UiO : Faculty of Mathematics and Natural Sciences
University of Oslo

Log-linearity for Cox's regression model

Thesis for the Degree Master of Science

Zaki Amini

Master's Thesis, Spring 2015

Abstract

Cox's regression model is one of the most applied methods in medical research. This method finds also applications in other fields such as econometrics, demography, insurance etc. This method is based on two crucial assumptions that (i) the method assumes log-linearity in covariates, and (ii) that the hazard ratio of two individuals are proportional. In survival analysis data, both numeric and binary covariates are typically encountered. There is no issue with the log-linearity assumption when working with binary covariates, however, the issue may arise when numeric covariates are involved. This thesis, thus, studies methods that are used to check assumption (i). For this purpose, there have been proposed a number of graphical procedures and formal test procedures in the literatures. This thesis in particularly aims to give a systematic review of the various test procedures and formal tests, and also to assess how the test procedures perform.

All the proposed test procedures will be illustrated using publicly available data. To study the performance of these procedures, both real and simulated data (using the Monte Carlo method) will be used. In the simulation studies, first we must find a general formula for how to generate survival data on the computer. That is done through the fundamental relation between hazard rate and survival function. It is shown how the Weibull distribution function can be used to generate appropriate survival data on the computer.

KEY WORDS: Cox's regression model; Survival analysis; Hazard rate; Censoring; Local test statistics; Fractional polynomials; P-spline; Martingale residuals; Monte-Carlo.

Preface

This thesis was written as part of my master's degree, in the master program "Modeling and Data analysis" offered by the University of Oslo. I have really learned a lot working on this thesis - learned much more than I would ever imagine when I started working on this. In particular, throughout this thesis I learned much about a proper academic writing (e.g. proper use of sources etc.) both in terms of content and grammar.

First and foremost, I want to thank my amazing supervisor, Ørnulf Borgan, for giving me a very interesting and challenging thesis. He was always very helpful giving me useful advice and feedback, and I always felt I was welcome to drop by his office and also to send him e-mails. Without his guidance, this thesis would be an impossible task. I really appreciated the amount of feedback I received from him.

Moreover, I also want to thank Kjersti Nygaard, Maja Alvilde Bratlien Larsen and my good friend Even Langfeldt Friberg who have given me useful comments on some parts of the thesis. I am really grateful for this. I also appreciate other fellow students, and in particular students at room B800, to have contributed to a great study environment (again especially thanks to Kjersti Nygaard for taking the initiative to social events). A special thanks to my family who always encouraged me to complete my degree.

Zaki Amini
May 2015

Contents

Chapter 1

Introduction

The main aim of regression models is to asses the effect of covariates on an outcome. All types of regression models have their own application areas. For example, in linear regression model, one assumes that the relation between the response variable and the predictor variables (covariates) is linear. Therefore linear regression is used to analyse data that has such properties.

In survival and event-history analysis, the response variable is called a *survival time*. By a survival time, we mean the time from an initiating event to the event of interest. Examples of survival times would be time from birth to death, time from marriage to divorce, or time from a patient gets cancer treatment until relapse or death. Here, the event of interest are birth, death and divorce. The event of interest typically occurs for some individuals under observation, but not for some other individuals. For example, in a lung cancer study, patients are asked to participate in a study to examine the effect of a drug on their survival from lung cancer. Some of the patients take part in the study until their death before the study terminates, then their survival times is considered as *uncensored*. Some others take part in the study, but before the study ends, they are lost to follow-up, then their survival time is considered as *censored*. The third category take part in the study, but after a while before the study terminates, they die of other causes than death (event of interest) due to lung cancer, then their survival times are also considered as censored. Thus, the data that arises at the end of the study is a mixture of complete and incomplete observations. Thus, to analyse such censored survival data requires different statistical tools than, for example, ordinary statistical methods such as linear regression. With censored survival data we are even not capable to calculate the sample mean of the data.

There has been developed several advanced methods to handle censored survival data, and **Cox's regression model** which is the focus of this thesis is one of the most used method in medical research. The method also finds applications in other fields such as demography, econometric, insurance, and reliability engineering. According to Van Noorden et al. [2014], Cox's original paper (Cox, 1972) is the second most cited paper in the history of statistics.

Cox regression assumes that the hazard rate for individual i takes the form

$$h(t|\mathbf{x}_i) = h_0(t) \exp\{\sum_{j=1}^{p} \beta_j x_{ij}\} \tag{1.1}$$

where $h_0(t)$ is called the baseline hazard, and β_j is the coefficient associated with the j-th covariate x_{ij}.

The Cox's regression model is based on two key assumptions: (i) that hazard rate assumes log-linearity in covariates, and (ii) that the hazard rates of two individuals are proportional.

Our task here is to check the first of these assumptions. This is the main objective of this master's thesis. In the statistical literatures, there have been suggested a number of graphical and formal tests for checking log-linearity assumption. In particular, we will try to provide a systematic review of the various graphical procedures and formal tests that have been proposed.

Chapter 2 reviews the background materials on survival analysis such as survival function, hazard rate, counting processes, Nelson-Aalen estimates, Kaplan-Meier estimates and Cox's regression. In addition, we will utilize the formal test statistics such as the Likelihood ratio test (LRT), the Wald test and the Score test to examine both simple null hypothesis and composite null hypothesis. To illustrate these procedures, we will be using publicly available data.

In Chapter 3, we will study all the proposed methods for checking log-linearity of the Cox regression. In particular,

- Methods that impose an additional term on (1.1), such as $\gamma_j g(x_j)$, where e.g. $g(x) = x^2$. And then testing the null hypothesis that $\gamma_j = 0$.
- Methods that impose one or more terms on (1.1) based on making cathegorical variables according to their quartiles. And then testing the null hypothesis that all the coefficients associated with the cathegorical covariates are zero.
- A more advanced method that extends (1.1) is based on fractional polynomials (FP) which allows to integrate logarithm, non-integer powers and possibly repeated powers of the covariates. And then testing the null hypothesis similarly as under the two simple methods.
- A more flexible approach than the FP method is to replace $\beta_j x_{ij}$ in (1.1) by $s(x_{ij})$ which is a linear combination of the B-spline basis functions, $f_j(x_i)$. That is, $s(x_{ij}) = \sum_{k=1}^{n} \gamma_k f_k(x_{ij})$. This method includes both plots and test procedures.
- The last method we will consider is the Martingale-based residuals method. This method is based on counting process and cumulative intensity processes. This method also includes both plots and tests procedures.

For illustrative purposes we will be using two publicly available datasets; the melanoma and the pbc datasets. The first one is described in section 2.3.1 whereas the last one is described in section 2.8.5. The melanoma dataset is used both in Chapter 2 and 3 for illustrative purposes, while the pbc dataset is used only in Chapter 3.

Chapter 4 presents checking the log-linearity assumption through simulation studies. In particular, we generate survival data through the Cox model (1.1) which is based on baseline hazard and hazard ratio. We will use parametric distributions such as Weibull distribution to formulate the survival times modeling, which is derived through baseline hazard $h_0(t)$. Next we will utilize the software **R** to generate experimental survival data in the computer. When we have the survival data at our disposal, the next step is to utilize the various tests procedures which we have developed in Chapter 3 for analyzing the generated data and draw conclusion based on how they perform.

In Chapter 5, we will summarize our findings and draw conclusion. Appendix part is reserved for part of the analysis that are not of very significant importance to understand. Thus, in this part we attach derivation of equations, extra figures and codes that are not considered being the main results. However, it can be of importance for justification purposes how the main results are obtained.

The statistical software **R** (R Development Core Team) will be used in all the analysis, both when using real datasets such as the melanoma and the pbc datasets, and also when generating survival data by means of the Monte-Carlo techniques.

Chapter 2

Background

2.1 Introduction

In this chapter we will review results that are of importance in our study. In particular, we will summarize survival function, hazard rate, counting processes, Nelson-Aalen estimator, Kaplan-Meier estimator, Cox regression and some formal tests as Wald test, Likelihood-ratio test and the Score test. The book by Aalen et al. [2008] will be used as a reference throughout this review. Further, the melanoma data will be used to illustrate our results. The dataset is described in section 2.3.1.

2.2 Survival function and cumulative hazard rate

Before we define the *survival function*, we need to clarify the term *survival time* T. One denotes the time from an initiating event to the event of interest as *survival time* T. The event of interest could be death, relapse, divorce, or failure as in reliability engineering. For instance, the *survival time* T may be the time from birth to death, time from the first birth to the second birth of a woman, time from entry to a study to relapse, or the time to failure of a component or a system.

The survival function denoted formally as $S(t) = Pr(T > t)$ is the probability that the survival time T is larger than time t. Or phrased slightly differently, the survival function specifies the probability that the event of interest has not occurred yet by time t. The survival curve is a probability curve which starts at 1 and decreases as time goes by. Later we will show that the survival function $S(t)$ can be estimated and plotted by means of the Kaplan-Meier estimator.

As we saw just above, the survival function $S(t)$ is an unconditional probability function, while the *hazard rate* function $h(t)$ is a conditional probability function. We assume that T is an absolutely continuous random variable, and we need to look at those individuals who experience the event of interest in a small time interval $[t, t + dt]$ given that the individuals haven't experienced the event of interest yet. Or $h(t)dt$ is the probability that the event of interest

occurs in a small time interval $[t, t+dt]$ given that it has not occurred earlier.

Mathematically, the hazard rate function can be expressed by

$$h(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} Pr(t \leq T \leq t + \Delta t | T \geq t) \tag{2.1}$$

where Δt is a small time step and we let it approach 0. We note that we also divide the expression on its interval length Δt.

The cumulative hazard rate is defined as

$$H(t) = \int_0^t h(s)ds \tag{2.2}$$

There are two important mathematical relations between the cumulative hazard rate and the survival function. The relations are derived as follows:

$$\begin{aligned} H'(t) = h(t) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} Pr(t \leq T \leq t + \Delta t | T \geq t) \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{Pr(t \leq T \leq t + \Delta t)}{P(T \geq t)} \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{S(t) - S(t + \Delta t)}{S(t)} \\ &= -\frac{S'(t)}{S(t)} \end{aligned} \tag{2.3}$$

Thus $-\frac{S'(t)}{S(t)} = h(t)$, assuming that $S(0) = 1$, then by integrating both sides, we arrive at $-\log(S(t)) = \int_0^t h(s)ds + C$. Finally this implies that

$$S(t) = \exp\{-\int_0^t h(s)ds\} = \exp\{-H(t)\}, \text{ since } C = 0. \tag{2.4}$$

The survival function $S(t)$ may be estimated by Kaplan-Meier estimator. We will illustrate this in section 2.6. The estimation of $H(t)$ is done by the Nelson-Aalen estimator. In section 2.5 this will be discussed.

2.3 Data and Censoring

What makes survival analysis so special is that we can not use ordinary statistical methods due to *censoring*. In the study of survival data, one has to wait for the event to occur. When the study ends and the analysis begins, we commonly note that the event of interest has occurred for some individuals and for some others not. We then have two types of data; complete and incomplete data. The latter is called *censoring* in survival and event history analysis.

In the study of survival data, it is common to encounter the concepts *right-censoring* and *left-truncation*. In particular, *right-censoring* is almost inevitable. Right-censoring occurs when the event of interest has not occurred at the end of the study. However, in real-life studies, right-censoring can also occur when

an individual withdraws from the study or is lost to follow-up. Formally, we denote $\tilde{T}_i$ be the observed survival time for individual i , which is either the true survival time T_i or the the censored survival time C_i. Then the right-censored survival time occurs when $C_i < T_i$, in which the true survival time is "to the right" of observed time. Thus, the observation from an individual is the pair $(\tilde{T}_i, \delta_i)$, where the censoring indicator δ_i is defined by

$$\delta_i = \begin{cases} 1, & \text{if} \quad \tilde{T}_i = T_i \\ 0, & \text{if} \quad \tilde{T}_i = C_i \text{ in which case } \tilde{T}_i < T_i \end{cases}$$

In some studies patients may come under observation after the initiating event. For example, in a study of myocardial infarction only those who survive and reach the hospital can be included in the study. Those who do not survive are therefore not included in the study. The data arising here is left-truncated. There are subtypes of right-censoring and truncation which we do not consider here.

