Topic 4

Survival Analysis

Peter Tea

Department of Mathematics and Statistics
University of Ottawa

Contents

1 Introduction

Survival analysis is used when the outcome variable of interest is the *waiting time* until an event occurs. The waiting time is usually referred to as the *survival time*: both these terms are used interchangeably. Examples of such events that may be of interest include: death, a physiological response to a therapy treatment, employment, etc.

In survival studies, the interest lies within the *survival* of the subjects, where survival generally means the event not yet occurring. The survival of the subjects, and likewise the risk of having the event occur, is studied by tracking the subjects through a longitudinal type study, where data is gathered on the subjects repeatedly over some study time duration. Moreover, in these survival studies, we will assume that the event being studied can only occur once for each subject (that is, death can only occur once for each subject).

1.1 Censored Data

Censoring is a common attribute for survival data, and is present when the subjects are effectively *lost* from the study. For example, some studies can only collect data within a limited duration of time. Consequently, events that are not observed within this study period will not be recorded. Therefore, survival times that occur after the study has already finished are *incomplete* in the sense that the researcher will not know the exact survival time of the subject. These unknown survival times are deemed *right-censored*.

Despite not knowing the true survival times of right-censored subjects, these censored survival times can still provide some useful information that a survival model can use. Specifically, with right-censored data, it is known that the true survival time is at least equal to or greater than the last observed survival time. Furthermore, it should be noted that there exist other forms of survival time censoring, such as left or interval censoring, which a survival model can use as well. However, right-censoring is the most common type of censoring encountered in survival analysis.

It should also be noted that by including censored data in our analysis, it is assumed that censoring is non-informative and independent of the event of interest. In other words, it is assumed that the censoring of a subject provides no additional information about the subject's risk of having the event occur. This assumption is made to avoid any bias in the survival model.

1.2 Survival Analysis vs. Other Models

Logistic regression, much like survival analysis, is applied to categorical response variables (for example, dead vs. not dead). However, unlike survival models, logistic regression does not consider the timing of occurrence for the event and thus cannot describe the time-to-event process in its model. Additionally, logistic regression cannot incorporate censored

data in its model estimation. Due to these limitations, a fitted logistic regression model on survival data would lose a lot of information and is thus not recommended in this situation.

1.2.1 Example

In a study, researchers are interested in studying the effects of some potential risk factors (for example, smoking status or obesity) on the presence or absence of a disease (for example, lung disease). So, while logistic regression can be used to study how the risk factors affect the presence or absence of a disease, survival analysis allows the researcher to also study how these risk factors affect the time until a subject contracts the disease. For survival analysis, the response variable can be thought of as the combination of time and an event occurring.

Survival analysis is also comparable to ordinary linear regression analysis, in the sense that it includes both a response variable, the time until an event occurs, and a set of covariates. However, due to censored data, survival analysis is usually preferred over linear regression. Some other apparent differences between the two models are as follows:

1. The response variable, time-to-event, can only be positive and thus the underlying distribution is skewed. Therefore, the underlying distribution for the response variable is not normal.

2. Ordinary regression would focus on the expected time that an event occurs. However, researchers may instead be interested in determining the probability of surviving past a certain time, which ordinary regression can not handle.

2 Some Survival Definitions

In this chapter, we introduce some principal definitions required to describe survival models.

2.1 Survival Time

The **survival time** is defined as the waiting time until the occurrence of an event. Additionally, it should be noted that the survival time is a strictly positive, continuous random variable. Let T represent a subject's survival time:

$$T \geq 0$$

Let $f(t)$ denote the probability density function of the survival time and $F(t)$ denote its cumulative distribution function (c.d.f.). $F(t)$ represents the probability that the event has occurred by some stated time value, t:

$$F(t) = Pr(T \leq t)$$

The c.d.f. can also be thought of as the probability of observing a survival time less than or equal to some specified time, t.

2.2 Survival Function

The **survival function** gives the probability of surviving (that is, the event not occurring) beyond a specified time, t. This function is denoted as S(t), and can be thought of as the complement to the c.d.f. of the survival time variable, T.

$$S(t) = P(T > t) = \int_t^\infty f(u)du \tag{1}$$

$$S(t) = 1 - F(t) \tag{2}$$

First note that:

$$\begin{aligned}
\rho(Y_{i1}, Y_{i2}) &= \frac{Cov(Y_{i1}, Y_{i2})}{sd(Y_{i1}) \cdot sd(Y_{i2})} \\
Cov(Y_{i1}, Y_{i2}) &= \rho(Y_{i1}, Y_{i2}) \cdot sd(Y_{i1}) \cdot sd(Y_{i2}) \\
&= (p) \cdot \sqrt{\sigma^2} \cdot \sqrt{\sigma^2} \\
&= p \cdot \sigma^2
\end{aligned}$$

And now using the definition of covariance and its linearity and symmetry properties we get:

$$\begin{aligned}
Var(Y_{i1} - Y_{i2}) &= Cov(Y_{i1} - Y_{i2}, Y_{i1} - Y_{i2}) \\
&= Cov(Y_{i1}, Y_{i1}) + Cov(Y_{i1}, -Y_{i2}) + Cov(-Y_{i2}, Y_{i1}) + Cov(-Y_{i2}, -Y_{i2}) \\
&= Var(Y_{i1}) + Var(Y_{i2}) - 2Cov(Y_{i1}, Y_{i2}) \\
&= \sigma^2 + \sigma^2 - 2(p \cdot \sigma^2) \\
&= 2\sigma^2(1 - p)
\end{aligned}$$

Moreover, the survival function can also be interpreted as the probability of observing a survival time greater than some specified time, t. For example, S(t=100 years) would represent the probability of surviving beyond 100 years.

Below, we illustrate some other properties of the survival function.

All survival functions are monotonically decreasing as the value of t increases.

At t = 0, then:

$$S(0) = 1$$

This implies that at the beginning of the study, the event of interest has yet to occur. However, as t $\to \infty$, then:

$$\lim_{t \to \infty} S(t) = 0$$

This implies that given enough time in the study, the event of interest will occur for all subjects (that is, no subjects will survive). This information is presented visually in Figure 1.

