Pseudo-observations in survival analysis

Master of Science Thesis

Autumn 2012/Spring 2013

Rikke Nørmark Mortensen

AALBORG UNIVERSITET

Department of Mathematical Sciences Aalborg University

Institut for Matematiske Fag
Fredrik Bajers Vej 7G
9220 Aalborg Øst
http://www.math.aau.dk

Title: Pseudo-observations in survival analysis

Project period: September 1st 2012 to June 3th 2013

Author:

Rikke Nørmark Mortensen

Supervisor : Poul Svante Eriksen

Copies: 4

Pages: 69

Deadline: June 3th 2013

Synopsis:

Survival data is a special type of data, which especially is characterised by the occurrence of censored data. The censoring mechanism means that standard statistically methods often are not suitable for analysing this type of data. Instead a range of statistical methods are developed to handle this type of data. A prominent role is played by the Cox proportional hazards model. This model is a regression model used to relate a set of covariates to the hazard function. In this project a new method for analysing survival data is studied. The method is based on pseudo-observations known from the jackknife theory. Pseudo-observations address one of the main problem with survival data, i.e. not having appropriate observations for all individual in the study population. The pseudo-observation approach allows for analysis of survival data by standard statistically methods. In this project the potential and efficiency of pseudo-observations for regression analysis of survival data is considered. Further, the method is compared to the traditional Cox proportional hazards model.

Dansk resume

Overlevelses data betegner en speciel type data som findes i mange empiriske studier, så som i epidemiologi, økonomiske studier og demografiske studier. Overlevelses data betegner en type data som måler tiden frem til en given hændelse. Variablen man er interesseret i er tiden frem til hændelses målt fra et givet begyndelses tidspunkt. Dette betyder at overlevelses data er opsamlet sekventielt, og dette har betydning for strukturen af dataet. Specielt vil denne opsamlings metode betyde at censureret hændelsestidspunkter kan forekomme for dele af studiepopulationen. En censurering betyder at den faktiske hændelsestid ikke er kendt. I stedet har man observeret en censureringstid som enten er større eller mindre end den faktiske hændelsestid. Den mest almindelige censureringsmekanisme er højre censurering, hvilket vil sige at tidspunktet for hændelses er større end det observerede censureringstidspunkt. Strukturen af overlevelses data betyder at traditionelle statistiske metoder ofte ikke kan anvendes på overlevelses data. Specielt er manglen på fuld information for dele af studiepopulationen et problem. En række statistiske modeller er udviklet specielt til at håndtere denne type data. En populær regressions model for overlevelses data er den proportionelle hazards Cox model. Dette er en semi-parametrisk regressionsmodel, brugt til at analysere effekten af en række kovariater på hazard funktionen.

I dette speciale er en ny metode til at analysere overlevelses data studeret. Denne metode bygger på pseudo-observationer kendt fra jackknife metoden. For hvert individ i studie populationen er en mænge af pseudo-observation udregnet, hvilket giver et fuldstændigt data sæt. Dette fuldstændige data sæt giver mulighed for at analysere overlevelses data med standard statistiske modeller, så som generaliseret lineær modeller. Dette giver mulighed for mere generelle analyser af overlevelses data end standard overlevelses modeller tillader. I dette speciale et regressions modeller baseret på pseudo-observationer sammenlignet med den traditionelle Cox model.

Preface

This Master of Science thesis is written by Rikke Nørmark Mortensen in fall 2012 and spring 2013 during the 9'th and 10'th semester of Mathematics at the Institute for Mathematical Sciences at Aalborg University.

The reader of this thesis is assumed to possess the mathematical qualifications corresponding to completion of the bachelor education of Mathematical Sciences at Aalborg University.

I will like to thank Poul Svante Eriksen for supervision throughout the study period of this project.

Reading instructions

References throughout the report will be presented according to the Harvard method. Mathematical definitions, figures, tables etc. are enumerated in reference to the chapter i.e. the first definition in chapter 2 has number 2.1, the second has number 2.2 etc. A reference to A.1 refer to the appendix.

Contents

1 Introduction **2**

2 Basic quantities **4**

2.1 Continuous random variables 4

2.2 Discrete random variables 9

3 Counting processes **10**

3.1 The Nelson-Aalen estimator 13

3.2 The Kaplan-Meier estimator 15

4 Semiparametric proportional hazards models **16**

4.1 The Cox proportional hazards model 17

4.1.1 The partial likelihood 18

4.1.2 The partial likelihood for distinct event times 19

4.1.3 The partial likelihood when ties are present 22

4.1.4 A discrete model analogue to the Cox proportional hazards model 24

4.1.5 Estimation of the hazard function and the survival function 25

5 Pseudo-observations **28**

5.1 Properties of the pseudo-observations 32

5.2 Regression models based on pseudo-observations 36

5.3 Pseudo-residuals 42

6 Regression splines **46**

7 The new pseudo-observation **52**

7.1 Properties of the new pseudo-observation 54

7.2 Regression analysis based on the new pseudo-observation 56

8 Discussion and conclusion **62**

9 Bibliography **66**

A Appendix **68**

A.1 The cumulative incidence function 68

1 Introduction

Survival data is a special type of data which arises in a number of applied settings such as medicine, biology, epidemiology, economics, and demography. The term survival data is used for data which measures the time to some event of interest. In the simplest case the event of interest is dead; however, the event of interest may also cover events like the onset of some disease or other complications. In this project, the terms event, dead, and failure will synonymously be used as the occurrence of some event of interest.

Survival data possess a number of features which makes it differ from other types of data. The main different lies in the way the survival data is measured. In other types of data, the responses are measured instantaneously and independent of the size of the response variable. In survival data the response variable is the *event time*, which is measured sequentially from the beginning of the study. This means that for survival data the large responses take longer time to measure than smaller responses. This way of measuring has a number of consequences, which must be dealt with when the survival data is analysed.

One consequence of the way survival data is measured is the occurring of *censored* data. The event time is said to be censored if the time of the event is not observed directly, but all that is known is that the event occurred either before or after some observed time, called the *censoring time*. The most common censoring mechanism is *right censoring*, which means that the event occurs at some time point after the observed censoring time. In some situations, right censoring occurs if an individual simply has not experienced the event before the termination time of the study. In other situations right censoring occurs if an individual leave the study before it is completed. If this early exit from the study is due to reasons unrelated to the event of interest, the right censoring is said to be *independt* of the event time. Less common censoring mechanisms include *left censoring* and *interval censoring*. An individual is said to be left censored if the event time is known to occur prior to the observed censoring time. Interval censoring means that the event is known to occur at some time within an interval. In this project, the term censored data will almost always refere to right censoring. Though, som eof the methods might be generalised to left censoring of interval censoring.

Due to the structure of survival data, statistically models for analysing this type of data have been developed as an independent area within statistics. Broadly the methods for analysing survival data can be divided into *non-parametric methods*, *parametric methods*, and *semi-parametric methods*. Non-parametric methods serve to draw inference about the event time distribution based on observed, possible censored data. Non-parametric methods include the *Kaplan-Meier estimator* and the *Nelson-Aalen estimator* introduced in chapter 3. These methods may be of interest on their own right or they may serve as a precursor to more detailed analysis. Often the observed data contain some additional information about each individual, which can be used as covariates in a regression analysis. Parametric methods may be used for maximum likelihood estimation of the unknown parameters in the regression model. However, from a practical point of view, the main objective when modelling survival data is to assess the effect of the covariates on the outcome of the regression model. This outcome is often given as a function well suited to describe the information in survival data, these functions are discussed in chapter 2. The event time distribution is often of secondary interest. This means that specifying a full parametric model may be a too strict assumption for the purpose of the analysis. More efficient results can often be obtained by a semi-parametric model. A semi-parametric model is a model in which the effect of the covariates is assumed to be parametric, but the effect of the time variable is given by some unspecified function.

A broad range of semi-parametric models have been developed for the purpose of analysing survival data. One of the most popular models is the *Cox proportional hazards model*. This model is a flexible model which is well suited in many applied settings. Inference on this model is based on a *partial likelihood* approach, in which the problem with censored data is handle by putting most emphasis on the observed event times.

In spite of the variety of models for analysis survival data, the occurrence of censored data put a restriction on the possibilities within analysis for this type of data. One example is graphical methods, such as residual plots, which is inconvenient due to censored event times. If no censoring occurs in the data, standard statistically models could be used to analyse the data. Standard statistically models often allows for more general analysis than the methods of survival data. In this project, *jackknife pseudo-observations* are considered as a tool for analysing survival data. Pseudo-observations address one of the main problems with survival data, i.e. not having appropriate responses for all individuals in the study. Hence, this approach is a step in the direction of analysing censored survival data by standard statistically methods. The approach were first suggested by Andersen et al. [2003] for performing generalised linear regression analysis of survival data. The method is based on a set of pseudo-observations defined for each individual in the study. The pseudo-observation approach is a must general method which may be applied in a number of applications. The aim of this project is to illustrate the potential and efficiency of this approach for regression analysis. Further, the potential of the methods for topics related to the Cox proportional hazards model is likewise considered.

2 Basic quantities

This chapter is written based on Klein and Moeschberger [1997] and Hosmer and Lemeshow [1999].

In this chapter some basic quantities used to describe the distribution of survival data is considered. Let X be a nonnegative random variable denoting the time to some event. The distribution of X can be described by the *cumulative distribution function* $F(t) = \mathrm{P}(X \leq t)$ and when X is a continuous random variable, also by the *density function* $f(t) = \frac{\mathrm{d}}{\mathrm{d}t}F(t)$. However, other functions are better suited to describe the distribution of time-to-event data. For survival data, one is often interested in the probability of surviving beyond the time t, this probability is given by the *survival function*. The *hazard function* describes the risk of an event in the next instant, given that the event has not occurred prior to the time t. These four functions all characterise different features of the distribution of X and given one of them, one can uniquely determine the others. Other parameters of interest for describing survival data are the *mean survival time* and the related *restricted mean survival time*, these two parameters are intuitive appealing in applied settings.

2.1 Continuous random variables

In this section, the basic functions for describing the survival distribution of non-negative continuous random variable X are given.

The survival function

The survival function is the most basic quantity to describe survival data. It gives the probability of observing an event beyond the time t.

Definition 2.1 *The survival function of a non-negative random variable X is given by:*

$$S(t) = P(X > t), \quad t \geq 0.$$

The survival function is related to the cumulative distribution function and the density function in the following way

$$\begin{aligned} S(t) &= 1 - F(t) \\ &= \int_t^\infty f(u)\mathrm{d}u. \end{aligned} \tag{2.1}$$

The density function can hence be written in terms of the survival function as

$$f(t) = -\frac{\mathrm{d}}{\mathrm{d}t}S(t).$$

Equation (2.1) implies that

$$S(0) = \int_0^\infty f(u)\mathrm{d}u = 1, \tag{2.2}$$

and further

$$\begin{aligned} S(\infty) &= \lim_{y\to\infty} S(y) \\ &= \lim_{y\to\infty} \int_y^\infty f(u)\mathrm{d}u \\ &= 0. \end{aligned} \tag{2.3}$$

From (2.2) and (2.3) it is seen that the survival function is a decreasing function taken values in the range of zero to one. Furthermore, equation (2.3) implies that an individual eventually is expected to experience the event if this just live long enough. In many contexts where survival data appear this assumption is reasonable, for instance if the event of interest is dead, which will occur for everybody in time. However, in some settings it is unlikely that all individual in the study eventually will experience the event, for instant if the event of interest is the time of start smoking, which may or may not occur to everybody.

Many different types of survival functions occur, however due to the above discussion, they all have the same basic shape. Figure 2.1 shows the survival curve from the Weibull distribution with scale parameter $\lambda = 1$ and three different values of the shape parameter α. The rate of change in the survival function indicates the risk of an event over time. In the figure it appears that subjects with a Weibull survival function with $\alpha = 1.1$ have a more favourable survival rate in the beginning of the study period, whereas, subjects with a Weibull survival function with $\alpha = 0.5$ have a more favourable survival rate at the ending of the study period. Though a distinctly comparison the three survival rates are difficult as the survival curves tends to coincide.

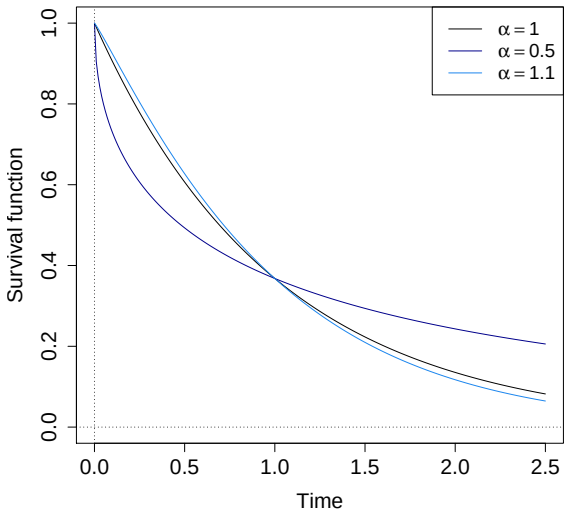

Figure 2.1: The Weibull survival functions with three different values of the shape parameter α.

The survival function gives the initial probability of an individual to survive from the time origin to a time point beyond the time t. Hence, the changes in the risk of an event with time are not captured in the survival function; another function which more properly describes this is the hazard function.

The hazard function

The hazard function describes the instantaneous risk of an event at time t, given that the event has not occurred prior to time t. That is, the hazard function gives conditional information on how the risk of an event changes with time.

Definition 2.2 *The hazard function for a random variable X is defined as*

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq X < t + \Delta t | X \geq t)}{\Delta t}$$

By the definition of $h(\cdot)$, the quantity $h(t)\Delta t$ may be considered as an approximate conditional probability of an event in the interval $[t, t + \Delta t)$. However, the hazard function itself is not a probability, but may rather be considered as the rate for which the risk of an event changes with time. The values of the hazard function can vary between zero and infinity, and the shape of $h(\cdot)$ can possess many different forms, reflecting the changes in the risk of an event with time.

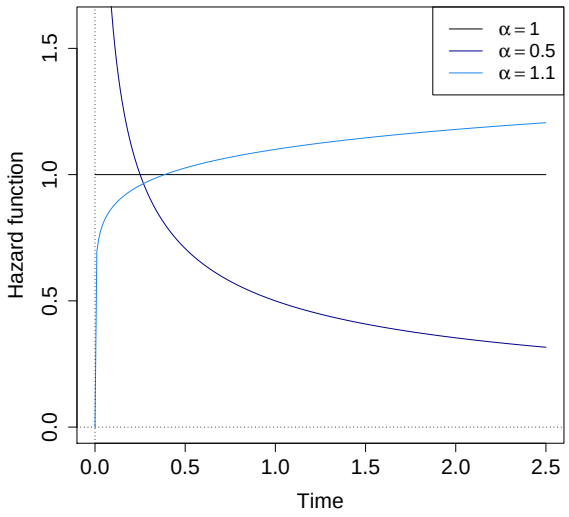

Figure 2.2: The hazard function from the Weibull distribution for three different values of the shape parameter α.

Figure 2.2 shows the three Weibull hazard functions corresponding to the survival functions in figure 2.1. In the figure it is seen that the tendency from the survival curves in figure 2.1 is much more clear from the hazard function. Individuals with Weibull hazard function with $\alpha = 1.1$ has an increased risk of the event over time, whereas, individuals with Weibull hazard function with $\alpha = 0.5$ have a decreasing risk of event over time. Hence the survival function and the hazard function differently describes the same aspects of the data. Though, the hazard function is often more informative about the event pattern in the data than the survival function, as the hazard function gives conditional information on the events.

The density function of X is defined as the derivative of the cumulative distribution function, that is

$$\begin{aligned} f(t) &= \frac{\mathrm{d}}{\mathrm{d}t}F(t) \\ &= \lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{\mathrm{P}(t \leq X < t+\Delta t)}{\Delta t}. \end{aligned}$$

From this it is seen that the hazard function can be written as

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{\mathrm{P}(t \leq X < t+\Delta t)}{\Delta t \cdot \mathrm{P}(X \geq t)} \\ &= \frac{f(t)}{S(t)} \qquad (2.4) \\ &= -\frac{\mathrm{d}}{\mathrm{d}t} \ln\left[S(t)\right]. \qquad (2.5) \end{aligned}$$

A function related to the hazard function is the cumulative hazard function.

Definition 2.3 *Let X be a random variable with hazard function $h(\cdot)$, the cumulative hazard function of X is defined by*

$$H(t) = \int_0^t h(u)\mathrm{d}u.$$

From a practical point of view, the hazard function $h(\cdot)$ is often of main interest, as this function is more intuitive clear. However, the cumulative hazard function $H(\cdot)$ is often easier to estimate from a given data set.

By equation (2.5) and the fundamental theorem of calculus, the cumulative hazard function and the survival function is related in the following way

$$H(t) = -\ln\left[S(t)\right], \tag{2.6}$$

and hence

$$S(t) = \exp\left[-H(t)\right] \tag{2.7}$$

$$= \exp\left[-\int_0^t h(u)\mathrm{d}u\right]. \tag{2.8}$$

Combining equation (2.4) with equation (2.8) one gets that the density function can be written in terms of the hazard function as

$$f(t) = h(t)\exp\left[-\int_0^t h(u)\mathrm{d}u\right]. \tag{2.9}$$

The mean survival time

Another parameter of interest is the *mean survival time* $\mu = \mathbb{E}[X]$, which may be written as

$$\mu = \int_0^\infty S(t)\mathrm{d}t,$$

where $S(\cdot)$ is the survival function of X. This parameter is intuitive appealing as it gives the expected lifetime for an individual with survival function $S(\cdot)$.

When analysing survival data right censoring often occur. This means that the tail of the survival time distribution may be difficult to estimate and hence an estimate of μ may be heavily biased. A parameter related to the mean survival time is the restricted mean survival time.

Definition 2.4 *The restricted mean survival time for a random variable X is defined by*

$$\mu_\tau = \mathbb{E}[\min(X, \tau)],$$

for $\tau > 0$.

This parameter gives the expected lifetime over the interval $[0, \tau]$. In similarity with the mean survival time, the restricted mean may be written as

$$\mu_\tau = \int_0^\tau S(t)\mathrm{d}t.$$

The restricted mean survival time is less sensitive to the occurrence of right censoring in a given sample than the overall mean μ.

2.2 Discrete random variables

In this section, the discrete analogue to the functions described in the previous section is given. Let X be a discrete random variables taking values $t_0 < t_1 < t_2 < \cdots$ and let $p(\cdot)$ denote the probability function

$$p(t_i) = \mathrm{P}(X = t_i), \quad i = 0, 1, 2, \ldots$$

The discrete survival function is then given by

$$S(t) = \sum_{t_i > t} p(t_i). \tag{2.10}$$

The hazard function at time t_i is given by the conditional probability of failing at time t_i, given survival until time t_i, that is

$$h(t_i) = \mathrm{P}(X = t_i | X \geq t_i) = \frac{p(t_i)}{S(t_{i-1})},$$

whith the convenience that $S(t_0) = 1$. Note that $p(t_i) = S(t_{i-1}) - S(t_i)$ from which it follows that

$$h(t_i) = 1 - \frac{S(t_i)}{S(t_{i-1})}.$$

The discrete survival function (2.10) may be written as a product of conditional survival functions

$$S(t) = \prod_{t_i \leq t} \frac{S(t_i)}{S(t_{i-1})},$$

and it follows that the survival function may be written in terms of the hazard function

$$S(t) = \prod_{t_i \leq t} \left[1 - h(t_i)\right].$$

3 Counting processes

This chapter is written based on Fleming and Harrington [1991], Andersen et al. [1993], and Klein and Moeschberger [1997].

Survival data consist of observations gathered over a period of time, and it is natural to model this type of data as a stochastic process. *Counting process* methods provide exact ways for studying incomplete data. In this chapter an introduction to counting process theory is given. The main object of this chapter is to define non-parametric estimators of the cumulative hazard function and the survival function.

Definition 3.1 *A stochastic process $N(t)$, $t \geq 0$ is called a counting process if it fulfils the following properties: $N(0) = 0$; $N(t) < \infty$ a.s. and the sample paths of $N(t)$ are with probability one right-continuous and piecewise constant with jump of size $+1$.*

Suppose a right censored sample with n individuals is given. Let $T_j = \min(X_j, C_j)$ be the study time for individual $j = 1, \ldots, n$ and let $\delta_j = \mathbb{1}[X_j \leq C_j]$. Here the event time X_j and the censoring time C_j are assumed to be independent, continuous random variables. The process $N_j(t) = \mathbb{1}[T_j \leq t, \delta_j = 1]$ is then a counting process defined for each individual j. Summing over $N_j(t)$ one gets a counting process

$$N(t) = \sum_{j=1}^{n} N_j(t), \tag{3.1}$$

which counts the number of events occurring prior to and including the time t.

The focus in this chapter is restricted to counting processes defined for right censored samples, though the theory may be applied in a more general setting. For a given right censored sample, the sample paths of the counting process $N(\cdot)$ given in (3.1) describes the times of events. Further, the difference $N(t) - N(s)$ is the number of events in the interval

$(s, t]$. However, at a given time t additional information on the sample may available, such as knowledge on censoring prior to t. The *history* of a counting process at time t is the accumulated knowledge of the sample prior to and including the time t, and is denoted $\boldsymbol{F}_t$. The history is assumed to be increasing, that is $\boldsymbol{F}_s \subseteq \boldsymbol{F}_t$, for all $s < t$.

Let $N(\cdot)$ be a given counting process and let t^- denote the time just prior to but not including the time t, then the quantity $\mathrm{d}N(t)$ is defined as

$$\mathrm{d}N(t) = N\left([t + \mathrm{d}t]^-\right) - N\left(t^-\right) \quad \mathrm{d}t > 0.$$

That is, $\mathrm{d}N(\cdot)$ is the change in the counting process over the interval $[t, t + \mathrm{d}t)$. If $\mathrm{d}t$ is sufficiently small then the quantity $\mathrm{d}N(t)$ is a zero-one random variable, meaning that either an event occur in the interval $[t, t + \mathrm{d}t)$ or no event occur in the interval.