2.3.1 Melanoma data

In this thesis, the melanoma dataset will be used for illustrative purposes. In the period 1962-77 a total of 205 patients with malignant melanoma (cancer of the skin) were operated at Odense University hospital in Denmark. The tumor was completely removed, including the skin within 2.5 cm around it. This was historically a clinical study with the objective of assessing the effect of risk factors (covariates) on survival. A number of covariates were recorded at the operation. Among the covariates considered to be of significance were *sex* and *age at operation* of the patients. Other clinical characteristics covariates such as *tumor* width and location on the body were considered as well as some histological classification including tumor thickness, growth patterns, types of malignant cells and ulceration. The latter covariate is scored as "present" if the surface of melanoma viewed in a microscope shows signs of ulcers and as "absent" otherwise.

We note that the survival time is only known for those patients who died before the end of 1977. The rest of the patients were considered as *censored*. The covariates we will use in our illustrations are coded as follows:

- status: 1=death from disease, 2=censored, 4=death from other cause
- lifetime: life time from operation in years
- ulcer: ulceration (1=present, 2=absent)
- tumor thickness in mm
- sex: 1=female, 2=male
- age at operation in years

The data are further described in Andersen et al. [1993, page 11].

2.4 Counting processes

In this section we describe how *counting processes* with their *intensity processes* are derived from the survival times. Further, we only consider the *censored* survival times.

Informally, a *counting process* is a process that counts the number of occurrences of an event over time. Examples of counting processes can be counting the number of time a person wakes up during night, getting the number of children in a family or counting deaths in patient groups.

We denote $\tilde{T}_i$ to be the right censored survival time of an individual i and δ_i denotes the indicator function that $\tilde{T}_i$ corresponds to the occurrence of the event. More precisely,

$$\delta_i = \begin{cases} 1 \text{ if } \tilde{T}_i = T_i \text{ (observed actual survival time)} \\ 0 \text{ if } \tilde{T}_i < T_i \text{ (observed right-censored survival time)} \end{cases}$$

Before we go further, we need to define a concept that is called *independent right-censoring*. Formally, the independent right-censoring is defined as

$$P(t \leq \tilde{T}_i < t + dt, \delta_i = 1 | \tilde{T}_i \geq t, past) = P(t \leq T_i < t + dt | T_i \geq t). \tag{2.5}$$

Expression (2.5) means that an individual who is still at risk at time t has the same risk of experiencing the event of interest in the small time interval $[t, t + dt]$ as it would be the case in the situation without *censoring*.

A more feasible way of expressing the independent right censoring is through the counting process $N_i(t)$ and the intensity process $\lambda_i(t)$. The counting process may be expressed by $N_i(t) = I(\tilde{T}_i \leq t, \delta_i = 1), i = 1, \ldots, n$, while the intensity process is expressed by

$$\lambda_i(t)dt = P(dN_i(t) = 1 | past) \tag{2.6}$$

where $dN_i(t)$ is the number of jumps of the process in the small time interval $[t, t + dt]$ for individual i. The intensity process $\lambda_i(t)$ is interpreted as the conditional probability that an event occurs in $[t, t + dt)$ for individual i, given all the events has been observed prior to this time interval, divided by the length of the interval dt. We note that each individual i has its own intensity process.

Thus, by (2.5) and (2.6), the intensity process $\lambda_i(t)$ of $N_i(t)$ takes the form

$$\lambda_i(t)dt = P(dN_i(t) = 1 | past) = P(t \leq \tilde{T}_i < t + dt, \delta_i = 1 | past), \tag{2.7}$$

where $\lambda_i(t) = 0$ whenever $\tilde{T}_i < t$. Finally, we have independent right-censoring when combining $h_i(t)$ for each individual i, and (2.5) so that the intensity process of $N_i(t)$ takes the form

$$\lambda_i(t) = h_i(t) Y_i(t), \tag{2.8}$$

where $Y_i(t) = I\{\tilde{T}_i \geq t\}$ is the risk indicator for individual i.

If we assume that $h_i(t) = h(t)$ for all i, the *aggregated* counting process given by $N(t) = \sum_{i=1}^{n} N_i(t) = \sum_{i=1}^{n} I\{\tilde{T}_i \leq t, \delta_i = 1\}$ has the intensity process $\lambda(t) = \sum_{i=1}^{n} \lambda_i(t) = Y(t)h(t)$, where $Y(t) = \sum_{i=1}^{n} Y_i(t)$ is the number at risk just before time t.

2.5 The Nelson-Aalen estimator

The cumulative hazard rate $H(t) = \int_0^t h(s)ds$ may be estimated by the Nelson-Aalen-estimator. The Nelson-Aalen-estimator is a non-parametric estimator that is used to estimate the cumulative hazard from censored survival data. The estimator does not require any distributional assumptions. One important use of this estimator is to check graphically the fit of parametric models.

To estimate the cumulative hazard rate $H(t) = \int_0^t h(s)ds$, we introduce the decomposition

$$\begin{aligned} dN(t) &= \lambda(t)dt + dM(t) \\ &= Y(t)dH(t) + dM(t), \end{aligned} \tag{2.9}$$

where the terms on the right hand side are the *signal* and *noise* part, respectively. While the term $dN(t)$ on the left hand side is the observation part.

From the estimating equation $dN(t) = Y(t)d\hat{H}(t)$ and assuming $Y(t) > 0$, we find that $d\hat{H}(t) = \frac{dN(t)}{Y(t)}$. Thus by integration, the Nelson-Aalen estimator takes the form

$$\hat{H}(t) = \int_0^t \frac{dN(s)}{Y(s)} = \sum_{T_j \leq t} \frac{1}{Y(T_j)}, \tag{2.10}$$

where the sum is over the jump times $T_1 < T_2 < ...$ of $N(t)$.

One may show that the Nelson-Aalen estimator, $\hat{H}(t)$, is unbiased with variance that may be estimated by

$$\hat{\sigma}^2(t) = \int_0^t \frac{dN(s)}{Y(s)^2} = \sum_{T_j \leq t} \frac{1}{Y(T_j)^2}. \tag{2.11}$$

Equation (2.11) may be used to construct a 95% confidence interval. After a little calculation, we may get the confidence interval: $\hat{H}(t) \pm 1.96\hat{\sigma}(t)$. The log-transformed confidence interval is given by $\hat{H}(t)e^{\pm 1.96\hat{\sigma}(t)/\hat{H}(t)}$.

2.5.1 Illustration of the Nelson-Aalen estimator

We use the melanoma data from subsection 2.3.1 to illustrate the Nelson-Aalen-estimator. When we interpret the Nelson-Aalen estimator, we mainly focus on the slope of the curve. The upper left curve is the plot of the the Nelson-Aalen estimate for females with 95% confidence interval, and the upper right curve is the Nelson-Aalen estimate for males with 95% confidence interval. We notice that the cumulative hazard rate of both genders look fairly linear. This implies

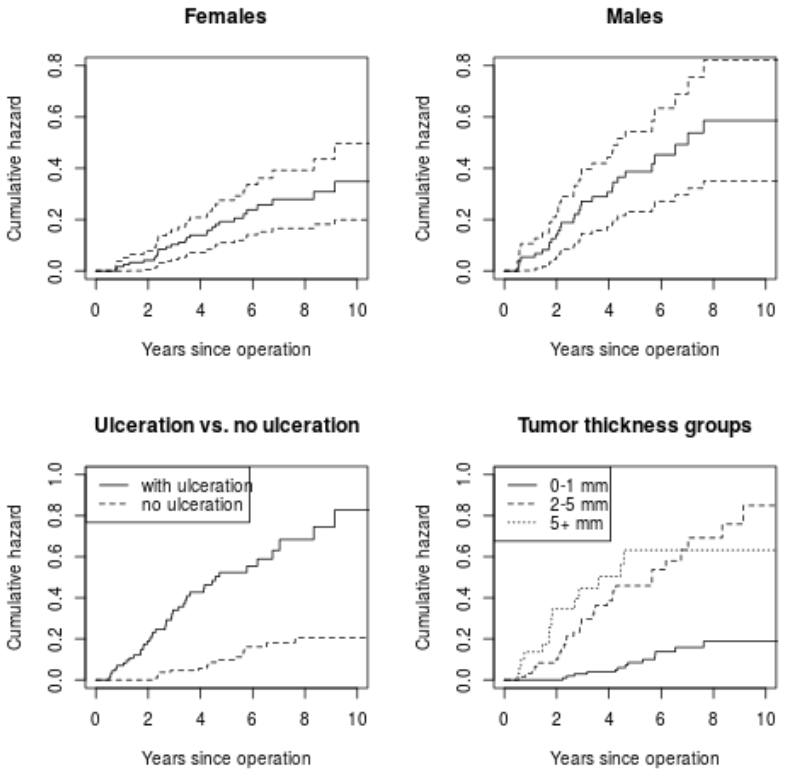

Figure 2.1: Nelson-Aalen estimate for patients with melanoma: the two upper curves are the cumulative hazard curves of females and males, respectively, while the two lower curves are patients with and without ulceration (left) and the thickness groups (right).

that the hazard rate of both genders are approximately constant. However, the male hazard rate curve seems to be steeper than the females cumulative hazard rate curve. This may be interpreted as males mortality in comparison to female mortality are higher after operation.

The lower left curve of figure 2.1 is a plot of Nelson-Aalen estimate for patients with ulceration present and ulceration absent. The curve clearly indicates that patients with ulceration present have much higher hazard rate compared to those without ulceration. Notice that the curves are displayed without confidence interval due to neatness, otherwise it would be looking messy.

Finally, the lower right curve of figure 2.1 displays plot of the Nelson-Aalen estimate based on the tumor thickness group. The estimates for the two groups with thickest tumor (2-5 mm and 5 +mm) are fairly linear and parallel until about four years after operation, then the cumulative hazard rate of thickness group 2-5 mm are continuing to be linear. This may imply that the cumulative hazard rate of this particular group is constant. However, the cumulative hazard curve of 5 +mm thickness group are constant after four years. This can be due to few people remaining at this stage. So according to this figure, the larger the size of the tumor thickness are the higher hazard rate becomes. We may conclude that the size of the tumor thickness has substantially effect on the hazard rate, and this in turn means that the risk of dying after operation depends on the size of the tumor thickness group.

2.6 The Kaplan-Meier estimator

As we mentioned earlier, the Kaplan-Meier estimator may be used to estimate the survival function $S(t)$. Before we go further to describe the Kaplan-Meier estimator, we assume that our results are only valid for the right censored data. How to handle data with tied events, we refer to section 3.2.2 in Aalen et al. [2008].

To estimate the survival function we have a sample of n individuals from the population. As in the case of the Nelson-Aalen estimator, from section 2.5, we let $N(t)$ counts the occurrences of an event in the time interval $[0, t]$ and we also let $Y(t)$ be the number of risk, as in the case of Nelson-Aalen estimator. Further, when the occurrences of an event is observed, we write the ordered times as $T_1 < T_2 < ....$

Thus, the Kaplan-Meier estimator for the survival function is given by

$$\hat{S}(t) = \prod_{T_j \leq t} \{1 - \frac{1}{Y(T_j)}\}. \tag{2.12}$$

The Greenwood's formula is used to estimate the variance of the Kaplan-Meier estimator. That is given by

$$\hat{\tau}(t)^2 = \hat{S}(t)^2 \sum_{T_j \leq t} \frac{1}{Y(T_j)\{Y(T_j) - 1\}} \tag{2.13}$$

Similar to the Nelson-Aalen-estimator, the Kaplan-Meier estimator is approximately normally distributed. A 95% confidence interval for $S(t)$ is constructed by $\hat{S}(t) \pm 1.96\hat{\tau}(t)$.

Using the log-minus-log transformation to construct the confidence interval, the normal approximation is improved. Therefore a 95% confidence interval for $S(t)$ with log-minus-log transformation is given by $\hat{S}(t)e^{\pm 1.96\hat{\tau}(t)/(\hat{S}(t)log\hat{S}(t))}$.

2.6.1 Illustration of the Kaplan-Meier estimator

The melanoma data is used further to illustrate the Kaplan-Meier estimator. Figure 2.2 displays estimate of Kaplan-Meier estimate for different risk factors. The two upper curves estimate the survival functions for females and males as indicated on the plots.

