laplace transform. As our Z value is unobservable, we require alternate ways of presenting the survival [3], density functions for t, as well as the mean and variance equations for Z. The Laplace transform provides a solution for this, we have that the distribution of our random variable Z can be uniquely specified the Laplace transform below,

$$\mathcal{L}(c) = E[exp(-cZ)]. \tag{70}$$

We derive the survival function for the frailty distribution by taking the expectation of the conditional survival function given in equation (69). We see that the expectation of this is equal to the output of the Laplace transform, with $c = H_0(t)$. Therefore, we have the survival and by $f(x) = h(x)S(x)$, we also have the density functions as follows,

$$S(t) = E[S(t; Z)] = \mathcal{L}(H_0(t)) \quad \text{and} \quad f(t) = -h_0(t)\mathcal{L}'(H_0(t)), \tag{71}$$

as $\mathcal{L}(H_0(t)) = -Z\mathcal{L}(H_0(t))$. In addition, from (70) we can clearly see that $\mathcal{L}(0) = 1$. We can obtain the expectation of Z by taking the negative of the derivative of $\mathcal{L}(0)$, i.e. $E[Z] = -\mathcal{L}'(0)$. Further, we find the second derivative to be $\mathcal{L}''(0) = E[Z^2]$ and following this pattern we find the k_{th} derivative of the Laplace transform to be $\mathcal{L}^k(c) = (-1)^k E[Z^k]$. Using this we derive the expectation and variance of the frailty distribution as follows,

$$E[Z] = -\mathcal{L}'(0) \quad \text{and} \quad Var(Z) = -\mathcal{L}''(0) - (\mathcal{L}'(0))^2. \tag{72}$$

Gamma version. We may choose the Z variables to follow a gamma distribution, with parameters α and λ as defined in chapter 6. We find in this case the Laplace transform is given by

$$\mathcal{L}(c) = \left(\frac{\lambda}{\lambda + c}\right)^{\alpha} \tag{73}$$

and as the expectation of frailty must be 1, so we apply a restriction of $\gamma = \lambda$ so that Z follows a $gamma(\lambda, \lambda)$ distribution. Therefore $E[Z] = 1$ and $Var(Z) = \sigma^2 = \lambda^{-1}$. The hazard and survival functions are as follows,

$$h(t) = \frac{h_0(t)}{1 + \frac{(H_0(t)}{\lambda}} \quad \text{and} \quad S(t) = \mathcal{L}(H_0(t)) = \left(1 + \frac{H_0(t)}{\lambda}\right)^{-\lambda}. \tag{74}$$

Log-normal version. Alternately, the log-normal distribution can also be used and in this case we apply the transformation $W = e^Z$, where W is a normally distributed variable. In particular, W is assumed to be an independent sample from a distribution with a mean of zero and an unknown variance. Despite this, we find the Laplace transform and survival and density functions to be difficult to obtain in closed form. [3]

12.2 Multivariate Frailty Models

We require a multivariate (shared) frailty model when we have subjects that are clustered into groups or recurrent events times are clustered for each individual. [19]

Shared clustering in groups. In extension from the 'clustering' defined in the introduction, we find the sub-grouping of data in studies according to particular characteristics is often the result of natural causes, such as grouping according to family, or region of residence. In this case we find survival of individuals belonging to the same subgroup to be dependent on one another. This is the most common case of shared frailty, however this is also needed in the case of recurrent events.

Recurrent events. We experience recurrent events when we observe the same type of event in one subject multiple times over the duration of the study. When examining these events, correlation may be caused by heterogeneity across subjects, as in the case of recurrent events each observation for a subject is taken from the same person. Meaning the usual covariates that would lead to unobserved heterogeneity are not present, making a correlation between the occurrence and timing of events in one individual apparent. Alternatively, correlation could also be caused by event dependence, which is when the occurrence of one event makes the occurrence of the next event more or less likely. Such as in the case of cancer, after experiencing one tumour, the occurrence of others often becomes more likely, making a correlation within the subject whether this be correlation is negative or positive, as demonstrated in this example.

Suppose our data is grouped into i clusters, and we have m_i individuals in the i_{th} cluster. We adapt the proportional hazards model to accommodate for this clustering and as first proposed by Vaupel et al. [22], we have the frailty model as follows

$$h_{ij}(t; \mathbf{x}_{ij}, Z_i) = h_0(t)exp(\boldsymbol{\beta}\mathbf{x}_{ij} + \sigma Z_i) \tag{75}$$

where $1 \leq j \leq m_i$ and $1 \leq i \leq n$, where $\boldsymbol{\beta}$ again denotes the px1 vector of regression coefficients and $\mathbf{x}_{ij}$ the covariate vector for the j_{th} individual in the i_{th} cluster. We have $Z_1, ..., Z_n$ as the frailties for each cluster, along with their standard error σ. Here we assume that the Z_i's are independent samples following some parametric distribution, with mean 0 and variance 1 [16]. For easier interpretation we write the model in the form

$$h_{ij}(t; \mathbf{x}_{ij}, u_i) = h_0(t) u_i exp(\boldsymbol{\beta}^T \mathbf{x}_{ij}) \tag{76}$$

where the variables u_i are assumed to be an independent sample from a distribution with a mean of zero and an unknown variance. Therefore, when we have $u_i > 1$, subjects in this cluster are at a greater risk of failure than they would be when using a typical Cox model. Conversely, for $u_i < 1$, the risk of failure decreases for those in the i_{th} cluster.

The Laplace transform. As seen in the univariate case, we can write the survival function in terms of the Laplace transform, to account for the unobservable Z_i's. We find the joint distribution of the survival times of individuals within a group is given by

$$\begin{aligned} S(\mathbf{y}_{i1}, ..., \mathbf{y}_{in_i}) &= P(\mathbf{Y}_{i1} > \mathbf{y}_{i1}, ..., \mathbf{Y}_{in_i} > \mathbf{y}_{in_i}) \\ &= \mathcal{L}\left(\sum_{j=1}^{n_i} H_0(\mathbf{y}_{ij}) exp(\boldsymbol{\beta}^T \mathbf{x}_{ij}) \right) \end{aligned} \tag{77}$$

where $\mathcal{L}(c) = E[exp(-cU)]$ for frailty U.

Gamma version. Again, we assume the u_i variables follow a gamma distribution, but in this case we take an expectation of 1 and a variance of $\theta = \frac{1}{\lambda}$. Continuing with $\alpha = \lambda$, we have the density function as

$$f(u) = \frac{u^{(\frac{1}{\theta}-1)} exp(-\frac{u}{\boldsymbol{\theta}})}{\Gamma(\frac{1}{\boldsymbol{\theta}}) \boldsymbol{\theta}^{\frac{1}{\theta}}}. \tag{78}$$

We find that large θ values reflect a greater frailty between groups in comparison to smaller values, as well as a stronger correlation within clusters [16]. We have the joint survival function for all m_i individuals in the i_{th} cluster is given by

$$\begin{aligned} S(\mathbf{y}_{i1}, ..., \mathbf{y}_{in_i}) &= P(\mathbf{Y}_{i1} > \mathbf{y}_{i1}, ..., \mathbf{Y}_{in_i} > \mathbf{y}_{in_i}) \\ &= \left(1 + \boldsymbol{\theta} \sum_{j=1}^{n_i} H_0(\mathbf{y}_{ij}) exp(\boldsymbol{\beta}^T \mathbf{x}_{ij}) \right)^{-\frac{1}{\theta}}. \end{aligned} \tag{79}$$

We can develop the frailty model further to accommodate for time dependent covariates by replacing Z_i with $Z_i(t)$ in (equation above), but that goes beyond the scope of this report. In addition, there may be more than one type of frailty to accommodate for, in this case we can extend Z_i to $Z_i + Y_i$ when we have two types of frailty in each cluster.

12.3 Nested Frailty Model

This model is used when we have a hierarchical clustering of data, for instance we may have regions in the UK where patients reside and within this we may have the hospital they attended for a given treatment. These present as two random effects, but the influence of these is not completely separate from one another. To formulate this in the case of two nested random effects, let us consider the Z_i independent clusters, each with n_i sub-clusters within them Therefore we have the nested frailty model as follows

$$h_{ijk}(t; \mathbf{x}_{ijk}, u_i, v_{ij}) = u_i v_{ij} h_0(y) exp(\boldsymbol{\beta}^T \mathbf{x}_{ijk}). \tag{80}$$

We now have both the cluster random effect u_i and subcluster random effect $v_i j$, which both are independently and identically gamma-distributed.