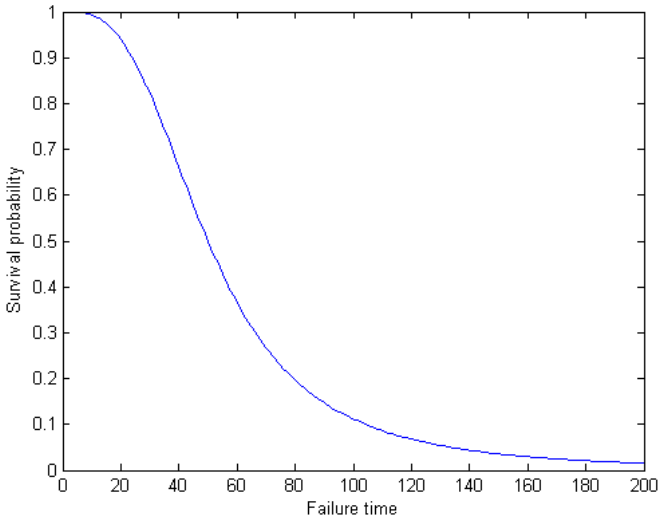

Figure 1: An example of an expected survival function curve. The area under the curve to the right of time t represents the proportion of individuals in the population who have survived up to time t.

It can also be shown that the expected value of T (that is, the mean survival time) is related to the survival function as follows:

$$E[T] = \int_0^\infty S(t)dt$$

It should be noted that there may exist some events where:

$$\lim_{t \to \infty} S(t) \neq 0$$

This implies that given enough time, some events may never occur. For example, consider a study where the event of interest is marriage. If there exist some subjects who wish to remain single, then these subjects will never have the event occur. If this is the case, then the probability distribution function will no longer integrate to 1, and the expected waiting time will be undefined.

2.3 Hazard Function

The **hazard function**, also known as the *conditional failure rate*, gives the instantaneous rate for the event of interest to occur. More importantly, the hazard function describes the conditional probability that survival ends after time t, given that the subject has already survived up to time t.

The hazard function, h(t), is formally defined as:

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr\{t \leq T < t + \Delta t | T \geq t\}}{\Delta t} \tag{3}$$

- Δt denotes the width of the time interval
- The numerator gives the conditional probability that the event will occur in the time interval $[t, t + \Delta t)$, given that the subject has already survived up to time, t.

Evaluating the limit gives an instantaneous rate of event occurrence per unit of time.

It can be shown (through Baye's theorem) that:

$$\begin{aligned} Pr\{t \leq T < t + \Delta t | T \geq t\} &= \frac{[(Pr(t \leq T < t + \Delta t) \cap (T \geq t)]}{Pr(T \geq t)} \\ &= \frac{Pr(t \leq T < t + \Delta t)}{Pr(T \geq t)} \\ &= \frac{F(t + \Delta t) - F(t)}{S(t)} \\ &= \frac{f(t)\Delta t}{S(t)} \end{aligned}$$

Therefore, the hazard function can be rewritten as:

$$h(t) = \frac{f(t)}{S(t)} \tag{4}$$

The advantage of studying the instantaneous hazard function, as opposed to a survival function, is that the hazard allows the researcher to examine how the risk of an event occurring changes as a function of time.

2.3.1 Cumulative Hazard Function

As an extension to the hazard function, **the cumulative hazard function** is defined as:

$$H(t) = \int_0^t h(u)du \tag{5}$$

The cumulative hazard function, H(t), provides the summed hazard up to time t. This function is able to describe the expected number of failures over the time interval [0,t].

2.4 Relationship between the Survival and Hazard function

It can be shown (with the help of equation 4) that for the functions S(t) and h(t), there exist a defined relationship:

$$S(t) = exp\left\{-\int_0^t h(u)du\right\} \tag{6}$$

$$= exp\left\{-H(t)\right\} \tag{7}$$

Therefore, if we know the form of the survival function, S(t), then we can easily derive the corresponding hazard function, h(t), and vice versa.

3 Estimation of S(t)

Different approaches exist to estimate the survival function. In this chapter, we introduce both non-parametric and parametric approaches.

3.1 Kaplan-Meier Estimation

The Kaplan-Meier (K-M) estimator is a non-parametric method used to estimate the survival function. In its estimation, this estimator incorporates all observations: it uses both censored and uncensored data. Specifically, the data is used to estimate the *conditional probability* of survival at each observed survival time: the probability of surviving beyond t, given that the subject has already survived up to time t. The overall survival probability is then obtained by taking the product of each estimated conditional probability of survival.

The K-M estimator is defined as follows:

$$\hat{S}(t) = \prod_{t_i \leq t} \left[\frac{n_i - d_i}{n_i}\right] \qquad i = 1, 2, ..., n \tag{8}$$

where:

- n_i denotes the number of subjects at risk of dying at time, t_i (that is, the number of subjects remaining in the sample during the ith survival time)
- d_i denotes the number of subjects who experience the event of interest during the ith survival time
- $\frac{n_i - d_i}{n_i}$ represents the conditional survival probability during the ith survival time

Often, Kaplan-Meier curves are used to visualize differences in survival between two categorical groups of a covariate. These curves however are not commonly used for assessing the effect of continuous covariates. Instead, Cox regression (chapter 4) should be considered to study the effect of continuous (and even categorical) covariates.

3.1.1 Kaplan-Meier: How it Works

The K-M procedure works as follows:

1. Rank the observations based on survival times from least to greatest:

   $$t_{(1)} \leq t_{(2)} \leq \ldots \leq t_{(m)}$$

   where m denotes the number of unique observed survival times.

   We denote each time interval beginning at an observed time and ending just before the next ordered time. The intervals are indexed by the rank order.

2. For each ith survival time, determine the number of observations at risk n_i, the number of deaths d_i and the number of censored observations c_i.

   The hazard rate is estimated by taking the ratio of the number of events to the number of observations at risk. That is, the probability of *dying* at the ith ranked survival time is given by:

   $$\frac{d_i}{n_i}$$

   The estimated probability of surviving is thus given by the complement of the hazard:

   $$1 - \frac{d_i}{n_i} = \frac{n_i - d_i}{n_i}$$

   The K-M estimator during any point in time of the study is then obtained by multiplying the sequence of conditional survival probability estimators.

3.1.2 WHAS example

To illustrate the steps involved in K-M estimation, an example is provided below with the dataset coming from the Worcester heart attack study (WHAS). In this study, the researchers are interested in survival time (in days) from the subjects' first admission to a hospital, until death. For simplification, only 5 subjects of this study will be considered. This example was provided by (Hosmer, 2008).