Definition 3.2 *The intensity process $\lambda(\cdot)$ of a counting process $N(\cdot)$ is defined as*

$$\lambda(t) = \lim_{\mathrm{d}t \to 0} \frac{P\left(\mathrm{d}N(t) = 1 | \boldsymbol{F}_{t^-}\right)}{\mathrm{d}t}.$$

For a right censored sample, the probability of individual j failing in a small interval $[t, t + \mathrm{d}t)$ is given by

$$\mathrm{P}\left(t \leq T_j < t + \mathrm{d}t, \delta_j = 1 | \boldsymbol{F}_{t^-}\right), \quad j = 1, \ldots, n. \tag{3.2}$$

For $T_j < t$ the probability in (3.2) is obviously equal zero. For $T_j \geq t$, the probability is given by

$$\begin{aligned}
\mathrm{P}\left(t \leq T_j < t + \mathrm{d}t, \delta_j = 1 | \boldsymbol{F}_{t^-}\right) &= \mathrm{P}\left(t \leq X_j < t + \mathrm{d}t, C_j \geq t + \mathrm{d}t | X_j \geq t, C_j \geq t\right) \\
&= \mathrm{P}\left(t \leq X_j < t + \mathrm{d}t | X_j \geq t\right) \cdot \mathrm{P}\left(C_j \geq t + \mathrm{d}t | C_j \geq t\right) \\
&= \frac{[F(t + \mathrm{d}t) - F(t)]}{S(t)} \cdot \mathrm{P}\left(C_j \geq t + \mathrm{d}t | C_j \geq t\right) \\
&= \frac{f(t)\mathrm{d}t}{S(t)} \cdot \mathrm{P}\left(C_j \geq t + \mathrm{d}t | C_j \geq t\right) \\
&\approx h(t)\mathrm{d}t,
\end{aligned}$$

for $\mathrm{d}t$ sufficiently small. Her $h(\cdot)$ is the hazard function given in definition 2.2

Let $Y(t)$ be *the risk set* at time t, that is $Y(\cdot)$ is process describing the number of individuals at risk at some given time;

$$Y(t) = \sum_{j=1}^{n} \mathbb{1}[T_j \geq t].$$

Then for $\mathrm{d}t$ sufficiently small

$$\begin{aligned}\mathrm{P}\left(\mathrm{d}N(t)=1|\boldsymbol{F}_{t^-}\right) &= \mathbb{E}\left[\mathrm{d}N(t)|\boldsymbol{F}_{t^-}\right] \\ &= \mathbb{E}\left[\#\{j \ : \ T_j \in [t, t+\mathrm{d}t), \delta_j = 1\}\big|\boldsymbol{F}_{t^-}\right] \\ &= Y(t)h(t)\mathrm{d}t,\end{aligned}$$

where $N(\cdot)$ is the counting process in (3.1). Hence, for a right censored sample the intensity function is given by $\lambda(t) = Y(t)h(t)$.

Definition 3.3 *Let $N(\cdot)$ be a counting process with intensity function $\lambda(\cdot)$, the cumulative intensity process is then defined by*

$$\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u, \qquad t \geq 0.$$

The general theory of *martingales* is a concept which arises naturally in the context of counting processes. A martingale is a stochastic process with the property that the expected value of the next observation from the process given the history is equal to the present observation. A stochastic process $M(\cdot)$ is called a *martingale* if it fulfils

$$\mathbb{E}[M(t)|\boldsymbol{F}_s] = M(s), \quad \text{for all} \quad s \leq t. \tag{3.3}$$

Definition 3.4 *Let $N(\cdot)$ be a counting process with cumulative intensity function $\Lambda(\cdot)$, then the counting process martingale is defined as*

$$M(t) = N(t) - \Lambda(t).$$

The counting process martingale has the property that

$$\begin{aligned}\mathbb{E}[\mathrm{d}M(t)|\boldsymbol{F}_{t^-}] &= \mathbb{E}[\mathrm{d}N(t) - \mathrm{d}\Lambda(t)|\boldsymbol{F}_{t^-}] \\ &= \mathbb{E}[\mathrm{d}N(t)|\boldsymbol{F}_{t^-}] - \mathbb{E}[\lambda(t)|\boldsymbol{F}_{t^-}] \\ &= 0,\end{aligned} \tag{3.4}$$

where the last equality follows because $\lambda(t)$ has a fixed value given $\boldsymbol{F}_{t^-}$. The property in (3.4) is equivalent with the martingale property in (3.3), this can be seen by

$$\begin{aligned}\mathbb{E}[M(t)|\boldsymbol{F}_s] - M(s) &= \mathbb{E}[M(t) - M(s)|\boldsymbol{F}_s] \\ &= \mathbb{E}\left[\int_s^t \mathrm{d}M(u)\big|\boldsymbol{F}_s\right] \\ &= \int_s^t \mathbb{E}\left[\mathrm{d}M(u)|\boldsymbol{F}_s\right] \\ &= \int_s^t \mathbb{E}\left[\mathbb{E}[\mathrm{d}M(u)|\boldsymbol{F}_{u^-}]\big|\boldsymbol{F}_s\right] \\ &= 0.\end{aligned} \tag{3.5}$$

This means that the counting process martingale $M(\cdot)$ given in definition 3.4 is indeed a martingale.

In the general theory of martingales, a process $\tilde{X}(\cdot)$ is called the *compensator* of a given process $X(\cdot)$ if the process $X(\cdot)-\tilde{X}(\cdot)$ is a martingale. From the definition of $M(\cdot)$ in definition 3.4 it is seen that $\Lambda(\cdot)$ is the compensator of the counting process $N(\cdot)$. The counting process martingale can be considered as a zero mean noise which arises when the compensator $\Lambda(\cdot)$ is subtracted from the counting process $N(\cdot)$.

Another quantity related to counting processes is the *predictable variation process* of the counting process martingale. This quantity is defined as the compensator of the process $M^2(\cdot)$ and denoted $\langle M\rangle(\cdot)$. Consider the increment of the process $M^2(\cdot)$

$$\begin{aligned}\mathrm{d}M^2(t) &= M^2\left([t+\mathrm{d}t]^-\right) - M^2(t^-)\\ &= \left(M(t^-)+\mathrm{d}M(t)\right)^2 - M^2(t^-)\\ &= \left(\mathrm{d}M(t)\right)^2 + 2M(t^-)\mathrm{d}M(t).\end{aligned}$$

Since

$$\mathbb{E}\left[2M(t^-)\mathrm{d}M(t)|\boldsymbol{F}_{t^-}\right] = 2M(t^-)\mathbb{E}\left[\mathrm{d}M(t)|\boldsymbol{F}_{t^-}\right] = 0,$$

it follows that

$$\mathbb{E}\left[\mathrm{d}M^2(t)|\boldsymbol{F}_{t^-}\right] = \mathbb{E}\left[\left(\mathrm{d}M(t)\right)^2\big|\boldsymbol{F}_{t^-}\right].$$

Hence the increment of the compensator of $M^2(\cdot)$ is equal to the conditional variance of the increment of $M(\cdot)$, that is

$$\mathrm{d}\langle M\rangle(t) = \mathbb{E}\left[\left(\mathrm{d}M(t)\right)^2\big|\boldsymbol{F}_{t^-}\right] = \mathrm{Var}[\mathrm{d}M(t)|\boldsymbol{F}_{t^-}].$$

Note that the process $M^2(\cdot) - \langle M\rangle(\cdot)$ is a martingale by similar arguments as in (3.5). As noted previously, if the interval $[t+\mathrm{d}t)$ is sufficiently small, then $\mathrm{d}N(t)$ is a zero-one random variable, the variance of $\mathrm{d}M(t)$ is then given by

$$\begin{aligned}\mathrm{Var}[\mathrm{d}M(t)|\boldsymbol{F}_{t^-}] &= \mathrm{Var}[\mathrm{d}N(t) - \mathrm{d}\Lambda(t)|\boldsymbol{F}_{t^-}]\\ &= \mathrm{Var}\left[\mathrm{d}N(t) - \mathbb{E}[\mathrm{d}N(t)|\boldsymbol{F}_{t^-}]|\boldsymbol{F}_{t^-}\right]\\ &= \mathrm{Var}\left[\mathrm{d}N(t)|\boldsymbol{F}_{t^-}\right]\\ &= \mathrm{d}\Lambda(t)(1-\mathrm{d}\Lambda(t))\\ &\approx \mathrm{d}\Lambda(t).\end{aligned}$$

3.1 The Nelson-Aalen estimator

The theory of counting processes allows a relatively simple derivation of quantities based on censored data. One of these quantities is the *Nelson-Aalen estimator* of the cumulative hazard function $H(\cdot)$, given in definition 2.3.

Suppose a right censored sample is given, then the increment of $N(\cdot)$ given in (3.1) can be written as

$$\mathrm{d}N(t) = Y(t)h(t)\mathrm{d}t + \mathrm{d}M(t),$$

where $M(\cdot)$ is the counting process martingale corresponding to $N(\cdot)$.

Assuming $Y(t) > 0$, then this can be rewritten as

$$\frac{\mathrm{d}N(t)}{Y(t)} = h(t)\mathrm{d}t + \frac{\mathrm{d}M(t)}{Y(t)}. \tag{3.6}$$

The process $Y(\cdot)$ is predictable, meaning that it is fixed given the history just prior to time t. From equation (3.4) it follows that

$$\mathbb{E}\left[\frac{\mathrm{d}M(t)}{Y(t)}\Big|\boldsymbol{F}_{t^-}\right] = \frac{\mathbb{E}[\mathrm{d}M(t)|\boldsymbol{F}_{t^-}]}{Y(t)} = 0.$$

This means that the last term on the right-hand side of equation (3.6) can be considered as a zero-mean noise given the history. The variance of this noise is given by

$$\mathrm{Var}\left[\frac{\mathrm{d}M(t)}{Y(t)}\Big|\boldsymbol{F}_{t^-}\right] = \frac{\mathrm{Var}[\mathrm{d}M(t)|\boldsymbol{F}_{t^-}]}{Y(t)^2} = \frac{\mathrm{d}\langle M\rangle(t)}{Y(t)^2}.$$

Let $J(t) = \mathbb{1}[Y(t) > 0]$ and define $0/0 = 0$, then integrating on both side of equation (3.6) gives

$$\int_0^t \frac{J(u)}{Y(t)}\mathrm{d}N(t) = \int_0^t J(u)h(u)\mathrm{d}u + \int_0^t \frac{J(u)}{Y(u)}\mathrm{d}M(u). \tag{3.7}$$

The integral

$$\hat{H}(t) = \int_0^t \frac{J(u)}{Y(t)}\mathrm{d}N(t) \tag{3.8}$$

is the Nelson-Aalen estimator of the cumulative hazard function $H(\cdot)$. This estimator is essentially the sum over event times up to and including time t, relative to the corresponding number of individuals at risk. For right-censored data, the integral

$$\int_0^t J(u)h(u)\mathrm{d}u$$

is the actual cumulative hazard function $H(\cdot)$, omitting the contributions of $h(\cdot)$ when the risk set is empty. The last integral on the right hand-side of equation (3.7) is a stochastic integral with respect to a martingale, and hence the integral itself is a martingale Andersen et al. [1993]. This integral can be considered as the statistically uncertainty in the Nelson-Aalen estimator.

3.2 The Kaplan-Meier estimator

An estimator of the survival function given in definition 2.1 is the *Kaplan-Meier estimator*, which may be derived from the Nelson-Aalen estimator. This is based on the discrete representation of the survival function and the Nelson-Aalen estimator. For a discrete random variable the survival function is given by $S(t) = \prod_{t_j \leq t}(1-h(t_j))$. The Kaplan-Meier estimator is then given by

$$\hat{S}(t) = \prod_{T_j \leq t} \left[1 - \mathrm{d}\hat{H}(T_j)\right],$$

where $\hat{H}(\cdot)$ is the Nelson-Aalen estimator.

The Kaplan-Meier estiamtor is a wildly used non-parametric estimator of the survival function. It may be thought of as an conditional survival function, resulting from a partitioning of the time scale and estimating the survival function on each partitioning. If no censoring occurs in the data, the Kaplan-Meier estimator reduces to one minus the empirical distribution function. It can be shown that the Kaplan-Meier estimator has non-negative bias, which converges to zero at an exponential rate for n approaching infinity Fleming and Harrington [1991].

An estimate of the variance of the Kaplan-Meier estimator is given by the Greenwood estimate

$$\hat{\mathrm{var}}[\hat{S}(t)] = \left(\hat{S}(t)\right)^2 \int_0^t \frac{\mathrm{d}N(s)}{Y(s)\left[Y(s) - \Delta N(s)\right]},$$

where $N(\cdot)$ is the counting process in (3.1) and $\Delta N(s) = N(s) - N(s^-)$.

4

Semiparametric proportional hazards models

Unless otherwise stated this chapter written based on Klein and Moeschberger [1997] and Kalbfleisch and Prentice [2002].

In this chapter regression methods for survival data is considered. As noted in chapter 2 functions like the hazard function and the survival function are well suited for describing the information of interest in survival data. Hence, regression models for survival data is cantered on these functions and especially the hazard function is frequently used as the outcome in the models.

A class of models used to analyse the effect of a set of covariates on the survival probability is the family of *semiparametric hazard models*. Suppose a right censored sample of n independent and identical distributed (i.i.d.) individuals are given. The data then consist of the triplets $(T_j, \delta_j, \boldsymbol{Z}_j)$, $j = 1, \ldots, n$, where $T_j = \min(X_j, C_j)$, $\delta_j = \mathbb{1}[X_j \leq C_j]$, and $\boldsymbol{Z}_j = [Z_{j1}, \ldots, Z_{jp}]^\top$ is a vector of covariates. In the following assume furthermore that the event times X_j and the censoring times C_j are continuous independent random variables. Let $h(t|\boldsymbol{Z}_j)$ denote the conditional hazard function for an individual with covariates $\boldsymbol{Z}_j$. The class of semiparametric hazard models is given by the family of functions which can be written as

$$h\left(t|\boldsymbol{Z}_j\right) = h_0(t)c\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right), \quad j = 1, \ldots, n. \tag{4.1}$$

Here $h_0(\cdot)$ is an unspecified non-negative function called the *baseline hazard function*, $c(\cdot)$ is a non-negative function of the covariates called the *link function*, and $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^\top$ is a vector of unknown parameters. Note that the distribution corresponding to the baseline hazard function is left unspecified.

One or more components of the covariate vector $\boldsymbol{Z}_j$ may depend on the time t, in which case the covariates are said to be *time-dependent*. If the value of the covariate vector is constant over time, then $\boldsymbol{Z}_j$ is said to be a *fixed-covariate* vector. In the rest of this project it is assumed that the covariates do not depend on time, and the case of time-depend covariates will not be comment further.

4.1 The Cox proportional hazards model

One important case of the model (4.1) is the *Cox proportional hazards model*, where the link function $c(\cdot)$ is given by

$$c\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right) = \exp\left[\sum_{l=1}^{p} \beta_l Z_{jl}\right] = \exp\left[\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right].$$

The model (4.1) then becomes

$$h\left(t|\boldsymbol{Z}_j\right) = h_0(t) \exp\left[\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right], \tag{4.2}$$

which can be rewritten as

$$\log\left[\frac{h(t|\boldsymbol{Z}_j)}{h_0}\right] = \boldsymbol{\beta}^\top \boldsymbol{Z}_j.$$

Hence the Cox proportional hazards model may be considered as a linear model in the logarithm of the ratio $\frac{h_j(t|\boldsymbol{Z}_j)}{h_0(t)}$. Each β_h is then interpreted as the change in the log of the ratio $\frac{h_j(t|\boldsymbol{Z}_j)}{h_0}$ per unit change in Z_h, assuming all other covariates constant. Note that the baseline hazard function $h_0(\cdot)$ may be regarded as the conditional hazard function for an individual with all covariates constant equal zero.

A key feature of the Cox proportional hazards model is that the hazard function of two individuals with covariates $\boldsymbol{Z'}$ and $\boldsymbol{Z''}$ respectively are proportional, this is seen by

$$\frac{h(t|\boldsymbol{Z'})}{h(t|\boldsymbol{Z''})} = \frac{h_0(t) \exp\left[\boldsymbol{\beta}^\top \boldsymbol{Z'}\right]}{h_0(t) \exp\left[\boldsymbol{\beta}^\top \boldsymbol{Z''}\right]} = \exp\left[\boldsymbol{\beta}^\top \left(\boldsymbol{Z'} - \boldsymbol{Z''}\right)\right], \tag{4.3}$$

which is constant over time. The ratio in (4.3) is called the *relative risk* (hazard ratio) of an event for an individual with covariates $\boldsymbol{Z'}$ compared to an individual with covariates $\boldsymbol{Z''}$.

Note that the relative risk in (4.3) may be written as

$$\frac{h(t|\boldsymbol{Z'})}{h(t|\boldsymbol{Z''})} = \exp\left[\beta_h(Z'_h - Z''_h)\right] \exp\left[\sum_{l \neq h} \beta_l (Z'_l - Z''_l)\right].$$

This means that when Z'_h and Z''_h differ by a unit and $Z'_l = Z''_l$, $l \neq h$, the exponential of each β_h is the relative risk of an event.

From the relations given in chaper 2, it follows that the conditional density function and the conditional survival function corresponding to the Cox proportional hazards model are given by

$$f\left(t|\boldsymbol{Z}\right) = \lambda_0(t) \exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}\right) \exp\left[-\exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}\right) \int_0^t \lambda_0(u)\mathrm{d}u\right]$$

$$S(t|\boldsymbol{Z}) = \exp\left[-\exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}\right) \int_0^t \lambda_0(u)\mathrm{d}u\right].$$

Inference on the unknown parameters in ordinary generalised linear models is often based on the method of maximum likelihood estimation (MLE). However, for semiparametric models it is not possible to specify an explicit expression of the full likelihood function, as the distribution corresponding to the baseline hazard function is not specified. Instead the estimation of the unknown parameters $\boldsymbol{\beta}$ are obtained by maximisation of the *partial likelihood*. Before any further discussion on estimation of the unknown parameters in the model (4.2) some notion on the partial likelihood in a general setting will be introduced.

4.1.1 The partial likelihood

This section is written based on Cox [1975].

Suppose a given data set consists of sampled values from a random vector $\boldsymbol{Y}$ with density function $f(y, \boldsymbol{\theta}, \boldsymbol{\beta})$. The parameter $\boldsymbol{\theta}$ is considered to be a nuisance parameter and $\boldsymbol{\beta}$ is the parameter of primary interest. Suppose that a one-to-one transformation exist between $\boldsymbol{Y}$ and a set of random variables $A_1, B_1, \ldots, A_k, B_k$. Let $A^{(i)} = (A_1, \ldots, A_i)$ and $B^{(i)} = (B_1, \ldots, B_i)$, $i = 1, \ldots, k$ and assume that the joint density of $A^{(k)}$ and $B^{(k)}$ can be written as

$$\prod_{i=1}^{k} f(b_i|b^{(i-1)}, a^{(i-1)}, \boldsymbol{\theta}, \boldsymbol{\beta}) \prod_{i=1}^{k} f(a_i|b^{(i)}, a^{(i-1)}, \boldsymbol{\beta}). \tag{4.4}$$

The information on $\boldsymbol{\beta}$ based on the first term of (4.4) may be inextricably linked with the nuisance parameter $\boldsymbol{\theta}$. In this case inference on $\boldsymbol{\beta}$ may be based solely on the second term of (4.4) called the *partial likelihood* of $\boldsymbol{\beta}$ based on $\{A_i\}$ in the sequence $\{A_i, B_i\}$.

The partial likelihood of $\boldsymbol{\beta}$ is hence given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} f(a_i|b^{(i)}, a^{(i-1)}, \boldsymbol{\beta}). \tag{4.5}$$

The partial likelihood (4.5) is not a likelihood in the traditional sense. However, as seen in the following, the partial likelihood function possesses some of the same properties as an ordinary likelihood function and hence the partial likelihood function may be treated as an ordinary likelihood.

Let $H_i = (B^{(i)}, A^{(i-1)})$ and consider the *score components* of (4.5) given by

$$U_i = \frac{\partial}{\partial \boldsymbol{\beta}} \ln \left[f(A_i|H_i, \boldsymbol{\beta}) \right] = \left[\frac{\partial}{\partial \beta_l} \ln \left[f(A_i|H_i, \boldsymbol{\beta}) \right] \right]_{1 \times p} \quad i = 1, \ldots, k.$$

Assume that the usual regular conditions hold for the conditional density function of A_i given $H_i = h_i$, then it follows that $\mathbb{E}[U_i|H_i = h_i] = \mathbf{0}$ and hence

$$\mathbb{E}\left[U_i\right] = \mathbb{E}\left[\mathbb{E}[U_i|H_i]\right] = \mathbf{0}. \tag{4.6}$$

The total *score function* for the partial likelihood is given by

$$U(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \ln[L(\boldsymbol{\beta})] = \sum_{i=1}^{k} U_i.$$

From the property in (4.6) it follows that $\mathbb{E}[U(\boldsymbol{\beta})] = \mathbf{0}$. Furthermore, suppose that $H_j = h_j$ is given then for all $i < j$ it follows that U_i is fixed and this implies that

$$\mathbb{E}\left[U_i U_j^\top\right] = \mathbb{E}\left[\mathbb{E}\left[U_i U_j^\top | H_j\right]\right] = \mathbb{E}\left[U_i \mathbb{E}\left[U_j^\top | H_j\right]\right] = 0.$$

Hence, the score components $U_1, \ldots, U_k$ have zero mean and are uncorrelated. The variance of U_i is given by

$$\begin{aligned} \text{Var}\left[U_i\right] &= \mathbb{E}\left[U_i U_i^\top\right] \\ &= -\mathbb{E}\left[\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \ln\left[f(A_i | H_i, \boldsymbol{\beta})\right]\right]. \end{aligned}$$

Let

$$\begin{aligned} I_i &= -\mathbb{E}\left[\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \ln\left[f(A_i | H_i, \boldsymbol{\beta})\right]\right] \\ &= -\left[\mathbb{E}\frac{\partial^2}{\partial \beta_g \partial \beta_h} \ln\left[f(A_i | H_i, \boldsymbol{\beta})\right]\right]_{p \times p}. \end{aligned}$$

It then follows that the total score $U(\boldsymbol{\beta})$ has zero mean and covariance matrix given by the *expected Fisher information*

$$I(\boldsymbol{\beta}) = -\mathbb{E}\left[\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \ln\left[L(\boldsymbol{\beta})\right]\right] = \sum_{i=1}^{k} I_i.$$

Under the conditions that the U_i, $i = l, \ldots, k$ possess some degree of independence and the variances I_i are not too disparate, one may apply the central limit theorem to $U(\boldsymbol{\beta})$ as $k \to \infty$. This implies that the distribution of $U(\boldsymbol{\beta})$ approach a normal distribution with mean zero and covariance matrix $I(\boldsymbol{\beta})$.