We see that as years pass, the survival probabilities decay exponentially for both gender. However, the male survival probabilities curve seems to be more linear than the female survival probabilities curve in the first 8 years after operation. We may interpret that females seems to have higher survival probabilities than males after operation. For instance, five years after operation the survival probability for males is roughly 0.70, while it is 0.80 for females. After 7 and 9 years the survival probabilities for both genders seems to be constant. However, we should not put much emphasis on this part of the curves since few people remain on this stage.

The lower-left corner of figure 2.2 displays the survival probability curves for patients with ulceration present and ulceration absent. The curves clearly indicate that patients with ulceration have much lower survival probabilities than those who have no ulceration.

Finally, the lower-right corner of figure 2.2 indicates plot of the Kaplan-Meier estimates according to the tumor thickness groups. As indicated in the plot, the curves for the patients with the tumor thickness groups 2-5 mm and 5 +mm are fairly parallel until about 5 years, but then the patients group with thickest tumor (5+mm) have constant survival probability. Again as indicated above, few people remain in the study 8 years after operation.

Patients with tumor thickness group 0-1 mm, not surprisingly, have much higher survival probabilities. According to figure 2.2 (lower-right corner), the chance of survival is 100% in the first four years for this particular group. So the straightforward interpretation is that patients with small size of tumor thickness have much higher survival probabilities compared to those with larger size.

2.7 Cox regression

The main purpose of Cox's regression model or regressions in general is to assess the effect of covariates. It is therefore important to make some comments

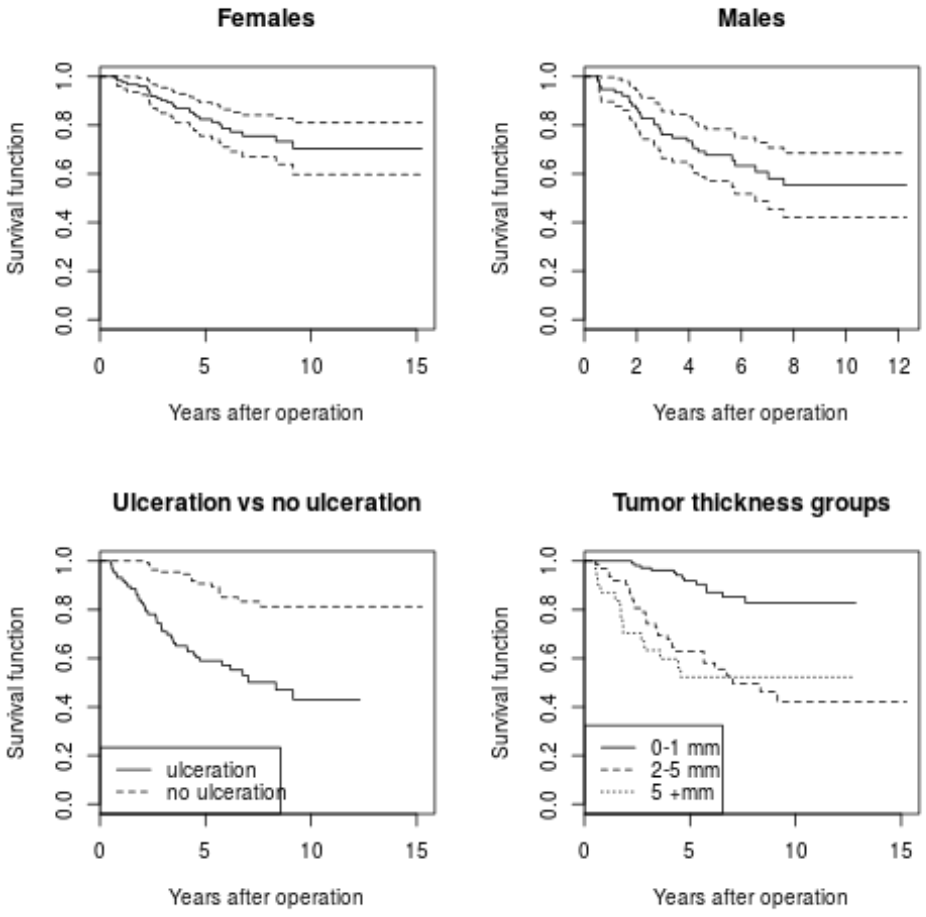

Figure 2.2: Kaplan-Meier estimate: the two upper curves illustrate the survival probabilities for females and males, respectively. The two lower curves: the lower-left corner plot displays survival probabilities for patients with ulceration present and ulceration absent, while the lower-right plot illustrates the survival probabilities for patients according to the thickness groups.

on the covariates before we study further.

Throughout our study we assume that the covariates are *predictable* and *fixed*. This means that the covariates should be measured in advanced and remain fixed throughout the study.

We now consider one counting process for each individual under study. Then we have $N_1(t), N_2(t), \ldots, N_n(t)$ with $N_i(t)$ counting the number of times the event of interest occurs for individual i in the interval $[0, t]$.

For survival data, we write

$$N_i(t) = \begin{cases} 1 & \text{if by time } t \text{ the event of interest has occurred for individual } i \\ 0 & \text{otherwise} \end{cases}$$

The vector of covariates for individual i is given by
$\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$. The intensity process of the counting process $N_i(t)$ may be expressed as

$$\lambda_i(t) = Y_i(t)h(t|\mathbf{x}_i), \tag{2.14}$$

where

$$Y_i(t) = \begin{cases} 1 & \text{if the individual } i \text{ is at risk for the event of interest just before time } t \\ 0 & \text{otherwise} \end{cases}$$

and $h(t|\mathbf{x}_i)$ is defined as the conditional hazard rate of individual i given the values of the covariates.

In equation (2.14), we assume that censoring and truncation are independent as we discussed earlier under the study of Nelson-Aalen and Kaplan-Meier estimators. This assumption implies that censoring may depend on the information in the past, but not in the future events.

In order to obtain a Cox's regression model, we need to specify how $h(t|\mathbf{x}_i)$ depends on $\mathbf{x}_i$. In the next subsection, we explore this.

2.7.1 Cox model

For Cox regression, the hazard rate for individual i is given by

$$h(t|\mathbf{x}_i) = h_0(t) \exp\{\boldsymbol{\beta}^T \mathbf{x}_i\} \tag{2.15}$$

where $h_0(t)$ is the *baseline* hazard rate that is left unspecified and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of regression coefficients which describe the effect of covariates. Thus we have a model based on both a parametric and a non-parametric part (baseline). Therefore (2.15) is called *semi-parametric*.

An example of relative risk is in order. We consider two individuals indexed 1 and 2, and we assume that all the components of the vector covariates $\mathbf{x}_1$ and $\mathbf{x}_2$ are the same except for the jth component where $\mathbf{x_{2j}} = \mathbf{x_{1j}} + 1$. Then the hazard ratio takes the form

$$\begin{aligned}\frac{h(t|\mathbf{x_2})}{h(t|\mathbf{x_1})} &= \frac{h_0(t)\exp\{\boldsymbol{\beta}^T\mathbf{x_2}\}}{h_0(t)\exp\{\boldsymbol{\beta}^T\mathbf{x_1}\}} \\ &= \exp\{\boldsymbol{\beta}^T(\mathbf{x_2}-\mathbf{x_1})\} \\ &= \exp\{\beta_j\}. \end{aligned} \tag{2.16}$$

Thus e^{β_j} is the hazard ratio for one unit increase in the jth covaraite when all other covariates remains fixed. This implies that the hazard rate of two individuals are proportional (i.e. does not depend on time).

2.7.2 Partial likelihood and estimation of β

Due to the baseline hazard $h_0(t)$, the ordinary maximum-likelihood estimation does not work for the Cox regression model, therefore the Cox's partial likelihood may be utilized instead.

The Cox's partial likelihood is, according to Aalen et al. [2008], expressed in the form

$$L(\boldsymbol{\beta}) = \prod_{T_j} \frac{\exp\{\boldsymbol{\beta}^T\mathbf{x}_{i_j}\}}{\sum_{l\in\mathscr{R}_j}\exp\{\boldsymbol{\beta}^T\mathbf{x}_l\}} \tag{2.17}$$

where $\mathscr{R}_j = \{l|Y_l(T_j)=1\}$ is the risk set at T_j and the index i_j means the individual who experiences an event at time T_j.

Furthermore, one may show that in large samples, $\hat{\boldsymbol{\beta}}$ is approximately multivariate normally distributed around the true value $\boldsymbol{\beta}$, and the covariance matrix for $\hat{\boldsymbol{\beta}}$ may be estimated by $I(\hat{\boldsymbol{\beta}})^{-1}$, where $I(\boldsymbol{\beta}) = \{-\frac{\partial^2}{\partial\beta_h\partial\beta_j}\log L(\boldsymbol{\beta})\}$ is the observed information matrix.

One may obtain a 95% confidence interval for the relative risk $\exp\{\beta_j\}$ by transforming the limits of standard confidence interval for β_j:

$$\exp\{\hat{\beta}_j \pm 1.96\ \mathrm{se}(\hat{\beta}_j)\}. \tag{2.18}$$

2.8 Hypothesis testing

In this subsection we discuss the hypothesis testing in two cases:

(i) when assuming that $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, and

(ii) when assuming $H_0 : \boldsymbol{\beta_1} = \boldsymbol{\beta}_{10}$ where $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{pmatrix}$.

Usually, we specify $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_{10}$ as zero. In section 2.8.1, we will discuss the case (i), and in section 2.8.3 we elaborate the case (ii). Further, these two null hypothesis are called *simple* and *composite*, respectively.

2.8.1 The simple hypothesis testing

The simple null hypothesis specifies $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, vs. the alternative hypothesis that at $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$.

To this end, we may apply the usual likelihood based test statistics:

- The likelihood ratio test statistic is given by

$$\chi^2_{LR} = 2\{\log L(\hat{\boldsymbol{\beta}}) - \log L(\boldsymbol{\beta_0})\} \tag{2.19}$$

- The score test statistic is expressed in the form

$$\chi^2_{SC} = \mathrm{U}(\boldsymbol{\beta}_0)^T I(\boldsymbol{\beta}_0)^{-1} \mathrm{U}(\boldsymbol{\beta}_0) \tag{2.20}$$

  where $\mathrm{U}(\boldsymbol{\beta}_0) = \frac{\partial}{\partial \beta} \log L(\boldsymbol{\beta})$ is the vector of score functions and $I(\boldsymbol{\beta}_0)$ is the information matrix.

- Finally, the Wald test statistic takes the form

$$\chi^2_W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \mathrm{I}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \tag{2.21}$$

  where $\mathrm{I}(\hat{\boldsymbol{\beta}})$ is the estimate of the information matrix for $\boldsymbol{\beta}$.

  All the test statistics described above are chi-squared distributed with p degree of freedom (df). We will illustrate the test statistics on melanoma data in the next subsection.

2.8.2 Illustration of the simple null hypothesis

We illustrate the Cox-regression model by analyzing the melanoma data further. The covariates we will consider may be explicitly specified as follows:

$$x_{i1} = \begin{cases} 0 & \text{if individual } i \text{ is a female} \\ 1 & \text{if if individual } i \text{ is a male} \end{cases}$$

$$x_{i2} = \begin{cases} 0 & \text{if individual } i \text{ is ulceration present} \\ 1 & \text{if individual } i \text{ is ulceration absent} \end{cases}$$

For the third covariate, we let $x_{i3} = \textit{tumor thickness}$ as a numeric covariate.

Table 2.1 summarize the Cox's regression analysis. The estimated regression coefficients $\hat{\beta}_j$ correspond to the covariates x_{i1}, x_{i2} and x_{i3}. We note that in practice, using for example, the statistical software **R**, we do not need to explicitly define the covariates the way we have done here. For example, since the covariates *sex* and *ulceration* are binary variables, we declare them as "factor" in **R**, then the software defines them internally.

When we interpret the results of the Cox's regression analysis, we usually focus on the hazard ratios e^{β_j}. For instance, the estimate of the hazard ratio for *sex* is $e^{\hat{\beta}_j} = 1.583$ which may be interpreted as mortality rate of a male is 58.3%

Table 2.1: *Estimated Cox's regression coefficients with their standard errors, standardized Z values and p-values for the Melanoma data*

j	Covariate x_j	$\hat{\beta}_j$	$e^{\hat{\beta}_j}$	se$(\hat{\beta})$	Z	p-value
1	sex	0.4595	1.5833	0.2668	1.723	0.0850
2	Ulceration	-1.1668	0.3114	0.3115	-3.746	0.0002
3	Thickness	0.1135	1.1201	0.0379	2.990	0.000278

Table 2.2: Estimated hazard ratios with 95% confidence intervals (CI) based on a Cox's regression analysis of the patients with malignant melanoma.