12.4 Other Frailty Models

We find there are some limitations of the shared frailty model [22]. In particular, that the model makes the unobserved factors the exact same within clusters, which may not be representative of the group. Also the models only produce a positive correlation of survival within clusters, which may not always be the case. For instance, clustering subjects by their hospital may not mean they are positively correlated, as more time and attention being put on one patient may make the performance of another patient with less care worse in comparison. Therefore, we shall introduce two frailty models that provide solutions to these issues.

Joint frailty model. In an attempt to overcome these hurdles a joint frailty model was proposed, for which the frailty effect between each pair of subjects is specified by two associated random variables, these are also known as correlated failure models. In this setting, each person would be assigned a different random variable, which means there is no longer any shared frailty. These two random variables are then associated by the means of a joint distribution.

Additive frailty model. This model provides another solution to the issues faced in the shared frailty model. This model is not based on the proportional hazard assumptions, instead the frailty acts additively on the baseline hazard function and this type of model is useful when we repeated observations, for instance if we have subjects with the same condition receiving different treatments. This model allows us to examine heterogeneity between trials as well as examining the effect of different treatments on survival.

13 Frailty Models in R

We shall demonstrate frailty models in R by the means of the 'kidney' data set, as shown in figure 26 in A.2, which is an example of recurrent events. In this study kidney patients had a catheter inserted and the time taken for the catheter to be removed, either due to infection or censoring was recorded. We have exactly two observations for each subject, with the time taken until infection or censoring recorded twice as two separate readings under the same id. Each observation has its corresponding 'id' number according to the subject, we also have a record of each subjects age, sex and disease type, given by 'disease'. We have four types of disease, denoted as GN, AN, PKD and other. Here we would expect to find a correlation between the time taken for catheter removal in the two observations of the same individual. However, we cannot base this off intuition only, therefore we will be using R to evaluate how well this data is suited to a frailty model and adjusting this model to accommodate for frailty.

Looking for evidence of frailty. We build the gamma frailty model based on the Cox model, where we construct the Cox model like usual but add in the 'frailty(id)' covariate. This means we take into account the dependence between observations linked to the same 'id', as shown in the 'kidney.f' model, coded in figure 27 in A.2. From the summary of this model, we can assess whether there is sufficient evidence for a frailty effect, from the 'variance of random effect' output. From the output in figure 28, we see the value is very small at $7.6e^{-}5$, suggesting that there is not sufficient evidence that 'id' has a statistically significant effect on event time, meaning the frailty model is insufficient in this case.

Improving the gamma frailty model. Despite this result, we can alter the model to see if there is evidence of frailty when taking less covariates into account. To decide which covariates to remove, we focus on the summary 'kidney.f' and see the estimate of each parameter remains the same in the two models, as well as the standard error. However we shall focus on the added standard error estimates, shown in this model by 'robust se', which are computed using the Lin-Wei method. We find these estimates to not be a great deal larger than the standard error value for the 'sex' and 'age' coefficients, thus indicating that the frailty effect within 'id' clusters is small for these coefficients. However, for the 'disease' coefficient we find the standard error estimate be at least 0.1 greater than the original error in all three cases, suggesting the effect of clustering within 'id' is relatively large, which makes this model inadequate as a frailty model. Therefore, we shall improve our model by removing the 'disease' coefficient, giving us the 'kidney.f2' model (figure 29 in A.2), where the variance for random effects is now given as 0.412 by figure 30. Meaning that in 'kidney.f2' there is indeed a statistically significant correlation for observations in the same 'id' group. We find the estimation of each parameter β is given also by this code, meaning we can construct the frailty model for 'kidney' using the model 'kidney.f2' as follows

$$h_i(t) = Zh_0(t)exp(0.005age_i - 1.587sex_i) \tag{81}$$

where frailty Z corresponds to the participants id and is unobservable.

14 Conclusion

The intention of this report has been both to provide an understanding of survival analysis, as well as discussing how we build a model within survival analysis and finding the most adequate model in the case of the Cox proportional hazards model. We have also discussed frailty models and their relation to survival analysis, as well as furthering our model building analysis by formulating a frailty model for a given set of data in R. Although we have covered a great breadth of information, there are still some areas that could have been explored further but were not due to the restrictions of time and length of the report. Therefore, further investigations I would like to explore include applying the model building guide to larger data sets than 'ovarian', as the lack of uncensored patients here meant we lacked a lot of thorough evidence that is required to build an adequate model. In further study I would like to take a greater look at residuals, as there are different types that I did not cover in this report, such as Cox-Snell and Martingale residuals. Lastly, I would explore fitting other types of models using R, whether this be parametric regression models, where I can use my knowledge of AFT to fit models for data which violate the proportional hazards assumption. Or whether this be fitting other types of frailty models, such as nested, joint and additive frailty models.

15 References

1. O.O. Aalen. *Heterogeneity in Survival Analysis.* Section of Medical Statistics, University of Oslo, Norway, 1988.

2. B. Altshuler, *Theory for the measurement of competing risks in animal experiments.* Mathematical Biosciences 6, 1970.

3. T.A. Balan, H. Putter. *A tutorial on frailty models.* Statistical Methods in Medical Research, 29(11), pp. 3424–3454. 2020.

4. M. Bradburn, T. Clark, S. Love et al. *Survival Analysis Part III: Multivariate data analysis – choosing a model and assessing its adequacy and fit.* Br J Cancer 89, 605–611 (2003).

5. N. Breslow. *Covariance analysis of censored survival data.* Biometrics 30, 89-99. 1974.

6. G. Brostrom, *Event History Analysis with R.* Chapman Hall/CRC The R Series, CRC Press, Boca Raton, FL, 2012.

7. D. Collett, *Modelling Survival Data in Medical Research, Texts in Statistical Science.* Chapman Hall / CRC, Boca Raton – London, 2015 (3rded.).

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

9. R. C. Elandt-Johnson and N. L. Johnson, *Survival Models and Data Analysis.* John Wiley Sons, New York, 1980.

10. I. Etikan, K. Bukirova, M. Yuvali. *Choosing statistical tests for survival analysis.* Biom Biostat Int J. 2018;7(5):477-481.

11. M.K. Goel, P. Khanna, J. Kishore. *Understanding survival analysis: Kaplan-Meier estimate.* Int J Ayurveda Resm, 2010.

12. F. Harrell, *The PHGLM Procedure.* In SAS Supplemental Library User's Guide, Version 5. Cary, NC: SAS Institute Inc. 1986.

13. P. Hougaard, *Fundamentals of survival data.* Biometrics 55, 1999.

14. J. D. Kalbfleisch and R. L. Prentice, *The Statistical Analysis of Failure Time Data.* Wiley Series in Probability and Mathematical Statistics, John Wiley Sons, New York, 2002 (2nd ed.).

15. E. L. Kaplan and P. Meier, *Nonparametric estimation from incomplete observations.* Journal of the American Statistical Association 53, 1958.

16. J.P. Klein, M.L. Moeschberger. *Survival analysis: techniques for censored and truncated data.* Vol. 2. New York: Springer, 2003.

17. D.G. Kleinbaum, M. Klein. *Survival Analysis: A Self-Learning text* Springer, New York, NY, 2012.

18. P. Laake, M. Fagerland. (2015). *Statistical Inference. Research in Medical and Biological Sciences: From Planning and Preparation to Grant Application and Publication.* 379-430. Oslo, Norway, 2015.

19. M. Mills, *Introducing Survival and Event History Analysis.* SAGE Publ., London, 2011.

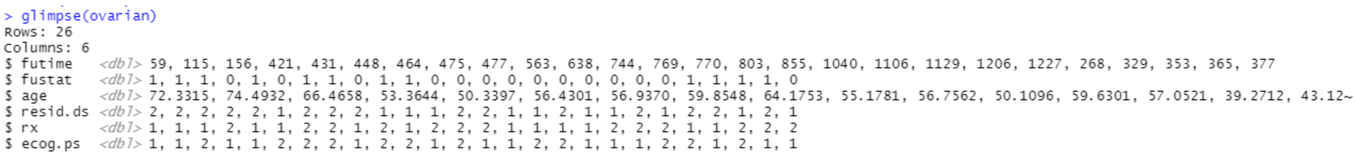
20. D. Schoenfeld, *Partial residuals for the proportional hazards regression model.* Biometrika, Volume 69, Issue 1, Pages 239–241. Massachusetts, USA, 1982.