Subject	Time	Censor
1	6	1
2	44	1
3	21	0
4	14	1
5	62	1

Table 1: Data for 5 subjects in the WHAS study

From the dataset above (Table 1) we first rank the observed survival times to build time intervals, where each interval begins at an observed time and ends just before the next ordered time. For example, subject 1 has the smallest survival time (6 days) and is used to define the following interval:

$$I_0 = \{t : 0 \leq t < 6\} = [0, 6)$$

Subject 4 has the second ranked ordered time (14 days) and is thus used to define the next survival time interval of interest:

$$I_1 = \{t : 6 \leq t < 14\} = [6, 14)$$

Using the same approach, the remaining time intervals are obtained as follows:

Index	Interval
I_0	$[0, 6)$
I_1	$[6, 14)$
I_2	$[14, 21)$
I_3	$[21, 44)$
I_4	$[44, 62)$
I_5	$[62, \infty)$

We then estimate the probability of surviving through each of these 6 intervals. For example, the probability of surviving in I_0 (the probability of surviving at each t in I_0) is:

$$\hat{S}(t) = 1.0$$

(since the first observed death occurs at t = 6, which is not included in the interval I_0).

The probabilities of survival for the remaining intervals are obtained by taking the product of each sequential probability of survival. These calculations are presented in Table 2.

Interval	n_i	d_i	$\frac{n_i - d_i}{n_i}$	Survival Estimate $\hat{S}(t)$
I_0	$n_0 = 5$	$d_0 = 0$	$\frac{5}{5}$	1.0
I_1	$n_1 = 5$	$d_1 = 1$	$\frac{4}{5}$	$\hat{S}(6) = 1.0 \text{ x } \frac{4}{5} = 0.8$
I_2	$n_2 = 4$	$d_2 = 1$	$= \frac{3}{4}$	$\hat{S}(14) = 1.0 \text{ x } \frac{4}{5} \text{ x } \frac{3}{4} = 0.6$
I_3	$n_3 = 3$	$d_3 = 0$	$\frac{3}{3}$	$\hat{S}(21) = 1.0 \text{ x } \frac{4}{5} \text{ x } \frac{3}{4} \text{ x } \frac{3}{3} = 0.6$
I_4	$n_4 = 2$	$d_4 = 1$	$\frac{1}{2}$	$\hat{S}(44) = 1.0 \text{ x } \frac{4}{5} \text{ x } \frac{3}{4} \text{ x } \frac{3}{3} \text{ x } \frac{1}{2} = 0.3$
I_5	$n_5 = 1$	$d_5 = 1$	$\frac{0}{1}$	$\hat{S}(62) = 1.0 \text{ x } \frac{4}{5} \text{ x } \frac{3}{4} \text{ x } \frac{3}{3} \text{ x } \frac{0}{1} = 0.0$

Table 2: This table illustrates the steps needed in order to calculate the K-M estimator of survival in each of the 6 defined time duration periods.

Note that for I_3 there were no deaths observed in this time interval, however one subject was still lost due to censoring. The probability of surviving through this time interval is 1.0 since although one subject was censored, we assume that this subject survived beyond its

last recorded survival time. This example purposefully illustrates that censored observations affect the observed number of subjects at risk, but do not affect the number of observed deaths. It should also be noted that the survival probability is constant between observed failure times.

3.2 Log-Rank Test

The log-rank test is used to test whether there exist any difference in the estimated survival curves for 2 or more groups in the sample. For example, we may wish to compare survival curves between: gender, treatment groups (eg. treated vs. placebo), different environments, etc. As such, the log-rank test can be thought of as being the analogue to a t-test or an ANOVA, where the survival of different groups are being compared.

The log-rank test tests the null hypothesis of no difference in survival between two or more independent groups. When the null hypothesis is true, then the survival curves are identical to one another. Formally, the hypotheses that we wish to test can be written as follows:

$$H_0 : S_1(t) = S_2(t) = \ldots = S_k(t) \text{ for all } t \leq \tau$$
$$H_1 : S_i(t) \neq S_j(t) \text{ for at least one pair i, j and some } t \leq \tau$$

where τ represents the largest observed survival time and k represents the number of groups being compared.

The calculations used in the log-rank test is based on the contingency table of group by status at each observed survival time, as shown in Table 3.

Event \ Group	1	0	Total
Die	d_{1i}	d_{0i}	d_i
Not Die	$n_{1i} - d_{1i}$	$n_{0i} - d_{0i}$	$n_i - d_i$
At Risk	n_{1i}	n_{0i}	n_i

Table 3: Contingency table used to test whether the Survival function is equal in two groups at the ith observed survival time. The two groups in question are denoted as "1" and "0" respectively. The event occurring is denoted as "Die" and the event not yet occurring is denoted as "Not Die".

The notation in Table 3 can be summarized as follows:

- The number of subjects at risk, at the observed survival time t_i, for group "0" is denoted by n_{0i} and for Group "1" by n_{1i}
- The number of observed deaths in each of these two groups is denoted by d_{0i} and d_{1i} respectively

- The total number at risk is denoted by n_i
- The total number of deaths is denoted by d_i

The log-rank test works by calculating, at each survival time, the observed and expected number of events in one of the groups of interest. For example, the expected number of deaths in group "1" (denoted as $\hat{e}_{1i}$) is calculated as follows:

$$\hat{e}_{1i} = \frac{n_{1i} \cdot d_i}{n_i}$$

Then, an overall summary statistic is obtained by summing the results across all the observed survival time points. The computed test statistic has degrees of freedom equal to k-1, where k represents the number of comparison groups.

The test statistic (for group 1) is then obtained as follows:

$$Q = \frac{\left[\sum_{i=1}^{m} w_i(d_{1i} - \hat{e}_{1i})\right]^2}{\sum_{i=1}^{m} w_i^2 \hat{v}_{1i}}$$

where:

- m is the last ordered survival time. The observed ordered survival times are denoted $t_{(i)}$ where $i = 1, ..., m$
- $\hat{v}_{1i}$ is the estimator for the variance of the expected number of deaths.
- w_i are weights whose values depend on the specific test. For example, the log-rank test uses $w_i = 1$ where each survival time is given equal weights. The Wilcoxon test, however, uses $w_i = n_i$ thus giving more weight to survival times with a larger number of subjects at risk.

At the level of α, we reject H_0 if:

$$Q \geq \chi^2_{\alpha(k-1)}$$

That is, the statistic Q is obtained using the chi-square distribution with k-1 degrees of freedom.