Covariate	Hazard ratio	95% CI
Sex	1.583	$[0.940, 2.671]$
Ulceration	0.311	$[0.170, 0.573]$
Thickness	1.120	$[1.030, 1.207]$

larger than mortality rate of a female. On the other hand, the p-value associated with *sex* is 0.085 which means no significance. This may be interpreted as male and female mortality are not significantly different after the operation even though the cumulative hazard ratio indicates that they are different.

For the covariate *ulceration*, the estimated relative risk (or hazard ratio) $e^{\hat{\beta}_j} = 0.312$. This may be interpreted as a patient without ulceration, the mortality that is only 31% of the mortality rate for a patient with ulceration. The p-value is 0.0002 which is statistically significant. By this we may conclude that *ulceration* has an effect for the mortality rate, which in turns means that this covariate is of importance for our analysis.

Last but not least, the numeric covariate tumor *thickness* has the estimated relative risk $e^{\hat{\beta}_j} = 1.120$ which means for one millimeter increase in tumor thickness, the hazard ratio increases by 12.0%. The p-value is about 0.0003 which is clearly significant.

The 95% confidence intervals associated with the three estimated regression coefficients $e^{\hat{\beta}_j}$ are given in table 2.2. Table 2.3 gives the summary of the test statistics of section 2.8.1 with their degree of freedom(df) and p-values. All the p-values indicate significance so we may reject the null hypothesis that all the covariates have no effect on the mortality (i.e. $\boldsymbol{\beta} = \mathbf{0}$).

2.8.3 Testing a composite null hypothesis

In section 2.8.1 under the null hypothesis, we assumed that all the $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, which may not be very realistic. In this section, we will test if a subset of the $\beta_j's$ takes specific values. That is, in a composite null hypothesis often one is

Table 2.3: The estimated of likelihood based test statistic with their degree of freedom(df) and p-values.

Test	Value	df	p-value
Likelihood ratio	39.39	3	1.44E-08
Wald	37.75	3	3.19E-08
Score	44.96	3	9.45E-10

interested to test the null hypothesis that q of the regression coefficients are zero. In particular, $H_0 : \boldsymbol{\beta_1} = \boldsymbol{\beta}_{10}$ versus the alternative hypothesis that $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_{10}$, where the vector $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T, \text{ and } \boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_q)^T \text{ is a } q\text{-vector, while } \boldsymbol{\beta}_2 = (\beta_{q+1}, \ldots, \beta_p)^T$$

is the $(p-q)$-vector of the remaining β_j's. In some literatures this is refered to as *local-tests*.

We may apply the likelihood based test statistics for this purpose. When formulating this one needs to work with a partitioned information matrix $\boldsymbol{I}$. That is, we consider a $p \times p$ information matrix $\boldsymbol{I}$, partitioned as

$$\boldsymbol{I} = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix} \text{ and the inverse } \boldsymbol{I^{-1}} = \begin{pmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{pmatrix} \tag{2.22}$$

where I_{11} and I^{11} are $q \times q$ matrices, while I_{22} and I^{22} are $(p-q) \times (p-q)$ correspond to the second partial derivatives of the minus log-likelihood with respect to $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$, respectively. The matrices I_{12} and I^{12} in addition to I_{21} and I^{21} are $q \times (p-q)$ and $p \times (p-q)$ mixed matrices of second partial derivatives.

Thus the likelihood based test statistics may be adjusted for this purpose. We may formulate them in the following manner:

The *Wald-test* statistic takes the form

$$\chi^2_W = (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})^T (I^{11})^{-1} (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \tag{2.23}$$

where $(I^{11})^{-1}$ is the $q \times q$ inverse matrix of I^{11}. In Appendix A, we show that $(I^{11})^{-1} = I_{11} - I_{12} I_{22}^{-1} I_{21}$. The test statistic measures how far the estimated coefficients $(\hat{\boldsymbol{\beta}}_1)$ are from $\boldsymbol{\beta}_{10}$. For large samples, (2.23) is χ^2-distributed with q degree of freedom under the null hypothesis.

If $\beta^\star$ is the maximum partial likelihood estimator under the null hypothesis, the *likelihood ratio* test statistic takes the form

$$\chi^2_{LR} = 2\{\log L(\hat{\boldsymbol{\beta}}) - \log L(\boldsymbol{\beta}^\star)\} \tag{2.24}$$

And this is approximately chi-squared distributed with q df under H_0.

Table 2.4: Estimated Cox's regression coefficients with standardized Z-values, p-values and confidence interval(CI) for the melanoma data.

Covaraite	$\hat{\beta}_j$	$e^{\hat{\beta}_j}$	se($\hat{\beta}_j$)	Z	p-value	CI
Ulcer	-1.218	0.296	0.309	-3.94	8.12E-5	$[0.16, 0.54]$
Thickn	0.112	1.121	0.036	3.16	0.0016	$[1.04, 1.20]$

The *score test*-statistic is given by

$$\chi^2_{SC} = U_1(\boldsymbol{\beta}^\star)^T I^{11}(\boldsymbol{\beta}^\star) U_1(\boldsymbol{\beta}^\star) \tag{2.25}$$

where $\boldsymbol{U}(\boldsymbol{\beta}) = \frac{\partial}{\partial \beta} \log L(\boldsymbol{\beta})$ and $\boldsymbol{U}(\boldsymbol{\beta}) = \begin{pmatrix} U_1(\boldsymbol{\beta}) \\ U_2(\boldsymbol{\beta}) \end{pmatrix}$. Finally, we note that $I^{11} = I_{11} - I_{12} I_{22}^{-1} I_{21}$ which is shown in appendix A, equation (A.11). The test statistic (2.25) in turn is approximately χ^2-distributed with q df under H_0.

2.8.4 Illustration of the composite null hypothesis

We illustrate the Cox-regression model for the composite hypothesis testing on the melanoma data. We specify the hypothesis as follows:

$H_0 : \beta_1 = 0$ versus the alternative hypothesis that $\beta_1 \neq 0$. We let β_1 correspond to the effect of the covariate *sex*. So under the null hypothesis, sex has no effect on mortality. In addition, we have β_2 and β_3 which correspond to the effect of the covariates *ulceration* and *thickness* respectively. More concrete, the $\boldsymbol{\beta}$-vector looks as $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_2^T)^T$, where $\boldsymbol{\beta}_2 = (\beta_2, \beta_3)^T$. In particular, we want to examine if *sex* has any effect on mortality/survival of the patients. Next we will perform the three test statistics that we discussed in the previous subsection.

The likelihood ratio(LR)-test:

To perform the likelihood ratio test, we fit two models: a *restricted* model where we add only covariates that we think are of importance and a $full$ model where we add all the covaraites. For this purpose, we fit a model based on ulceration and tumor thickness as numeric covariates. The summary of the estimated Cox's regression model under the restricted model is given in table 2.4.

Imposing *sex* to the restricted model, we obtain a full model which gives the same output as table 2.1 under the simple hypothesis testing. But we note that the results of the test statistics are different which we explain below.

To perform the LR-test, we need to find the log-likelihood value under both models (i.e. restricted and full model). In the statistical software **R**, there exists routine for doing this. First one fits a model based on the covariates *ulceration* and *thickness*, and then specify the estimator of $\boldsymbol{\beta}$-vector under the null hypothesis. Next we fit a full model based on all the covariates where we adjust the

Table 2.5: The estimated of likelihood based test statistic with their degree of freedom(df) and p-values under the composite hypothesis.

Test	Value	df	p-value
Likelihood ratio	2.95	1	0.086
Wald	2.97	1	0.085
Score	3.02	1	0.082

full model by the $\boldsymbol{\beta}$-vector. By doing this, the log-likehood values are produced to be -264.98 and -263.51 under the two models, respectively. Now we find the LR-test statistic by (2.19). That is,

$$\begin{aligned}\chi^2_{LR} &= 2\{\log L(\hat{\boldsymbol{\beta}}) - \log L(\boldsymbol{\beta_0})\} \\ &= 2\{-263.51 - (-264.98)\} = 2.95, \text{ with df} = 3 - 2 = 1.\end{aligned}$$

The p-value is 0.086 which is clearly not significant. The **R**-code for how to obtain all the three test statistics is given in appendix B.1.

The Wald-test statistic:

To perform the Wald test statistic, we need the information matrix and the covariance matrix based on $\hat{\boldsymbol{\beta}}$. These are calculated in appendix A, section A.2.

The Wald chi-squared test statistic which is given by (2.23), can be calculated. For convenience we recall it here and it gives:

$$\chi^2_W = (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})(I^{11})^{-1}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})^\top = 0.46 \times 14.05 \times 0.46 = 2.97.$$

The p-value is 0.085 wich is clearly insignificant. This indicates that the Wald test is similar to the likelihood ratio test.

The Score-test statistic:

The Score test statistic is calculated similarly to the two other tests. That is, under the null hypothesis β_1 which corresponds to the effect of $x_1 = sex$ is $\beta_{10} = 0$. The initial $\boldsymbol{\beta}$-vector is thus given as $\boldsymbol{\beta}_{init} = (0.000, -1.218, 0.114)^T$. We impose this vector to the full model. The procedure is done in **R** and is described in B.1. We get $\chi^2_{sc} = 3.02$. The p-value corresponding to this is 0.082 which is accordance to other tests (i.e. the likelihood ratio test and the Wald test). Table 2.5 summarize the three tests we have been through in this section under the composite hypothesis.

So far we have utilized only the melanoma data set to illustrate our methodology. In chapter 3 (next chapter), we will utilize it for further analysis, but to make our work more interesting we will also utilize a dataset on patients with primary biliary cirrhosis which is described next.

2.8.5 PBC data

Primary biliary cirrhosis (pbc) is rare but a severe liver disease of unknown origin, with widespread presence of about 50-cases per million population. In 1974-1984, the Mayo Clinic in the USA conducted a double-blinded randomized trial in pbc comparing the drug D-penicillamine (DPCA) with placebo. There were 424 patients who were qualified for the clinical trial, but only 312 of them agreed to participate.

A number of covariates were recorded for each of the 312 patients. More specifically, the two demographic covariates age and sex, and the two biochemical covariates albumin and bilirubin. Also some covariates were recorded, but we will not consider them here. Bilirubin is a red bile pigment derived from the degradation of hemoglobin during the destruction of red blood cells. A large concentration of bilirubin may be a sign of bad liver function. Albumin is a protein produced in the liver. The protein has some essential functionalities. Redcued albumin, for instance, may cause the liver be damaged. Since the covariate *treatments* is shown to not having any significant effect on mortality, we will therefore disregard it from the analysis.

By the date of 1986, 125 of the 312 patients had died; only 11 deaths that had not occurred due to pbc, 8 patients lost to follow up (censoring), and 19 patients had gone through liver transplantation.

The data set is taken from Fleming and David [1991, Appendix D, detail description given in section 0.2]

2.8.6 Summary

When we analyzed the melonama data graphically in section 2.5.1 and 2.6 (by means of Nelson-Aalen estimator or Kaplan-Meier estimator), we found small differences between the cumulative hazard rates of males and females. We found the estimated cumulative hazard rate for patients with ulceration present much higher than patients with ulceration absent. Finally, the estimated cumulative hazard rates for tumor *thickness* groups indicated clearly differences among three different tumor thickness groups based on the size of the tumor thickness. The thickest group had the highest estimate of the cumulative hazard rate.

Under the Cox's regression model, we tried to do the analysis more formally. In subsection 2.8.2, under the simple hypothesis, we rejected the null hypothesis that all the covriates have no effecs (i.e. $\boldsymbol{\beta} = 0$) on the mortality of the patients. In subsection 2.8.4 we tested the null hypothesis that the covariate *sex* had no impact on mortality of the patients assuming the other covariates (*ulceration* and *thickness*) had an effect on the hazard rate (or mortality rate). It turned out that sex does not have any statistically significant impact on the mortality of the patients in contradiction to the graphical analysis. However, *ulceration* and *thickness* viewed to be significantly important for the hazard rate, both in term of the graphical analysis and formal tests.