21. J.W. Vaupel, K.G. Manton, E. Stallard. *The impact of heterogeneity in individual frailty on the dynamics of mortality.* Demography 16, 439–454. 1979.

22. A. Wienke. *Frailty Models in Survival Analysis.* Chapman and Hall, London, 2010.

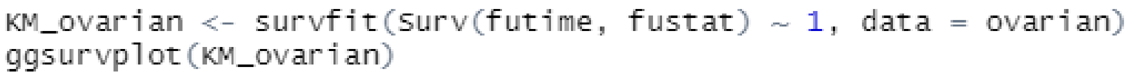
A R Code and Output

A.1 'ovarian' Data Set



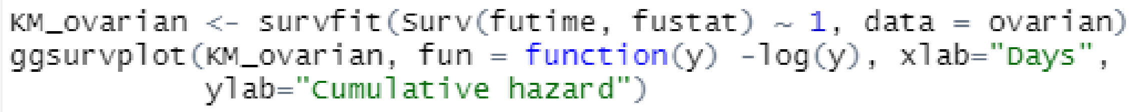
```
> glimpse(ovarian)
Rows: 26
Columns: 6
$ futime   <dbl> 59, 115, 156, 421, 431, 448, 464, 475, 477, 563, 638, 744, 769, 770, 803, 855, 1040, 1106, 1129, 1206, 1227, 268, 329, 353, 365, 377
$ fustat   <dbl> 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0
$ age      <dbl> 72.3315, 74.4932, 66.4658, 53.3644, 50.3397, 56.4301, 56.9370, 59.8548, 64.1753, 55.1781, 56.7562, 50.1096, 59.6301, 57.0521, 39.2712, 43.12~
$ resid.ds <dbl> 2, 2, 2, 2, 2, 1, 2, 2, 2, 1, 1, 1, 2, 2, 1, 1, 2, 1, 1, 2, 1, 2, 2, 1, 2, 1
$ rx       <dbl> 1, 1, 1, 2, 1, 1, 2, 2, 1, 2, 1, 2, 2, 2, 1, 1, 1, 1, 2, 2, 2, 1, 1, 2, 2, 2
$ ecog.ps  <dbl> 1, 1, 2, 1, 1, 2, 2, 2, 1, 2, 2, 1, 2, 1, 1, 2, 2, 1, 1, 1, 2, 2, 1, 2, 1, 1
```

Figure 7: Glimpse of the 'ovarian' data set



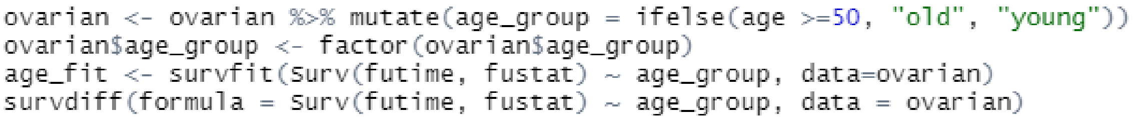
```
KM_ovarian <- survfit(Surv(futime, fustat) ~ 1, data = ovarian)
ggsurvplot(KM_ovarian)
```

Figure 8: Code for Kaplan Meier plot



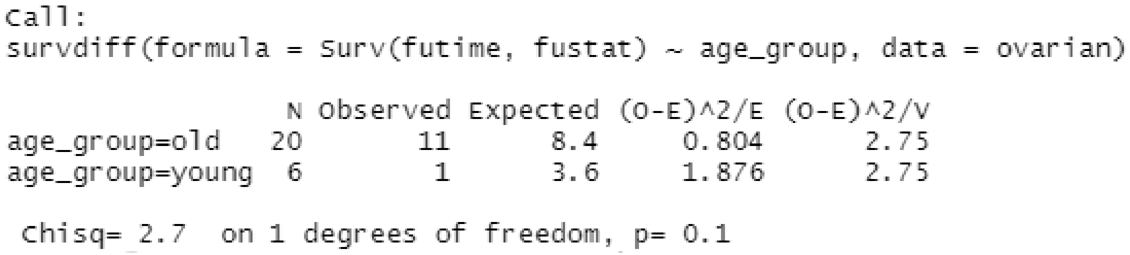
```
KM_ovarian <- survfit(Surv(futime, fustat) ~ 1, data = ovarian)
ggsurvplot(KM_ovarian, fun = function(y) -log(y), xlab="Days",
           ylab="Cumulative hazard")
```

Figure 9: Code for Nelson Aalen plot



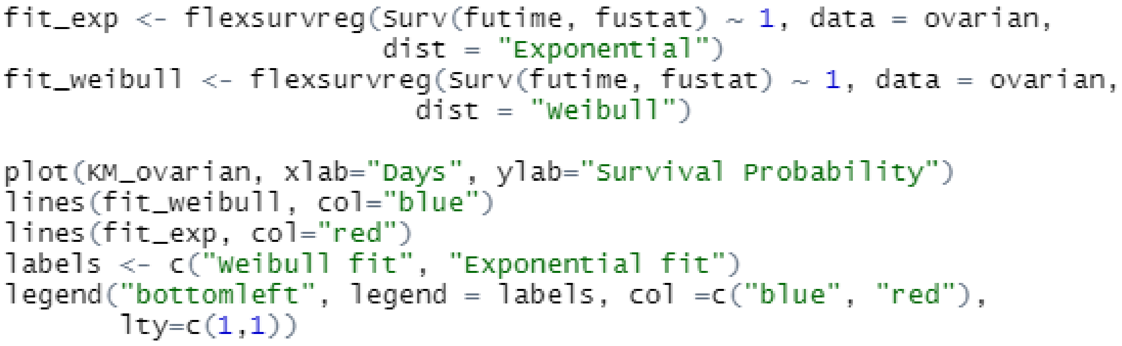
```
ovarian <- ovarian %>% mutate(age_group = ifelse(age >=50, "old", "young"))
ovarian$age_group <- factor(ovarian$age_group)
age_fit <- survfit(Surv(futime, fustat) ~ age_group, data=ovarian)
survdiff(formula = Surv(futime, fustat) ~ age_group, data = ovarian)
```

Figure 10: Chi squared version of the log-rank test code



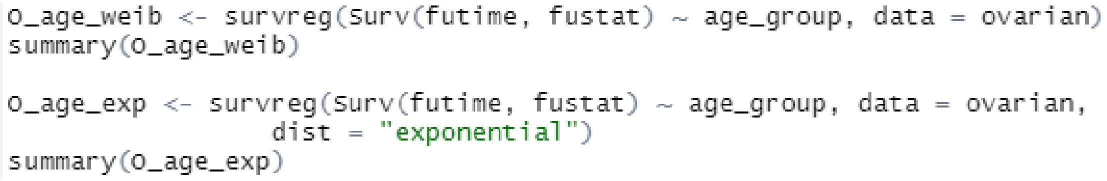
```
Call:
survdiff(formula = Surv(futime, fustat) ~ age_group, data = ovarian)

                 N Observed Expected (O-E)^2/E (O-E)^2/V
age_group=old   20       11      8.4     0.804      2.75
age_group=young  6        1      3.6     1.876      2.75

 Chisq= 2.7  on 1 degrees of freedom, p= 0.1
```