3.3 An example in R

An example here is provided in order to illustrate the process of obtaining and analysing survival curves in R. The following dataset (CNS lymphoma data) came from a clinical study conducted at Oregon Health Sciences University (OHSU). The patients in this study were treated for their CNS lymphoma condition with 2 different therapies. We would like to estimate the survival function for these patients and also, we wish to examine whether there exist any difference in survival based on the treatment group of the patient.

```
###Kaplan-Meier R codes
# "Survival_Time" is a column variable denoting the survival (in years)
# "Status" is a column indicating which subjects died (1) and which subjects are
censored (0)

> data <-read.csv("CNSlymphoma.csv", header = TRUE, sep = ",")

> library(survival)

> my.survival.model <- survfit(Surv(Survival_Time, Status)  ~ Treatment_Group,
data = data, type="kaplan-meier")
#   Fit a survival curve  K-M estimator, stratifying for the variable "Treatment group".

> library(survminer)
#   This package will create a much nicer K-M survival curve.

> ggsurvplot(my.survival.model, data = data,
           title = "Kaplan-Meier survival function of years
to death, stratifying by treatment group" ,
           conf.int=TRUE,
           risk.table=TRUE,
           legend.labs=c("Treatment Group 1", "Treatment Group 2"),
           legend.title="Group",
           main="Kaplan-Meier Curve for CNS lymphoma Survival",
           risk.table.height=.3)
```

The K-M curves are given in Figure 2. Visually, it appears that treatment group 1 is consistently above treatment group 2 in terms of survival probability throughout the entire study period. Therefore, we might conclude that treatment 1 was more effective than treatment 2. To formally test whether the survival is indeed different for these two treatment groups, we can run a log-rank test.

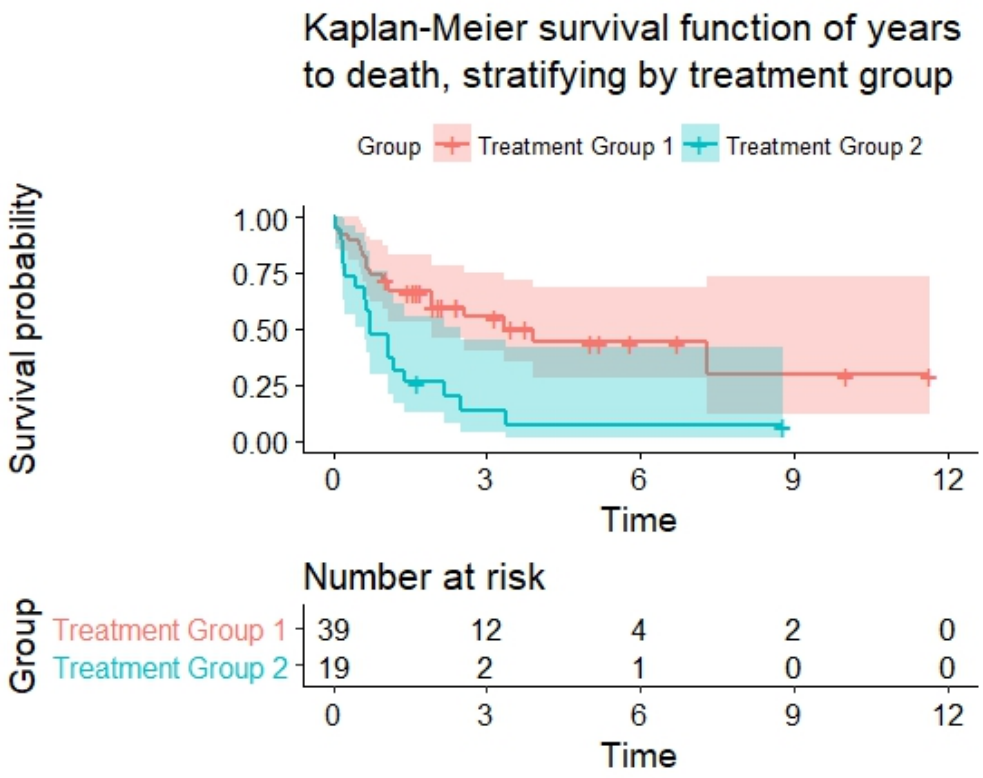

Figure 2: The Kaplan-Meier survival estimation and survival table for the 2 different treatment groups

```
> survdiff(Surv(Survival_Time, Status) ~ Treatment_Group, data=data)
# Compare the survival distributions of the 2 treatment groups (log-rank test)
#H_0: The survival curves are the same...

Call:
survdiff(formula = Surv(Survival_Time, Status) ~ Treatment_Group,
    data = data)

                    N Observed Expected (O-E)^2/E (O-E)^2/V
Treatment_Group=0 39       19    26.91      2.32      9.52
Treatment_Group=1 19       17     9.09      6.87      9.52

 Chisq= 9.5  on 1 degrees of freedom, p= 0.00203
```

From the R output, we can see that in using a log-rank test, which tests whether the survival curves are the same for both treatment groups, we reject the null hypothesis (at the level of $\alpha = 0.05$). Therefore, we conclude that the survival curve is not the same in both treatment groups.

3.4 Parametric estimation

Parametric methods can also be used to estimate the survival of subjects instead of the non-parametric K-M method. Parametric estimation is used when the underlying distribution of the survival time is known, which can then lead to more precise estimates of S(t).

For example, consider the case where the failure rate (i.e. the hazard rate) is constant over some time interval:

$$h(t) = \lambda$$

for some constant λ.
Now, with the hazard function set, we can derive the survival function using equation (6):

$$\begin{aligned} S(t) &= exp\left\{-\int_0^t h(u)du\right\} \\ &= exp\left\{-\int_0^t \lambda du\right\} \\ &= exp\left\{-\lambda t\right\} \end{aligned}$$

If a survival distribution has this survival function, then the survival distribution is deemed exponential with parameter λ.
It can be shown (with the help of equation 4) that the probability distribution function for the exponential distributed survival time is:

$$f(t) = \lambda exp\left\{-\lambda t\right\}$$

Some other known and popular survival distributions include: Weibull, Gompertz, log-logistic, log-normal or gamma. To determine whether a parametric approach is appropriate, it may be a good idea to first plot the hazard function for the observed time-to-event data and then determine whether or not the plot is consistent with a known parametric distribution.

4 Survival Regression Models

It should be noted that while the K-M estimator is effective in estimating the survival function and allows the analyst to compare survival between groups, it still does not allow the analyst to examine how continuous explanatory variables affect the survival of the subject.

Typically, we are often interested in modelling the relationship between the survival time, T, with a set of covariates. For example, we may be interested in determining whether age, gender, or different treatment groups affect the relative risk of survival. Regression methods encountered in survival analysis typically focuses on the hazard function, h(t), as the outcome of interest. However, due to equation (6), the inferences made on the hazard can easily be translated for the survival as well. A popular regression model used in such cases is the Cox regression model which allows us to model the relationship between the hazard function and its predictor variables.