Chapter 3

Checking log-linearty

3.1 Introduction

In this chapter, we will describe methods for checking log-linearity of numeric covariates when assuming that log-linearity is fine for other covariates. We will use both simple and advanced methods to examine this. Since the covariates tumor *thickness* from the melanoma data, and *bilirubin*, *age* and *albumin* from the pbc-data are numeric, they will be used to illustrate the methods.

We consider a Cox's regression model with fixed covariates of the form

$$h(t|\mathbf{x}) = h_0(t)\exp\{\boldsymbol{\beta}^T\mathbf{x}\} \tag{3.1}$$

There are two key assumptions for model (3.1):

(i) The model assumes log-linearity in covariates. That is,

$$\log h(t|x) = \log h_0(t) + \boldsymbol{\beta}^\top\mathbf{x} \tag{3.2}$$

where $\mathbf{x} = (x_1, \ldots, x_p)^\top$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$.

(ii) And model (3.1) assumes proportional hazards. It means that the hazard ratio of two individuals with the covariates $\mathbf{x_1}$ and $\mathbf{x_2}$ is expressed as

$$\frac{h(t|\mathbf{x_2})}{h(t|\mathbf{x_1})} = \exp\{\boldsymbol{\beta^T}(\mathbf{x_2} - \mathbf{x_1})\}, \tag{3.3}$$

which is independent of time.

We will only consider the assumption (i), i.e. the log-linearity assumption (3.2).

3.2 Simple methods

3.2.1 Model extension with a function

To check log-linearity for a numeric covariate, say x_1, we assume that log-linearity is in order for other covariates. For this purpose, we will specifically look at two models:

(i) Generally, for a numeric covariate x_1, we have a model of the form

$$h(t|\mathbf{x}) = h_0(t)\exp\{\beta_1 x_1 + \boldsymbol{\beta_2^T x_2}\} \tag{3.4}$$

where $\mathbf{x} = (x_1, \mathbf{x_2}^T)^T$ and $\mathbf{x_2} = (x_2, \ldots, x_p)^T$. The $\boldsymbol{\beta}$-vector corresponding to the $\mathbf{x}$-vector is given by $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_2^T)^T$, where $\boldsymbol{\beta}_2$ in turn is given by $\boldsymbol{\beta}_2 = (\beta_2, \ldots, \beta_p)^T$.

If we increase x_1 by one unit value $(x_1 + 1)$, we obtain a Cox's regression model $h_0(t)\exp\{\beta_1(x_1+1) + \boldsymbol{\beta_2^T x_2}\}$. The hazard ratio corresponding to $x_1 + 1$ and x_1 yields e^{β_1}, which corresponds to a log-linear effect of x_1. This means that one unit increase in a numeric covariate x_1 should have the same effect whatever the value of x_1 is and whatever the values of the other covariates are if no interaction effects are included.

(ii) Now we want to look at another model of the form

$$h(t|\mathbf{x}) = h_0(t)\exp\{\beta_1 x_1 + \gamma_1 g(x_1) + \boldsymbol{\beta}_2^\top \mathbf{x_2}\} \tag{3.5}$$

where β_1 corresponds to x_1 as before, and $\boldsymbol{\beta}_2 = (\beta_2, \ldots, \beta_p)^\top$ corresponds to $\mathbf{x}_2 = (x_2, \ldots, x_p)^\top$. Moreover, we impose γ_1 that corresponds to the effect of the function $g(x_1)$. Increasing x_1 by one unit, the Cox's regression model yields

$$h(t|\mathbf{x}) = h_0(t)\exp\{\beta_1(x_1+1) + \gamma_1 g(x_1+1) + \boldsymbol{\beta}_2^\top \mathbf{x_2}\} \tag{3.6}$$

The hazard ratio for this takes the form

$$\exp\{\beta_1 + \gamma\{g(x_1+1) - g(x_1)\}\} \tag{3.7}$$

This implies that the hazard ratio for the difference between $g(x_1)$ and $g(x_1+1)$ does not give log-linear effect; the effect of x_1 is not log-linear.

Imposing $\gamma_1 g(x_1)$ in model (3.4), we arrive at (3.5) which is more complex model. The task now is to examine if the complex model (3.5) gives a better fit or not. For this purpose, we test a null hypothesis of the form:

$$H_0 : \gamma_1 = 0 \text{ vs. the alternative hypothesis } H_A : \gamma_1 \neq 0. \tag{3.8}$$

If we fail to reject the null hypothesis, we return back to our original model (3.4), i.e. model with the log-linear effect of x_1.

Example 1: Illustration using melanoma data

We will utilize the local test statistics (the composite tests) similarly to the tests we performed in section 2.8.3. As we asserted under the null hypothesis that $\gamma_1 = 0$, this may imply that we first fit a model based on the covariates *thickness*, *sex* and *ulceration*. The summary and the interpretation of the Cox's regression model is in accordance with table 2.1.

Next we fit another model, where we add all the covariates. Note that we may specify the function $g(x_1)$ as, for example, $g(x_1) = x_1^2$. The trick

Table 3.1: Estimated Cox's regression coefficients with standard errors, standardized z value, p-values and test statistic for the melanoma data. Testing if there is a log-linear effect for the tumor thickness covariate.

Covariate	β_j	$\exp(\beta_j)$	$se(\beta_j)$	z	p-value
Thickness^2	-0.004	0.9963	0.0080	-0.4570	0.6479
Thickness	0.1675	1.1182	0.1231	1.3610	0.1736
Sex	0.4175	1.5181	0.2816	1.4820	0.1382
Ulceration	-1.1192	0.3265	0.3275	-3.4180	0.0006

Test statistic	Value	df	p-value
Likelihood ratio	0.22	1	0.639
Wald	0.21	1	0.646
Score	0.21	1	0.646

now is to make a new variable (covariate) in the melanoma dataset. That is done by squaring the covariate *thickness*. Now the full model is based on the covariates $thickness^2, thickness, sex$ and $ulceration$. We specify a $\boldsymbol{\beta}$-vector prior to fitting the full model, then we use the vector as the initial $\boldsymbol{\beta}$-vector when fitting the full model. In particular, the $\boldsymbol{\beta}$ initial vector becomes $\boldsymbol{\beta} = (0.000, 0.113, 0.460, -1.167)^T$. Next we impose the $\boldsymbol{\beta}$-vector as initial vector on the full model. The details on how it is done, we refer to the **R**-code given in appendix B.1. Summary of the results for this model is given in table 3.1. Thus we may interpret that we failed to reject the null hypothesis, and therefore we sustain our log-linear model, (3.4).

Example 2: Illustration using pbc-data

Similar to Example 1, the function $g(x_1) = x_1^2$ may take the covariates $x_1 = bilirubin, x_1 = albumin$ and $x_1 = age$ where we model each of these once at a time. Table 3.2 summarize the results of the models where we assert under the null hypothesis that $\gamma_1 = 0$. To this end, the local-test statistic is utilized similarly to earlier, and it reveals that we reject the null hypothesis for the case *bilirubin*, but we fail to reject the null hypothesis when testing the covariates *albumin* and *age*. Thus, under the first case (i.e. when testing *bilirubin*) we found $\gamma_1 \neq 0$. This means that the complex model 3.5, i.e. $h(t|\mathbf{x}) = h_0(t)\exp\{\beta_1 x_1 + \gamma_1 g(x_1) + \boldsymbol{\beta}_2^\top \mathbf{x_2}\}$ fits better than the simple model, so the effect of bilirubin is not log-linear. Since we did not reject the null hypothesis for the cases *albumin* and *age*, hence we may conclude that there are log-linear effect of x_1 for *albumin* and *age*, but not for the covariate *bilirubin*.

Table 3.2: Estimated Cox's regression coefficients with standard errors, standardized z-value, p-values and the local-test statistic for the pbc data. Testing if the covariates *bilirubin,albumin,age* have log-linear effects, respectively. Notice that the test statistic values and p-values associated with three likelihood based test are give in parenthesis.

Covariate	$\hat{\beta}_j$	$e^{\hat{\beta}_j}$	se($\hat{\beta}_j$)	z	p-value
Bilirubin^2	-0.010	0.990	0.002	-4.164	3E-5
Bilirubin	0.362	1.436	0.051	7.024	2.2E-12
Age	0.033	1.033	0.009	3.685	0.000
Albumin	-1.260	0.284	0.226	-5.568	2.6E-8
Sex	-0.362	.696	0.257	-1.411	0.158
Age^2	-0.000	0.999	0.001	-0.340	0.734
Bilirubin	0.138	1.148	0.014	9.912	≈ 0
Age	0.057	1.058	0.072	0.784	0.433
Albumin	-1.481	0.227	0.221	-6.690	≈ 0
Sex	-0.576	0.562	0.252	-2.281	0.0226
Albumin^2	0.294	1.342	0.279	1.055	0.291
Bilirubin	0.140	1.151	0.014	9.920	≈ 0
Age	0.032	1.033	0.009	3.508	0.000452
Albumin	-3.366	0.035	1.793	-1.877	0.060
Sex	-0.542	0.582	0.248	-2.182	0.029

Test statistic	Value	df	p-value
Likelihood ratio	(21.52, 0.12, 1.04)	1	(5.6E-6, 0.729, 0.308)
Wald	(19.07, 0.12, 1.11)	1	(2.5E-5, 0.729, 0.292)
Score	(19.85, 0.12, 1.12)	1	(1.6E-5, 0.729, 0.290)

3.2.2 Model extension with quartiles

An alternative approach to the method from the previous subsection, is to work with quartiles of the the numeric covariate x_1. We may define indicator functions for the covariate x_1 according to its quartiles:

$$I_1(x_1) = \begin{cases} 1 \text{ if } Q_1 \leq x_1 < Q_2 \\ 0 \text{ otherwise} \end{cases}$$
$$I_2(x_1) = \begin{cases} 1 \text{ if } Q_2 \leq x_1 < Q_3 \\ 0 \text{ otherwise} \end{cases}$$
$$I_3(x_1) = \begin{cases} 1 \text{ if } x_1 \geq Q_3 \\ 0 \text{ otherwise} \end{cases}$$

where the Q_i for $i = 1, 2, 3$ indicates the quartiles of x_1.

Thus, the Cox's regression model when imposing the indicator functions takes the form

$$h(t|\mathbf{x}) = h_0(t)e^{\beta_1 x_1 + \gamma_1 x_1 I_1(x_1) + \gamma_2 x_1 I_2(x_1) + \gamma_3 x_1 I_3(x_1) + \boldsymbol{\beta}_2^\top \mathbf{x_2}} \tag{3.9}$$

where $\boldsymbol{\beta}_2 = (\beta_2, \ldots, \beta_p)^\top$ is associated with $\mathbf{x}_2 = (x_2, \ldots, x_p)^\top$ as before.

We want to test $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0$ vs. the alternative hypothesis that at least one of the γ's are non zero. If we fail to reject the null hypothesis, then we return back to the original model (3.4), otherwise model (3.9) will indicate non-log-linear effect of x_1. We will illustrate this by Example 3 and Example 4 for the melanoma and pbc data sets.

Example 3: Illustration using melanoma data

Table 3.3 summarize the results of estimated the Cox's regression model (3.9) for the melanoma data. All the p-values of the likelihood-based test statistics fail to reject the null hypothesis. By this we may conclude that imposing the quartiles of the covariate x_1 (*thickness*) on the original model, (3.4), we achieve model (3.9). However, since $\gamma_1 = \gamma_2 = \gamma_3 = 0$ which in turn means that we return back to model (3.4). For more details on how the procedures are done by means of the software, we refer to the **R**-code given in appendix B.1.

Example 4: Illustration using pbc-data

Similarly as Example 3, but now we work with the pbc-data. We summarize our analysis for all the three numeric covariates (i.e. *bilirubin, age, albumin*) in table 3.4. For the bilirubin covariate, we clearly reject the null hypothesis $\gamma_1 = \gamma_2 = \gamma_3 = 0$. For this we conclude that imposing quartiles of *bilirubin* covariate on model (3.4), results model (3.9) which does not give log-linear effect of x_1. In contrast, when we analyzed the covariates age and albumin in similar manner as for the covariate bilirubin, we ended up to not rejecting the null hypothesis. We thus interpret that imposing quartiles of covariates age and albumin to model 3.4, the log-linearity effect of x_1 still remains.