Figure 11: Chi squared version of the log-rank test output



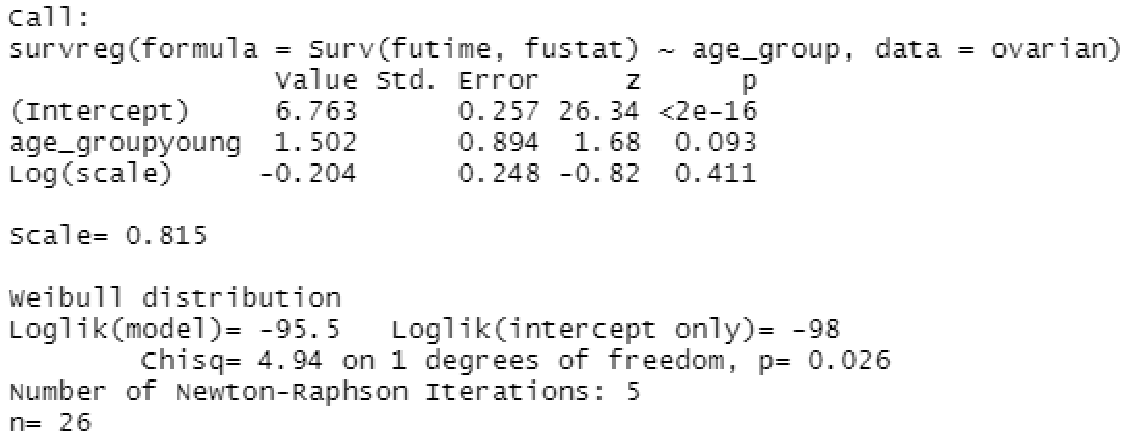
```
fit_exp <- flexsurvreg(Surv(futime, fustat) ~ 1, data = ovarian,
                       dist = "Exponential")
fit_weibull <- flexsurvreg(Surv(futime, fustat) ~ 1, data = ovarian,
                           dist = "Weibull")

plot(KM_ovarian, xlab="Days", ylab="Survival Probability")
lines(fit_weibull, col="blue")
lines(fit_exp, col="red")
labels <- c("Weibull fit", "Exponential fit")
legend("bottomleft", legend = labels, col =c("blue", "red"),
       lty=c(1,1))
```

Figure 12: Code for fitting of exponential and Weibull models



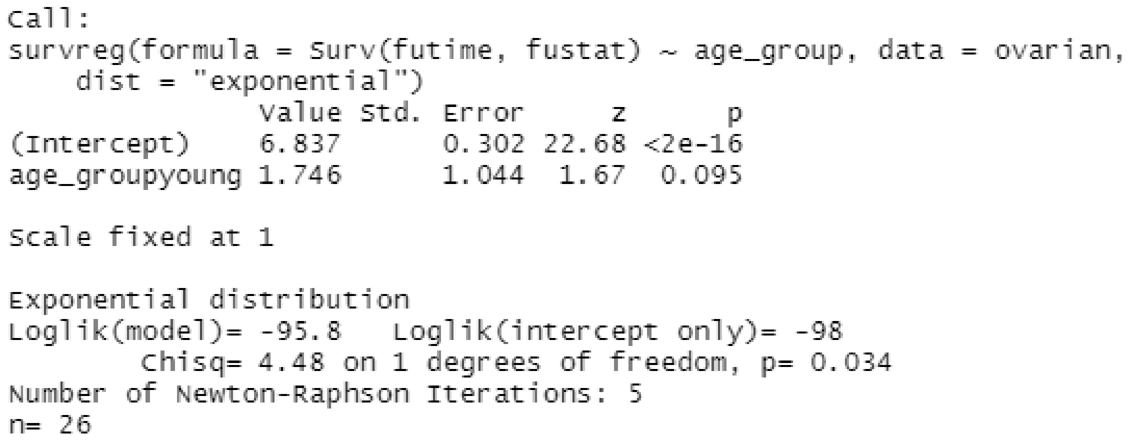
```
O_age_weib <- survreg(Surv(futime, fustat) ~ age_group, data = ovarian)
summary(O_age_weib)

O_age_exp <- survreg(Surv(futime, fustat) ~ age_group, data = ovarian,
                     dist = "exponential")
summary(O_age_exp)
```

Figure 13: Code for fitting exponential and Weibull models, split into 'old' and 'young' individiuals



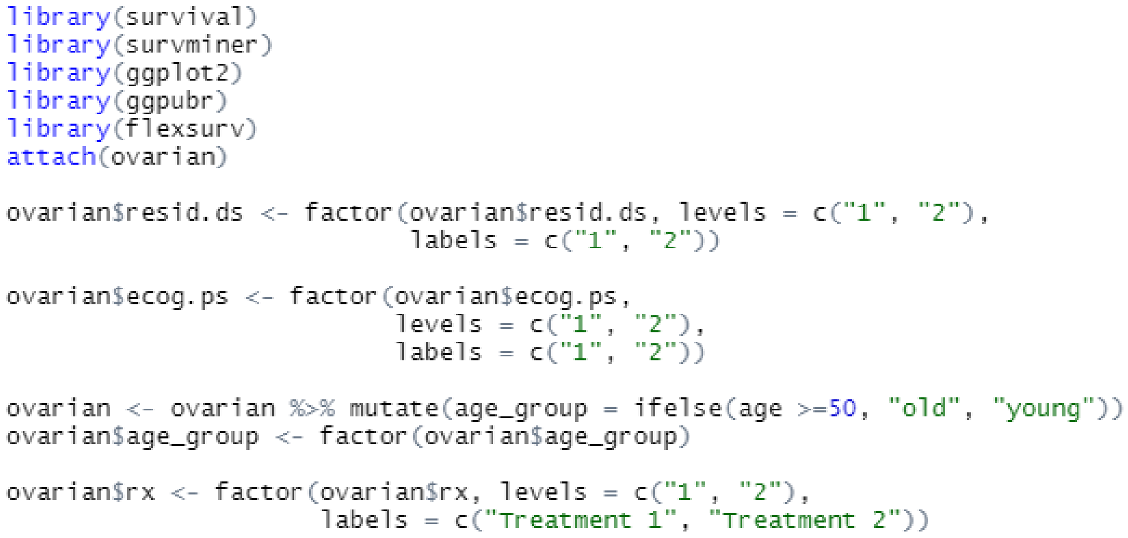
```
Call:
survreg(formula = Surv(futime, fustat) ~ age_group, data = ovarian)
                Value Std. Error     z      p
(Intercept)     6.763      0.257 26.34 <2e-16
age_groupyoung  1.502      0.894  1.68  0.093
Log(scale)     -0.204      0.248 -0.82  0.411

Scale= 0.815

Weibull distribution
Loglik(model)= -95.5   Loglik(intercept only)= -98
        Chisq= 4.94 on 1 degrees of freedom, p= 0.026
Number of Newton-Raphson Iterations: 5
n= 26
```

Figure 14: Summary of Weibull model for 'old' and 'young' individuals



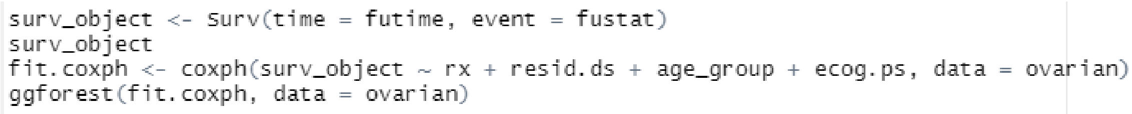
```
Call:
survreg(formula = Surv(futime, fustat) ~ age_group, data = ovarian,
    dist = "exponential")
               Value Std. Error     z      p
(Intercept)    6.837      0.302 22.68 <2e-16
age_groupyoung 1.746      1.044  1.67  0.095

Scale fixed at 1

Exponential distribution
Loglik(model)= -95.8   Loglik(intercept only)= -98
        Chisq= 4.48 on 1 degrees of freedom, p= 0.034
Number of Newton-Raphson Iterations: 5
n= 26
```

Figure 15: Summary of exponential model for 'old' and 'young' individuals



```
library(survival)
library(survminer)
library(ggplot2)
library(ggpubr)
library(flexsurv)
attach(ovarian)

ovarian$resid.ds <- factor(ovarian$resid.ds, levels = c("1", "2"),
                           labels = c("1", "2"))

ovarian$ecog.ps <- factor(ovarian$ecog.ps,
                          levels = c("1", "2"),
                          labels = c("1", "2"))

ovarian <- ovarian %>% mutate(age_group = ifelse(age >=50, "old", "young"))
ovarian$age_group <- factor(ovarian$age_group)

ovarian$rx <- factor(ovarian$rx, levels = c("1", "2"),
                     labels = c("Treatment 1", "Treatment 2"))
```

Figure 16: Set up code for 'ovarian'



```
surv_object <- Surv(time = futime, event = fustat)
surv_object
fit.coxph <- coxph(surv_object ~ rx + resid.ds + age_group + ecog.ps, data = ovarian)
ggforest(fit.coxph, data = ovarian)
```