4.1 An introduction to modelling hazards

In survival analysis, the hazard can be modelled as a function of both time and of a set of predictor variables (along with its corresponding unknown parameters). This survival regression model can then be used to either:

i. Predict survival times

ii. Compare survival times among different groups

The most common form of regression modelling of the hazard function is the proportional hazard form:

$$h(t, x, \beta) = h_0(t)r(x, \beta) \tag{9}$$

Specifically, in this model, the hazard function is written as a product of 2 functions where:

- X is a (n x p) matrix representing the k different covariates
- β is a (p x 1) matrix representing the p different parameters in the model. Here, $p = k + 1$.
- $h_0(t)$ is a function of time and represents the baseline hazard function that describes the hazard risk for subjects when $x_1 = x_2 = \ldots = x_k = 0$ (that is, when all k covariates are equal to zero)
- $r(x, \beta)$ represents how the hazard changes as a function of covariates. This function affects the baseline hazard across all values of t

Thus, the hazard function is described as a multiplicative relationship between the baseline hazard function and a function of the covariates. In this form, a one unit change in the predictor variable causes a proportional change in the hazard.

It should be noted that when $r(x, \beta) = 1$, then $h_0(t)$ is equal to the hazard function. Therefore, $h_0(t)$ is often referred to as the *baseline hazard function* which is analogous to the intercept in an ordinary regression model.

Many survival regression models are deemed semi-parametric meaning that these models are based on a parametric regression model, but its regression structure does not make any assumption about the probability distribution of event times (Allison, 2012). That is, there are no assumptions of the distribution of the baseline hazard function.

With the hazard function set, the survival function can now be derived using equation (6).

$$\begin{aligned} S(t,x,\beta) &= exp\left\{-\int_0^t h(u,x,\beta)du\right\} \\ &= exp\left\{-r(x,\beta)\int_0^t h_0(u)du\right\} \\ &= S_0(t)^{-r(x,\beta)} \end{aligned}$$

where:

- $S_0(t) = e^{-H_0(t)}$ denotes the baseline survival function.
- x denotes the k covariates
- β denotes the p the regressor coefficients

4.2 Cox Proportional Hazards Model

It has been proposed by Cox that the model (9) can be parametrized with $r(x,\beta) = exp\{x'\beta\}$. In doing so, the hazard function can now be modelled with the following equation:

$$h(t,x,\beta) = h_0(t)exp\{x'\beta\} \tag{10}$$

It should be noted that this regression model is still semi-parametric since it does not make any explicit specification of its model error component. That is, the Cox proportional hazard model (Cox PH) does not make any assumptions on its baseline hazard function, however the model does assume that the covariates have a log-linear effect on the hazard function. Therefore, an advantage in using the Cox PH model is that we can fit survival models without assuming the underlying distribution (such is the condition required in parametric estimation).

With the hazard function set, under the Cox PH model, the survival function can now be explicitly defined as:

$$S(t,x,\beta) = S_0(t)^{-exp(x'\beta)} \tag{11}$$

4.2.1 Parameter Estimation

The parameters of the Cox PH model (11) can be estimated through *partial maximum likelihood*. The parameter estimates from the partial likelihood function have the same distributional properties as the full maximum likelihood estimates.

Let

- t_i denote the observed survival time
- x_i denote the predictor variable
- c_i denote the censoring variable

for $i = 1, 2, \ldots, n$

To derive the likelihood function, we first note the contributions of both censored and uncensored survival times in our model:

- When $c_i = 1$, the subject's survival time is uncensored (that is, it is known that the survival time occurred at time, t_i). Thus, these observations necessarily contribute the p.d.f. f(t) to the likelihood model.
- When $c_i = 0$, the subject's survival time is censored (that is, it is only known that the survival time is atleast time, t_i). Thus, these observations contribute the survival function S(t) to the likelihood model (since $S(t) = Pr(T > t)$).

Assuming independent observations, the likelihood function, $l(\beta)$, can be written as:

$$l(\beta) = \prod_{i=1}^{n} f(t, \beta, x_i)^{c_i} \prod_{i=1}^{n} S(t, \beta, x_i)^{1-c_i}$$

By using the partial likelihood function, the maximum partial likelihood estimator can be estimated using some numerical method, for example the Newton-Raphson method (Ekman, 2017).

4.3 Hazard ratios

Given covariate values $x = a$ and $x = b$, the hazard ratio (HR) for these covariates can be defined as the ratio of their respective hazard functions:

$$HR(t, x = a, x = b) = \frac{h(t, x = a, \beta)}{h(t, x = b, \beta)} \tag{12}$$

$$\begin{aligned} HR(t, x = a, x = b) &= \frac{h_0(t) \cdot r(x = a, \beta)}{h_0(t) \cdot r(x = b, \beta)} \\ &= \frac{r(x = a, \beta)}{r(x = b, \beta)} \\ &= \frac{exp\{a\beta\}}{exp\{b\beta\}} \\ &= exp\{\beta(a - b)\} \end{aligned}$$

Thus, the hazard ratio depends only on the function of covariates and not on the baseline hazard function or the survival time. The hazard ratio is a comparative measure of survival experience over the *entire* study time period *(Hosmer, 2008)*.

4.4 Interpretation of the model

To understand how the covariates affect the hazard function, we must first determine an appropriate transformation of the hazard function to induce linear coefficients in the regression model. Like generalized linear models, we are looking for a proper link function. The appropriate link function for the proportional hazard model is the **natural logarithm** transformation.

The transformation works as follows:

$$h(t, x, \beta) = h_0(t) \cdot exp\{x'\beta\}$$

$$ln[h(t, x, \beta)] = ln\left[h_0(t) \cdot exp\{x'\beta\}\right]$$

$$= ln[h_0(t)] + x'\beta$$

Consider now when the log-hazard function changes in the covariate value from x = a to x = b.

$$ln[h(t, x = a, \beta)] - ln[h(t, x = b, \beta)] = [ln[h_0(t)] + (a)\beta] - [ln[h_0(t)] + (b)\beta]$$

$$ln\left[\frac{h(t, x = a, \beta)}{h(t, x = b, \beta)}\right] = (a)\beta - (b)\beta$$

$$= \beta(a - b)$$

Following exponentiation of this result, we obtain:

$$\frac{h(t, x = a, \beta)}{h(t, x = b, \beta)} = e^{\beta(a-b)}$$

The significance of this result is better understood if we also consider the hazard ratio. By definition, the parts on the left side of the result represents the hazard ratio comparing covariate values x = a to x = b. Therefore, we conclude that:

$$HR(t, x = a, x = b, \beta) = e^{\beta(a-b)}$$

Thus, the hazard ratio is related to the regression coefficients: The measure of effect is the hazard ratio.