4.1.2 The partial likelihood for distinct event times

In this section the partial likelihood for the Cox proportional hazards model in (4.2) will be discussed. The partial likelihood function for this model was first proposed by Cox [1972]. The basic notion used in the construction of the partial likelihood for this model is that the time interval between two successive event times provides no information on the effect of the covariates on the hazard of event. Furthermore, the construction of this partial likelihood assumes that the survival times are measured in continuous time eliminating the possibility of ties in the data. In practice, however, data is often sampled at discrete time which may result in several individuals having the same survival time. Hence modifications of the partial

likelihood are needed in order to account for possible ties in the data. In this section the case of tied data will be ignored; approximation of the partial likelihood for tied data will be discussed in the next section. As discussed in the previous section, the partial likelihood is constructed by considering only the part of (4.2) which solely depends on $\boldsymbol{\beta}$; in this case the baseline hazard function $h_0(\cdot)$ is considered as a nuisance parameter.

Suppose a given right censored sample consist of n individuals, from which k individuals have an observed event time and $n-k$ individuals have a censored event time. Let $t_1 < t_2 < \cdots < t_k$ denote the ordered observed event times, and let R_i denote the *risk set* at time t_i, $i = 1, \ldots, k$. That is R_i is a set consisting of all the individuals with event or censoring time greater than or equal to t_i. Hence, R_i is given by

$$R_i = \{j \ : T_j \geq t_i\}.$$

The partial likelihood function is constructed by considering the probability of individual i failing in the small interval $[t_i, t_i + \mathrm{d}t_i)$ conditioned on the risk set. That is, the probability of the individual failing in the interval $[t_i, t_i + \mathrm{d}t_i)$ is the individual actually observed failing, conditioned on one individual from the risk set fails in the interval $[t_i, t_i + \mathrm{d}t_i)$. In the notation of section 4.1.1, let B_i contain information on censoring in the interval $[t_{i-1}, t_i)$ and the information that one individual from the risk set fails in the small interval $[t_i, t_i + \mathrm{d}t_i)$. Let A_i be the information that subject i fails in the interval $[t_i, t_i + \mathrm{d}t_i)$.

The ith term in the partial likelihood (4.5) is given by

$$L_i(\boldsymbol{\beta}) = f\left(a_i | b^{(i)}, a^{(i-1)}, \boldsymbol{\beta}\right). \tag{4.7}$$

The conditioning on $b^{(i)}, a^{(i-1)}$ gives information on all censoring and failure times prior to the time t_i, and the information that an event occur in the interval $[t_i, t_i + \mathrm{d}t_i)$. For $\mathrm{d}t_i$ sufficient small the term (4.7) then becomes

$$\begin{aligned} L_i(\boldsymbol{\beta}) &= \frac{h\left(t_i | \boldsymbol{Z}_i\right) \mathrm{d}t_i}{\sum_{j \in R_i} h\left(t_i | \boldsymbol{Z}_j\right) \mathrm{d}t_i} \\ &= \frac{h_0(t_i) \exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_i\right)}{\sum_{j \in R_i} h_0(t_i) \exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right)} \\ &= \frac{\exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_i\right)}{\sum_{j \in R_i} \exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right)}. \end{aligned} \tag{4.8}$$

Then each individual with an observed event time contribute to the partial likelihood function with the probability given in (4.8), so that

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_i\right)}{\sum_{j \in R_i} \exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right)}. \tag{4.9}$$

Note that the partial likelihood depends on the observed event times t_i through the order of them, not on the actual value of them.

The estimates of the parameters $\boldsymbol{\beta}$ are found by maximising the partial likelihood function, which is equivalent to maximisation of the logarithm of the partial likelihood. The *log-partial likelihood* is given by

$$\begin{aligned} l(\boldsymbol{\beta}) &= \log\left[L(\boldsymbol{\beta})\right] \\ &= \sum_{i=1}^{k} \boldsymbol{\beta}^\top \boldsymbol{Z}_i - \sum_{i=1}^{k} \log\left[\sum_{j \in R_i} \exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right)\right]. \end{aligned} \tag{4.10}$$

The maximum of (4.10) can then be obtained by solving the p equations $U(\boldsymbol{\beta}) = \mathbf{0}$, where the score function is given by

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{k} \left(\boldsymbol{Z}_i - \mathcal{E}_i(\boldsymbol{\beta})\right), \tag{4.11}$$

for

$$\mathcal{E}_i(\boldsymbol{\beta}) = \sum_{j \in R_i} \frac{\boldsymbol{Z}_j \exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right)}{\sum_{l \in R_i} \exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_l\right)}.$$

This means that $\mathcal{E}_i(\boldsymbol{\beta})$ is the expectation of $\boldsymbol{Z}_i$ with respect to the distribution

$$p_i(\boldsymbol{\beta}) = \frac{\exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right)}{\sum_{l \in R_i} \exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_l\right)},$$

on the risk set R_i at time t_i.

The *observed Fisher information* is the matrix given by

$$\mathcal{I}(\boldsymbol{\beta}) = -\left[\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} l(\boldsymbol{\beta})\right]_{p \times p}$$

where the (g, h)'th element is given by

$$\begin{aligned} I_{g,h}(\boldsymbol{\beta}) = &\sum_{i=1}^{k} \frac{\sum_{j \in R_i} Z_{jg} Z_{jh} \exp\left[\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right]}{\sum_{j \in R_i} \exp\left[\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right]} \\ &- \sum_{i=1}^{k} \left[\frac{\sum_{j \in R_i} Z_{jg} \exp\left[\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right]}{\sum_{j \in R_i} \exp\left[\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right]}\right] \left[\frac{\sum_{j \in R_i} Z_{jh} \exp\left[\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right]}{\sum_{j \in R_i} \exp\left[\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right]}\right]. \end{aligned}$$

The discussion in section 4.1.1 suggests that the partial likelihood (4.9) may be treated as an ordinary likelihood on the assumption that certain conditions are fulfilled. The significance of the estimated values of the unknown parameters $\boldsymbol{\beta}$ based on the partial likelihood may then be assessed by standard large-sample likelihood theory. Let $\hat{\boldsymbol{\beta}}$ denote the MLE of the unknown parameters $\boldsymbol{\beta}$ obtained by maximisation of the partial likelihood. Counting process theory can be used to show that $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$ with an asymptotic zero

mean normal distribution and an estimated covariance matrix given by $I(\boldsymbol{\beta})^{-1}$ Kalbfleisch and Prentice [2002], that is

$$\hat{\boldsymbol{\beta}} \overset{d}{\approx} N(\boldsymbol{\beta}, I(\boldsymbol{\beta})^{-1}). \tag{4.12}$$

Under the usual regularity conditions, the observed Fisher information then converges a.s. to the expected Fisher information, hence the observed Fisher information may be used as an estimator for the covariance matrix of $\hat{\boldsymbol{\beta}}$.

Three main test exists to test the global null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$. The *Wald test* is based on the asymptotic normal distribution of $\hat{\boldsymbol{\beta}}$. The asymptotic result (4.12) implies that the quadratic form $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)I(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top$ has an asymptotic chi-squared distribution with p degrees of freedom when $\boldsymbol{\beta}_0 = \boldsymbol{\beta}$. The Wald statistic is given by

$$\chi^2_W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)I(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top,$$

which has an asymptotic chi-squared distribution with p degrees of freedom under the null hypothesis.

The *score test* is based on the assumption of asymptotic normal distribution for the score function (4.11) associated with the partial likelihood. As discussed in section 4.1.1 the score function has an asymptotic normal distribution when certain regularity and independence conditions are fulfilled. The score statistic is given by

$$\chi^2_S = U(\boldsymbol{\beta}_0)I^{-1}(\boldsymbol{\beta}_0)U(\boldsymbol{\beta}_0)^\top,$$

which likewise has an asymptotic chi-squared distribution with p degrees of freedom under the null hypothesis. The score statistic has the advantage that it can be computed without calculating the MLE $\hat{\boldsymbol{\beta}}$.

The third test is the *likelihood ratio test*, given by

$$\chi^2_{\text{LR}} = -2\left[l(\boldsymbol{\beta}_0) - l(\hat{\boldsymbol{\beta}})\right].$$

A Taylor series expansions of $l(\boldsymbol{\beta}_0)$ around $\hat{\boldsymbol{\beta}}$ may be applied to show, that this test statistic has an asymptotic chi-squared distribution with p degrees of freedom under the null hypothesis Azzalini [2002].

4.1.3 The partial likelihood when ties are present

Often survival times are recorded to the nearest day, week, or month, meaning that the survival data is sampled at discrete time rather than continuous time. This way of gathering the survival data often causes ties in the data, i.e. several individuals have the same event time. The construction of the partial likelihood in (4.9) assumes that no ties are present in the data, and hence this likelihood must be modified in order to handle ties. Several modifications of

the partial likelihood (4.9) has been proposed; of these an exact partial likelihood function due Kalbfleisch and Prentice [2002], an approximation due to Breslow [1974], and an approximation due to Efron [1977] are notable. Before any further discussion on these modifications of the partial likelihood some more notations are needed.

Let $D_i = \{i_1, \ldots, i_{d_i}\}$ be the set containing all individuals with an event at time t_i and let Q_i be the set of the $d_i!$ permutations of the elements of D_i. Further, let $P = (p_1, \ldots, p_{d_i})$ be an element of Q_i and let $R_i(P, r)$ denote the risk set $R_i - \{p_1, \ldots, p_{r-1}\}$, for $1 \leq r \leq d_i$.

The modification of the partial likelihood function due to Kalbfleisch and Prentice [2002] is constructed by taking the average of all likelihood functions that arises when breaking the ties in all possible ways. The contribution at each t_i to the likelihood is given by

$$\frac{1}{d_i!} \exp\left(\boldsymbol{\beta}^\top \boldsymbol{s}_i\right) \sum_{P \in Q_i} \prod_{r=1}^{d_i} \left[\sum_{l \in R_i(P,r)} \exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_l\right) \right]^{-1}.$$

Here $\boldsymbol{s}_i$ be the sum over the covariates for individuals with an event at time t_i, that is $\boldsymbol{s}_i = \sum_{j \in D_i} \boldsymbol{Z}_j$. The corresponding partial likelihood is then given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \exp\left(\boldsymbol{\beta}^\top \boldsymbol{s}_i\right) \sum_{P \in Q_i} \prod_{r=1}^{d_i} \left[\sum_{l \in R_i(P,r)} \exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_l\right) \right]^{-1}. \tag{4.13}$$

This likelihood function provides an exact partial likelihood function when ties are present. However, the computations of this likelihood can be very time consuming when the number of ties in the data is high.

An approximation of the exact likelihood is given by Breslow [1974] and defined as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp\left(\boldsymbol{\beta}^\top \boldsymbol{s}_i\right)}{\sum_{j \in R_i} \exp\left[\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right]^{d_i}}. \tag{4.14}$$

This likelihood is a fairly simple approximation of the exact partial likelihood, and by default it is used by many statistical packages.

The approximation due to Efron [1977] is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp\left(\boldsymbol{\beta}^\top \boldsymbol{s}_i\right)}{\prod_{j=1}^{d_i} \left[\sum_{h \in R_i} \exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_h\right) - \frac{j-1}{d_i} \sum_{l \in D_i} \exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_l\right) \right]}. \tag{4.15}$$

This approximation is closer to the exact likelihood than the approximation in (4.14), however when the number of ties in the data is small, the two approximations give similar results.

When no ties are present, that is when $d_i = 1$ for all i, then the likelihoods (4.13), (4.14), and (4.15) are all equal to the partial likelihood in (4.9).

4.1.4 A discrete model analogue to the Cox proportional hazards model

The main interest when using the Cox proportional hazards model to analyse survival data is often the relative risk of an event. This means that the focus is often restricted to estimation of the parameters $\boldsymbol{\beta}$. However, in some studies it may be of interest to estimate the survival information of an individual with covariates $\boldsymbol{Z}_j$. In order to obtain this information, estimate of the baseline hazard function, the baseline cumulative hazard function, or the baseline survival function is needed. Given an estimate of any of these three functions, estimates of the other functions may then be determined by the relationships given in chapter 2. In the next section an estimation method for estimating these functions are presented. The method relies on a discrete model analogue to the Cox proportional hazards model, hence before the estimation method is presented, this discrete model will be considered.

The discrete model considered in this section is based on the survival function for the Cox proportional hazards model applied to a discrete model. The survival function for the Cox proportional hazards model is given by

$$S(t|\boldsymbol{Z}) = S_0(t)^{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z})}, \tag{4.16}$$

where $S_0(t)$ is a baseline survival function.

Assume X is a discrete random variable taken values $0 < t_1 < t_2 < \cdots$. If the hazard function corresponding to a discreate baseline survival function $S_0(\cdot)$ at time t_i has value $h(t_i) = \mathrm{P}(X = t_i | X \geq t_i)$, then

$$S_0(t) = \prod_{t_i \leq t} (1 - h(t_i)). \tag{4.17}$$

Inserting this discrete baseline survival function, the survival function relation (4.16) becomes

$$S(t|\boldsymbol{Z}) = \prod_{t_i \leq t} (1 - h(t_i))^{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z})}. \tag{4.18}$$

Let $h(t_i|\boldsymbol{Z})$ denote the discrete hazard function for an individual with covariates $\boldsymbol{Z}$, then

$$\begin{aligned} 1 - h(t_i|\boldsymbol{Z}) &= \mathrm{P}(T > t_i | T \geq t_i, \boldsymbol{Z}) \\ &= \frac{S(t_i|\boldsymbol{Z})}{S(t_{i-1}|\boldsymbol{Z})} \\ &= \frac{\prod_{t_j \leq t_i} (1 - h(t_j))^{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z})}}{\prod_{t_j \leq t_{i-1}} (1 - h(t_j))^{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z})}} \\ &= (1 - h(t_i))^{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z})}. \end{aligned}$$

It follows that a discrete analogue of the Cox proportional hazards model is given by

$$h(t_i|\boldsymbol{Z}) = 1 - (1 - h(t_i))^{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z})}. \tag{4.19}$$

Let

$$\begin{aligned} \mathrm{d}H(t|\boldsymbol{Z}) &= H([t+\mathrm{d}t]^-\big|\boldsymbol{Z}) - H(t^-|\boldsymbol{Z}) \\ &= \mathrm{P}\left(X \in [t, t+\mathrm{d}t)|\boldsymbol{Z}\right). \end{aligned}$$

For $\mathrm{d}t$ sufficiently small, the model (4.19) may then be written as

$$\mathrm{d}H(t|\boldsymbol{Z}) = 1 - (1 - \mathrm{d}H_0(t))^{\exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}\right)}. \tag{4.20}$$

If $\mathrm{d}t$ is sufficiently small and $\mathrm{d}H_0$ is replaced with a continuous hazard function $h_0(t)\mathrm{d}t$, then the model (4.20) gives the Cox proportional model (4.2).

4.1.5 Estimation of the hazard function and the survival function

Suppose a sample of n i.i.d. individuals are given and let $t_1, \ldots, t_k$ be the event times. Let D_i denote the set of individuals failing at time t_i and let d_i be the number of individuals in D_i. Suppose m_i individuals are censored in the interval $[t_i, t_{i+1})$, $i = 0, \ldots, k$, where $t_0 = 0$ and $t_{k+1} = \infty$ and let $t_{i1}, \ldots, t_{im_i}$ be the censoring times in the interval. The likelihood of the data can then be written as

$$L = \prod_{i=1}^{k} \left[\prod_{l \in D_i} \left[S(t_i^-|\boldsymbol{Z}_l) - S(t_i|\boldsymbol{Z}_l) \right] \prod_{l=1}^{m_i} S(t_{il}|\boldsymbol{Z}_l) \right]. \tag{4.21}$$

Maximisation of this likelihood with respect to the baseline survival function is obtained by letting $S_0(t) = S_0(t_i)$ for $t_i \leq t < t_{i+1}$. That is, the survival between two successive event times is assumed constant. Assuming constant survival function between two successive event times gives a discrete model in which the cumulative baseline hazard function is a sum of discrete hazard components, that is

$$H_0(t) = \sum_{j|t_j \leq t} (1 - \alpha_j). \tag{4.22}$$

With the assumption of a discrete model where $S_0(t) = S_0(t_i)$ for $t_i \leq t < t_{i+1}$, the likelihood function (4.21) may be modified

$$L = \prod_{i=1}^{k} \left[\prod_{l \in D_i} \left[S(t_{i-1}|\boldsymbol{Z}_l) - S(t_i|\boldsymbol{Z}_l) \right] \prod_{l \in B_i} S(t_i|\boldsymbol{Z}_l) \right], \tag{4.23}$$

where B_i denotes the set of individuals censored in the interval $[t_i, t_{i+1})$.

From (4.22) it follows that

$$\mathrm{d}H_0(t) = 1 - \alpha_i, \quad \text{for} \quad t = t_i.$$

That is, $\alpha_i = 1 - \mathrm{d}H_0(t_i)$ may be regarded as the probability of an individual surviving through the interval $[t_i, t_{i+1})$. Hence, the baseline survival function can then be written as a product of

α_i, $i = 1, \ldots, k$. The MLE of $S_0(t)$ can then be obtained by the MLE of α_i and the likelihood (4.23) may be maximised with respect to α_i rather than $S_0(t)$.

The relation between the survival function and the hazard function for the model (4.20) may be expressed by the product

$$\begin{aligned} S(t_i|\boldsymbol{Z}) &= \prod_{j|t_j \leq t_i} (1 - \mathrm{d}H(t_j|\boldsymbol{Z})) \\ &= \prod_{j|t_j \leq t_i} (1 - \mathrm{d}H_0(u|\boldsymbol{Z}))^{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z})} \\ &= \prod_{j|t_j \leq t_i} (1 - (1 - \alpha_j))^{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z})} \\ &= \prod_{j|t_j \leq t_i} \alpha_j^{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z})}. \end{aligned} \tag{4.24}$$

Using (4.24) the term $\prod_{l \in D_i} [S(t_{i-1}|\boldsymbol{Z}_l) - S(t_i|\boldsymbol{Z}_l)]$ in the likelihood function (4.23) can be written as

$$\begin{aligned} \prod_{l \in D_i} [S(t_{i-1}|\boldsymbol{Z}_l) - S(t_i|\boldsymbol{Z}_l)] &= \prod_{l \in D_i} \left[\prod_{j|t_j \leq t_{i-1}} \alpha_j^{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z}_l)} - \prod_{j|t_j \leq t_i} \alpha_j^{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z}_l)} \right] \\ &= \prod_{l \in D_i} \left[\left(1 - \alpha_i^{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z}_l)}\right) \prod_{j|t_j \leq t_{i-1}} \alpha_j^{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z}_l)} \right]. \end{aligned} \tag{4.25}$$

Inserting the survival function (4.24) and the expression (4.25) the likelihood in (4.23) becomes

$$\begin{aligned} L &= \prod_{i=1}^{k} \left[\prod_{l \in D_i} \left[\left(1 - \alpha_i^{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z}_l)}\right) \prod_{j|t_j \leq t_{i-1}} \alpha_j^{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z}_l)} \right] \prod_{l \in B_i} \left[\prod_{j|t_j \leq t_i} \alpha_j^{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z}_l)} \right] \right] \\ &= \prod_{i=1}^{k} \left[\prod_{j \in D_i} \left(1 - \alpha_i^{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z}_j)}\right) \prod_{l \in R_i - D_i} \alpha_i^{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z}_l)} \right]. \end{aligned} \tag{4.26}$$

Assume the unknown parameters $\boldsymbol{\beta}$ are estimated. The MLE of α_m, $m = 1, \ldots, k$ may then be found by maximising the log likelihood function corresponding to (4.26). Taking the logarithm of this likelihood the log likelihood function is

$$l = \sum_{i=1}^{k} \left[\sum_{j \in D_i} \log \left(1 - \alpha_i^{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z}_j)}\right) + \sum_{l \in R_i - D_i} \exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_l\right) \log\left(\alpha_i\right) \right] \tag{4.27}$$

Differentiating (4.27) with respect to α_m, $m = 1, \ldots, k$ gives

$$\frac{\partial}{\partial \alpha_m} \log(L) = - \sum_{j \in D_m} \frac{\exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right)}{1 - \alpha_m^{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z}_j)}} \cdot \frac{\alpha_m^{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z}_j)}}{\alpha_m} + \sum_{l \in R_m} \frac{\exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right)}{\alpha_m} - \sum_{l \in D_m} \frac{\exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right)}{\alpha_m}$$

The MLE of α_m may then be obtained as a solution to the equation

$$\sum_{j \in D_m} \frac{\exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right)}{1 - \alpha_m^{\exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right)}} = \sum_{l \in R_m} \exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right). \tag{4.28}$$

When there are no ties in the data, that is when $d_m = 1$ the solution of (4.28) is given by

$$\hat{\alpha}_m = \left(1 - \frac{\exp\left(\hat{\boldsymbol{\beta}}^\top \boldsymbol{Z}_m\right)}{\sum_{l \in R_m} \exp\left(\hat{\boldsymbol{\beta}}^\top \boldsymbol{Z}_l\right)}\right)^{\exp\left(-\hat{\boldsymbol{\beta}}^\top \boldsymbol{Z}_m\right)},$$

where $\hat{\boldsymbol{\beta}}$ is the MLE of the unknown parameter $\boldsymbol{\beta}$. When ties occur in the data, no closed form solution exists to the equation (4.28), and hence numerical methods must be applied to solve the equation.

Given the estimates $\hat{\alpha}_m$, $m = 1, \ldots, k$, the estimate of the baseline hazard function is then

$$\hat{h}_0(t) = 1 - \hat{\alpha}_m. \tag{4.29}$$

for $t_m \leq t < t_{m+1}$. And the baseline survival function $S_0(t)$ can be estimated by

$$\hat{S}_0(t) = \prod_{m | t_m \leq t} \hat{\alpha}_m. \tag{4.30}$$

The estimated value of $S_0(t)$ is zero for $t \geq t_k$, unless censored survival times occur at times greater than t_k, in which case $\hat{S}_0(t)$ is undefined for $t > t_k$.