Figure 17: Log-rank 'forest' code

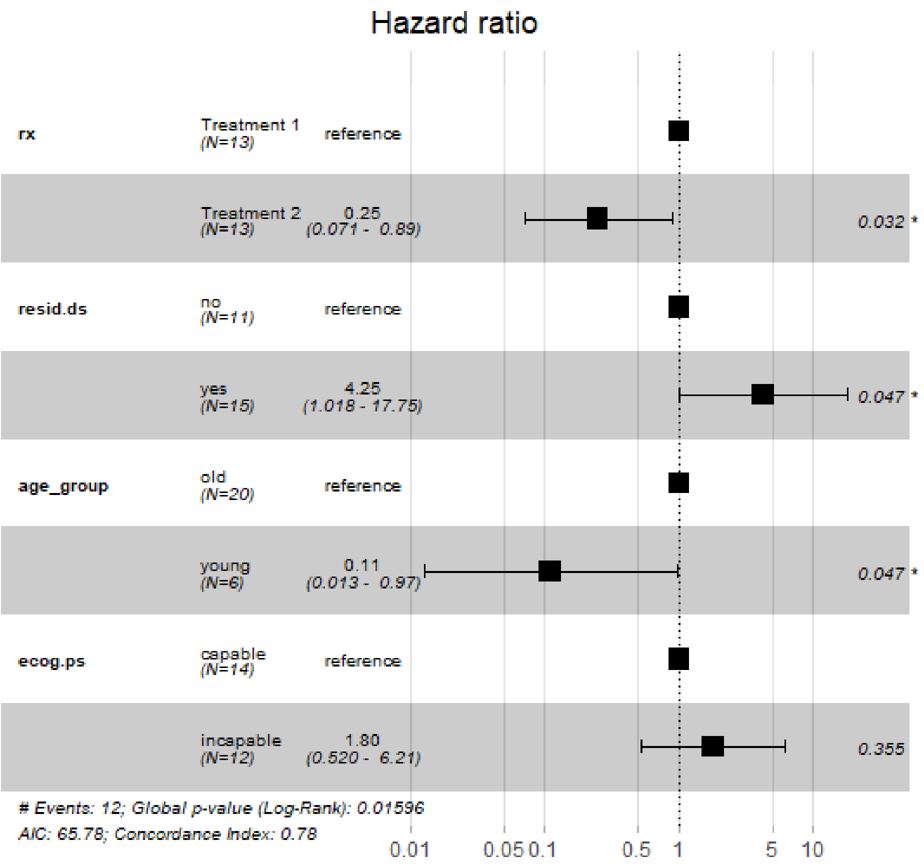
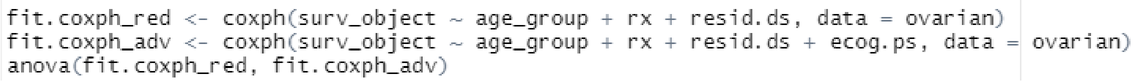
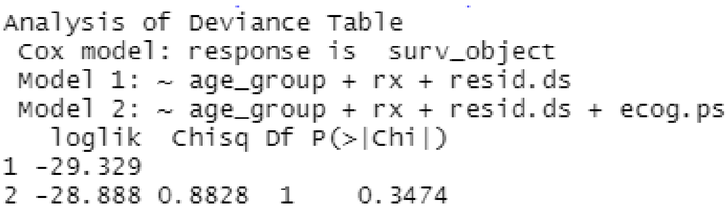


Figure 18: Log-rank 'forest'



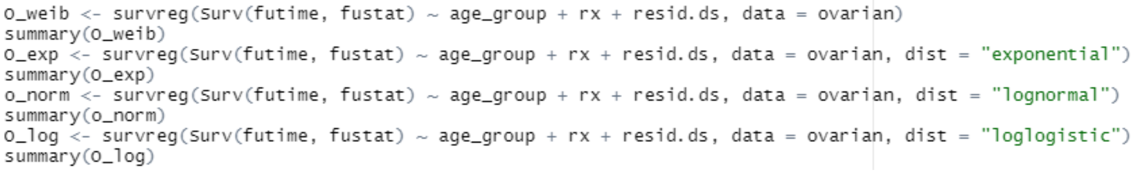
```
fit.coxph_red <- coxph(surv_object ~ age_group + rx + resid.ds, data = ovarian)
fit.coxph_adv <- coxph(surv_object ~ age_group + rx + resid.ds + ecog.ps, data = ovarian)
anova(fit.coxph_red, fit.coxph_adv)
```

Figure 19: Likelihood Ratio test code



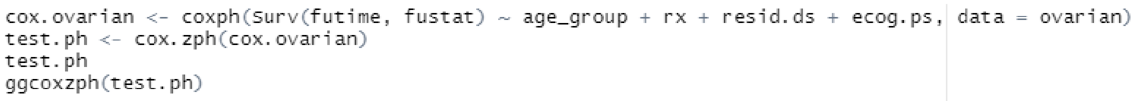
```
Analysis of Deviance Table
 Cox model: response is  surv_object
 Model 1: ~ age_group + rx + resid.ds
 Model 2: ~ age_group + rx + resid.ds + ecog.ps
   loglik  Chisq Df P(>|Chi|)
1 -29.329
2 -28.888 0.8828  1    0.3474
```

Figure 20: Likelihood Ratio test output



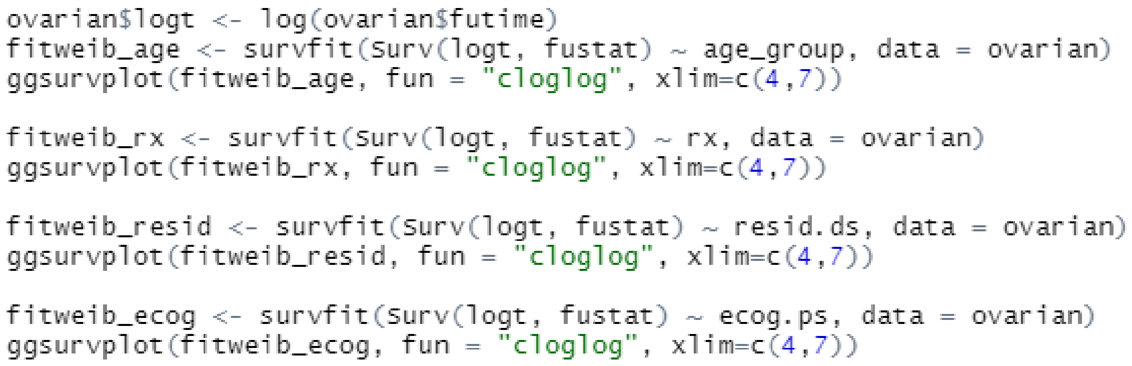
```
o_weib <- survreg(Surv(futime, fustat) ~ age_group + rx + resid.ds, data = ovarian)
summary(o_weib)
o_exp <- survreg(Surv(futime, fustat) ~ age_group + rx + resid.ds, data = ovarian, dist = "exponential")
summary(o_exp)
o_norm <- survreg(Surv(futime, fustat) ~ age_group + rx + resid.ds, data = ovarian, dist = "lognormal")
summary(o_norm)
o_log <- survreg(Surv(futime, fustat) ~ age_group + rx + resid.ds, data = ovarian, dist = "loglogistic")
summary(o_log)
```

Figure 21: Code for p-values of each covariate and model for each distribution



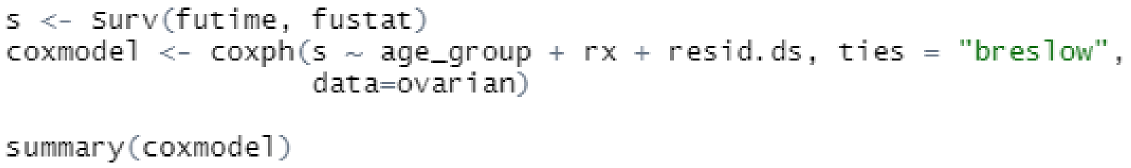
```
cox.ovarian <- coxph(Surv(futime, fustat) ~ age_group + rx + resid.ds + ecog.ps, data = ovarian)
test.ph <- cox.zph(cox.ovarian)
test.ph
ggcoxzph(test.ph)
```