4.4.1 Categorical Covariate

It can be shown that if the covariate is dichotomous (that is the covariate, is coded 0 or 1) then:

$$HR(t, 1, 0, \beta) = e^{\beta \cdot (1-0)} = e^{\beta}$$

where e^{β} represents the hazard ratio (or alternatively, β represents the the log hazard ratio.)

For example, consider a study where the obtained hazard ratio compares the hazard rates between males (coded "1") and females (coded "0"). If $HR = 2$, then this would indicate that at any time during the study, the rate of death among males is 2x greater than the rate of death for females.

4.4.2 Continuous Covariate

Consider the hazard ratio which compares the hazard rate between subject i with the covariate $x_1 = x + 1$ and subject j with the covariate $x_1 = x$. The hazard ratio between subject i and subject j can be written as:

$$\begin{aligned} HR &= \frac{h_i(t)}{h_j(t)} \\ &= exp\{\beta(x + 1 - x)\} \\ &= e^{\beta} \end{aligned}$$

The interpretation works as follows: For a unit increase in the predictor variable, the hazard ratio changes by a factor or proportion of $exp\{\beta\}$.

It can be noted that if we expect a covariate to increase the relative survival of a subject, then its coefficient value should be negative (which results in a hazard ratio value less than 1). An increase in survival is associated with a decrease in hazard rate.

4.5 Assumptions

The Cox PH model makes two main assumptions:

1. The hazard ratio is constant over time

   It is assumed that the covariates of the hazard function is independent of the survival time. That is, the hazard ratio should not vary over time. In other words, the covariates should have the same effect, at all durations in time, on the hazard ratio and thus there should be no interaction between the covariates and survival time. This if often referred to as the **proportional hazard assumption**.

2. There exist a log-linear relationship between the hazard and its covariates

$$ln[h(t, x, \beta)] = ln[h_0(t)] + x'\beta$$

Another assumption in survival analysis is that time, T, is continuous. However, sometimes in research the time to event is not measured on a continuous scale, but rather on a discrete scale. Thus, outcome events may appear to occur simultaneously in the same time interval. When the number of tied events is large, the obtained regression coefficients may be biased. Some alternative methods exist for dealing with such ties by adjusting the partial likelihood function. These methods include: Breslow approximation, Efron approximation, exact partial likelihood.

4.5.1 Model Diagnostics

To determine whether the proportional hazard assumption is valid, the analyst should examine whether the estimated hazard curves for each level of the covariate are equidistant over time. If the covariate is categorical, then this can be done by plotting the log[-log S(t)] vs time for two groups. If the assumption is valid then the two survival curves should be approximately parallel and should never cross. Unfortunately, this diagnostic is limited to categorical covariates and if we wish to further assess the proportional hazard assumption for continuous covariates, then Shoenfeld residuals should instead be analysed.

4.5.2 Residuals

Often in verifying regression model adequacy, residuals of the model are examined. The residuals can usually be thought of as the difference between the observed value of the outcome variable and the value predicted by the model.

However, a Cox PH model includes censored observations and the use of the partial maximum likelihood function, which makes the analysis of the model diagnostics much more complicated. Therefore, finding residuals may be difficult.

Despite these issues, different residuals have been developed to help with validating model adequacy. Some of these residuals are briefly introduced below:

- The Cox-Snell residuals can be examined to help draw conclusions on the overall fit of the model.
- The Martingale residuals can be examined to help draw conclusions on the functional relationship between a covariate and the hazard (for example, a log-linear relationship)

- The Schoenfeld residuals can be examined to help draw conclusions on the proportional hazards assumption

4.5.3 Shoenfeld Residuals

In this subsection, we will further introduce the Schoenfeld residual and see how it can be used to test one of the main assumptions of the Cox PH model.

The Schoenfeld residuals are calculated for each covariate at each observed survival time. For the kth covariate and the ith survival time, the Schoenfeld residual is obtained as follows:

$$\hat{r}_{ik} = c_i(x_{ik} - \sum_{i=1}^{j\varepsilon R_i} x_{kj} \cdot p(\hat{\beta}_k, x_{jk})) \qquad i = l, 2, ..., n$$

where

- $p(\hat{\beta}_k, x_{jk}) = \dfrac{exp(\hat{\beta}_k \cdot x_{jk})}{\sum_{j\varepsilon R_i} exp(\hat{\beta}_k \cdot x_{jk})}$ represents the probability of the event occurring for the ith subject
- t_i represents survival time
- x_i represents the model covariates
- C_i is the censoring indicator

$$C_i = \begin{cases} 1 \text{ for uncensored observations} \\ 0 \text{ for censored observations} \end{cases}$$

In words, Shoenfeld residuals represent the difference between the observed covariate value and its average value. These residuals are not defined for censored survival times.

If the proportional hazards assumption holds, then we should observe that the Schoenfeld residuals are not correlated with survival time. Thus, the null hypothesis can be written as H_0: The correlation between the Schoenfeld residuals and ranked failure times is zero (meaning that proportional hazards assumption is valid).

4.6 Time-Dependent Covariates

If there is evidence that the proportional hazard assumption is not valid, then there exist some strategies to help circumvent this problem. For example, the analyst can fit a time $\cdot$ covariate interaction term in the model.

Another solution is stratification, where the time-dependent variable is treated as a stratification variable. These stratified covariates are usually included in the model to account for any nuisance confounding variable. In stratifying the variable, the interpretation of the stratifying parameter estimate in terms of the hazard ratio should be avoided (since the baseline hazard function will change).

In other scenarios, the covariate may be fixed throughout the entire study period, however the covariates' effect on survival may vary over time. Such a covariate displays a time-dependent effect and may be dealt with by using a piecewise Cox model where we divide the study period into distinct intervals. Then, a different Cox PH model is fitted in each interval to compare the coefficients for each covariate across the different time intervals.

4.7 Significance of the model

A partial likelihood ratio test can be used to assess the overall fit of the model. Its test statistic, G, is calculated by comparing the log partial likelihood of the model containing the covariates to the log partial likelihood of the model not containing the covariates.