By equation (2.6) the cumulative baseline hazard function $H_0(t)$ can be estimated by

$$\hat{H}_0(t) = -\ln\left(\hat{S}_0(t)\right) = -\sum_{m | t_m \leq t} \ln\left(\hat{\alpha}_m\right). \tag{4.31}$$

The estimates (4.29), (4.30), and (4.31) can then be used to estimate the survival information for an individual with covariates $\boldsymbol{Z}$. An estimate of the hazard function for an individual with covariates $\boldsymbol{Z}$ is given by

$$\hat{h}(t|\boldsymbol{Z}) = \hat{h}_0(t) \exp\left[\hat{\boldsymbol{\beta}}^\top \boldsymbol{Z}\right], \tag{4.32}$$

where $\hat{h}_0(t)$ is the estimate of $h_0(t)$ given in (4.29). Furthermore, the corresponding cumulative hazard function can be estimated by integrating both sides of equation (4.32), that is

$$\int_0^t \hat{h}(u|\boldsymbol{Z}) \mathrm{d}u = \exp\left(\hat{\boldsymbol{\beta}}^\top \boldsymbol{Z}\right) \int_0^t \hat{h}_0(t) \mathrm{d}u,$$

from which it follows that

$$\hat{H}(t|\boldsymbol{Z}) = \exp\left(\hat{\boldsymbol{\beta}}^\top \boldsymbol{Z}\right) \hat{H}_0(t).$$

From equation (2.7) it follows that the survival function for the individual with covariates $\boldsymbol{Z}$ can be estimated by

$$\hat{S}(t|\boldsymbol{Z}) = \left(\hat{S}_0(t)\right)^{\exp(\hat{\boldsymbol{\beta}}^\top \boldsymbol{Z})}.$$

5 Pseudo-observations

Unless otherwise stated is this chapter written based on Andersen et al. [2003], Klein and Andersen [2005], and Andersen and Perme [2010].

In this chapter *jackknife pseudo-observations* are considered as a method for analysing survival data. The method were first proposed by Andersen et al. [2003] as an approach for performing generalised linear regression analysis on survival data. The theory of pseudo-observations for regression analysis is based on the pseudo-observations known from the jackknife method Miller [1974]. The jackknife is a non-parametric method used to study the precision of some estimate of an unknown population parameter. Pseudo-observations for regression analysis are an extension of this method, where estimators based on the entire sample are used to perform regression analysis on the individual level. Pseudo-observations address one of the main problems with survival data, i.e. not having appropriate responses for all individuals in the study. This means that the pseudo-observations may also be used for graphical assessment of the model assumptions in a regression analysis.

The basic idea of the pseudo-observations is simple. Let $X_1, \ldots, X_n$ be i.i.d. copies of a random variable X. Furthermore, let $\theta = \theta(X)$ be a parameter of the form

$$\theta = \mathbb{E}[\phi(X)] = \int \phi(x) \, \mathrm{d}F_X(x),$$

where $\phi(\cdot)$ is some function of X. The function $\phi(\cdot)$ and thereby θ may be multivariate.

Suppose an unbiased (or approximately unbiased) estimator $\hat{\theta} = \hat{\theta}(\boldsymbol{X})$ of θ is available, that is

$$\mathbb{E}[\hat{\theta}] = \int \hat{\theta}(\boldsymbol{x}) \mathrm{d}F_X(\boldsymbol{x}) = \theta.$$

Here the notion $\hat{\theta}(\boldsymbol{X})$ indicates that the estimator is based on the entire sample $\boldsymbol{X} = \{X_1, \ldots, X_n\}$. For each individual in the study, the pseudo-observations know from the jackknife theory are defined in terms of the estimator $\hat{\theta}$.

Definition 5.1 *Let $X_1, \ldots, X_n$ be i.i.d. random variables and let $\hat{\theta}(\boldsymbol{X})$ be an unbiased (or approximately unbiased) estimator of the parameter $\theta = \mathbb{E}[\phi(X)]$. For each X_j the pseudo-observation is defined by*

$$\hat{\theta}_j(\boldsymbol{X}) = n\hat{\theta}(\boldsymbol{X}) - (n-1)\hat{\theta}^{-j}(\boldsymbol{X}), \quad j = 1, \ldots, n,$$

where $\hat{\theta}^{-j}(\cdot)$ is an estimator similar to $\hat{\theta}(\cdot)$ based on the observations $i \neq j$.

Note that if the estimator $\hat{\theta}$ is chosen as the sample mean, then the j'th pseudo-observation is simply given by $\phi(X_j)$. In general, the pseudo-observation $\hat{\theta}_j$ may be considered as the contribution of subject j to the estimate of $\mathbb{E}[\phi(X)]$ based on a sample of size n.

The intuition for performing regression analysis based on the pseudo-observations relies on some appealing features concerning the expected value of the pseudo-observations. Let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ be i.i.d. covariates, where each $\boldsymbol{Z}_j = [Z_{j1}, \ldots, Z_{jp}]^\top$. Then it is easily seen that $\hat{\theta}$ is also an unbiased estimator of θ with respect to the joint distribution of X and $\boldsymbol{Z}$. Furthermore,

$$\theta = \int \phi(x) \mathrm{d}F_X(x) = \int\int \phi(x) \mathrm{d}F_{X\boldsymbol{Z}}(x, \boldsymbol{Z}) = \int \mathbb{E}[\phi(X)|\boldsymbol{Z} = \boldsymbol{z}] \mathrm{d}F_{\boldsymbol{Z}}(\boldsymbol{z}), \tag{5.1}$$

which means that θ may be represented as the marginal mean of the conditional expectation of $\phi(X)$ given $\boldsymbol{Z}$. Replacing the distribution $F_{\boldsymbol{Z}}(\cdot)$ in (5.1) with the empirical distribution $\hat{F}_{\boldsymbol{Z}}(\cdot)$, one can interpret $\hat{\theta}$ as an estimator of the average of $\mathbb{E}[\phi(X)|\boldsymbol{Z}]$.

Define the random variables

$$\theta_j(\boldsymbol{Z}_j) = \mathbb{E}_X[\phi(X_j)|\boldsymbol{Z}_j], \quad j = 1, \ldots, n, \tag{5.2}$$

and let $\tilde{\theta}(Z)$ denote the average of $\theta_j(\boldsymbol{Z}_j)$, that is

$$\tilde{\theta}(Z) = \frac{1}{n} \sum_i \theta_i(\boldsymbol{Z}_i).$$

Then

$$\mathbb{E}_{X\boldsymbol{Z}}[\tilde{\theta}(Z)] = \mathbb{E}_{\boldsymbol{Z}}[\tilde{\theta}(Z)] = \frac{1}{n} \sum_i \mathbb{E}_{\boldsymbol{Z}}[\theta_i(\boldsymbol{Z}_i)] = \theta,$$

which means that $\tilde{\theta}(Z)$ is also an unbiased estimator of θ with respect to the joint distribution $F_{X\boldsymbol{Z}}(\cdot, \cdot)$. Consider now the *leave-one-out* estimator

$$\tilde{\theta}^{-j}(Z) = \frac{1}{n-1} \sum_{i \neq j} \theta_i(\boldsymbol{Z}_i),$$

which is based on the observations $i \neq j$. Then $\theta_j(\boldsymbol{Z}_j)$ in (5.2) may be represented as

$$\theta_j(\boldsymbol{Z}_j) = n\tilde{\theta}(Z) - (n-1)\tilde{\theta}^{-j}(Z).$$

Similar calculations applied to $\hat{\theta}(\boldsymbol{X})$ defines the pseudo-observation $\hat{\theta}_j(\boldsymbol{X})$ from definition 5.1,

$$\hat{\theta}_j(\boldsymbol{X}) = n\hat{\theta}(\boldsymbol{X}) - (n-1)\hat{\theta}^{-j}(\boldsymbol{X}).$$

Since $\mathbb{E}_{X\boldsymbol{Z}}[\hat{\theta}(\boldsymbol{X})] = \theta = \mathbb{E}_{X\boldsymbol{Z}}[\tilde{\theta}(Z)]$ it follows that $\hat{\theta}_j(\boldsymbol{X})$ has the same expectation as $\theta_j(\boldsymbol{Z}_j)$ with respect to $F_{X\boldsymbol{Z}}(\cdot,\cdot)$, that is

$$E_{X\boldsymbol{Z}}[\hat{\theta}_j(\boldsymbol{X})] = \mathbb{E}_{X\boldsymbol{Z}}[\theta_j(\boldsymbol{Z}_j)]. \tag{5.3}$$

It follows that the pseudo-observation $\hat{\theta}_j(\boldsymbol{X})$ and the conditional mean $\theta(\boldsymbol{Z}_j)$ estimates the same parameter in the sense of (5.3).

Example 5.1 (Survival probabilities) Let $X_1, \ldots, X_n$ be n i.i.d. survival times with survival function $S(t_0) = \mathbb{E}[\mathbb{1}[X_j > t_0]]$ a time t_0. In this example the function of interest $\phi(\cdot)$ is given by

$$\phi(X_j) = \phi_{t_0}(X_j) = \mathbb{1}[X_j > t_0], \quad j = 1, \ldots, n,$$

and the parameter θ is the survival function $S(\cdot)$ evaluated at time t_0.

The survival function may be estimated by the Kaplan-Meier estimator $\hat{S}(\cdot)$, which is an approximately unbiased estimator of $S(\cdot)$. The j'th pseudo-observation is then given by

$$\hat{S}_j(t_0) = n\hat{S}(t_0) - (n-1)\hat{S}^{-j}(t_0), \tag{5.4}$$

where $\hat{S}^{-j}(\cdot)$ is the Kaplan-Meier estimator of $S(\cdot)$ based on the observations $i \neq j$.

A multivariate version of this is to study a grid of fixed time points $t_1, \ldots, t_k$ simultaneously. In this case

$$\phi(X_j) = [\phi_{t_1}(X_j), \ldots, \phi_{t_k}(X_j)] = [\mathbb{1}[X_j > t_1], \ldots, \mathbb{1}[X_j > t_k]],$$

with parameters

$$\theta = [\theta(t_1), \ldots, \theta(t_k)] = [S(t_1), \ldots, S(t_k)].$$

□

When θ is a multivariate parameter of dimension k, then k pseudo-observations are defined for each individual j as follows

$$\hat{\theta}_{jl} = n\hat{\theta}(t_l) - (n-1)\hat{\theta}^{-j}(t_l), \quad l = 1, \ldots, k.$$

Example 5.2 (The restricted mean survival time) Let the setting be as in example 5.1. The restricted mean survival time is defined as $\mu_\tau = \mathbb{E}[\min(X, \tau)]$, for $\tau > 0$. In this example the function $\phi(\cdot)$ is given by

$$\phi(X) = \phi_\tau(X) = \min(X, \tau),$$

and the parameter of interest is $\theta = \mu_\tau$. The restricted mean may be written in terms of the survival function $\mu_\tau = \int_0^\tau S(t)\mathrm{d}t$ and μ_τ may be estimated by

$$\hat{\mu}_\tau = \int_0^\tau \hat{S}(t)\mathrm{d}t,$$

where $\hat{S}(\cdot)$ is the Kaplan-Meier estimator. The j'th pseudo-observation is then given by

$$\hat{\mu}_{\tau j} = n\int_0^\tau \hat{S}(t)\mathrm{d}t - (n-1)\int_0^\tau \hat{S}^{-j}(t)\mathrm{d}t = \int_0^\tau \hat{S}_j(t)\mathrm{d}t, \quad j = 1, \ldots, n, \tag{5.5}$$

where $\hat{S}_j(\cdot)$ is the pseudo-observation given in example 5.1.

□

Example 5.3 (Competing risks cumulative incidences function) Consider a competing risks analysis, that is an analysis in which a subject may fail from any one of K causes. This may be represented by a *latent failure time approach*, in which it is assumed that there are K potential failure times $\tilde{X}_1, \ldots, \tilde{X}_K$ for each individual. One observes $X = \min(\tilde{X}_1, \ldots, \tilde{X}_K)$ and a variable $\epsilon = r$ if $X = \tilde{X}_r$, $r = 1, \ldots, K$.

The competing risks probabilities may be summarised by either the *cause specific hazard function* or the *cumulative incidence function*. The cause specific hazard function for cause r is defined by

$$h_r(t) = \lim_{\Delta t \to 0} \frac{\mathrm{P}(t \leq \tilde{X}_r < t + \Delta t \mid X \geq t)}{\Delta t}, \quad r = 1, \ldots, K.$$

This function gives the instantaneous risk of failure due to cause r, given that no failure has occurred prior to time t. The cumulative incidence function for cause r is defined by

$$F_r(t) = \mathbb{E}[\mathbb{1}[X \leq t, \epsilon = r]] = \int_0^t h_r(u)\exp\left[-\int_0^u \sum_{i=1}^K h_i(v)\mathrm{d}v\right]\mathrm{d}u. \tag{5.6}$$

This function gives the probability of failing due to cause r prior to time t in the presence of all causes of failure.

In this example the parameter of interest is the cumulative incidence function for cause r $\theta = F_r(t)$ and the function $\phi(\cdot)$ is given by $\phi(X) = \phi_{tr}(X) = \mathbb{1}[X \leq t, \epsilon = r]$.

Suppose a right-censored sample with n individuals are given. For each individual j, the data consists of the triplet $(T_j, \delta_j, \epsilon_j)$, where T_j is the study time, δ_j indicates whatever the j'th individual are censored or not, and ϵ_j indicate which competing risks caused the event. Let $Y(t)$ denote the risk set at time t and let $N_r(t)$ be a counting process given the number of individuals failing due to cause r prior to time t, that is

$$N_r(t) = \sum_{j=1}^n \mathbb{1}[T_j \leq t, \delta_j = 1, \epsilon_j = r].$$

The cumulative incidence function (5.6) may be estimated by the *Aalen-Johansen estimator* given by

$$\hat{F}_r(t) = \int_0^t \prod_{v<u} \left(1 - \frac{\sum_{r=1}^K \mathrm{d}N_r(v)}{Y(v)}\right) \mathrm{d}\hat{H}_r(u),$$

where $\hat{H}_r(\cdot)$ is the Nelson-Aalen estimator for the cumulative cause-r specified hazard function given by

$$\hat{H}_r(t) = \int_0^t \frac{\mathrm{d}N_r(u)}{Y(u)}.$$

The Aalen-Johansen estimator is an approximately unbiased estimator of $F_r(t)$ Andersen et al. [1993]. The j'th pseudo-observation corresponding to $F_r(\cdot)$ at time t_0 is then given by

$$\hat{F}_{jr}(t_0) = n\hat{F}_r(t_0) - (n-1)\hat{F}_r^{-j}(t_0), \quad r = 1, \ldots, K. \tag{5.7}$$

□

Before any further comments on regression analysis with pseudo-observations, some general properties of the pseudo-observations are considered. In the following let $T_j = \min(X_j, C_j)$ be the study time for individual j and let $\delta_j = \mathbb{1}[X_j \leq C_j]$.

5.1 Properties of the pseudo-observations

Studies of the pseudo-observations and their properties have so far been restricted to the case of the survival function (example 5.1), the restricted mean survival time (example 5.2), and the competing risks cumulative incidence function (example 5.3).

In a study of pseudo-observations in the context of the jackknife method, Tukey [1958] conjectured that the pseudo-observations may be treated as though they are i.i.d. Graw et al. [2009] showed that this is indeed the case when the pseudo-observations are calculated based on the competing risks cumulative incidence function and when n approaches infinity. In the case of no competing risks, the estimated cumulative incidence function $\hat{F}_r(t)$ reduces to $1 - \hat{S}(t)$ Fleming and Harrington [1991] and hence the property of i.i.d. also holds for pseudo-observations defined for the survival function and the restricted mean survival time. The results have not yet been proven in a general setting. Furthermore, it follows directly from the definition of the pseudo-observation and the unbiased assumption of $\hat{\theta}$ that each pseudo-observation is an (approximately) unbiased estimator of θ.

Figure 5.1 shows how the pseudo-observation for the survival function in example 5.1 may change over time. Note that the pseudo-observation is defined for all individuals at all time

points no matter study time or censoring status for each individual. Figure 5.1 (a) shows the pseudo-observation for an individual in a data set where no censoring occur. In this case the formula for $\hat{S}_j(t)$ in (5.4) simplifies to the indicator function $\hat{S}_j(t) = \mathbb{1}[X_j \geq t]$. Hence the pseudo-observation is equal to one while subject j is still alive and then jumps down to zero when he fails. Figure 5.1 (b) and (c) shows the pseudo-observation for an individual with event time $X_j = 1$ and an individual with censoring time $C_j = 1$, respectively, in a data with roughly 25% censoring ($n = 250$). As the pseudo-observation for each individual is calculated using the Kaplan-Meier estimator based on the entire sample, the value of the pseudo-observation changes at each event time in the data and it is constant between two successive event times. It is observed that the pseudo-observation for an individual still at risk at a given time point is above one. This is caused by the lowering of the risk set before T_j when omitting individual j in $\hat{S}_{-j}(\cdot)$. The risk set is decreasing over time, this means that the difference between $\hat{S}(\cdot)$ and $\hat{S}_{-j}(\cdot)$ is increasing with t and hence the value of the pseudo-observation is likewise increasing. For an individual with observed event time (figure 5.1 (b)) the end of follow up causes a jump in the value of the pseudo-observation, which is due to omitting the event in $\hat{S}_{-j}(\cdot)$. In figure 5.1 (c) the censoring of the event time causes a turning point in the value of the pseudo-observation. As the value of the pseudo-observation is constant between two successive event times, the turning point is at the first event time after the censoring time $C_j = 1$.

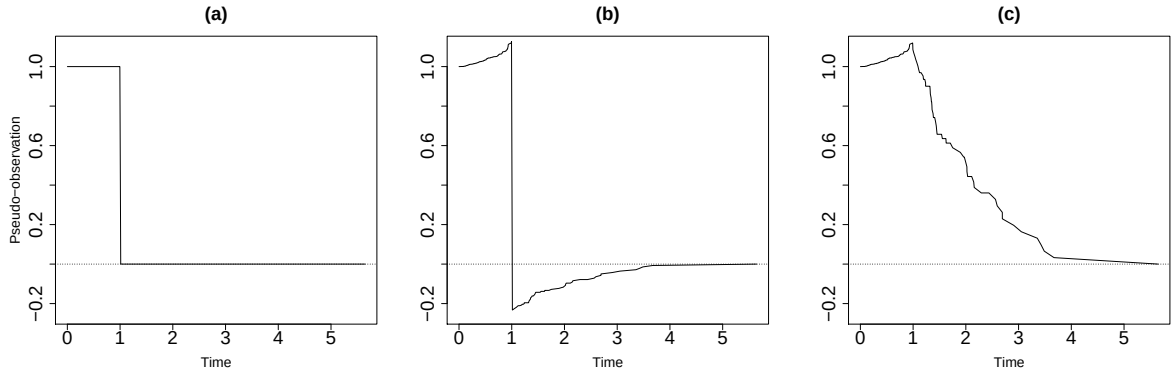

Figure 5.1: The pseudo-observation for the survival function over time. (a) The pseudo-observation for an individual with event time $X_j = 1$ in a data set with no censoring. (b) The pseudo-observation for an individual with event time $X_j = 1$ in a censored data set. (c) The pseudo-observation for an in individual with censored event time $C_j = 1$.

Figure 5.2 shows the pseudo-observations for the restricted mean survival time in example 5.2 for all individuals in a data set with no censoring. The dotted line shows the truncated time $\tau = 4$ and the dashed line shows the equality between the pseudo-observation and the observed event time. When no censoring occur in the data, the pseudo-observation for the survival function reduces to $\hat{S}_j(t) = \mathbb{1}[X_j > t]$. Hence in this case, equation (5.5) implies that the pseudo-observation for the restricted mean is equal to the observed event time when $X_j \leq \tau$ and equal to τ otherwise.

Figure 5.3 shows plots of the pseudo-observation for the restricted mean survival time

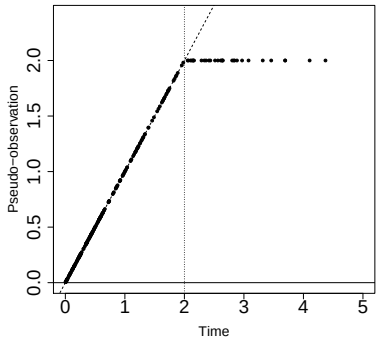

Figure 5.2: The pseudo-observations for the restricted mean survival time for all individuals in a data set with no censoring. The dotted line marks the truncated time and the dashed line marks the equality between the pseudo-observation and the observed event time.

for all individuals in the data with various degree of censoring (top=25%, middle=50%, bottom=75%) and various values of τ (left=0.5, middle=1, right=2). In a censored data set, the pseudo-observation for the survival function is above one before an event or censoring time, see figure 5.1 (b) and (c). For an individual with censored event time, this pseudo-observation remain positiv after the time of censoring, while this pseudo-observation for an individual with observed event time become negative after the time of event. Hence for $C_j \leq \tau$, it follows from equation (5.5) that the pseudo-observation for the restricted mean survival is greater than the actual event time for censored individuals. In the case of 25% censoring, the pseudo-observations for individuals with observed event time seems to follow the actually observed event time quite well. In the case of 50% and 75% censoring the pseudo-observations are mostly below the true event time. The somewhat remarkable picture in the case of 75% censoring with estimated survival times below zero does of course not make sense from a practical perspective. However, 75% censoring are contrary unlikely in an empirical setting.

Figure 5.4 shows the pseudo-observation for the cause-1 cumulative incidence function $F_1(\cdot)$ in the case where no censoring occur in the data. The plots show how the pseudo-observation may develop over time for a single individual. Figure 5.4 (a) shows the pseudo-observation for an individual failing due to cause 1 at time $X_j = 1$ and figure 5.4 (b) shows the pseudo-observation for an individual failing due to cause 2. Similar to the pseudo-observation for the survival function, the pseudo-observation for $F_1(t)$ reduces to the indicator $\hat{F}_j(t) = \mathbb{1}[X_j \leq t, \epsilon = 1]$ in the case of no censoring, see appendix A.1.