Figure 22: Code for Schoenfeld residuals plot



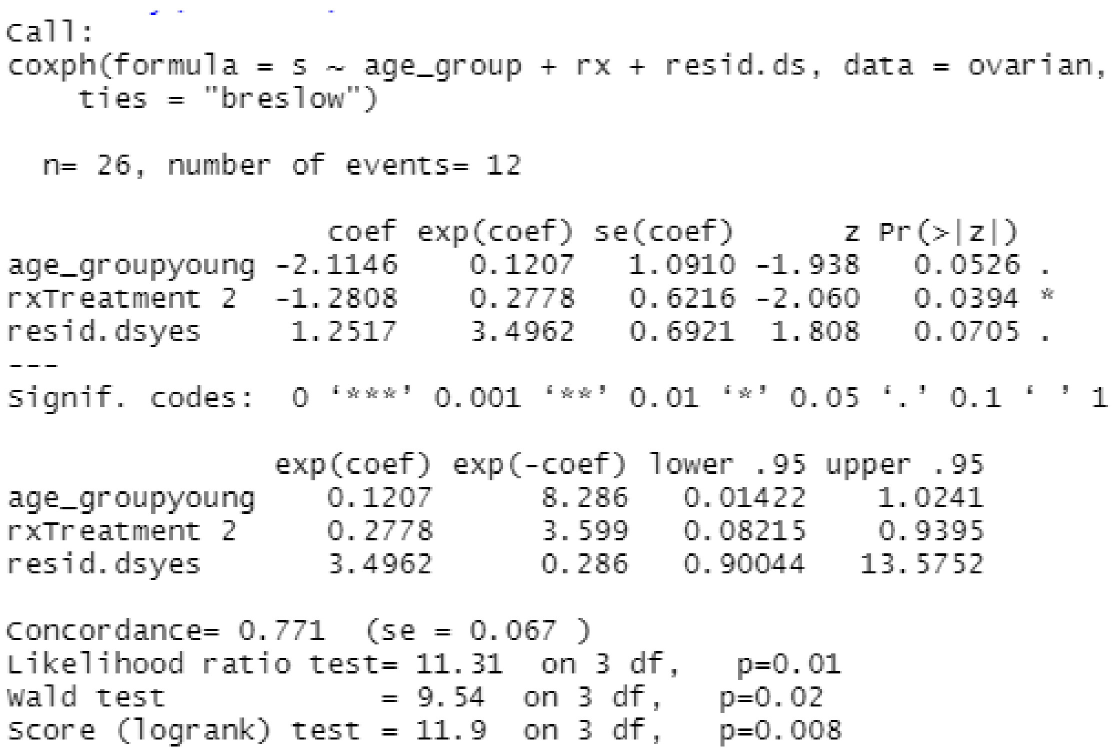
```
ovarian$logt <- log(ovarian$futime)
fitweib_age <- survfit(Surv(logt, fustat) ~ age_group, data = ovarian)
ggsurvplot(fitweib_age, fun = "cloglog", xlim=c(4,7))

fitweib_rx <- survfit(Surv(logt, fustat) ~ rx, data = ovarian)
ggsurvplot(fitweib_rx, fun = "cloglog", xlim=c(4,7))

fitweib_resid <- survfit(Surv(logt, fustat) ~ resid.ds, data = ovarian)
ggsurvplot(fitweib_resid, fun = "cloglog", xlim=c(4,7))

fitweib_ecog <- survfit(Surv(logt, fustat) ~ ecog.ps, data = ovarian)
ggsurvplot(fitweib_ecog, fun = "cloglog", xlim=c(4,7))
```

Figure 23: Code for log-logistic survival curve for each covariate



```
s <- Surv(futime, fustat)
coxmodel <- coxph(s ~ age_group + rx + resid.ds, ties = "breslow",
                  data=ovarian)

summary(coxmodel)
```

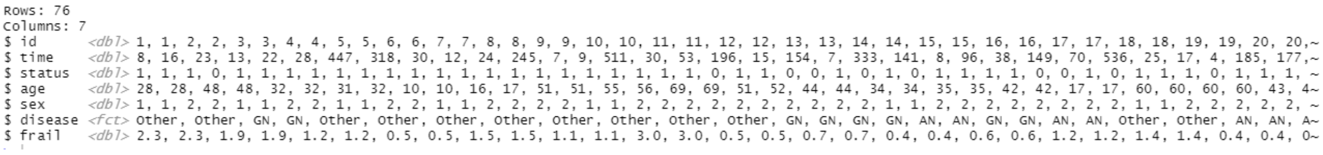
Figure 24: Code for generating confidence intervals for $\boldsymbol{\beta}$



```
Call:
coxph(formula = s ~ age_group + rx + resid.ds, data = ovarian,
    ties = "breslow")

  n= 26, number of events= 12

                   coef exp(coef) se(coef)      z Pr(>|z|)
age_groupyoung -2.1146    0.1207   1.0910 -1.938   0.0526 .
rxTreatment 2  -1.2808    0.2778   0.6216 -2.060   0.0394 *
resid.dsyes     1.2517    3.4962   0.6921  1.808   0.0705 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

               exp(coef) exp(-coef) lower .95 upper .95
age_groupyoung    0.1207      8.286   0.01422    1.0241
rxTreatment 2     0.2778      3.599   0.08215    0.9395
resid.dsyes       3.4962      0.286   0.90044   13.5752

Concordance= 0.771  (se = 0.067 )
Likelihood ratio test= 11.31  on 3 df,   p=0.01
Wald test            = 9.54  on 3 df,   p=0.02
Score (logrank) test = 11.9  on 3 df,   p=0.008
```

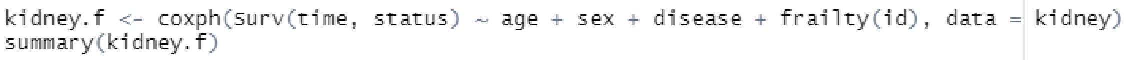
Figure 25: Output for generating confidence intervals of $\boldsymbol{\beta}$

A.2 'kidney' Data Set



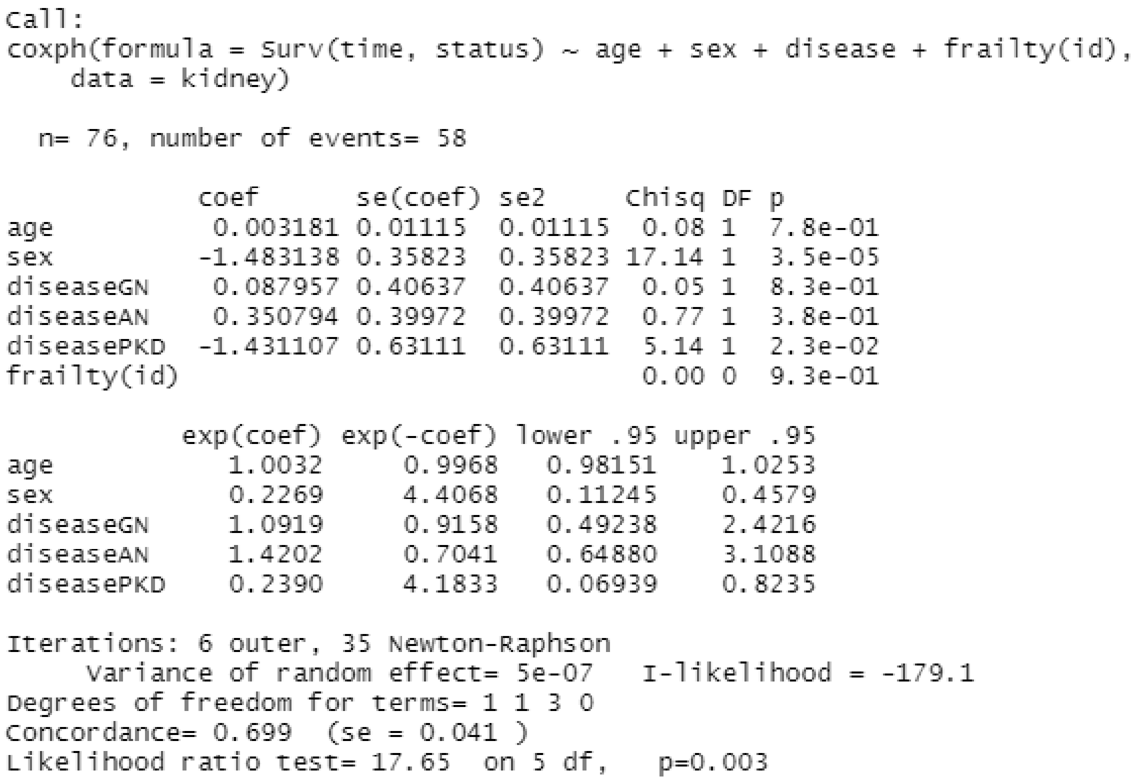
```
Rows: 76
Columns: 7
$ id      <dbl> 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9, 10, 10, 11, 11, 12, 12, 13, 13, 14, 14, 15, 15, 16, 16, 17, 17, 18, 18, 19, 19, 20, 20,~
$ time    <dbl> 8, 16, 23, 13, 22, 28, 447, 318, 30, 12, 24, 245, 7, 9, 511, 30, 53, 196, 15, 154, 7, 333, 141, 8, 96, 38, 149, 70, 536, 25, 17, 4, 185, 177,~
$ status  <dbl> 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, ~
$ age     <dbl> 28, 28, 48, 48, 32, 32, 31, 32, 10, 10, 16, 17, 51, 51, 55, 56, 69, 69, 51, 52, 44, 44, 34, 34, 35, 35, 42, 42, 17, 17, 60, 60, 60, 60, 43, 4~
$ sex     <dbl> 1, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 2, 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 2, 2, 2, 2, 2, ~
$ disease <fct> Other, Other, GN, GN, Other, Other, Other, Other, Other, Other, Other, Other, GN, GN, GN, GN, AN, AN, GN, GN, AN, AN, Other, Other, AN, AN, A~
$ frail   <dbl> 2.3, 2.3, 1.9, 1.9, 1.2, 1.2, 0.5, 0.5, 1.5, 1.5, 1.1, 1.1, 3.0, 3.0, 0.5, 0.5, 0.7, 0.7, 0.4, 0.4, 0.6, 0.6, 1.2, 1.2, 1.4, 1.4, 0.4, 0.4, 0~
```