Table 3.3: Estimated Cox's regression coefficients with standard errors, standardized z value, p-values and the test statistics for the melanoma data. Note that the first interval (i.e. $x_1 < Q_1$) in tumor thickness is the reference covariate, while thickness_2 corresponds to the interval between lower-quartile and median of thickness, thickness_3 corresponds to the interval between median and upper-quartile of thickness, and finally thickness_4 corresponds to the interval between upper-quartile and upward of thickness.

Covariate	$\hat{\beta}_j$	$\exp(\hat{\beta}_j)$	$se(\hat{\beta}_j)$	z	p-value
Thickness	0.284	1.329	0.361	0.788	0.431
Thickness_2	0.048	1.049	0.257	0.187	0.851
Thickness_3	0.054	1.055	0.289	0.185	0.853
Thickness_4	-0.139	0.870	0.325	-0.428	0.669
Sex	0.391	1.479	0.286	1.370	0.171
Ulceration	-0.932	0.394	0.332	-2.804	0.005

Test statistic	Value	df	p-value
Likelihood ratio	4.76	3	0.190
Wald	4.93	3	0.177
Score	5.06	3	0.167

3.3 Fractional polynomials

3.3.1 Introduction

Polynomials are often popular in statistical analysis, but unfortunately they are either limited in their range of curve shapes such as linear and quadratic polynomials, or they produce undesirable artifacts such as edge effects and waves. Fractional polynomials (FP) differ from these, and they offer flexibility in their parameterization. In particular, FPs allow to integrate logarithm, allow non-integer powers and allow possibly repeated powers. With the FP regression one obtains much wider range of curve shapes than can be obtained with regular polynomials. In the next subsection, we start gently describing the basic FP mathematically and then we go beyond formulating its full definition. FPs are introduced by Royston and Altman [1994] and modified by Sauerbrei and Royston [1999].

3.3.2 Model formulation

A regular polynomial of degree m may be expressed as

$$\gamma_1 x + \gamma_2 x^2 + \cdots + \gamma_m x^m \tag{3.10}$$

where we have dropped the intercept γ_0 due to the Cox's baseline, $h_0(t)$. The the basic FPs, without repeated powers, for $x > 0$ are based on the following functions: $\{x^{-2}, x^{-1}, x^{-1/2}, \log(x), x^{1/2}, x, x^2, x^3\}$, where the powers $-2 < -1 < \cdots < 3$ are pre-selected, according to Royston and Altman [1994], from the set $\mathscr{P} = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$. In general, the FPs with no repeated

Table 3.4: Estimated Cox's regression coefficients with standard errors, standardized z-value, p-values and the local test statistics for the pbc data. Testing if the quartiles of the covariates *bilirubin, age* and *albumin* have log-linear effect. Note that the reference covariate is in the first interval such that $x_i < Q_1, i = 1, 2, 3$, where Q_1 is the first quartile. For instance, bilirubin$< Q_1$ is the reference covariate, while bilirubin$_2$ corresponds to the interval between lower-quartile and median of bilirubin, bilirubin$_3$ corresponds to the interval between median and upper-quartile of bilirubin, and finally bilirubin$_4$ corresponds to the interval between upper-quartile and upward of bilirubin. Note that the likelihood-based test statistics with their p-values for all the three cases are given in parenthesis, on the bottom of the table.

Covariate	$\hat{\beta}_j$	$\exp(\hat{\beta}_j)$	$se(\hat{\beta}_j)$	z	p-value
Bilirubin	-0.207	0.811	0.376	-0.550	0.582
Bilirubin$_2$	0.583	1.791	0.283	2.057	0.040
Bilirubin$_3$	0.483	1.621	0.331	1.460	0.144
Bilirubin$_4$	0.347	1.415	0.363	0.957	0.3339
Age	0.037	1.038	0.009	3.989	8.6E-5
Albumin	-1.351	0.259	0.227	-5.960	2.6E-9
Sex	-0.339	0.712	0.255	-1.329	0.186
Age	-0.009	0.990	0.031	-0.306	0.760
Age$_2$	0.005	1.005	0.008	0.627	0.530
Age$_3$	0.009	1.009	0.010	0.864	0.388
Age$_4$	0.019	1.019	0.013	1.421	0.155
Bilirubin	0.142	1.153	0.015	9.584	≈ 0
Albumin	-1.572	0.208	0.233	-6.735	≈ 0
Sex	-0.592	0.553	0.257	-2.307	0.021
Albumin	-1.698	0.183	0.512	-3.319	0.001
Albumin$_2$	0.006	1.006	0.117	0.047	0.963
Albumin$_3$	0.018	1.018	0.140	0.130	0.896
Albumin$_4$	0.087	1.090	0.165	0.529	0.597
Bilirubin	0.140	1.151	0.014	9.842	≈ 0
Age	0.033	1.033	0.009	3.482	0.000
Sex	-0.506	0.603	0.256	-1.977	0.048

Test statistic	Value	df	p-value
Likelihood ratio	(19.92, 3.61, 0.92)	3	(0.000, 0.307, 0.820)
Wald	(20.22, 3.75, 0.94)	3	(0.000, 0.290, 0.820)
Score	(21.24, 3.77, 0.94)	3	(9.4E-5, 0.287, 0.820)

powers for powers $p_1 < p_2 < \cdots < p_m$ of degree m, may be expressed in the form

$$f_m(x;\boldsymbol{\gamma},p) = \gamma_1 x^{(p_1)} + \gamma_2 x^{(p_2)} + \cdots + \gamma_m x^{(p_m)} \tag{3.11}$$

Usually the degree $m = 1$ or $m = 2$ is sufficient for a good fit. In FPs $x^{(0)} = \log(x)$ rather than $x^{(0)} = 1$. In addition, $x^{(p)} = x^p$ when $p \neq 0$.

As an example for degree $m = 3$ and powers $p = \{-2, 0, 1/2\}$, the FP takes the form $f_3(x;\boldsymbol{\gamma},p) = \gamma_1 x^{-2} + \gamma_2 \log(x) + \gamma_3 x^{1/2}$.

The full definition of FP, when we take the possibly repeated powers into account, is given in the form

$$f_m(x;\boldsymbol{\gamma},p) = \sum_{j=1}^{m} \gamma_j G_j(x) \tag{3.12}$$

where $G_1(x) = x^{(p_1)}$ and $G_j(x)$ for $j = 2,\ldots,m$ of $x > 0$ is given by

$$G_j(x) = \begin{cases} x^{(p_j)} & \text{if } p_j \neq p_{j-1} \\ G_{j-1}(x)\log(x) & \text{if } p_j = p_{j-1} \end{cases} \tag{3.13}$$

An example to make the idea clear is as follows. A fractional polynomial of degree $m = 3$ with powers $p = \{-2, -2, 1\}$ has the functions $G_1(x) = x^{-2}, G_2(x) = x^{-2}\log(x)$, and $G_3(x) = x^{1/2}$. Thus it takes the form

$$f_3(x;\boldsymbol{\gamma},p) = \gamma_1 x^{-2} + \gamma_2 x^{-2}\log(x) + \gamma_3 x^{1/2} \tag{3.14}$$

where $\log(x)$, in the second term, is achieved due to repeated powers.

Finally, how to link the FPs within the Cox model? The point now is to link equation (3.14) with Cox model (3.4). By doing this we eventually achieve a model of the form

$$\begin{aligned} h(t|\mathbf{x}) &= h_0(t)\exp\{\beta_1 x_1 + f_3(x_1;\boldsymbol{\gamma},p) + \boldsymbol{\beta}_2^T\mathbf{x}_2\} \\ &= h_0(t)\exp\{\beta_1 x_1 + \gamma_1 x_1^{-2} + \gamma_2 x_1^{-2}\log(x_1) + \gamma_3\sqrt{x_1} + \boldsymbol{\beta}_2^T\mathbf{x}_2\} \end{aligned} \tag{3.15}$$

where $\beta_1 x_1$ and $\boldsymbol{\beta}_2^T\mathbf{x}_2$ are as before, while $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)$.

3.3.3 Model selection algorithm

A regression model with one FP with degree m is considered to have $2m$ degrees of freedom (df) where one degree for each coefficient (γ) and one for each power (p). Deviance or maximum log-likelihood is used as model selection criterion. Deviance is defined as minus twice maximum log-likelihood.

The best first-degree fractional polynomial model (FP1) for $G(x)$ is the model with the highest log-likelihood or equivalently with the lowest deviance among eight possible single powers models with one regressor ($x^{-2},\ldots,x^3$ from the set $\mathscr{P}$). The best second-degree fractional polynomial (FP2) model is the model with highest log-likelihood (or equivalently the lowest deviance) among 36 (28

with $p_2 \neq p_1$ and 8 with $p_2 = p_1$) possible models from the set $\mathscr{P}$, i.e. from the power combinations: $(x^{-2}, x^{-2}\log(x)), (x^{-2}, x^{-1}), \ldots, (x^3, x^3\log(x))$.

The FP2 with smallest deviance at α (a number between 0 and 1) level is favored compared to the best FP1 if the deviance difference exceeds the $100(1-\alpha)$-percentile of χ^2 with 2 df. Otherwise, the FP1 is chosen compared to a linear term if the corresponding deviance difference exceeds the $100(1-\alpha)$-percentile of χ^2 with 1 df. For details, we refer to Sauerbrei and Royston [1999].

The algorithm implemented in **R**-software (the *mfp* package) for fractional polynomials is denoted RA2, and it uses the so called *closed test procedure*. The procedure allows complexity of candidate models to increase progressively from a prespecified minimum model (a null model) to a prespecified fractional model (maximum model). The algorithm works as follows:

(i). First perform a *4 df*-test or equivalently a second-degree test ($m = 2$) at α level of the best-fitting FP2 model against the null model. If the test is not significant, drop the covariate and the procedure is terminated, otherwise, continue to the next step.

(ii). Then perform a *3 df*-test at α level of the best fitting FP2 model against the log-linear model (i.e. straight line), the final model is the log-linear model if the test is not significant, otherwise continue to the third step.

(iii). Finally, perform a *2 df test* at α level of the best FP2 model against the best FP1. If the test is significant the final model is the FP model with second-degree, $m = 2$, otherwise the FP-model with first degree, $m = 1$.

For details and for an alternative approach (Sequential procedure), we refer to Sauerbrei et al. [2005].

In this thesis, we are concerned with checking log-linearity of a numeric covariate, and hence step 2 of this algorithm is particularly of interest. In practice with the use of softwares (e.g. **R**), one only needs to specify which covariate(s) to model within FP, then the software is doing all the steps and model selection procedures internally, then the final model is produced. Finally, one may look at p-values to determine whether the test is significant or not. As usual, a rejection level of 5% is used for significance criterion.

Example 5: Illustration of FP using melanoma data

Table 3.5 summarize the results of the Cox's regression model within the FP framework. The p-value for testing log-linearity tumor thickness is 0.192 which is not significant. Thus, one may interpret that the model is log-linear due to log-linear effect of the covariate *thickness*. The p-value under the null model, on the other hand, is significant. The **R** software internally tries different exponents from the set $\mathscr{P}$, and finds the one that is most appropriate. For example, we use the function *fp* to fit the covariate *thickness* FP model within the Cox's regression framework, the software fits an appropriate model internally. The results of the final model is therefore the same as table 2.1 from section 2.8.2.

Table 3.5: The p-values are given according to the RA_2 algorithm for the covariate tumor thickness.

Covariate	Null (p-value)	Linear (p-value)	FP (p-value)	Power 2
Thickness	0.018	0.192	0.550	-0.5

Example 6: Illustration of FP using pbc-data

We will fit FP-Cox's regression models for the covariates *bilirubin*, *age* and *albumin* from the pbc-data. We will model one covariate at a time. The summary of the final three models for the transformed covariates are given in table 3.6. The result of each model is separated with others by a horizontal line. On the bottom of the table, the results of p-values the under null-, linear- and FP models are displayed.

We notice that since the exponent of FP for *bilirubin* is 0, this results logarithmic transformation of bilirubin, $\log(bil/10)$, where the number 10 in the denominator, inside the logarithm, appears due to rescaling. The p-value associated with the bilirubin covariate is 6.6E-9, which shows non-log-linearity effect of the covariate. The p-value under the FP model suggests that the FP regression model may fit better. Repeating the same procedure for the numeric covariates *age* and *albumin*, in similar fashion, it turns out that *age* and *albumin* have log-linear effect since the p-values associated with these two covariates, under the model selection, do not reject the null hypothesis. Notice that age is also divided on 100 due to rescaling.