Figure 5.5 shows the pseudo-observation corresponding to $F_1(\cdot)$ in a data set where censoring occur. The figure shows the pseudo-observation over time for three different cases. Figure 5.5 (a) shows the pseudo-observation over time for an individual censored at time $C_j = 2$. Figure (5.5) (b) and (c) shows the pseudo-observation for individuals with observed event time; an individual failing due to cause 1 and an individual failing due to cause 2, respectively. For all three cases the pseudo-observation is decreasing below zero at the beginning of the study. For the individual with an observed event time due to cause 2 the pseudo-observation con-

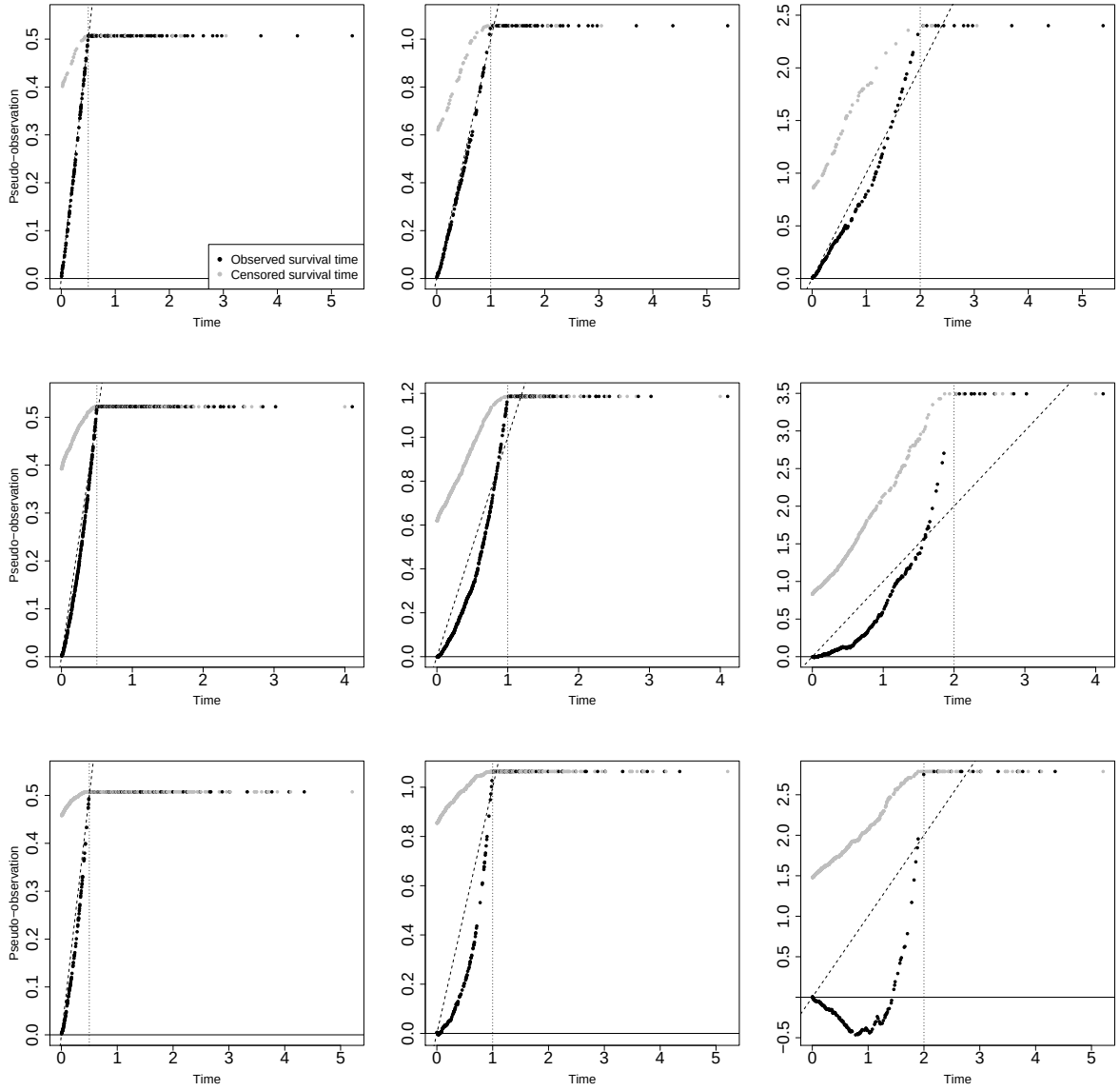

Figure 5.3: The pseudo-observation for the restricted mean for each individual with various degree of censoring (top=25%, middle=50%, bottom=75%) and various choices of τ (left=0.5, middle=1, right=2). Censored individuals are marked with grey dots and uncensored individuals are marked with black dots. The dotted line marks the truncated time and the line of equality is marked by the dashed line.

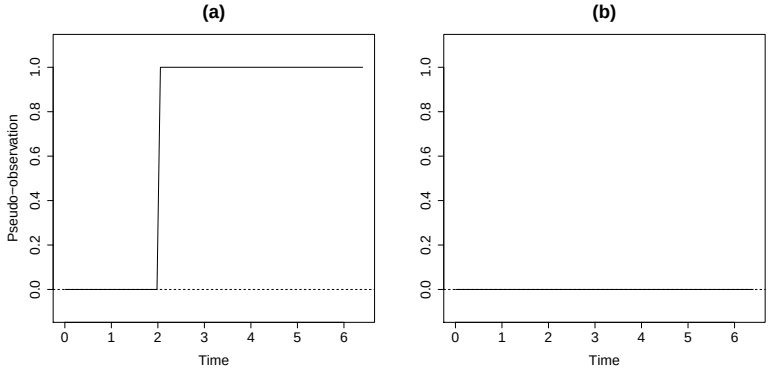

Figure 5.4: The pseudo-observation for the cumulated incidence function (for cause 1) over time in a data with no censoring. (a) The pseudo-observation for an individual failing due to cause 1. (b) The pseudo-observation for an individual failing due to cause 2.

tinue to decrease. For the individual with censored event time the pseudo-observation starts to increase after censoring time. As for the pseudo-observation for the survival function, the turning point is at the first event time occurring in the data after the observed censoring time. For the individual with observed event time due to cause 1 the pseudo-observation jumps to some value above one at the observed event time $X_j = 2$ and then continue to decrease.

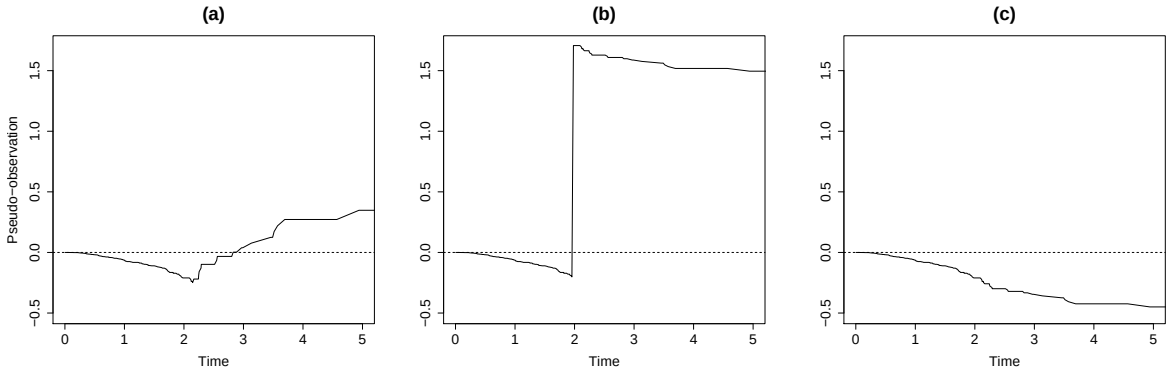

Figure 5.5: The pseudo-observation for the cumulative incidence function (for cause 1) in time in a censored data set. (a) The pseudo-observation for a censored individual. (b) The pseudo-observation for an individual failing due to cause 1. (c) The pseudo-observation for an individual failing due to cause 2.

5.2 Regression models based on pseudo-observations

When analysing survival data one is often interested in describing the survival experience of an individual based on some covariates. Traditional regression models for survival data is often based on the hazard function, which for many applied settings is a suitable approach. However, in some situations it might be desirable with more general regression models when analysing survival data. In this section, pseudo-observations are used for performing generalised linear regression analysis on survival data.

Consider the generalised linear model

$$g(\theta_j) = \alpha + \boldsymbol{\beta}^\top \boldsymbol{Z}_j, \tag{5.8}$$

where $g(\cdot)$ is some link function and θ_j is given in (5.2).

If a fully parametric model is specified, the model (5.8) may be fitted using standard maximum likelihood estimation. However, as noted previously a fully parametric model may not be appropriate in some applied settings. Andersen et al. [2003] suggested to replace the function $\phi(\cdot)$ by a pseudo-observation, and then estimate the unknown parameters by using the generalised estimation equation (GEE) based on the pseudo-observations.

Note that, for each individual j the parameter θ_j may be multivariate, that is $\theta_j = [\theta_{j1}, \ldots, \theta_{jk}]^\top$. Hence, for each θ_{jl}, $l = 1, \ldots, k$ one may specify a model of the form

$$g(\theta_{jl}) = \alpha_l + \boldsymbol{\beta}^\top \boldsymbol{Z}_j, \tag{5.9}$$

where the notation α_l indicates that the intercept may depend one the time t_l.

Allowing the intercept in the model (5.9) to depend on time gives $k + p$ parameters $\boldsymbol{\beta}^* = [\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_p]^\top$ to be estimated. When the GEE is used for estimation, the estimate of $\boldsymbol{\beta}^*$ is given as the solution to the equation

$$U(\boldsymbol{\beta}^*) = \sum_{j=1}^{n} \left[\frac{\partial}{\partial \boldsymbol{\beta}^*} g^{-1}(\alpha + \boldsymbol{\beta}^\top \boldsymbol{Z}_j) \right]^\top V_j^{-1} \left[\hat{\theta}_j - g^{-1}(\alpha + \boldsymbol{\beta}^\top Z) \right] = \sum_{j=1}^{n} U_j(\boldsymbol{\beta}^*) = \mathbf{0}, \tag{5.10}$$

where $g^{-1}(\alpha + \boldsymbol{\beta}^\top \boldsymbol{Z}_j)$ is short for the k-vetor with elements $g^{-1}(\alpha_l + \boldsymbol{\beta}^\top \boldsymbol{Z}_j)$, $l = 1, \ldots, k$. The matrix V_j is a $k \times k$ working covariance matrix of the pseudo-observations $\hat{\theta}_j$, which may account for the correlation inherent in the pseudo-observations defined for each individual.

Let $\hat{\boldsymbol{\beta}}^*$ denote the solution to (5.10). As the pseudo-observations are used for the estimation rather than the observed data, the usual asymptotic properties of the estimates following from the GEE Liang and Zeger [1986] do not directly apply to this setting. Graw et al. [2009] studied the asymptotic properties of $\hat{\boldsymbol{\beta}}^*$ obtained from the estimation equations (5.10), based on pseudo-observations corresponding to the cumulative incidence function. The results showed that the estimated regression parameters $\hat{\boldsymbol{\beta}}^*$ are asymptotic normally distributed and consistent estimates of $\boldsymbol{\beta}^*$. The mean of $\hat{\boldsymbol{\beta}}^*$ is $\boldsymbol{\beta}^*$ and the covariance matrix can be estimated by the sandwich estimator

$$\hat{\Sigma} = I(\hat{\boldsymbol{\beta}}^*)^{-1} \hat{\text{var}}(U(\hat{\boldsymbol{\beta}}^*)) I(\hat{\boldsymbol{\beta}}^*)^{-1},$$

where

$$I(\boldsymbol{\beta}^*) = \sum_{j=1}^{n} \left(\frac{\partial}{\partial \boldsymbol{\beta}^*} g^{-1}(\boldsymbol{\beta}^{*\top} Z_j) \right)^\top V_j^{-1} \left(\frac{\partial}{\partial \boldsymbol{\beta}^*} g^{-1}(\boldsymbol{\beta}^{*\top} \boldsymbol{Z}_j) \right), \quad \hat{\text{var}}(U(\boldsymbol{\beta}^*)) = \sum_{j=1}^{n} U_j(\boldsymbol{\beta}^*) U_j * (\boldsymbol{\beta}^*)^\top.$$

The central part of the proof by Graw et al. [2009] is the finding that

$$\mathbb{E}[\hat{F}_{jr}(t)|\boldsymbol{Z}_j] = g^{-1}(\boldsymbol{\beta}^{*\top}\boldsymbol{Z}_j) + o_p(1), \tag{5.11}$$

from which it follows that $\mathbb{E}[U(\boldsymbol{\beta}^*)] = \mathbf{0}$ for the true parameter $\boldsymbol{\beta}^*$.

The results found by Graw et al. [2009] relies on two assumptions; the censoring times C_j is independent of $(X_j, \delta_j, \boldsymbol{Z}_j)$ and only time points $t < \tau$ such that $S_C(\tau) > \nu > 0$ are considered. Here $S_C(\cdot)$ denotes the survival function for the censoring distribution. By arguments as in section 5.1, the asymptotic results also holds for pseudo-observations corresponding to the survival function in example 5.1 and the restricted mean survival time in example 5.2. The results have not been proven in a general setting.

Notice that once the pseudo-observations have been calculated, one might consider a number of models for analysing the data. However, the number and position of time points, for which the pseudo-observations are calculated, is a choice which must be made prior to the analysis. One time point is enough to obtain estimates of the regression parameter, however, more times point may be more efficient for capturing the trend in the event distribution. One choice is to place the time points at equally spaced percentiles of the data. This choice will place focus at regions with dense observed study times. Another choice is to place the time points at equally spaced study times, which will give a more varied representation of the observed data. However, the portion and position of censored data may also have some impact one the best suitable choice, cf. the plots in section 5.1. Klein and Andersen [2005] studied the number of time points in the context of competing risks. The results showed that increasing the number of equally spaced time points did not have a significant impact on the precision of the estimated regression parameters. Although this is confirmed in the example below, one may conjecture that the best choice is related to the structure of the observed data.

Example 5.4 (Regression analysis on the survival function) Consider the pseudo-observations for the survival function in example 5.1. In that example the function $\phi(\cdot)$ was given as the indicator function $\phi_{t_0}(X_j) = \mathbb{1}[X_j > t_0]$. Suppose one wishes to perform regression analysis on this function. The outcome of the regression analysis θ_j is then given by $S(t_0|\boldsymbol{Z}_j)$. Choosing the link function as the cloglog-function, one gets the model

$$\begin{aligned} \text{cloglog}(S(t_0|\boldsymbol{Z}_j)) &= \log\left(-\log\left(S(t_0|\boldsymbol{Z}_j)\right)\right), \quad j = 1, \ldots, n \\ &= \alpha_0 + \boldsymbol{\beta}^\top \boldsymbol{Z}_j, \end{aligned} \tag{5.12}$$

where $\alpha_0 = \log\left(H_0(t_0)\right)$ and $H_0(\cdot)$ is a cumulative hazard function. The model (5.12) correspond to the Cox proportional hazards model as this model has survival function given by

$$S(t|\boldsymbol{Z}_j) = S_0(t)^{\exp(\boldsymbol{\beta}^\top \boldsymbol{Z}_j)} = \exp\left(-H_0(t)\exp\left(\boldsymbol{\beta}^\top \boldsymbol{Z}_j\right)\right).$$

The model (5.12) may be extended to a multivariate model by simultaneous considering a grid of time points $t_1, \ldots, t_k$. A model is then specified at each time point t_l by

$$\log\left(-\log\left(S(t_l|\boldsymbol{Z}_j)\right)\right) = \alpha_l + \boldsymbol{\beta}^\top \boldsymbol{Z}_j, \quad l = 1, \ldots, k, \tag{5.13}$$

where $\alpha_l = \log(H_0(t_l))$.

To illustrate the use of pseudo-observations in regression analysis, the model (5.13) is fitted to a simulated data set. Event times ($n = 250$) were simulated from a Cox proportional hazards model with constant baseline hazard function $h_0(t) = 2.5$. For each individual a uniform distributed covariate $z_1 \sim \text{Uni}(-1, 1)$ and a normal distributed covariate $z_2 \sim \text{N}(0, 1)$ were simulated with parameters $\beta_1 = 3$ and $\beta_2 = -1$, respectively. Exponentially distributed data were superimposed to obtain roughly 25% censored data. For each individual a pseudo-observation were calculated at equally spaced time points within the range of the observed times. The number of time points were given as 5, 10, 20, and all observed times (250), respectively. The pseudo-observations for each individual, calculate at different time points, were assumed to be independent. 500 replicates of the simulated data were generated and for each replication the model (5.13) were fitted. For comparison the Cox proportional hazards model were likewise fitted to the data.

Table 5.1 shows the results of the analysis based on 10 time points. The table shows the average of the 500 estimated regression parameters and their corresponding standard deviations together with the average of the standard errors of the regression parameters. The GEE model based on the pseudo-observations seems to perform quite well, though with a small bias on the estimated regression parameters. Comparing with the traditional Cox proportional hazards model, both the standard deviation of the 500 replicated analysis and the average of the standard error of the regression parameters seems to be higher for the model based on the pseudo-observations. Though, the standard deviation and the standard errors are agreeable.

For comparison the average of the regression parameters and their standard errors for the analysis based on 5, 20, and all time points are shown i figure 5.6. Figure 5.6 (a) shows the estimates of the parameter $\beta_1 = 3$ and figure 5.6 (b) shows the estimate of the parameter $\beta_2 = -1$. These plots indicate that the performance of the model seems to be quite robust with respect to the number of time points. No precision seems to be gained by including all time points in the study.

Andersen et al. [2003] suggested to estimate the baseline hazard function by averaging over the point estimates $\hat{\alpha}_l$. The logarithm of the cumulated baseline hazard function can be written as $\log(H_0(t)) = \log(h_0(t)) + \log(t)$. The baseline hazard function can hence be estimated by

$$\hat{h}_0 = \exp\left[\frac{1}{k}\sum_{i=1}^{k}\log\left(\hat{H}_0(t_i)\right) - \log(t_i)\right].$$

The mean of the 500 estimated baseline hazard functions based on 10 time points was $\hat{h}_0(t) = 2.56$, which is in agreement with the true baseline hazard function $h_0(t) = 2.5$ (SD = 0.26).

	Pseudo-observations			The Cox proportional hazards model		
Parameters	**Est.**	$\mathbf{SD_{sim}}$	$\mathbf{SE_{est}}$	**Est.**	$\mathbf{SD_{sim}}$	$\mathbf{SE_{est}}$
β_1	3.06	0.25	0.25	3.02	0.20	0.20
β_2	−1.02	0.11	0.11	−1.00	0.09	0.09

Table 5.1: The results after fitting a GEE model based on 10 pseudo-observations for each individual and the Cox proportional hazards model. The results shows the average of the estimated regression parameters and their standard deviations of 500 replicates of the data. Furthermore, the averages of the standard errors of the regression parameters are likewise given.

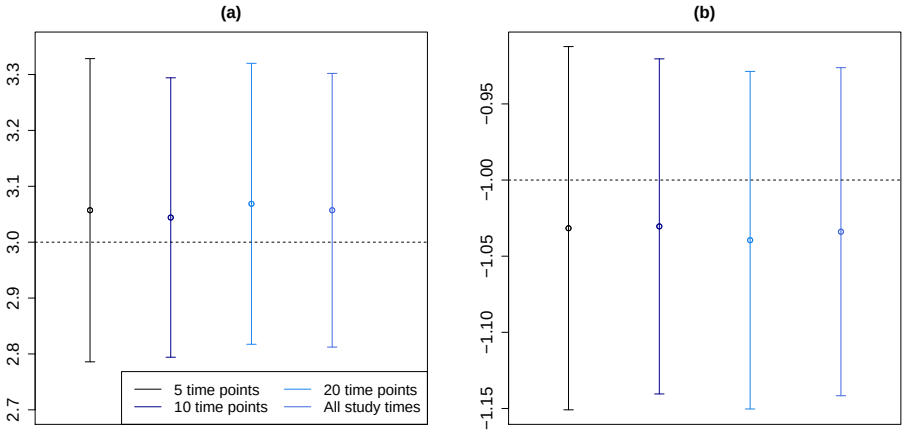

Figure 5.6: The results after fitting a GEE model based on pseudo-observations. The plots show the average of the estimated regression parameters and their corresponding standard errors for models based on 5, 10, and 20 equally spaced time points. Further, estimates for a model based on all observed study times (250) are likewise given. (a) Shows the estimates of the parameter $\beta_1 = 3$. (b) Shows the estimates of the parameter $\beta_2 = -1$.

□

The results from the example above show that the GEE model based on the pseudo-observations are not quite competitive to the traditional Cox proportional hazards model. However, the strength of the method based on the pseudo-observations arises in situations where no standard models exist. An example is regression analysis on the restricted mean survival time μ_τ. From standard regression models on the hazard function it is possible to assess the effect of the covariates on the restricted mean survival time by the relations given in chapter 2. However, often the assumed model relating the covariates to the hazard function gives rises to complicated relations between μ_τ and the covariates. Pseudo-observations allow for direct modelling between the covariates and restricted mean survival time.

Correlation structure in the GEE model

In this project the GEE model is used for the regression analysis. However, notice that once the pseudo-observations have been calculated, one might consider analysing the data by other

models for longitudinal data. An important property of the GEE model is the incorporation of the correlation structure inherent in the pseudo-observations for each individual. Further, one may even use some incorrect working covariance matrix, but the resulting regression parameters are still consistent and asymptotic normally distributed.