Specifically, the test statistic is calculated as follows:

$$G = 2\left\{L_p(\hat{\beta}) - L_p(0)\right\}$$

where:

- $L_p(\hat{\beta})$ denotes the log partial likelihood of the model which includes the predictor variables
- $L_p(0)$ denotes the log partial likelihood of the model which does not include the predictor variables.

Under the null hypothesis ($H_0 : \beta = 0$), it can be shown that:

$$G \sim \chi^2_k$$

where k is the number of regressors in the fitted model.

It should also be noted that with the Cox PH model, an analogue to the R^2 value (which are frequently interpreted in ordinary regression models) have been developed. These R^2 values are deemed pseudo-R-squared values since these values are calculated using some modifications in its equation. However the interpretation of these values are still very similar to the interpretations for the ordinary R^2 values. The formula for these new R^2 values is as follows:

$$R^2 = 1 - exp\left\{\frac{2}{n}(L_p(0) - L_p(\hat{\beta}))\right\}$$

4.8 An example in R

As seen in the previous R example, the same dataset (CNS lymphoma data) will be used to illustrate the use of a Cox PH model.

The explanations for the variables in the dataset can be viewed in Table 4.

Covariate	**Explanation**
TreatmentGroup	Two different treatment groups coded "1" and "0"
Sex	"1" = female, "0" = male
Age	Recorded age when the subject received the treatment
KarnofskyScore	Karnofsky performance score before the treatment. Values can range from 0 - 100.
Radiation	What was the amount of radiation that the patient received? "1" if amount of radiation ¿ 4000. "0" otherwise.
ChemoPrior	Has the patient received a chemotherapy treatment before? "1" = yes, "0" = no.
LesionNumber	"0" = Single lesion, "1" = Multiple lesions
LesionLocation	"0" = Superficial lesion, "1" = Deep lesion
LesionType	"0" = Supra lesion, "1" = Infra lesion, "2" = Both
Procedure	"1" = Subtotal resection, "2" = Biopsy, "3" = Other

Table 4: Explanation of the variables in the CNS lymphoma dataset. Data was obtained from (Tableman, 2003).

```
### PART2: Cox proportional Hazard

library(MASS)
# Need this package for model building...

model <- coxph(Surv(Survival_Time, Status) ~ as.factor(Treatment_Group) +
as.factor(Sex) + Age + Karnofsky_Score + as.factor(Lesion_Number) +
```

```
as.factor(Lesion_Location) + as.factor(Lesion_Type) + as.factor(Procedure) +
as.factor(Radiation) + as.factor(Chemo_Prior),
               data = data)
#     Fit the ENTIRE model...with all recorded covariates


fit.model<- step(model,  direction = "backward", trace=FALSE)
# Perform backwards elimination model fitting

summary(fit.model)

Call:
coxph(formula = Surv(Survival_Time, Status) ~ as.factor(Treatment_Group) +
    as.factor(Sex) + Age + Karnofsky_Score + as.factor(Radiation) +
    as.factor(Chemo_Prior), data = data)

  n= 58, number of events= 36

                               coef exp(coef) se(coef)      z Pr(>|z|)
as.factor(Treatment_Group)1  1.77843   5.92053  0.73085  2.433  0.01496 *
as.factor(Sex)1             -1.37781   0.25213  0.49563 -2.780  0.00544 **
Age                          0.03341   1.03397  0.01540  2.169  0.03007 *
Karnofsky_Score             -0.03680   0.96386  0.01180 -3.120  0.00181 **
as.factor(Radiation)1       -1.12510   0.32462  0.71500 -1.574  0.11559
as.factor(Chemo_Prior)1      1.15064   3.16020  0.50863  2.262  0.02368 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

                            exp(coef) exp(-coef) lower .95 upper .95
as.factor(Treatment_Group)1    5.9205     0.1689   1.41339   24.8004
as.factor(Sex)1                0.2521     3.9662   0.09544    0.6661
Age                            1.0340     0.9671   1.00323    1.0657
Karnofsky_Score                0.9639     1.0375   0.94183    0.9864
as.factor(Radiation)1          0.3246     3.0805   0.07994    1.3182
as.factor(Chemo_Prior)1        3.1602     0.3164   1.16619    8.5637

Concordance= 0.772  (se = 0.054 )
Rsquare= 0.397   (max possible= 0.987 )
Likelihood ratio test= 29.34  on 6 df,   p=5.235e-05
Wald test            = 28.57  on 6 df,   p=7.335e-05
Score (logrank) test = 34.17  on 6 df,   p=6.226e-06
```

The output provides the direct interpretation of the hazard ratio for each of the selected covariates. For example, the hazard is estimated to be about 5.92x greater for treatment group "1", when compared to treatment group "0". Additionally, when looking at the effects of the continuous covariate *KarnofskyScore*, we see that a one unit increase in this variable decreases the hazard multiplicatively by a factor of 0.963. In other words, a one unit increase in *KarnofskyScore* will decrease the hazard ratio by about 3.7 %.

We now move onto model diagnostics. First, we would like to test whether the proportional hazards assumption is met.

```
###Model Diagnostics

cox.zph(fit.model)
#   Tests the proportional hazards assumption for each covariate...
#   The Pearson product-moment correlation between the scaled Schoenfeld residuals
 and survival time is being tested.
#   H_0: There exists no correlation

                                rho   chisq     p
as.factor(Treatment_Group)1  0.14791 0.81579 0.366
as.factor(Sex)1              0.12016 0.56450 0.452
Age                         -0.04285 0.07386 0.786
Karnofsky_Score              0.00952 0.00254 0.960
as.factor(Radiation)1       -0.04032 0.05575 0.813
as.factor(Chemo_Prior)1     -0.01240 0.00520 0.942
GLOBAL                           NA 3.29046 0.772
```

Above, we are testing whether there exist any significant interaction between each covariate with log(time). For this test, the null hypothesis states that there exist no significant interaction and thus the proportional hazard assumption is valid.

Since none of the obtained p-values are significant (at the level of $\alpha = 0,05$), we can reasonably conclude that there is no significant interaction with time and thus the proportional hazard assumption is met for all the selected covariates in the model.