Figure 26: Glimpse of kidney data set



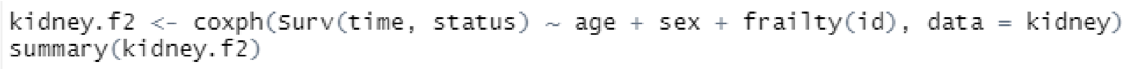
```
kidney.f <- coxph(Surv(time, status) ~ age + sex + disease + frailty(id), data = kidney)
summary(kidney.f)
```

Figure 27: First frailty model for kidney data



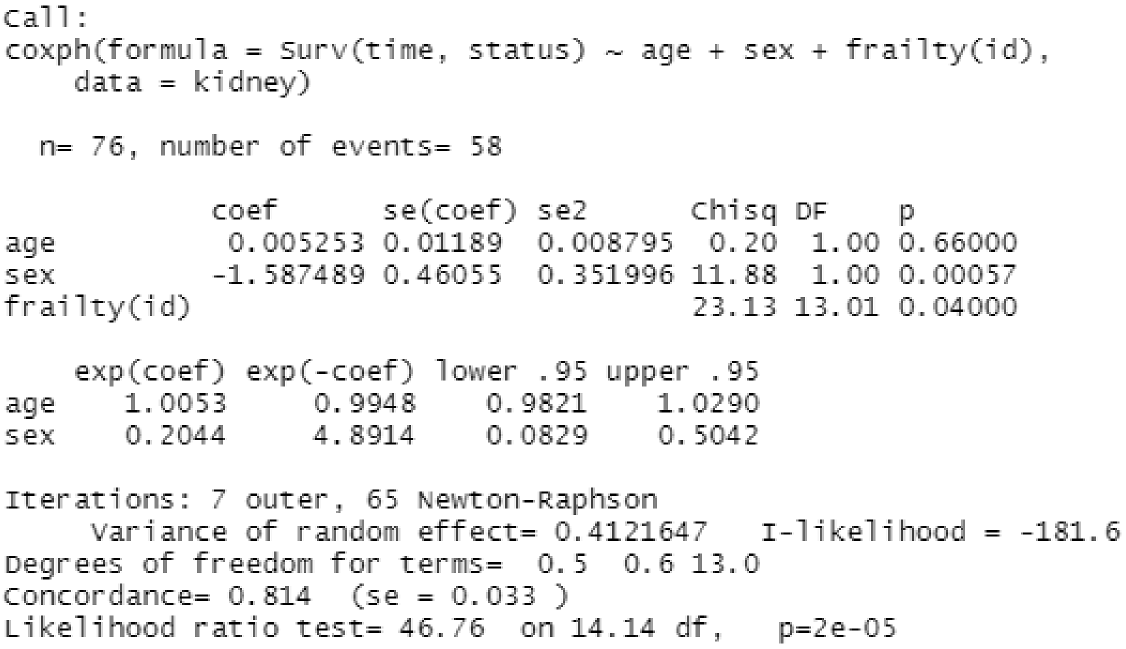
```
Call:
coxph(formula = Surv(time, status) ~ age + sex + disease + frailty(id),
    data = kidney)

  n= 76, number of events= 58

            coef      se(coef) se2     Chisq DF p
age          0.003181 0.01115  0.01115  0.08 1  7.8e-01
sex         -1.483138 0.35823  0.35823 17.14 1  3.5e-05
diseaseGN    0.087957 0.40637  0.40637  0.05 1  8.3e-01
diseaseAN    0.350794 0.39972  0.39972  0.77 1  3.8e-01
diseasePKD  -1.431107 0.63111  0.63111  5.14 1  2.3e-02
frailty(id)                             0.00 0  9.3e-01

           exp(coef) exp(-coef) lower .95 upper .95
age           1.0032     0.9968   0.98151    1.0253
sex           0.2269     4.4068   0.11245    0.4579
diseaseGN     1.0919     0.9158   0.49238    2.4216
diseaseAN     1.4202     0.7041   0.64880    3.1088
diseasePKD    0.2390     4.1833   0.06939    0.8235

Iterations: 6 outer, 35 Newton-Raphson
     Variance of random effect= 5e-07   I-likelihood = -179.1
Degrees of freedom for terms= 1 1 3 0
Concordance= 0.699  (se = 0.041 )
Likelihood ratio test= 17.65  on 5 df,   p=0.003
```

Figure 28: Summary of first frailty model for kidney data



```
kidney.f2 <- coxph(Surv(time, status) ~ age + sex + frailty(id), data = kidney)
summary(kidney.f2)
```

Figure 29: Second Frailty model for kidney data



```
Call:
coxph(formula = Surv(time, status) ~ age + sex + frailty(id),
    data = kidney)

  n= 76, number of events= 58

            coef      se(coef) se2      Chisq DF    p
age          0.005253 0.01189  0.008795  0.20  1.00 0.66000
sex         -1.587489 0.46055  0.351996 11.88  1.00 0.00057
frailty(id)                             23.13 13.01 0.04000

    exp(coef) exp(-coef) lower .95 upper .95
age    1.0053     0.9948    0.9821    1.0290
sex    0.2044     4.8914    0.0829    0.5042

Iterations: 7 outer, 65 Newton-Raphson
     Variance of random effect= 0.4121647   I-likelihood = -181.6
Degrees of freedom for terms=  0.5  0.6 13.0
Concordance= 0.814  (se = 0.033 )
Likelihood ratio test= 46.76  on 14.14 df,   p=2e-05
```

Figure 30: Summary of second frailty model for kidney data

B General Theory

B.1 Mortality Rate Curve

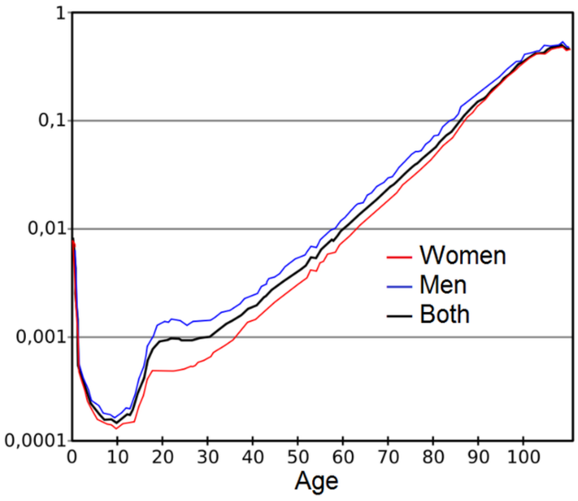


Figure 31: Mortality rate curve

B.2 The Likelihood Function Without Censoring

Recall that in probability theory we have $X_1, \dots, X_n$ random samples of a random variable X, where each sample has a distribution $f(x;\theta)$, which is known as the probability density function for continuous X or probability mass functions for discrete X. Here θ represents either a single parameter or a transposed vector of multiple parameters, $\boldsymbol{\theta}$, which $f(x;\boldsymbol{\theta})$ solely depends on. We use this distribution function to generate the likelihood function for our sample of random variables. In general, this is given by the product of the density functions for each of the random variable $X_1, ...X_n$, i.e. by the equation

$$L(\boldsymbol{\theta};x) = \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta}). \tag{82}$$