The GEE model estimates the parameters $\boldsymbol{\beta}^*$ by solving the equations given in (5.10). It is assumed that V_j can be expressed in terms of a *correlation matrix* R_j

$$V_j = A_j^{1/2} R_j A_j^{1/2},$$

where A_j is a diagonal matrix with elements $\text{Var}[\hat{\theta}_{jl}]$. Further, it is assumed that the variance $\text{Var}[\hat{\theta}_{jl}]$ can be written as a function of the mean value $\mu_{jl} = \mathbb{E}[\hat{\theta}_{jl}|\boldsymbol{Z}_j]$, that is $\text{Var}[\hat{\theta}_{jl}] = v(\mu_{jl})$, for some function $v(\cdot)$. The choice of the correlation matrix R_j will affect the efficiency of the estimates, and in general it is desirable to choose R_j close to the true correlation structure in the data. Fitzmaurice et al. [2008]

Klein and Andersen [2005] studied the choice of the working covariance matrix in the competing risks situation based on the pseudo-observation for the cause-r cumulative incidence function. In this study three different covariance matrices were considered. The simplest is the *independence*, that is the pseudo-observations for each individual, calculated at k different time points, are assumed to be independent and the correlation matrix R_j is chosen to be the $k \times k$ identity matrix I. The second choice of the working covariance matrix is the *exact* covariance matrix when no censoring occurs in the data. When no censoring occur in the data the pseudo-observation (5.7) reduces to the indicator function of a type r event occurring prior to time t, see figure 5.4. This means that

$$\text{Cov}\left(\hat{\theta}_{jl}, \hat{\theta}_{jm}\right) = F_r(t_l)\left[1 - F_r(t_m)\right], \qquad t_l \leq t_m.$$

Let $F_r(t_l|\boldsymbol{Z}_j) = g^{-1}(\alpha_l + \boldsymbol{\beta}^{*\top}\boldsymbol{Z}_j)$, it follows that an exact covariance matrix is given by

$$v_{jlm} = F_r(t_l|\boldsymbol{Z}_j)[1 - F_r(t_m|\boldsymbol{Z}_j)], \quad t_l \leq t_m, \tag{5.14}$$

for $j = 1, \ldots, n$ and $l, m = 1, \ldots, k$. The third covariance matrix considered by Klein and Andersen [2005] is a *common* covariance matrix V, which is defined by the sample covariance matrix with elements

$$v_{lm} = \frac{1}{n}\sum_{j=1}^{n}(\hat{\theta}_{jl} - \bar{\theta}_l)(\hat{\theta}_{jm} - \bar{\theta}_m), \tag{5.15}$$

where $\bar{\theta}_l$ is given by the sample mean

$$\bar{\theta}_l = \frac{1}{n}\sum_{j=1}^{n}\hat{\theta}_{jl}.$$

In terms of bias of the regression parameters Klein and Andersen [2005] found no systematic different between the three choices of working covariance matrices.

	The exact covariance matrix			The common covariance matrix		
Parameters	**Est.**	$\mathbf{SD_{sim}}$	$\mathbf{SE_{est}}$	**Est.**	$\mathbf{SD_{sim}}$	$\mathbf{SE_{est}}$
β_1	3.06	0.25	0.25	3.08	0.25	0.26
β_2	-1.02	0.11	0.11	-1.02	0.12	0.11

Table 5.2: The results of fitting a GEE model based on the pseudo-observations with exact correlation structure and the common correlation structure. The table shows the average of the 500 estimated regression parameters and the corresponding standard deviation. Further, the average of the 500 estimated standard error are likewise given.

Similar studies have not been conduct for regression analysis based on pseudo-observations for the survival function and the restricted mean. However as mentioned previously, the Kaplan-Meier estimator may be regarded as a special case of the Aalen-Johansen estimator with no competing risks. Hence, working covariance matrices similar to those mentioned above may be applied in regression analysis based on the pseudo-observations for the survival function and the restricted mean. In fact, when no censoring occur in the data, pseudo-observations for the survival function reduce to the indicator function $\hat{S}_j(t) = \mathbb{1}[X_j \geq t]$ and an exact covariance matrix similar to (5.14) may be chosen with elements

$$v_{jlm} = S_j(t_l|\boldsymbol{Z}_j)\left[1 - S_j(t_m|\boldsymbol{Z}_j)\right], \quad t_l \leq t_m. \tag{5.16}$$

Example 5.5 (Example 5.4 continued) In example 5.4 a model corresponding to the Cox proportional hazards model were introduced by fitting a GEE model based on the pseudo-observations for the survival function. In that example the pseudo-observations calculated for each individual were assumed to be independent. For comparison, consider now a GEE model with working covariance matrix given by the exact covariance matrix (5.16) and a common covariance matrix (5.15), respectively. Let the setting be as in example (5.4). In table 5.2 the mean of the estimated parameters and their standard deviation is given together with the mean of estimated standard errors of the parameters estimated by the sandwich estimator, see table 5.1 for the results based on the independence assumption. In agreement with Klein and Andersen [2005] no significant different is found between the three different covariance matrices.

□

5.3 Pseudo-residuals

Assessment of model adequacy is an important part of the data analysis. If the model assumptions are violated the results of the model are invalid. A residual plot is a very common method to assess the model assumptions for a fitted regression model. However, graphical methods

are often improper for time-to-event data due to censoring in the data. Pseudo-observations may be used in graphical evaluations of regression models for incomplete data. Perme and Andersen [2008] suggested to use *pseudo-residuals* to evaluate the model assumptions of the Cox proportional hazards model and the additive hazard model. These pseudo-residuals may also be used in the more general setting where regression models are based on pseudo-observations as discussed in section 5.2.

Checking the Cox proportional hazards model using pseudo-observations

Consider a Cox proportional hazards model as given in (4.2). The pseudo-observations for the survival function in example 5.1 may be used to assess the model assumptions for this model.

For each individual in the study, let $\hat{S}(t|\boldsymbol{Z}_j)$ denote the predicted survival function for the fitted Cox proportional hazards model. Perme and Andersen [2008] suggested to evaluate the assumptions of the Cox proportional hazards model by comparing the pseudo-observation with the predicted value of the survival function for each individual. That is, the pseudo-residuals are defined by

$$\hat{\epsilon}_j(t) = \frac{\hat{S}_j(t) - \hat{S}(t|\boldsymbol{Z}_j)}{\sqrt{\hat{S}(t|\boldsymbol{Z}_j)[1 - \hat{S}(t|\boldsymbol{Z}_j)]}}, \quad j = 1, \ldots, n. \tag{5.17}$$

If no censoring occur in the data, the pseudo-observation $\hat{S}_j(t)$ reduced to the indicator of $X_j > t$. Hence, the denominator of (5.17) is an estimate of the conditional standard error of $\hat{S}_j(t)$ given $\boldsymbol{Z}_j$ in the case of no censoring.

The pseudo-residuals defined in (5.17) may be used to check the proportional hazards assumption for the Cox proportional hazards model and for determining the functional form of a covariate which best explain the effect on the conditional hazard function. Suppose a Cox proportional hazards model (4.2) has been fitted to a given data set. To determine if the model is correctly specified, the residuals (5.17) are plotted against a covariate for each individual j at a number of fixed time points. If the model fits the data well, no trend should be seen in the residual plots.

Figure 5.7 shows the pseudo-residuals of a fitted Cox proportional hazards model for four different data sets. Each data set ($n = 1000$) were fitted using a single covariate $Z \sim$Unif$(-1, 1)$ with an assumed linear effect of Z. The four data sets were simulated from the model (4.2) with constant baseline hazard function $h_0(t) = 2.5$; one data set were constructed to meet the assumptions of the model and three data sets were constructed such that the model assumptions are violated. Exponential distributed data were superimposed to obtain roughly 25% censored data. The residuals were calculated at four different time points corresponding to the 20'th, 40'th, 60'th, and 80'th percentiles of the observed study times.

The top row of figure 5.7 shows the pseudo-residuals for a data set where the assumptions of the fitted model are met. The data were constructed to follow the model described above with

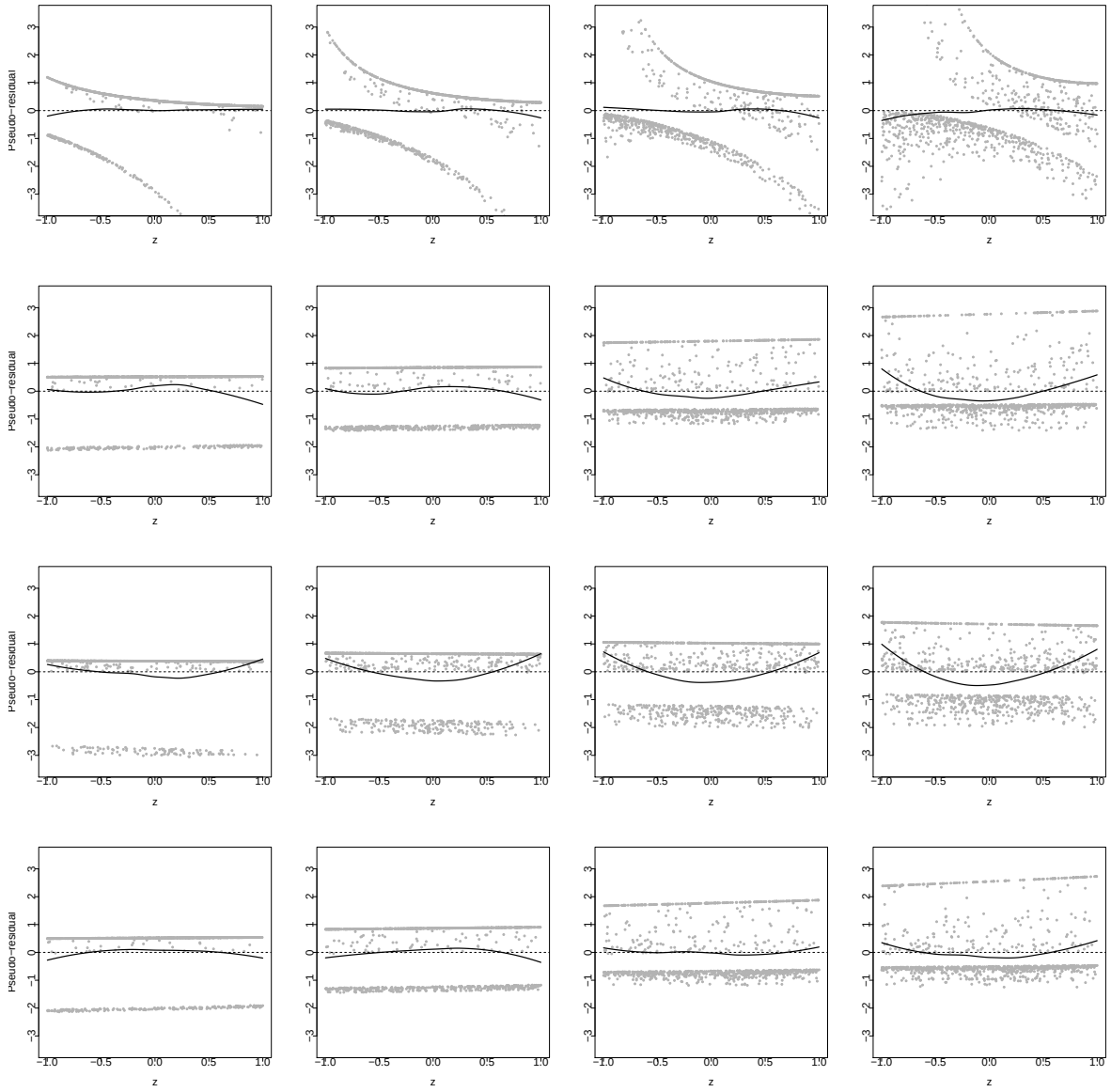

Figure 5.7: The pseudo-residuals for four different data sets calculated at four different time points. The top row shows the pseudo-residuals for a data set, where the assumptions of the model are met. The three other rows shows the pseudo-residuals for data sets where the assumptions of the model are violated. The columns are the pseudo-residuals calculated at four different time point.

parameter $\beta = -2$ and a linear effect of the covariates. The second row of the figure shows the pseudo-residuals for a data set where the true regression parameter β is constructed as a function of the time t; $\beta(t) = 2$ for $t \leq \tau$ and $\beta(t) = -2$ for $t > \tau$. The value of τ were chosen as the expected median of the event times. This construction implies that the proportional hazards assumption of the model (4.2) is violated. The effect of Z is assumed linear. The third row of the figure shows the pseudo-residuals for a data set where the parameter $\beta = -2$ is constant in time, but the effect of the covariates is quadratic rather than linear. The residuals

plotted in the bottom row of figure 5.7 are calculated based on a data set where both the effect of Z is quadratic and the regression parameter is a function of time as described above.

The grey dots in figure 5.7 represent the residuals and the black curve is the smoothed average of the residuals calculated by using local polynomial regression fitting. For all 16 plots, the residuals tend to fall in three groups. The top group consists of the individuals still at risk at the time at which the pseudo-observations are calculated. The pseudo-observation for these individuals is above one and increases with time until either event or censoring, see figure 5.1 (b) and (c). The bottom group of residuals belongs to the individuals with an event time prior to the time at which the pseudo-observations are calculated. The pseudo-observations for these individuals are negative, see figure 5.1 (b) and hence the pseudo-residuals for these individuals are likewise negative. The group in the middle consists of the individuals with a censoring time prior to the time point at which the pseudo-observations are calculated. The pseudo-observations for these individuals remain positive, see figure 5.1 (c), and hence the pseudo-residuals of these individuals will be in between the pseudo-residuals for the individuals still at risk and the pseudo-residuals for the failed individuals.

The trend in the data is illustrated by the smoothed average given by the black curves in the residual plots. The plots in the top row of figure 5.7 were based on a data set were the assumptions of the model are meet. For all of the four time points, the smoothed average seems to be rather horizontal and hence the assumptions of constant parameter and no non-linear effect of the covariates are well reflected in the plots. The second row of the figure were based on the data set with the regression parameter given as a function of time t. The change form positive β to negative β at the expected median event time is seen in the change of the smoothed curve from the two leftmost plots to the two rightmost plots. The third row in the figure was based on a data set with a quadratic effect of the covariates. This quadratic effect of the covariates is quite obvious from the residual plots at all four time points. Further, the shape of the curve is fairly similar at all four time points; this reflects the constant parameter over time. At the bottom row of the figure, the smoothed curves indicate both the quadratic form of the covariate and the functional form of the true regression parameter. Though, the quadratic form of Z is not as obvious as seen in the third row of the figure.

From this example it appears that the pseudo-residuals provide a useful method for assessment of the model assumptions in a Cox proportional hazards model.

6 Regression splines

This chapter is written based on Green and Silverman [1994] and Hastie et al. [2001].

In the previous chapter pseudo-observations were used for generalised linear regression analysis of survival data containing censored event times. The method were used to define a model corresponding to the traditional Cox proportional hazards model, though the strengthen of this method appears mainly in situations where no standard regression models exist. In this chapter *spline functions* are used for estimation of the intercept in the regression model, when the intercept is assumed to be a function of the time t.

A spline is a function constructed based on polynomial segments which satisfy certain conditions of smoothness. Let $[a, b]$ be an interval on $\mathbb{R}$ and let $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_r\}$ be real numbers, called *knots*, satisfying

$$a < \tau_1 < \cdots < \tau_r < b.$$

Further, let a be denoted by τ_0 and b denoted by τ_{r+1}.

Definition 6.1 *A function $s_{\boldsymbol{\tau}} : [a, b] \to \mathbb{R}$ is said to be a spline function of order d if it fulfils*

- *$s_{\boldsymbol{\tau}}(t)$ is a polynomial of order d on each interval $[\tau_i, \tau_{i+1}]$, $i = 0, 1, \ldots, r$.*
- *$s_{\boldsymbol{\tau}}(t) \in C^{d-1}[a, b]$*

A spline function or order d and with r knots have $(r+1)(d+1)$ parameters to be estimated. The second condition of definition 6.1 means that the spline $s_{\boldsymbol{\tau}}(\cdot)$ and its derivatives up to the $d-1$'th order are continuous at all points on the interval $[a, b]$ and in particular at all knots $\boldsymbol{\tau}$. This means that d constrains are put on each knot, leaving $(r+1)(d+1) - rd = r + d + 1$ degrees of freedom to determine the spline function.

In figure 6.1 a spline function of order 3 with three knots is plotted. A spline function of order 3 is called a *cubic spline*. The vertical lines in the plot indicate the intervals for which each polynomial segment is defined.

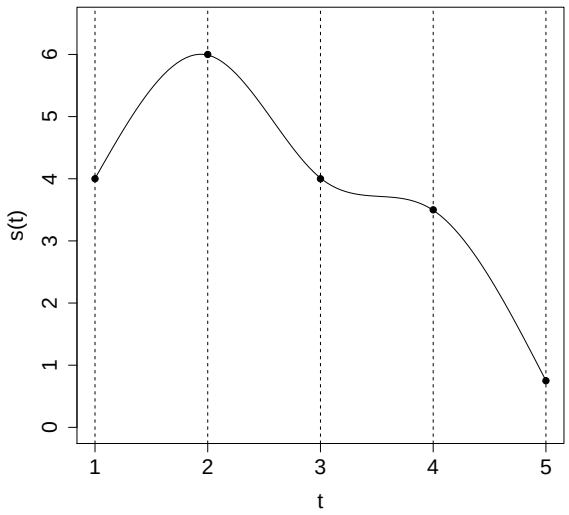

Figure 6.1: A cubic spline function with three knots. The vertical lines indicate the intervals at which each polynomial segment is defined.

The piecewise polynomial structure of the spline functions makes them suitable for flexible estimation of unknown functions. Suppose a data set is given, the parameters in the spline function may then be chosen in some proper way for capturing important patterns in the data, and hence estimate the function underlying the data. This idea may be transferred to regression models for estimation of non-parametric relations in a model.

Suppose a data set consists of n i.i.d. individuals is given. In section 5.2 pseudo-observations, calculated at k different time points, were used to fit a model of the form

$$g(\theta_{jl}) = \alpha_l + \boldsymbol{\beta}^\top \boldsymbol{Z}_j, \quad j = 1, \ldots, n, \quad l = 1, \ldots, k, \tag{6.1}$$

where θ_{jl} is given in (5.2) and $g(\cdot)$ is some link function.

The notation α_l in (6.1) indicates that the intercept of the model may depend on the time t_l. In practice this is implemented by including an indicator of each t_l in the covariate vector $\boldsymbol{Z}_j$. In example 5.4 this method where used to obtain point estimates of the baseline hazard function $h_0(\cdot)$ in the Cox proportional hazards model. However, a more flexible method for estimating $h_0(\cdot)$ is to base the estimation on spline functions, which also allows for estimating the value of $h_0(\cdot)$ at any given time t.

The idea may be stated in a general setting by considering the semi-parametric model

$$g(\theta_{jl}) = \varphi(t_l) + \boldsymbol{\beta}^\top \boldsymbol{Z}_j, \tag{6.2}$$

where $\varphi(\cdot)$ is some unknown function. Suppose $\varphi(\cdot)$ may be estimated by some spline function $s_{\boldsymbol{\tau}}(\cdot)$. The idea is then to write $s_{\boldsymbol{\tau}}(\cdot)$ as a linear combination of some *basis splines*, that is, a set of spline function spanning the space of splines of the same order and with the same knots as $s_{\boldsymbol{\tau}}(\cdot)$. The model is then linear in these basis functions and may be fitted by ordinary estimation methods. This approach is known as *regression spline approximation*. In this report the GEE is used for estimation.

A common basis used is the *B-spline basis*. In addition to the r knots given above, the construction of a B-spline basis of order d requires $2(d+1)$ additional knots. The B-spline basis is defined recursively by

$$B_{i,h+1}(t) = \frac{t-\tau_i}{\tau_{i+h}-\tau_i} B_{i,h}(t) + \frac{\tau_{i+h+1}-t}{\tau_{i+h+1}-\tau_{i+1}} B_{i+1,h}(t),$$

$$B_{i,0}(t) = \begin{cases} 1 & \tau_i \leq t < \tau_{i+1} \\ 0 & \text{otherwise,} \end{cases},$$

where $h = 0, \ldots, d$ and $i = 0, \ldots, r + 2(d+1) - 1$.

A spline function of order d can then be written as

$$s_{\boldsymbol{\tau}}(t) = \sum_{i=0}^{r+d} \gamma_i B_{i,d}(t) = \boldsymbol{\gamma}^\top \boldsymbol{B}(t),$$

where $\boldsymbol{B}(t) = [B_{0,d}(t), \ldots, B_{r+d,d}(t)]^\top$ is the B-spline basis and $\boldsymbol{\gamma} = [\gamma_0, \ldots, \gamma_{r+d}]^\top$ are parameters to be estimated.

The number and positions of the knots are important for the complexity of the spline function. In regions with dense knot spacing the spline function is fairly complex, whereas at regions with few knots, the spline function tends to be more smooth. One may choose either equally spaced observations or equally spaced percentiles of the observed data for the knot positions. However, several general methods exist for choosing the number and positions of the knots.

Consider now the unknown function $\varphi(\cdot)$ from the model (6.2). This function may be estimated by a linear combinations of B-spline basis functions, that is

$$\varphi(t) \approx \boldsymbol{\gamma}^\top \boldsymbol{B}(t). \tag{6.3}$$

The regression model of interest is then given by

$$g(\theta_{jl}) = \boldsymbol{\gamma}^\top \boldsymbol{B}(t_l) + \boldsymbol{\beta}^\top \boldsymbol{Z}_j. \tag{6.4}$$

If $\varphi(\cdot)$ is a continuous function a spline approximation $\boldsymbol{\gamma}^\top \boldsymbol{B}(\cdot)$ always exists, such that

$$\sup \left| \varphi(t) - \boldsymbol{\gamma}^\top \boldsymbol{B}(t) \right| \to 0, \quad \text{for} \quad r \to \infty,$$

where r is the number of knots de Boor [1978].

When the approximation sign in (6.3) is replaced by a equality sign, the asymptotic results regarding the regression parameters found by Graw et al. [2009] may directly be applied to the model (6.4). When the relation in (6.3) is truly an approximation some care must be taken. However, when the approximation in (6.3) is reasonable the results seem likewise to be applicable. The central part of the proof by Graw et al. [2009] rely on the property

$$\mathbb{E}[\hat{\theta}_{jl}|\boldsymbol{Z}_j] = g^{-1}(\varphi(t) + \boldsymbol{\beta}^\top \boldsymbol{Z}_j) + o_p(1). \tag{6.5}$$

When the error term resulting from the approximation of the function $\varphi(\cdot)$ by the spline function $\boldsymbol{\gamma}^\top \boldsymbol{B}(t)$ is negligible, it is reasonable to assume that the property (6.5) also holds when $\varphi(\cdot)$ is replaced by $\boldsymbol{\gamma}^\top \boldsymbol{B}(\cdot)$.

Let $\hat{\boldsymbol{\gamma}}$ denote the estimate of the parameters $\boldsymbol{\gamma}$ in (6.3) and let $\hat{\Sigma} = \text{Cov}(\hat{\boldsymbol{\gamma}})$ be an estimated covariance matrix of $\hat{\boldsymbol{\gamma}}$. A pointwise variance function of the estimate $\hat{s}_{\boldsymbol{\tau}}(\cdot) = \hat{\boldsymbol{\gamma}}^\top \boldsymbol{B}(\cdot)$ is then given by

$$v(t) = \text{Var}[\hat{s}_{\boldsymbol{\tau}}(t)] = \boldsymbol{B}^\top(t)\hat{\Sigma}\boldsymbol{B}(t).$$

In section 4.1.5 a maximum likelihood approach was used to estimate the baseline hazard function for the Cox proportional hazards model. The approach was based on a discrete analogue to the traditional Cox proportional hazards model. In the example below a method based on pseudo-observations and cubic splines are suggested as an alternative method for estimating the baseline hazard function.