We can also look at Schoenfeld residual plots for each covariate to further investigate the validity of the proportional hazard assumption.

```
par(mfrow = c(3, 2))
plot(cox.zph(fit.model))
#  Plots the scaled Schoenfeld residuals.
#   Each plot is of scaled Schoenfeld residuals against transformed time for each
covariate in a model fit).
# The solid line is a smoothing spline fit to the plot, with the broken
 lines representing a  2-standard-error band around the fit.
```

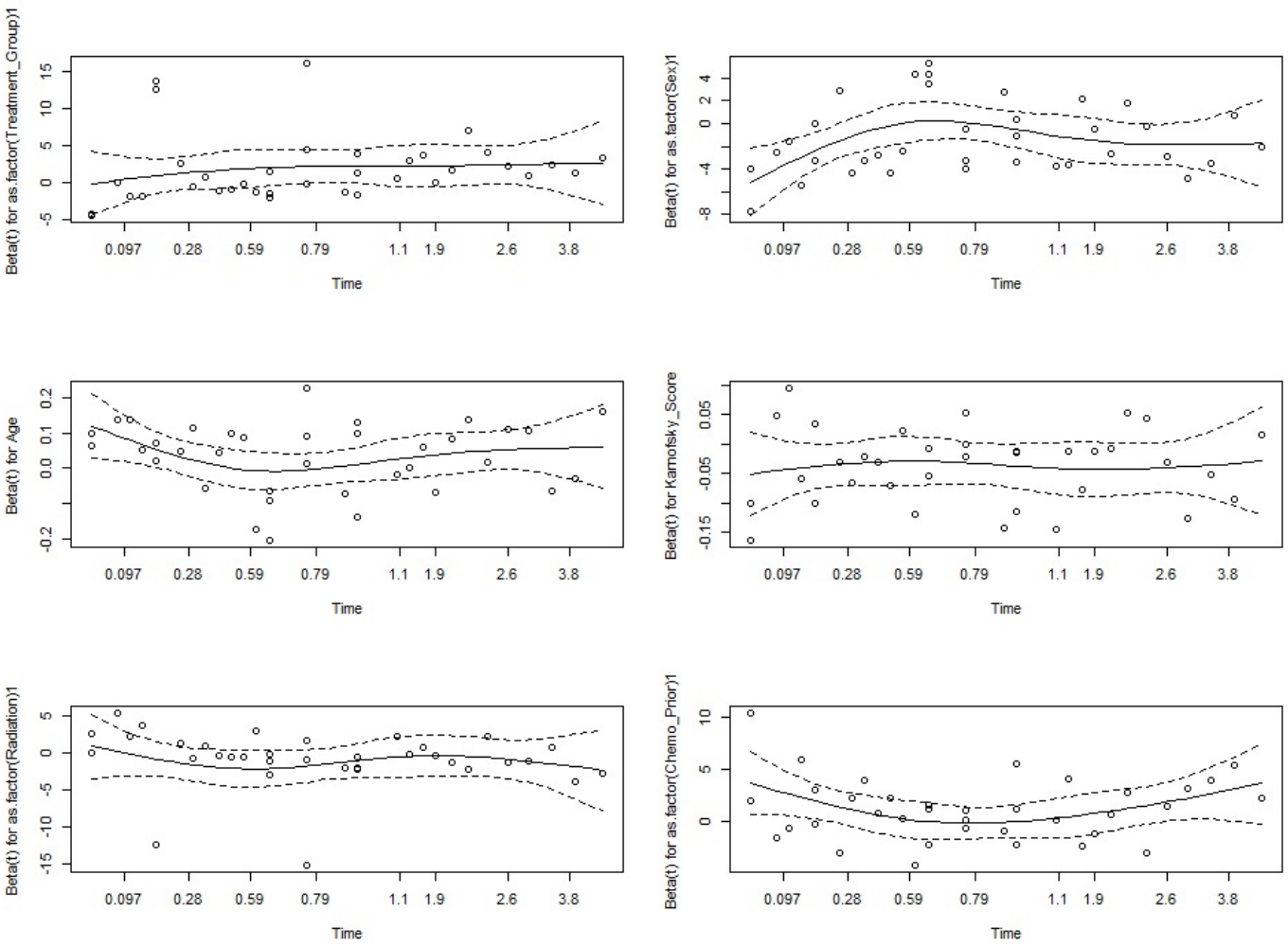

Figure 3: The schoenfeld plots for each fitted covariate in the cos proportional hazard model.

From these plots (found on Figure. 3), we can see that there are no systematic departures from a horizontal line. That is, the residuals exhibit a random pattern at each failure time. Thus, the proportional hazard assumption is valid.

When performing Cox regression, it is also important to check for multicollinearity in the regressors. Multicollinearity occurs when the covariates have high correlations among themselves. This results in unreliable and unstable estimates of regression coefficients.

```
library(car)
#Checking for multicollinearity
vif(fit.model)
#VIF >10 indicates high correlation with the other regressors.
```

```
as.factor(Treatment_Group)             as.factor(Sex)                      Age
                  4.192127                  1.723898                 1.591115
      as.factor(Radiation)     as.factor(Chemo_Prior)
                  4.081378                  1.821504
```

From the output, it appears that the inflation of the standard errors of the parameter estimates are not of significant concern (since none of the VIF values are greater than 10).

5 Sources

Allison, D. Paul. Survival Analysis. Statistical Horizons. University of Pennsylvania. URL: https://statisticalhorizons.com/wp-content/uploads/2012/01/Allison_SurvivalAnalysis.pdf

Ekman, Anna. (2017) Variable selection for the Cox proportional hazards model: A simulation study comparing the stepwise, lasso and bootstrap approach. Umea University.

Fox, John. Weisberg, Sanford. (2011). Cox Proportional-Hazards Regression for Survival Data in R. McMaster University.
URL: https://socserv.socsci.mcmaster.ca/jfox/Books/Companion/appendix/ Appendix-Cox-Regression.pdf

Hosmer, David W. Lemeshow, Stanley. May, Susanne. (2008). Applied survival analysis: regression modeling of time-to-event data. John Wiley & Sons, Inc.

Madigan, David. (2004). Introduction to Survival Analysis. Columbia University. URL: http://www.stat.columbia.edu/ madigan/W2025/notes/survival.pdf

Rodriguez, G. (2007). Lecture Notes on Generalized Linear Models. Chapter 7 Survival Models. URL: http://data.princeton.edu/wws509/notes/

Stevenson, Mark. (2009). An Introduction to Survival Analysis. EpiCentre, IVABS, Massey University.

Tableman, Mara. Kim, S. Jong. (2003). Survival Analysis Using S: Analysis of Time-to-Event Data. Chapman and Hall/CRC.

Zhang, Hui Hong. (2015). Checking proportionality for Coxs regression model. Faculty of Mathematics and Natural Sciences, University of Oslo.