In the case of survival data, we will assume T is a continuous random variable which we use to denote survival time. We have $T_1, ...T_n$ times of failure recorded from n individuals and their density is given by $f(t;\theta)$. Using the assumption that each failure time is independent of each other, we have the likelihood function

$$L(\theta;t) = \prod_{i=1}^{n} f(t_i;\theta) \tag{83}$$

Example. Suppose the random variables are samples that follow a *gamma*(α, λ) distribution with known parameter $\alpha > 0$. Then for general t, we have the density function

$$f(t;\lambda) = \frac{\lambda^\alpha t^{\alpha-1} e^{-\lambda t}}{\Gamma(\alpha)} \tag{84}$$

where θ represents our one parameter of interest, λ. We find the maximum likelihood estimator using the method outlined below, we will see that this is given by the value of θ that solves

$$\sum_{i=1}^{n} \frac{\partial \log f(t_i;\theta)}{\partial \theta} = 0 \tag{85}$$

Thus, in the case of the gamma distribution, the likelihood function is given by

$$L(\lambda;t) = \prod_{i=1}^{n} f(t_i;\lambda) = \prod_{i=1}^{n} \frac{\lambda^\alpha t_i^{\alpha-1} e^{-\lambda t_i}}{\Gamma(\alpha)}. \tag{86}$$

B.2.1 Derivation of Maximum Likelihood Estimators

To derive the estimator $\hat{\theta}$ for a particular random variable X, we take the logs of the likelihood function to obtain

$$l(\boldsymbol{\theta};\boldsymbol{x}) = \log\{L(\theta;\boldsymbol{x})\} = \sum_{i=1}^{n} \log\{f(x_i;\theta)\} \tag{87}$$

We then find the partial derivatives of this equation with respect to each parameter of $\boldsymbol{\theta}$ when dealing with multiple parameters, or simply find the partial derivative with respect to θ when dealing with one unknown parameter. In the case of multiple parameters, we have

$$\frac{\partial l(\theta; \boldsymbol{x})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{\partial \log\{f(x_i; \theta)\}}{\partial \boldsymbol{\theta}} \tag{88}$$

for each θ. This is known as the score equation, $U(\theta)$, or equations when dealing with multiple parameters. Solving these gives us the maximised values of $\boldsymbol{\theta}$, which are the maximum likelihood estimates of $\boldsymbol{\theta}$, as required.

Example. Using our likelihood function for $gamma(\alpha, \lambda)$, where we have a single unknown parameter λ, we find the log likelihood to be

$$l(\lambda; T) = (\alpha - 1) \sum_{i=1}^{n} \log(t_i) - n \log(\Gamma(\alpha)) + n\alpha \log(\lambda) - \lambda n\bar{t} \tag{89}$$

where $\bar{t}$ denotes the mean of t_i. The partial derivative with respect to λ is given as follows

$$\frac{\partial l(\lambda; T)}{\partial \lambda} = \frac{n\alpha}{\lambda} - n\bar{t} \tag{90}$$

Lastly, setting this equal to zero and solving for λ, we have the maximum likelihood estimates, $\hat{\lambda}$, for the gamma distribution as $\hat{\lambda} = \frac{\alpha}{\bar{t}}$. We can check this is truly a maximum for the likelihood by taking the second partial derivative, which we find to be

$$\frac{\partial^2 l(\lambda; T)}{\partial^2 \lambda} = \frac{n\alpha}{\lambda^2} \tag{91}$$

and as λ, n and α are strictly positive, this is always positive and hence λ in indeed a maximum.

B.2.2 Asymptotic Expectation and Variance

We use the term asymptotic here to describe an approximation that we can take when we have a sufficiently large sample size, n. As we know the maximum likelihood estimator is a function of our given data $X = (X_1, .., X_n)$, we have a sampling distribution. Thus for sufficiently large n, we are able to approximate our distribution for our n data, this is given by

$$\hat{\theta} \sim N(\theta, J^{-1}) \tag{92}$$

where J denotes the fisher information matrix, $I_n(\theta)$, which is given as follows

$$I_n(\theta) = E\left[-\frac{\partial^2 l(\theta)}{\partial \theta^2}\right]. \tag{93}$$

From this, we can say that maximum likelihood estimator is that it is asymptotically unbiased, as we have, $\lim_{n\to\infty} E(\hat{\theta}) = \theta$. This distribution also tells us the asymptotic variance for $\hat{\theta}$, which is given by

$$Var(\hat{\theta}) \approx J^{-1} = \frac{1}{I_n(\theta)}. \tag{94}$$

Continuing with our example of the $gamma(\alpha, \lambda)$ distribution, we have fisher information

$$I_n(\lambda) = -E\left[\frac{\partial^2 l(\lambda; t)}{\partial \lambda^2}\right] = \frac{n\alpha}{\lambda^2} \tag{95}$$

and therefore, the asymptotic variance is given by

$$Var(\hat{\lambda}) \approx \frac{1}{I_n(\lambda)} = \frac{\lambda^2}{n\alpha} \tag{96}$$

.

B.3 Cox PH Model for Comparison of Two Groups

We we have one covariate in our data, like the case of clinical trials where we have a group of patients receiving a new drug and another group of patients receiving a placebo. We have the hazard rate $h_N(t)$ for the patients receiving the new drug and $h_P(t)$ for the control group receiving a placebo drug at time t. In this instance, we formulate the proportional hazard between the two groups of individuals as follows

$$h_N(t) = \psi h_P(t) \tag{97}$$

where $t \in [0, \infty)$ and ψ is a positive constant, denoting the hazard ratio. We find that when $\psi < 1$, the hazard rate at t is greater for the group receiving the new drug than the placebo, and conversely when $\psi > 0$ the hazard is greater for the placebo group.

Example - Log linear case. As we require a non-negative hazard rate, we resolve this by setting our hazard ratio to $\psi = e^(\beta)$, where β denotes an unknown parameter. This is the log-linear case and we indicate which group each individual is in by introducing an indicator Z of the form

$$Z = \begin{cases} 0 & \text{Group 1} \\ 1 & \text{Group 2} \end{cases} \tag{98}$$

Therefore, $z_i = 0$ for participants in group 1 and $z_i = 1$ for those in group 2. We set $\psi = e^{\beta}$, where some $\beta \in \mathbb{R}$, and have the hazard model for the i_{th} individual, given by

$$h_1(t) = e^{\beta z_i} h_0(t). \tag{99}$$

Finally, we have the hazard rate $h_0(t)$ for subjects in group one and $e^{\beta} h_0(t)$ for subjects in group 2.

School of Mathematics
FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

UNIVERSITY OF LEEDS

Academic integrity statement

You must sign this (typing in your details is acceptable) and include it with each piece of work you submit.

I am aware that the University defines plagiarism as presenting someone else's work, in whole or in part, as your own. Work means any intellectual output, and typically includes text, data, images, sound or performance.

I promise that in the attached submission I have not presented anyone else's work, in whole or in part, as my own and I have not colluded with others in the preparation of this work. Where I have taken advantage of the work of others, I have given full acknowledgement. I have not resubmitted my own work or part thereof without specific written permission to do so from the University staff concerned when any of this work has been or is being submitted for marks or credits even if in a different module or for a different qualification or completed prior to entry to the University. I have read and understood the University's published rules on plagiarism and also any more detailed rules specified at School or module level. I know that if I commit plagiarism I can be expelled from the University and that it is my responsibility to be aware of the University's regulations on plagiarism and their importance.

I re-confirm my consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the EU/EEA) to monitor breaches of regulations, to verify whether my work contains plagiarised material, and for quality assurance purposes.

I confirm that I have declared all mitigating circumstances that may be relevant to the assessment of this piece of work and that I wish to have taken into account. I am aware of the University's policy on mitigation and the School's procedures for the submission of statements and evidence of mitigation. I am aware of the penalties imposed for the late submission of coursework.

Name Faye Williams

Student ID 201308646