At the end, we illustrate how the FP transformation of the covariate bilirubin, when linked to the Cox's regression, may look like. The final result takes the form

$$h(t|\mathbf{x}) = h_0(t)\exp\{0.994\log(bil/10) - 1.197alb + 0.034age - 0.242sex\} \quad (3.16)$$

where bil (bilirubin), alb (albumin), age and sex are the covariates.

3.4 Penalized splines

3.4.1 Introduction

A more flexible approach than the fractional polynomials method, for checking log-linearity of numeric covariates, is *penalized-smoothing spline*. In this section, we first start to give a very quick background of B-splines, then we will describe the penalized smoothing splines. Eventually, we will demonstrate this method using the melanoma and pbc datasets. We use Eilers and Marx [1996] as a reference.

Table 3.6: Estimated Cox's regression coefficients with standard errors, standard z-test statistic with their p-values, and also p-values of the test statistics under the null, linear and FP models for the pbc-data. Notice that *bil* and *alb* stand for the covariates bilirubin and albumin, respectively.

Covariate	$\hat{\beta}_j$	$\exp(\hat{\beta}_j)$	$se(\hat{\beta}_j)$	z	p-value
$\log(bil/10)$	0.994	2.701	0.097	10.206	≈ 0
Albumin	-1.197	0.302	0.227	-5.278	1.3E-7
Age	0.034	1.035	0.009	3.827	0.000
Sex	-0.242	0.785	0.253	-0.955	0.339
Bilirubin	0.138	1.148	0.014	9.921	≈ 0
Albumin	-1.483	0.227	0.222	-6.677	2.4E-11
Age/100	3.232	25.334	0.921	3.509	0.000
Sex	-0.559	0.572	0.249	-2.247	0.025
Bilirubin	0.014	0.115	0.014	9.921	≈ 0
Albumin/10	-0.148	3.63E-7	0.222	-6.677	2.44E-11
Age	0.032	0.103	0.009	3.509	0.000
Sex	-0.558	0.572	0.249	-2.247	0.025

Covariate	Null (p-value)	Linear (p-value)	FP (p-value)	Power 2
Bilirubin	2.9E-22	6.6E-9	0.113	0
Age	0.010	0.760	0.560	0.5
Albumin	1.34E-8	0.625	0.596	0

3.4.2 The Cox model with penalized splines

Basis-splines or B-splines consist of piecewise polynomials connected at a number of *knots*. Cubic B-splines are the most frequently used basis functions. A linear combination of B-splines basis functions give a smooth curve. That is,

$$s(x_i) = \sum_{j=1}^{n} \gamma_j f_j(x_i) \tag{3.17}$$

where n is the number of basis functions $f_j(x_i)$, while γ_j is the associated coefficients.

This method is incorporated within the Cox regression model. In particular, to apply this method for checking log-linearity for a numeric covariate, say x_1, we may now fit a penalized smoothing spline $s(x_1)$ for this covariate assuming that the log-linearity assumption holds for the remaining covariates. We thus modify the Cox model (3.1) slightly. That is,

$$h(t|\mathbf{x}) = h_0(t) \exp\{s(x_1) + \boldsymbol{\beta}^T \mathbf{x_2}\} \tag{3.18}$$

Next we may see if the spline estimate becomes fairly linear. When the effect of a numeric covariate is not log-linear, we may transform the covariate.

If we let the number of knots to be large, then the fitted curves will give more variation than is justified by the data. To prevent overfitting, a penalty term as the second derivative $s(x_1)$, is imposed. This is obtained by maximizing the penalized log partial likelihood

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}) - \theta \int \{s''(x)\}^2 dx \tag{3.19}$$

where $\boldsymbol{\beta}$ is the coefficient vector associated with the $\boldsymbol{\beta}$-vector, whereas $\boldsymbol{\gamma}$ is the coefficients corresponding to the spline functions in (3.17). The penalty term includes a smoothing parameter θ which controls the penalty applied to the curvature in $s(x)$ via its second derivative which in turn determines the behavior of the fitted estimate $\hat{s}(x)$.

This method is implemented in existing software packages such as in **R**. In particular, *pspline* function is implemented in the *survival* package. When a numeric covariate is specified within this function, then the software does all the analysis internally, and then the final results are produced. This function selects, by default, the optimal smoothing through the degree of freedom (df). That is, the standard implementation of spline function uses $df = 4$. An alternative implementation for model selection (Govindarajulu et al. [2009]) is based on minimizing Akaike's information criterion (AIC) to select df, if df is not specified. The AIC starts with a default of 15 spline terms in B-spline basis expansion. AIC then selects the optimal smoothing parameter which is used in penalized partial likelihood fit, which is equivalent to selecting the optimal df. The choosing knots are evenly spaced across the range of $f_j(x_i)$. For each specified penalized numeric covariate, the software produces both test for linear and non-linear part of $s(x_i)$. For our purpose, we need to look at the non-linear

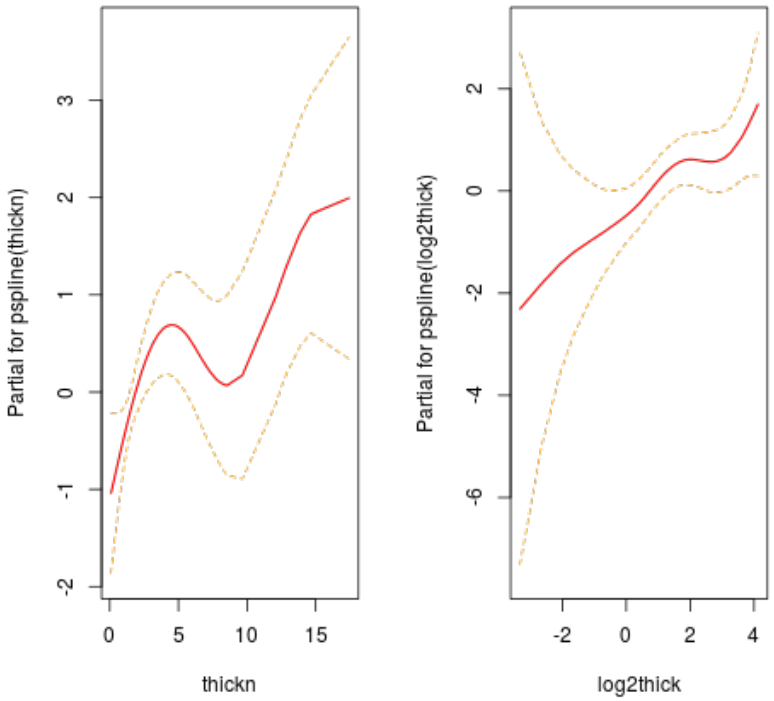

Figure 3.1: The penalized spline plots of tumor thickness and log tumor thickness by base 2 in a Cox model with sex and ulceration as the other covariates.

part to determine whether the test is significant or not. Additionally, the function *termplot* is utilized to plot the spline estimate.

Now let us illustrate the method by looking at two examples where the numeric covariate tumor thickness is used from the melanoma data, and bilirubin, albumin and age also as numeric covariates are used from the pbc dataset.

Example 7: checking log-linearity of tumor thickness

Figure 3.1 displays two plots; the left panel is the plot of penalized-splines versus tumor thickness with their confidence interval (the dashed line), and the right panel is the plot of transformed tumor thickness versus the log-transformed tumor thickness covariate with their confidence interval. The left panel looks clearly non-log-linear. The p-value (0.04) is given in table 3.7 along with the plot suggests the log-transformed of the tumor thickness covariate. The log by base 2 is a good choice for this purpose. The plot on the right panel is the log-transformed. It looks fairly log-linear. Now, the p-value is 0.41 which does not reject the null hypothesis that the log-transformed of the covariate is log-linear.

Table 3.7: Estimated Cox's regression coefficients with their standard error, χ^2-test, degree of freedom (df) and the p-values. Notice that the table displays the results of two models: a Cox model with untransformed tumor thickness and sex and ulceration as the other covariates (upper part), and the log-transformed tumor thickness with sex and ulceration as the two other covariates. The results of the two models is separated by the horizontal line (lower part).

Covariate	β_j	se(β_j)	se2	χ^2	df	p-value
pspline(thickn), linear	0.115	0.038	0.038	8.94	1.00	0.003
pspline(thickn), nonlinear				8.29	2.93	0.038
Sex	0.496	0.286	0.283	3.02	1.00	0.082
Ulceration	-0.927	0.331	0.328	7.85	1.00	0.005
pspline(log2thick), linear	0.400	0.130	0.130	9.47	1.00	0.002
pspline(log2thick), nonlinear				2.88	2.99	0.410
Ulceration	-0.878	0.329	0.328	7.14	1.00	0.008
Sex	0.462	0.284	0.280	2.65	1.00	0.100

Example 8: checking log-linearity of bilirubin, albumin and age

In figure 3.3, we have plotted the penalized splines estimate versus the covariates bilirubin, albumin and age in a Cox model with sex the as other covariate. Bilirubin does not seem to have log-linear effect, whereas albumin and age seem to have log-linear effect. However, the log-linearity assumption for age seems to be violated from around year 67 upwards. We may look at the p-values to determine formally whether the covariates are log-linear or not. Table 3.8 summarize the log-linear tests for these covariates. The p-values associated with the covariates bilirubin, albumin and age are 3.8E-7, 0.69 and 0.055, respectively. We may conclude that bilirubin is clearly not log-linear, and the log-linearity assumption holds for the covariate albumin. The covariate age is on the borderline for being significantly non-linear.

The lower part of the table summarize the results of the Cox model where the covariate bilirubin is log-transformed by base 2, and the other covariates are untransformed. Now the p-values corresponding to these covariates are 0.51, 0.69 and 0.089, respectively, which do not reject the null hypothesis that there is a log-linear effect of the covariates.

3.5 Martingale residuals

3.5.1 Introduction

An alternative approach to the methods we have been through so far, is the so-called martingale-based residuals method. This method may be constructed via the counting process $N_i(t)$ for $i = 1, \ldots, n$, the estimated cumulative intensity process $\Lambda_i(t)$, and the cumulative of the baseline hazard $H_0(t)$ from the Cox model (3.1). After we have described the method, then we demonstrate

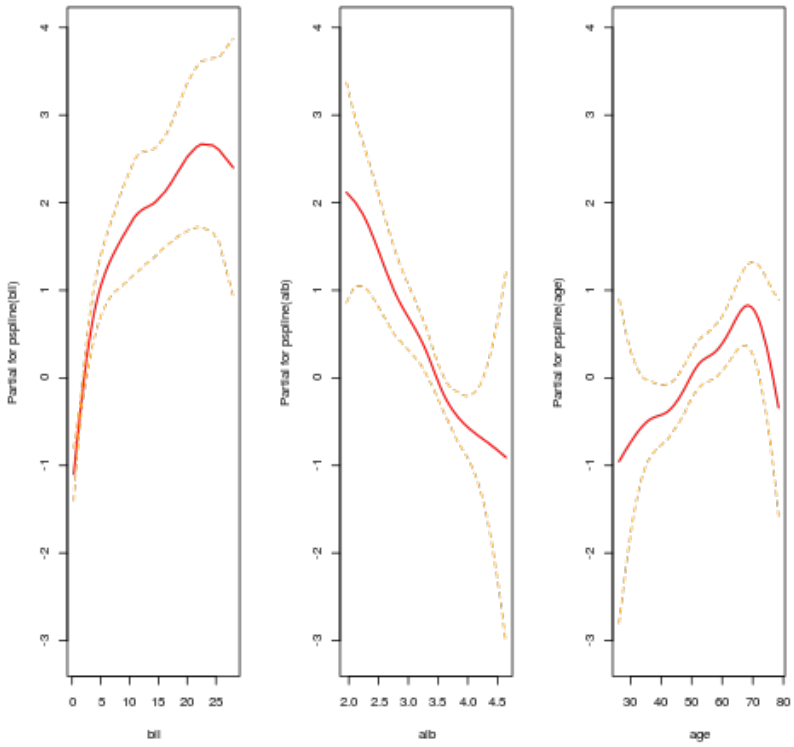

Figure 3.2: The penalized spline plots of the numeric covariates bilirubin, albumin and age in a Cox model with sex as the binary covariate.