Example 6.1 (Estimation of $h_0(\cdot)$ in the Cox proportional hazards model) In example 5.4 a model corresponding to the Cox proportional hazards model were introduced by fitting a GEE model based on the pseudo-observation given in (5.4). The model considered is given by

$$\text{cloglog}(S(t|\boldsymbol{Z}_j)) = \log\left(\int_0^t h_0(u)\text{d}u\right) + \boldsymbol{\beta}^\top \boldsymbol{Z}_j. \tag{6.6}$$

Often the interest of the Cox proportional hazards model is restricted to the parameters $\boldsymbol{\beta}$. Though, it may be of interest to estimate the baseline hazard function $h_0(\cdot)$, as this will give information on the survival experience for each individual in the sample.

Andersen et al. [2003] suggested to estimate the baseline hazard function at different time points, by allowing the intercept of the model (6.1) to depend on time, see example 5.4. However, this approach only allows one to estimate the baseline hazard function at the time points for which the pseudo-observations are calculated. A different approach is to base the estimation of $h_0(\cdot)$ on spline functions.

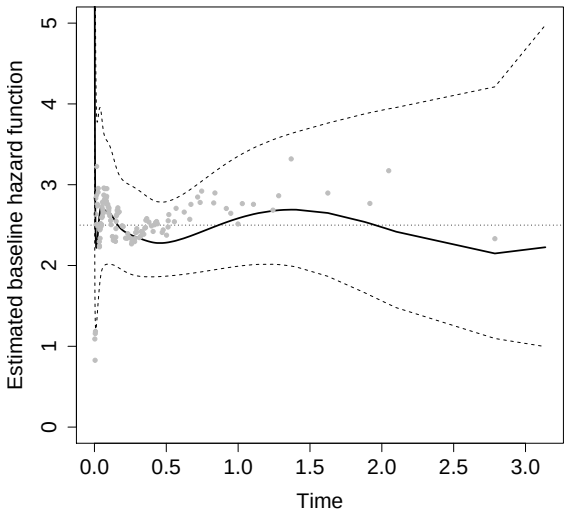

Figure 6.2: The baseline hazard function estimated by a regression spline. The solid line indicates the estimate of the baseline hazard function and the dotted lines indicate af 95% CI. The dots indicate the point estimates of the baseline hazard function proposed by Andersen et al. [2003].

In this example, cubic regression splines are used to estimate the logarithm of $\int_0^t h_0(u)\mathrm{d}u$ from which $h_0(\cdot)$ can be deduced. The model considered is given by

$$\text{cloglog}(S(t|\boldsymbol{Z}_j)) = \boldsymbol{\gamma}^\top \boldsymbol{B}(t) + \boldsymbol{\beta}^\top \boldsymbol{Z}_j, \tag{6.7}$$

where

$$\boldsymbol{B}(t) = \sum_{i=0}^{r+3} B_{i,3}(t).$$

Survival data ($n = 250$) were simulated from a Cox regression model with constant baseline hazard function $h_0(t) = 2.5$. Two covariates were considered; a uniform distributed covariate z_1 over the interval $[-1, 1]$ and a standard normal distributed covariate z_2, with parameters $\beta_1 = -1$ and $\beta_2 = 1$, respectively. Exponentially distributed data were superimposed to obtain roughly 25% censoring. A GEE model based on the pseudo-observations were fitted with the pseudo-observations calculated at 125 equally spaced time points.

The logarithm of the cumulated baseline hazard function was estimated by including B-spline basis functions in the GEE model. The knots used in the estimation were given by the 25'th, 50'th, and the 75'th quantile of the study times. For comparison the point estimates of the baseline hazard function suggested by Andersen et al. [2003] is likewise calculated. Figure 6.2 shows the estimated baseline hazard function with a 95% pointwise confidence interval (CI) indicated by dotted lines. The dots indicates the point estimates of the baseline hazard function suggested by Andersen et al. [2003].

The spline-estimate of the baseline hazard function seems to be agreeable with the true baseline hazard function $h_0(t) = 2.5$. Though, the estimate is quite fluctuating at the beginning of the time scale. In example 5.4 it was seen that the pseudo-observations tend to produce large standard errors on the regression parameters. A similar trend is seen in figure 6.2 with a large 95% CI on the estimated baseline hazard function. The trend is especially clear in the region with few observations.

□

7 The new pseudo-observation

In chapter 5 jackknife pseudo-observations were introduced as a method for performing generalised linear regression analysis on survival data. The results of the small simulation study in example 5.4 showed that though the bias of the estimated regression parameters is quite small, the parameters have a large variability. A significant part of this variability is related to the variability in the pseudo-observations, when they are considered as functions over time. In this chapter a different estimator is proposed for performing regression analysis on the survival function. The estimator is defined to more precisely resemble the information in the observed data. The hope is that this information will give more efficient regression estimates.

Consider the pseudo-observation for the survival function in example 5.1 given by

$$\hat{S}_j(t) = n\hat{S}(t) - (n-1)\hat{S}_{-j}(t), \quad j = 1, \ldots, n \tag{7.1}$$

where $\hat{S}(\cdot)$ is the Kaplan-Meier estimator.

Each pseudo-observation (7.1) may be considered as the contribution of individual j to the Kaplan-Meier estimate of the survival function $S(\cdot)$. Hence, each pseudo-observation is to some extend an estimate of the survival probability for each individual. However, as seen in figure 5.1, each pseudo-observation takes value outside the range of $[0, 1]$ when censoring occur in the data set. From this it appears that the information given by the observed data is not fully utilised in the definition of the pseudo-observation. For an individual with an observed event time, the survival probability is estimated by some number above one prior to the observed event time, hereafter the survival probability is estimated by some number below zero. The same holds for an individual with a censored event time; prior to this observed censored time the survival probability is estimated by some number above one.

A more proper utilisation of the observed data is to estimate the survival probability by conditioning on the observed data. Figure 7.1 shows a plot of an estimator of the probability

$\mathrm{P}(X_j > t | X_j \geq T_j)$, where X_j is the event time for individual j and T_j is the observed study time. The estimators used in the plots are given by

$$\hat{\theta}_j(t) = \begin{cases} 1 & \text{for} \quad t < T_j \\ 0 & \text{for} \quad T_j \leq t \end{cases} \tag{7.2}$$

when T_j is an observed event time, and

$$\hat{\theta}_j(t) = \begin{cases} 1 & \text{for} \quad t < T_j \\ \frac{\hat{S}(t)}{\hat{S}(T_j)} & \text{for} \quad T_j \leq t \end{cases} \tag{7.3}$$

when T_j is a censored event time. Here $\hat{S}(\cdot)$ is the Kaplan-Meier estimator.

Figure 7.1 (a) is a plot of the estimator (7.2) over time and figure 7.1 (b) is a plot of the estimator (7.3) over time. Both the estimators stay inside the valid range of $[0, 1]$. Further, the estimator (7.2) utilises the observed information for an individual with observed event time, by estimating the survival probability by one prior to the observed event time and after the event has occurred, the survival probability is estimated by zero. For an individual with censored event time, the information that the event happens at some time after the observed time is utilised by estimating the survival probability by one prior to this censored time. After the observed censored time, the survival probability is estimated by some number between zero and one.

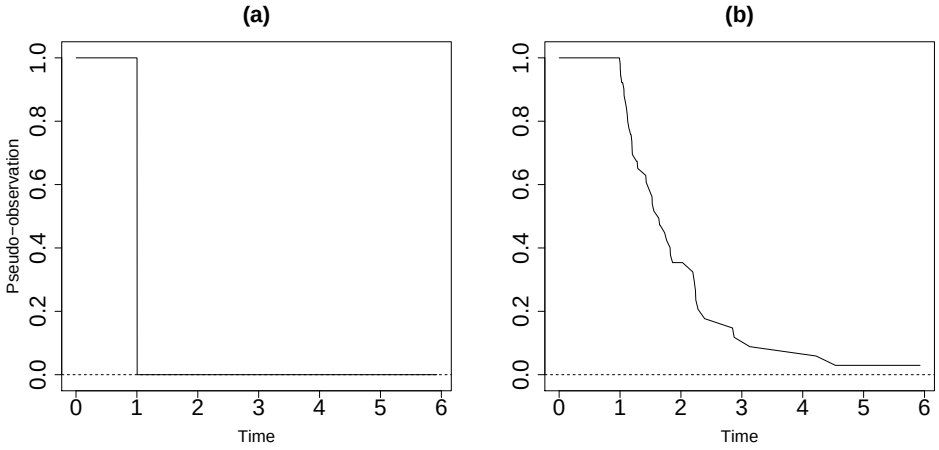

Figure 7.1: The new pseudo-observation for a single individual over time. (a) The new pseudo-observation for an individual with observed event time $X_j = 1$. (b) The new pseudo-observation over time for an individual with observed censored event time $C_j = 1$.

In chapter 5 the pseudo-observation (7.1) were used for generalised linear regression analysis of censored survival data. The idea is that once the pseudo-observations have been calculated one may use the pseudo-observations for a various number of analysis. However, calculation of the pseudo-observations (7.1) is computational heavy. Calculation of the pseudo-observations for a sample with n individuals, requires calculations of the Kaplan-Meier estimator for $n + 1$ samples. For comparison, the proposed estimator in (7.3) only requires calculation of the Kaplan-Meier estimator for a single sample.

7.1 Properties of the new pseudo-observation

In this section, some properties regarding the distribution of the *new pseudo-observations* (7.2) and (7.3) are considered.

First consider the estimator (7.2) for an individual with observed event time, this estimator may be written in terms of the indicator function $\hat{\theta}_j(t) = \mathbb{1}[X_j > t]$. The mean value of this estimator is then given by

$$\begin{aligned}\mathbb{E}[\hat{\theta}_j(t)] &= \mathbb{E}[\mathbb{1}[X_j > t]] \\ &= S_X(t).\end{aligned}$$

The variance of this estimator is given by

$$\mathrm{Var}[\hat{\theta}_j(t)] = S_X(t)\left[1 - S_X(t)\right] = S_X(t)F_X(t).$$

The estimator (7.3) is a bit more complicated as this estimator depends on both the event time distribution and the censoring distribution. Let C_j denote the censoring time for individual j and suppose this is independent of the event time X_j. Further, let $\hat{S}_{X(n)}$ denote the Kaplan-Meier estimator of the event times based on a sample with n individuals. The mean value of the estimator (7.3) is then given by

$$\begin{aligned}\mathbb{E}[\hat{\theta}_j(t)] &= \mathbb{E}[\mathbb{1}[X_j \wedge C_j > t]] + \mathbb{E}\left[\mathbb{1}[C_j \leq t \wedge X_j]\frac{\hat{S}_{X(n)}(t)}{\hat{S}_{X(n)}(C_j)}\right] \\ &= \mathbb{E}[\mathbb{1}[X_j > t]\mathbb{1}[C_j > t]] + \mathbb{E}\left[\mathbb{1}[C_j \leq t \wedge X_j]\frac{\hat{S}_{X(n)}(t)}{\hat{S}_{X(n)}(C_j)}\right] \\ &= S_X(t)S_C(t) + \mathbb{E}\left[\mathbb{1}[C_j \leq t \wedge X_j]\frac{\hat{S}_{X(n)}(t)}{\hat{S}_{X(n)}(C_j)}\right]. \qquad (7.4)\end{aligned}$$

Consider the second term of the right hand side of (7.4). The Kaplan-Meier estimator is an uniformly consistent estimator of the survival function on the interval $[0, t]$ for all $t < u$, where $u = \sup\{t : \mathrm{P}(X > t) > 0\}$ Fleming and Harrington [1991]. This means that for $t < u$ the second term of (7.4) can be rewritten as

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}[C_j \leq t \wedge X_j]\frac{\hat{S}_{X(n)}(t)}{\hat{S}_{X(n)}(C_j)}\right] &= \int_0^t \int_c^\infty \frac{\hat{S}_{X(n)}(t)}{\hat{S}_{X(n)}(c)} \mathrm{d}F_X(x)\mathrm{d}F_C(c) \\ &= \int_0^t \frac{\hat{S}_{X(n)}(t)}{\hat{S}_{X(n)}(c)} \int_c^\infty 1 \mathrm{d}F_X(x)\mathrm{d}F_C(c) \\ &\xrightarrow[n\to\infty]{} \int_0^t \frac{S_X(t)}{S_X(c)} \int_c^\infty 1 \mathrm{d}F_X(x)\mathrm{d}F_C(c) \\ &= \int_0^t \frac{S_X(t)}{S_X(c)} S_X(c)\mathrm{d}F_C(c) \\ &= S_X(t)F_C(t). \qquad (7.5)\end{aligned}$$

It follows that the mean value of $\hat{\theta}_j(t)$ in (7.3) converges in probability to the survival function of the event times:

$$\mathbb{E}[\hat{\theta}_j(t)] \xrightarrow[n\to\infty]{} S_X(t)S_C(t) + S_X(t)F_C(t) = S_X(t). \tag{7.6}$$

That is, the estimator $\hat{\theta}_j(t)$ in (7.3) is an asymptotic unbiased estimator of the survival function $S_X(t)$.

The variance of the estimator $\hat{\theta}_j(t)$ in (7.3) can be written as

$$\mathrm{Var}[\hat{\theta}_j(t)] = \mathbb{E}\left[\hat{\theta}_j(t)^2\right] - \left(\mathbb{E}[\hat{\theta}_j(t)]\right)^2.$$

The first term on the right hand side of this variance is given by

$$\begin{aligned}\mathbb{E}\left[\hat{\theta}_j(t)^2\right] &= \mathbb{E}\left[\mathbb{1}[X_j \wedge C_j > t]\right] + \mathbb{E}\left[\mathbb{1}[C_j \le t \wedge X_j]\frac{\hat{S}^2_{X(n)}(t)}{\hat{S}^2_{X(n)}(C_j)}\right] \\ &\quad +2\mathbb{E}\left[\mathbb{1}[X_j \wedge C_j > t]\mathbb{1}[C_j \le t \wedge X_j]\frac{\hat{S}^2_{X(n)}(t)}{\hat{S}^2_{X(n)}(C_j)}\right] \\ &= S_X(t)S_C(t) + \int_0^t\int_c^\infty \frac{\hat{S}^2_{X(n)}(t)}{\hat{S}^2_{X(n)}(c)^2}\mathrm{d}F_X(x)\mathrm{d}F_C(c). \end{aligned} \tag{7.7}$$

By the consistency of the Kaplan-Meier estimator, the second term on the right hand side of (7.7) converges in probability to a term given by:

$$\int_0^t\int_c^\infty \frac{\hat{S}^2_{X(n)}(t)}{\hat{S}^2_{X(n)}(c)}\mathrm{d}F_X(x)\mathrm{d}F_C(c) \xrightarrow[n\to\infty]{} \int_0^t \frac{S^2_X(t)}{S^2_X(c)}S_X(c)\mathrm{d}F_C(c) = S^2_X(t)\int_0^t \frac{1}{S_X(c)}\mathrm{d}F_C(c).$$

It follows that the variance of the estimator $\hat{\theta}_j(t)$ converges in probability to:

$$\mathrm{Var}[\hat{\theta}_j(t)] \xrightarrow[n\to\infty]{} \left[S_X(t)S_C(t) + S^2_X(t)\int_0^t \frac{1}{S_X(c)}\mathrm{d}F_C(c)\right] - S^2_X(t). \tag{7.8}$$

This variance depends on the censoring distribution and hence a proper expression of the variance can only be obtained in some special cases.

An upper bound of the variance in the limit is given by

$$\begin{aligned}\lim_{n\to\infty}\mathrm{Var}[\hat{\theta}_j(t)] &= S_X(t)S_C(t) + S^2_X(t)\int_0^t \frac{1}{S_X(c)}\mathrm{d}F_C(c) - S^2_X(t) \\ &\le S_X(t)S_C(t) + S_X(t)F_C(t) - S^2_X(t) \\ &= S_X(t)F_X(t).\end{aligned}$$

Similar, a lower bound of the variance is found by

$$\begin{aligned}\lim_{n\to\infty}\mathrm{Var}[\hat{\theta}_j(t)] &= S_X(t)S_C(t) + S^2_X(t)\int_0^t \frac{1}{S_X(c)}\mathrm{d}F_C(c) - S^2_X(t) \\ &\ge S_X(t)S_C(t) + S^2_X(t)F_C(t) - S^2_X(t) \\ &= S_X(t)F_X(t)S_C(t).\end{aligned}$$

Consider now two pseudo-observations $\hat{\theta}_i(t)$ and $\hat{\theta}_j(t)$ for individual i and j, respectively. The two pseudo-observations are independent if $\mathbb{E}[\hat{\theta}_i(t)\hat{\theta}_j(t)] = \mathbb{E}[\hat{\theta}_i(t)]\mathbb{E}[\hat{\theta}_j(t)]$. In the case of two uncensored individuals it follows from the independence of X_i and X_j that the pseudo-observations are likewise independent. When subject i is uncensored and subject j is censored, the mean value $\mathbb{E}[\hat{\theta}_i(t)\hat{\theta}_j(t)]$ is given by

$$\mathbb{E}[\mathbb{1}[X_i > t]\mathbb{1}[X_j \wedge C_j > t]] + \mathbb{E}\left[\mathbb{1}[X_i > t]\mathbb{1}[C_j \leq t \wedge X_j]\frac{\hat{S}_{X(n)}(t)}{\hat{S}_{X(n)}(C_j)}\right]. \tag{7.9}$$

By arguing as above, the limit of (7.9) is given by

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}[\hat{\theta}_i(t)\hat{\theta}_j(t)] &= S_X^2(t)S_C(t) + S_X^2(t)F_C(t)\\ &= S_X^2(t).\end{aligned}$$

This means that in this case $\hat{\theta}_i(t)$ and $\hat{\theta}_j(t)$ $i \neq j$ are asymptotic independent. Consider now the case where both $\hat{\theta}_i(t)$ and $\hat{\theta}_j(t)$ are pseudo-observations for censored individuals, the mean value $\mathbb{E}[\hat{\theta}_i(t)\hat{\theta}_j(t)]$ is then given by

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}[X_i \wedge C_i > t]\mathbb{1}[X_j \wedge C_j > t]\right] + \mathbb{E}\left[\mathbb{1}[X_i \wedge C_i > t]\mathbb{1}[C_j \leq t \wedge X_j]\frac{S_{X(n)}(t)}{S_{X(n)}(C_j)}\right]\\ &+ \mathbb{E}\left[\mathbb{1}[X_j \wedge C_j > t]\mathbb{1}[C_i \leq t \wedge X_i]\frac{S_{X(n)}(t)}{S_{X(n)}(C_i)}\right]\\ &+ \mathbb{E}\left[\mathbb{1}[C_i \leq t \wedge X_i]\frac{S_{X(n)}(t)}{S_{X(n)}(C_i)}\mathbb{1}[C_j \leq t \wedge X_j]\frac{S_{X(n)}(t)}{S_{X(n)}(C_j)}\right].\end{aligned}$$

By similar arguments as a above, the limit of the mean value $\mathbb{E}[\hat{\theta}_i(t)\hat{\theta}_j(t)]$ is then given by

$$\begin{aligned}\lim_{n\to\infty}\mathbb{E}[\hat{\theta}_i(t)\hat{\theta}_j(t)] &= S_X^2(t)S_C^2(t) + S_X^2(t)S_C(t)F_C(t) + S_X^2(t)S_C(t)F_C(t) + S_X^2(t)F_C^2(t)\\ &= S_X^2(t)[1 + F_C^2(t) - 2F_C(t)] + 2S_X^2(t)F_C(t)S_C(t) + S_X^2(t)F_C^2(t)\\ &= S_X^2(t) + 2S_X^2(t)F_C^2(t) - 2S_X^2(t)F_C(t) + 2S_X^2(t)F_C(t)S_C(t)\\ &= S_X^2(t).\end{aligned}$$

From this is follows that the two pseudo-observations $\hat{\theta}_i(t)$ and $\hat{\theta}_j(t)$, $i \neq j$ are asymptotically independent.

7.2 Regression analysis based on the new pseudo-observation

The idea introduced by Andersen et al. [2003] of performing regression analysis based on the *old pseudo-observation* (7.1) may be adapted to the new pseudo-observation in (7.2) and (7.3).

Consider the generalised linear model

$$g\left(\mathbb{E}[\mathbb{1}[X_j > t]\big|\boldsymbol{Z}_j]\right) = \alpha + \boldsymbol{\beta}^\top \boldsymbol{Z}_j,$$

	New pseudo-observations			Old pseudo-observation		
Parameters	**Est.**	$\mathbf{SD_{sim}}$	$\mathbf{SE_{est}}$	**Est.**	$\mathbf{SD_{sim}}$	$\mathbf{SE_{est}}$
$\boldsymbol{\beta_1}$	2.79	0.21	0.20	3.06	0.25	0.25
$\boldsymbol{\beta_2}$	−0.92	0.09	0.09	−1.02	0.11	0.11

Table 7.1: The results after fitting a GEE model based on the new pseudo-observation and the old pseudo-observation, respectively. The results shows the average of the estimated regression parameters and their standard deviations of 500 replicates of the data. Furthermore, the averages of the standard errors of the regression parameters are likewise given.

where $g(\cdot)$ is some link function and the intercept α may depend on the time t. The idea is to replace the function $\phi_t(X_j) = \mathbb{1}[X_j > t]$ by the pseudo-observation when fitting the model. Graw et al. [2009] showed that the GEE model gives consistent and normally distributed estimates of the regression parameters when the function $\phi(\cdot)$ is replaced by the pseudo-observation (7.1). An essential part of the proof by Graw et al. [2009] rely on the property

$$\mathbb{E}[\hat{S}_j(t)|\boldsymbol{Z}_j] = g^{-1}(\alpha + \boldsymbol{\beta}^\top \boldsymbol{Z}_j) + o_p(1). \tag{7.10}$$

For the asymptotic results to hold for regression analysis based on the new pseudo-observation, one might be able to argue similar as in Graw et al. [2009]. However, at the due date of this project I had not been able to show the property (7.10) for the new pseudo-observation.

Example 7.1 (Regression analysis on the survival probability) In example 5.4 a model corresponding to the Cox proportional hazards model were defined by fitting a GEE model based on the pseudo-observation (7.1) with link function given as the cloglog-function. A similar model is obtained by regression analysis based on the new pseudo-observation. In this example, the performance of the new pseudo-observation is compared to the old pseudo-observation by fitting a GEE model with link function given as the cloglog-function.

Survival data ($n = 250$) were simulated from a Cox proportional hazards model with constant baseline hazard function $h_0(t) = 2.5$. In the data, two covariates were included; a uniform distributed covariate z_1 over the interval $[-1, 1]$ and a standard normal distributed covariate z_2, with parameters $\beta_1 = 3$ and $\beta_2 = -1$, respectively. Exponential distributed data were superimposed to obtain roughly 25% censoring. The GEE model were fitted by calculating the pseudo-observations at 10 equally spaced study times, and the pseudo-observations defined for each individual were supposed to be independent. Each analysis were replicated 500 time. The average of the estimated regression parameter and the corresponding standard deviations were calculated. The average of the estimated standard errors of the parameters was likewise calculated based on the sandwich estimator. The results of the analysis are given in table 7.1.

The results show a considerable large bias on the estimated regression parameters based on the new pseudo-observation compared to the results based on the old pseudo-observation. However, the variability of the estimates based on the new pseudo-observation is considerable smaller than the estimates based on the old pseudo-observation.

	The exact covariance matrix			The common covariance matrix		
Parameters	**Est.**	**SD_{sim}**	**SE_{est}**	**Est.**	**SD_{sim}**	**SE_{est}**
β_1	2.78	0.20	0.20	2.80	0.21	0.20
β_2	-0.92	0.10	0.09	-0.92	0.09	0.09

Table 7.2: The results after fitting a GEE model based on the new pseudo-observation. The results shows the estimates based on a model with the exact covariance matrix and the common covariance matrix, respectively.

In example 5.5 two addtional working covariance matrices were considered for the GEE model based on the old pseudo-observation. The covariance matrices considered were a covariance matrix, which is the exact covariance when no censoring occurs in the data and a common working covariance matrix V as given in (5.15). For an individual with observed event time, the new pseudo-observation may be written in terms of the indicator function $\hat{\theta}_j(t) = \mathbb{1}[X_j > t]$, and hence an exact covariance matrix of the pseudo-observations is given by a matrix with elements

$$v_{ilm} = S_i(t_l|\boldsymbol{Z}_j)[1 - S_i(t_m|\boldsymbol{Z}_j)], \quad t_l < t_m. \tag{7.11}$$

In table 7.2 the results of fitting the same data to a GEE model based on the new pseudo-observation with working covariance matrix given by the exact covariance (7.11) and the common covariance given in (5.15) is given. The table shows the average of the estimated parameters and their standard deviation of 500 replicates of the analysis. The average of the standard errors of the parameters estimated by the sandwich estimator is likewise given. Similar to the results found in example 5.5, no systematic different were found when using different covariance matrices.

□

Example 7.2 (Estimation of $h_0(\cdot)$ in the Cox proportional hazards model) In example 6.1 cubic regression splines were used to estimate the the baseline hazard function in the Cox proportional hazards model. The model considered were given by

$$\text{cloglog}(S(t|\boldsymbol{Z})) = s_{\boldsymbol{\tau}}(t) + \boldsymbol{\beta}^\top \boldsymbol{Z}, \tag{7.12}$$

where

$$s_{\boldsymbol{\tau}}(t) = \sum_{i=0}^{r+3} \gamma_i B_{i,3}(t),$$

is a cubic spline function.

In this example the model (7.12) is fitted based on the new pseudo-observations. Survival data ($n = 250$) were simulated from a Cox regression model with constant baseline hazard function $h_0(t) = 2.5$. Two covariates were considered; a uniform distributed covariate z_1 over

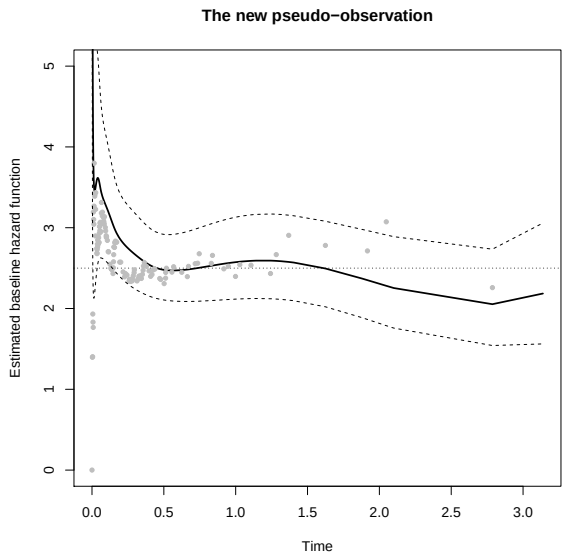

Figure 7.2: The baseline hazard function estimated by the regression spline. The solid line indicates the estimate of the baseline hazard function and the dotted lines indicate af 95% CI. The dots indicate the point estimates of the baseline hazard function proposed by Andersen et al. [2003].

the interval $[-1, 1]$ and a standard normal distributed covariate z_2, with parameters $\beta_1 = -1$ and $\beta_2 = 1$, respectively. Exponentially distributed data were superimposed to obtain roughly 25% censoring. A GEE model based on the pseudo-observation calculated at 125 equidistance time points. The knots used in the estimation of the spline function were given by the 25'th, 50'th, and the 75'th quantile of the study times. For comparison the point estimates of the baseline hazard function suggested by Andersen et al. [2003] is likewise calculated. Figure 7.2 shows the estimated baseline hazard function based on the new pseudo-observation. A 95% pointwise CI is indicated by dotted lines. The dots indicates the point estimates of the baseline hazard function suggested by Andersen et al. [2003].

Similar to example 6.1, the estimated baseline hazard function seems to be agreeable with the true baseline hazard function $h_0(t) = 2.5$. For comparison with the results based on the old pseudo-observation see figure 6.2. However, the 95% CI of the estimated baseline hazard function based on the new pseudo-observation is much smaller than the 95% CI of the estimate based on the old pseudo-observation. The trend is especially apparent at the end of the time scale where few observations occur. This observation is in tune with the results found in example 7.1.

□

Example 7.3 (The resricted mean survival time) In example 5.2 an estimator of the restricted mean survival function $\mathbb{E}[\min(X_j, \tau)]$ for $\tau > 0$ were found by integrating over the pseudo-observation (7.1). A similar estimator is obtained by integrating over the estimator

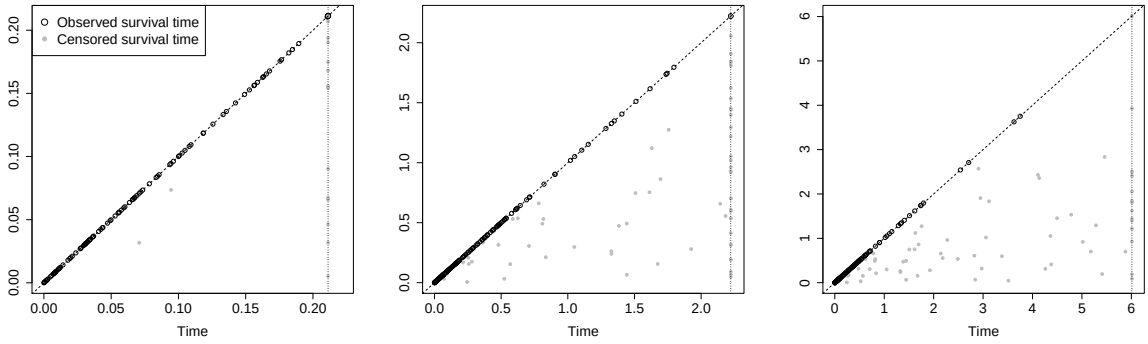

Figure 7.3: Crude restricted mean survial time

(7.2) for individuals with observed event time and the estimator (7.3) for individuals with observed censored event time.

For an individual with observed event time, the estimator (7.2) may be written in terms of the indicator function $\hat{\theta}_j(t) = \mathbb{1}[X_j > t]$, and hence an estimate of the restricted mean survival time is simply given by the true restricted survival time

$$\hat{\mu}_\tau = \min(X_j, \tau).$$

For an individual with observed censored event time, the estimate of the restricted mean function is given by

$$\begin{aligned}\hat{\mu}_\tau &= \int_0^\tau \hat{\theta}_j(t)\mathrm{d}t \\ &= \int_0^{T_j} 1\mathrm{d}t + \int_{T_j}^\tau \frac{\hat{S}(t)}{\hat{S}(T_j)}\mathrm{d}t \\ &= T_j + \frac{1}{\hat{S}(T_j)}\int_{T_j}^\tau \hat{S}(t)\mathrm{d}t.\end{aligned}$$

In figure 7.3 the pseudo-observations for the conditional restricted mean survival time are plotted against the true restricted survival time for a simulated data set. The plot shows the estimated conditional restricted mean survival time for three different values of τ; the median of the observed study times (left), the 95th percentile of the observed study times (middle), and the maximum observed study time (right). For individuals with censored event time the estimated conditional restricted mean survival time is smaller than the true survival time in all three plots. For individuals with observed event time the estimated value of the conditional restricted mean survival time follows the true survival time.

□

8 Discussion and conclusion

The inherent structure of survival data entail that standard statistical models cannot be used for analysis this type of data. Especially the occurrence of censored data is a technicality, which must be dealt with when analysing survival data. Survival data is often summarized through the hazard function and the survival function. These functions serve to describe important aspects of the survival experience in a given data set. The mean survival time is another parameter of great interest, which gives the expected time until some event occur. However, due to right censoring the mean survival time is often ill determined beyond a certain range. A related quantity is the restricted mean survival time which is less sensitive to right censoring. Traditional regression analysis of survival data is hence concentrated on relating covariates to these functions, with the hazard function playing the most prominent role. Often the functions themself are not of primary interest. From a practical point of view, one is often more interested in the effect of the covariates. From this perspective, semi-parametric models have been used as the primary tool for regression analysis of survival data; here the Cox proportional hazards model is one of the most popular methods.

In this project jackknife pseudo-observations have been studied as a tool for analysing survival data. Pseudo-observations adress the problem of not having appropriate observations for all individuals in the study. The approach is based on a set of pseudo-observations defined for each individual in the study. The pseudo-observations allow one to analyse the effect of a set of covariates on the event times by models more general than the standard models used for survival data. In section 5.2 the GEE approach based on pseudo-observations were used to fit a generalised linear model for functions of the event times. The advantage of this approach is that it allows one to model the event times by generalised linear models without specifying a full parametric model. The justification for the use of pseudo-observations in the GEE approach is based on some neat results found by Graw et al. [2009] concerning the estimated regression parameters.

In section 5.2, regression analysis based on pseudo observations were compared to the traditional Cox proportional hazards model. A small simulation study showed that the model

based on pseudo-observations is not competitive to this standard method of analysing survival data. The weakness of the pseudo-observation approach is mainly the relative large standard errors of the estimated regression parameters; this variation is related to the variability in the set of pseudo-observations. The question regarding the number of time points for which the pseudo-observations are calculated were also considered. The results showed that increasing the number of equidistant time points did not increase the performance of the model. However, in a general setting it is reasonable to believe that the number and position of the time points will affect the performance of the regression analysis. For the regression analysis to perform well, the survival structure in the true event distribution must be captured by the pseudo-observations. Hence, improper positions of the pseudo-observations will reasonable effect the efficiency of the analysis.

In chapter 6 cubic regression splines based on pseudo-observations were used to estimate the baseline hazard function of the Cox proportional hazards model. This is a simple method which allow for a flexible estimation of the baseline hazard function. A simple example with a constant baseline hazard function was considered. The results show that the approach is a reliable method for estimation of this hazard function. In general cubic regression splines are flexible functions and it is reasonable be believe that the method will perform well for more complex functions. In chapter 6 it was argued that when the error term resulting from the spline approximation of the unknown function is negligible, the resulting regression parameters will possess the same desirable properties found by Graw et al. [2009]. However, if the spline approximation introduce a large error term, a corresponding bias on the estimated parameters is expected.

In chapter 7 a new pseudo-observation was proposed for estimating survival probabilities. The motivation behind the definition of this new pseudo-observation is the large estimated variance obtained from regression analysis based the old pseudo-observation. The new pseudo-observation was defined to more proper capture the information given by the observed data. A small simulation study showed that the new pseudo-observation tend to give estimated regression parameters with less variability compared to the old pseudo-observation. Though, a large bias on the regression parameters resulting from this new pseudo-observation was observed. During the period of this project attempt has been made to prove asymptotic properties for the new pseudo-observation similar to those found by Graw et al. [2009]. However, a bias related to the censoring distributed was encountered. This suggests that the range of the new pseudo-observations might be inadequate for compensation of the uncertainty related to the survival experience for censored individuals. In contrast, the definition of the old pseudo-observation causes this estimator to take values outside the range of $[0, 1]$, from which the variability in the regression results is caused by. However, these unusually values seem to compensate each other to produce estimates capturing the information in the true underlying event distribution. This discussion suggests a trade-off between bias and variation on the estimated regression results. The old pseudo-observation seems to produce estimates which on average produce accurate estimates. However, the variability makes these estimates less reliable in a single sample. The potential of the new pseudo-observation rely on its ability to produce regression results with high reliability, though accounting for the bias.

One may consider to reduce the bias of the estimated regression parameters resulting from the new pseudo-observation by a two-step estimation method. The first step is to estimate some initial regression parameters based on the sample of individuals with observed event time only. The new pseudo-observation for individuals with observed event time is simply given by $\hat{\theta}_j(t) = \mathbb{1}[X_j > t]$. It follows directly from the results of Liang and Zeger [1986] that this estimator will produce consistent estimates of the true regression parameters, though a large variance is expected due to omitting censored individuals. The next step is then to predict the value of $S(\cdot|\boldsymbol{Z}_j)$ for censored individuals, based on these initial parameters. The new pseudo-observation for censored individuals given in (7.3) may then be modified as follows

$$\hat{\theta}_j(t) = \begin{cases} 1 & \text{for} \quad t < C_j \\ \frac{S(t|\boldsymbol{Z}_j)}{S(C_j|\boldsymbol{Z}_j)} & \text{for} \quad C_j \leq t \end{cases} .$$

The final regression analysis is then based on all individuals in the study, including censored individuals. Though, rather than using the estimator proposed in chapter 7, the analysis is based on the above estimator for censored individuals. This estimator will increase the computational burden, however, the bias of the resulting regression parameters is expected to decrease will maintaining the small variance. Due to time pressure, this method has not been considered in the study.

The generalisation of jackknife pseudo-observations for analysis of survival data a is relative new method with the first article printed in 2003 (Andersen et al. [2003]). The method is not yet fully developed, and more reseach is needed to answer a number of unanswered questions, like the best choice regarding the number and position of the time points for which the pseudo-observations are calculated. In this project, the efficiency of the pseudo-observations has been addressed together with some practical considerations for capable regression results. The idea of pseudo-observations for survival analysis has been proposed in a general setting, though studies of the pseudo-observations have so far been restricted to three special estimators. The results found in this project show that the pseudo-observations have a potential as a general method for analysing survival data in a number of ways. The potential of the pseudo-observation relies on their ability to generalise survival data by introducing proper observations for all individuals in the study. This allows for analysis of survival data by methods usually restricted to data with full information. The approach based on pseudo-observations is not competetive to traditional methods used to analyse survival data. However, the pseudo-observations alows for general statistical methods to supplement the tradtional survival data methods.

9 Bibliography

Andersen and Perme, **2010**. Per Kragh Andersen and Maja Pohar Perme. *Pseudo-observations in survival analysis.* Statistical Methods in Medical Research, 19(1), 71–99, 2010.

Andersen, Borgan, Gill, and Keiding, **1993**. Per Kragh Andersen, Ørnulf Borgan, Richard D. Gill, and Niels Keiding. *Statistical models based on counting processes.* Springer-Verlag, 1993. ISBN 0-387-97872-0.

Andersen, Klein, and Rosthøj, **2003**. Per Kragh Andersen, John P. Klein, and Susanne Rosthøj. *Generalised linear models for correlated pseudo-observations, with applications to multi-state models.* Biometrika, 90(1), 15–27, 2003.

Azzalini, **2002**. A. Azzalini. *Statistical Inference Based on the likelihood.* Chapman & Hall, 2002. ISBN 978-0-4126-0650-2.

Binder, Gerds, and Andersen, **2012**. Nadine Binder, Thomas A. Gerds, and Per Kragh Andersen. *Pseudo-observations for competing risks with covariates dependent censoring - research report 12/06*, Department of Biostatistics University of Copenhagen, 2012.

Breslow, **1974**. N. Breslow. *Covariance Analysis of Censored Survival Data.* Biometrics, 30 (1), 89–99, 1974.

Cox, **1972**. D.R. Cox. *Regression Models and Life-Tables.* Journal of the Royal Statistical Society. Series B (Methodological), 34(2), 187–220, 1972.

Cox, **1975**. D.R. Cox. *Partial likelihood.* Biometrika, 62(2), 269–276, 1975.

Boor, **1978**. Carl de Boor. *A practical guide to splines.* Springer, 1978. ISBN 0-387-90356-9.

Efron, **1977**. Bradley Efron. *The Efficiency of Cox's Likelihood Function for Censored Data.* Journal of the American Statistical Association, 72(359), 557–565, 1977.

Fitzmaurice, Davidian, Verbeke, and Molenberghs, **2008**. Garrett Fitzmaurice, Marie Davidian, Geert Verbeke, and Geert Molenberghs. *Longitudinal data analysis.* Chapman & Hall/CRC, 2008. ISBN 978-1-58488-658-7.

Fleming and Harrington, **1991**. Thomas R. Fleming and David P. Harrington. *Counting processes and survival analysis.* Chichester: Wiley, 1991. ISBN 0-471-52218-x.

Graw, Gerds, and Schumacher, **2009**. Frederik Graw, Thomas A. Gerds, and Martin Schumacher. *On pseudo-values for regression analysis in competing risks models.* Lifetime Data Analysis, 15(2), 241 –255, 2009.

Green and Silverman, **1994**. P.J. Green and B.W. Silverman. *Nonparametric regression and generalized linear models.* London: Chapman & Hall, 1994. ISBN 0-412-30040-0.

Hastie, Tibshirani, and Friedman, **2001**. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning : data mining, inference, and prediction.* New York: Springer, 2001. ISBN 0-387-95284-5.

Hosmer and Lemeshow, **1999**. David W. Hosmer and Stanley Lemeshow. *Applied survival analysis: regression modelling of time to event data.* Wiley, 1999. ISBN 0-471-15410-5.

Kalbfleisch and Prentice, **2002**. John D. Kalbfleisch and Ross L. Prentice. *The statistical analysis of failure time data.* Wiley-Interscience, 2002. ISBN 978-0-471-36357-6.

Klein and Andersen, **2005**. John P. Klein and Per Kragh Andersen. *Regression Modeling of Competing Risks Data Based on Pseudovalues of the Cumulative Incidence Function.* Biometrics, 61(1), 223–229, 2005.

Klein and Moeschberger, **1997**. John P. Klein and Melvin L. Moeschberger. *Survival analysis: techniques for censored and truncated data.* Springer-Verlag, 1997. ISBN 0-387-94829-5.

Liang and Zeger, **1986**. Kung-Yee Liang and Scott L. Zeger. *Longitudinal data analysis using generalized linear models.* Biometrika, 73(1), 13–22, 1986.

Miller, **1974**. Rupert G. Miller. *The jackknife - a review.* Biometrika, 61(1), 1–15, 1974.

Perme and Andersen, **2008**. Maja Pohar Perme and Per Kragh Andersen. *Checking hazard regression models using pseudo-observations.* Statistics in medicine, 27(25), 5309–28, 2008.

Tukey, **1958**. J. W. Tukey. *Bias and confidence in not quite large samples (abstract).* Annals of Mathematical Statistics, 29, 614, 1958.

A

Appendix

A.1 The cumulative incidence function

The section is written based on Binder et al. [2012].

In chapter 5 the Aalen-Johansen estimator were used to define pseudo-observations for the cumulative cause-r incidence function in a competing risks analysis. In section 5.1 it was claimed that in the case of no censoring, the pseudo-observation reduces to an indicator function $\hat{F}_{jr}(t) = \mathbb{1}[X_j \leq t, \epsilon_j = r]$. The justic for this is given below. Suppose a sample with n individuals are given and let $N_r(\cdot)$ be the counting process defined by

$$N_r(t) = \sum_{j=1}^{n} \mathbb{1}[T_j \leq t, \delta_j = 1, \epsilon_j = r],$$

where $T_j = \min(X_j, C_j)$, $\delta_j = \mathbb{1}[X_j \leq C_j]$, and ϵ_j indicates which competing risks caused the event. Further, let $Y(\cdot)$ denote the risk set and let $\hat{H}_r(\cdot)$ denote the Nelson-Aalen estimator for the cause-r specified hazard function.

Consider the Aalen-Johansen estimator of the cause-r cumulative incidence function given by

$$\begin{aligned}\hat{F}_r(t) &= \int_0^t \prod_{v<u} \left(1 - \frac{\sum_{r=1}^{K} \mathrm{d}N_r(v)}{Y(v)}\right) \mathrm{d}\hat{H}_r(u) \\ &= \int_0^t \hat{S}_X(u^-)\mathrm{d}\hat{H}_r(u),\end{aligned}$$

where $\hat{S}(\cdot)$ denote the Kaplan-Meier estimator of the survival function for all event times in the study and K is the number of competing risks. It can be shown that $\frac{Y(t)}{n} = \hat{S}_X(t^-)\hat{S}_C(t^-)$,

where $\hat{S}_C(\cdot)$ is the Kaplan-Meier estimator for the censoring disribution. From this it follows that the Aalen-Johansen estimator can be rewritten as

$$\hat{F}_r(t) = \int_0^t \hat{S}_X(u^-)\frac{\mathrm{d}N_r(u)}{Y(u)} = \int_0^t \hat{S}_X(u^-)\frac{\mathrm{d}N_r(u)}{n\hat{S}_X(u^-)\hat{S}_C(u^-)} = \frac{1}{n}\sum_{j=1}^{n}\int_0^t \frac{\mathrm{d}N_r(u)}{\hat{S}_C(u^-)}. \tag{A.1}$$

When no censoring occur in the data $\hat{S}_C(t) = 1$ and the Aalen-Johansen estimator in (A.1) reduces to

$$\hat{F}_r(t) = \frac{N_r(t)}{n}.$$