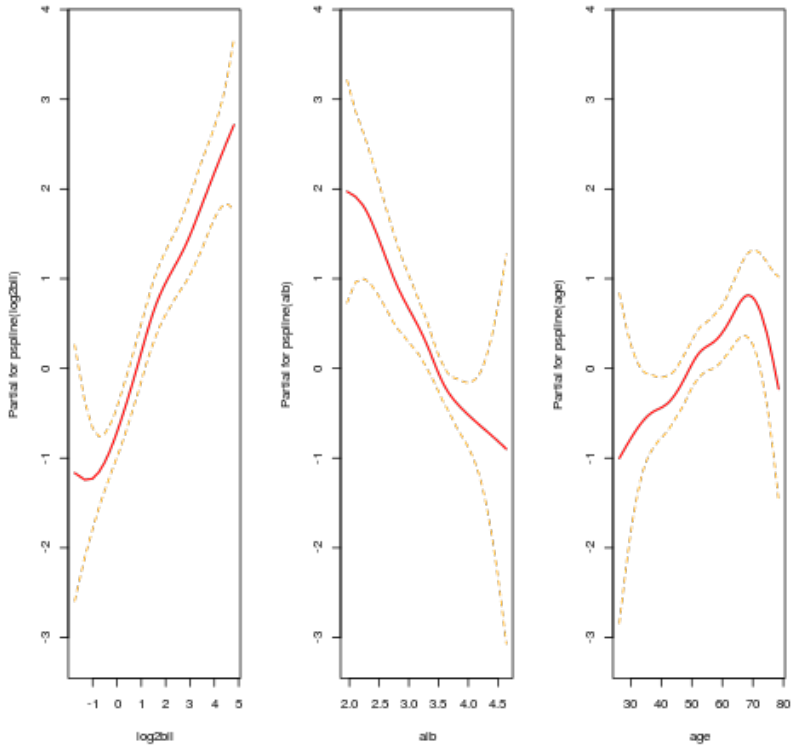

Figure 3.3: The penalized spline plot of log bilirubin, albumin and age in a Cox model with sex as the binary covariate.

Table 3.8: Estimated Cox regression coefficients with their standard error, χ^2-test, degree of freedom (df) and the p-values. Notice that the table displays the results of two models: the untransformed bilirubin, albumin and age as numeric covariate in a Cox model with sex as other covariate (the upper part of the table). The Cox model of the log-transformed bilirubin, albumin, age and sex as the covariates is placed in the lower part of the table.

Covariate	β_j	se(β_j)	Se2	χ^2	df	p-value
Bilirubin (linear)	0.145	0.017	0.0168	72.54	1.00	0
Bilirubin (nonlin)				32.75	3.02	3.8E-7
Albumin (linear)	-1.317	0.223	0.219	34.92	1.00	3.4E-9
Albumin (nonlin)				1.47	3.03	0.69
Age (linear)	0.035	0.0094	0.00927	13.99	1.00	1.8E-4
Age (nonlin)				7.66	3.04	0.055
Sex	-0.473	0.265	0.261	3.20	1.00	0.74

Covariate	β_j	se(β_j)	Se2	χ^2	df	p-value
Log2bilirubin (linear)	0.701	0.069	0.069	101.90	1.00	0
Log2bilirubin (nonlin)				2.36	3.06	0.51
Albumin (linear)	-1.2637	0.22260	0.21887	32.23	1.00	1.4E-8
Albumin (nonlin)				1.49	3.03	0.69
Age (linear)	0.036	0.0093	0.009	15.25	1.00	9.4E-5
Age (nonlin)				6.59	3.05	0.089
Sex	-0.4190	0.267	0.262	2.46	1.00	0.12

it by utilizing the melanoma and pbc data. This method is introduced by Lin et al. [1993].

3.5.2 Martingale residuals and model check

The cumulative baseline hazard, $H_0(t) = \int_0^t h_0(u)du$ of the Cox model (3.1), may be estimated by the Breslow estimator

$$\hat{H}_0(t) = \sum_{T_j \leq t} \frac{1}{\sum_{l \in \mathscr{R}_j} \exp\{\hat{\boldsymbol{\beta}}^T \mathbf{x}_l\}} \tag{3.20}$$

If the Cox model (3.1) is correctly specified, the processes

$$M_i(t) = N_i(t) - \Lambda_i(t) \tag{3.21}$$

are martingales, where $\Lambda_i(t)$ is the cumulative of the intensity processes which is given by

$$\Lambda_i(t) = \int_0^t \lambda_i(u)du = \int_0^t Y_i(u) \exp\{\boldsymbol{\beta}^T \mathbf{x}_i\} h_0(u)du \tag{3.22}$$

and $N_i(t)$ is the counting process for individual $i = 1, \ldots, n$.

We may estimate $\Lambda_i(t)$ by inserting $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ and $d\hat{H}_0(u)$ for $h_0(u)du$ where $\hat{H}_0(t)$ is the Breslow estimator given by (3.20). The estimated cumulative intensity process takes the form

$$\begin{aligned} \hat{\Lambda}_i(t) &= \int_0^t Y_i(u) \exp\{\hat{\boldsymbol{\beta}} \mathbf{x}_i^T\} d\hat{H}_0(u) \\ &= \sum_{T_j \leq t} \frac{Y_i(T_j) \exp\{\hat{\boldsymbol{\beta}} \mathbf{x}_i^T\}}{\sum_{l \in \mathscr{R}_j} \exp\{\hat{\boldsymbol{\beta}} \mathbf{x}_l^T\}} \end{aligned} \tag{3.23}$$

Then the martingale residuals processes are the difference between the counting processes and the estimated cumulative intensity processes, that is,

$$\hat{M}_i(t) = N_i(t) - \hat{\Lambda}_i(t) \tag{3.24}$$

Finally, the martingale residuals are (when evaluating the martingale residuals processes at τ) given by

$$\hat{M}_i = \hat{M}_i(\tau) = N_i(\tau) - \hat{\Lambda}_i(\tau) \tag{3.25}$$

where τ is the upper limit of study time.

The martingale residuals are interpreted as the difference between the observed and expected number of events for the *ith* individual. The martingale residuals have some properties that reminds us of the ordinary residuals from linear models. In particular, for any t,

$$\sum_{i=1}^{n} \hat{M}_i(t) = 0, \tag{3.26}$$

$$E\{\hat{M}_i(t)\} \approx cov\{\hat{M}_i(t), \hat{M}_j t\} \approx 0, \text{ for } i \neq j \tag{3.27}$$

in large samples.

To check log-linearity of a numeric covariate, or phrased differently, to check the functional form of a numeric covariate, Lin et al. suggests to work with the partial cumulative sums of the martingale-based residuals with respect to covariate values $\leq x$, rather than working directly with the raw martingale residuals. That is,

$$\begin{aligned} W_k(x) &= \sum_{i=1}^{n} I(x_{ik} \leq x)\hat{M}_i(\tau) \qquad (3.28)\\ &= \sum_{i=1}^{n} I(x_{ik} \leq x)N_i(\tau) - \sum_{T_j}\sum_{i\in\mathscr{R}_j} I(x_{ik} \leq x)\frac{\exp\{\hat{\boldsymbol{\beta}}^T\mathbf{x}_i\}}{\sum_{l\in\mathscr{R}_j}\exp\{\hat{\boldsymbol{\beta}}^T\mathbf{x}_l\}} \end{aligned}$$

Here $W_k(x)$ may be interpreted as the observed number of events minus the expected number of events for individuals with a value of the *k-th* covariate less than or equal to x (i.e. $x_{ik} \leq x$) when the model is correctly specified.

According to Lin et al. [1993], if the model is correctly specified, $W_k(x)$ is asymptotically distributed as a mean-zero Gaussian process. The limiting distribution is analytically intractable, but Lin et al. suggest to approximate it using Monte Carlo simulations. One may detect an unusual observed process $w_k(.)$ under model (3.1) by plotting it along with a few realizations from the simulated process $\hat{W}_k(.)$.

Furthermore, since $W_k(.)$ fluctuates randomly around zero under the null hypothesis, one may plot the process (3.28) as a function of numeric covariate versus x to measure the extremity of $w_k(.)$. A natural numerical measure would be $s_k = \sup_x |w_k(x)|$. A large value of s_k would indicate that the covariate has the wrong functional form. One may also obtain a formal test; the so called *supremum* test. The p-value $pr(S_k \geq s_k)$ may be approximated by $pr(\hat{S}_k \geq s_k)$, where $\hat{S}_k = \sup_x |\hat{W}_k(x)|$. Lin et al. suggests to consider an asymptotic approximation of $W_k(x)$ and to replace the stochastic process $dM_i(t)$ in this approximation by $G_i dN_i(t)$ where $\{G_l; l = 1, \ldots, n\}$ are sampled from standard normal distribution (keeping the data fixed) and $N_i(t)$ is the observed counting process. In practice, the computation is performed using the *timereg* package in **R**.

Example 9: Checking log-linearity of tumor thickness

Figure 3.4 displays the cumulative sum of the martingale residuals plotted versus tumor thickness in the Cox model with ulceration and sex as the other covariates. The grey curves in the background are 50 simulated processes. We notice that the solid curve starts at 0 from the left side and then jumps up and down and then finally stabilizes throughout until it reaches 0. So the curves start at zero and end at zero; that is what we theoretically expect. The p-value is 0.161 which is produced by the software (**R**). This is compared with 1000 simulated processes. The null hypothesis that the cumulative martingale residuals

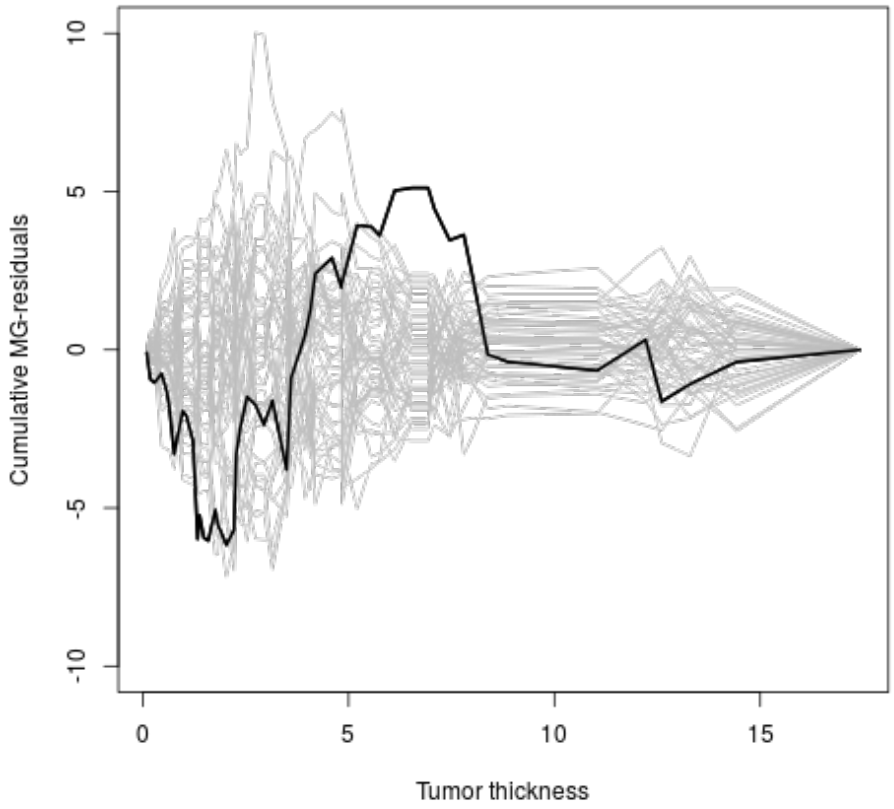

Figure 3.4: Plot of the cumulative martingale residuals versus tumor thickness, in a Cox model.

of the tumor thickness covariate ($W_k(.)$, from equation (3.28)) fluctuates randomly around zero, is not rejected. We may conclude that the functional form of the untransformed tumor thickness covariate seems fine, i.e. there is a log-linear effect of tumor thickness. Yet we note that the curve is far from being perfect. The logarithmic transformation of the covariate will improve the plot. However, since the p-value does not reject the null hypothesis, thus we do not consider the logarithmic transformation.

Example 10: Checking log-linearity of bilirubin, albumin and age

Figure 3.5 displays two plots of cumulative sums of the Martingale residuals against the numeric covariates bilirubin and $\log_2$ bilirubin in the Cox models, with albumin, age and sex as the other covariates. The upper plot is the delib-

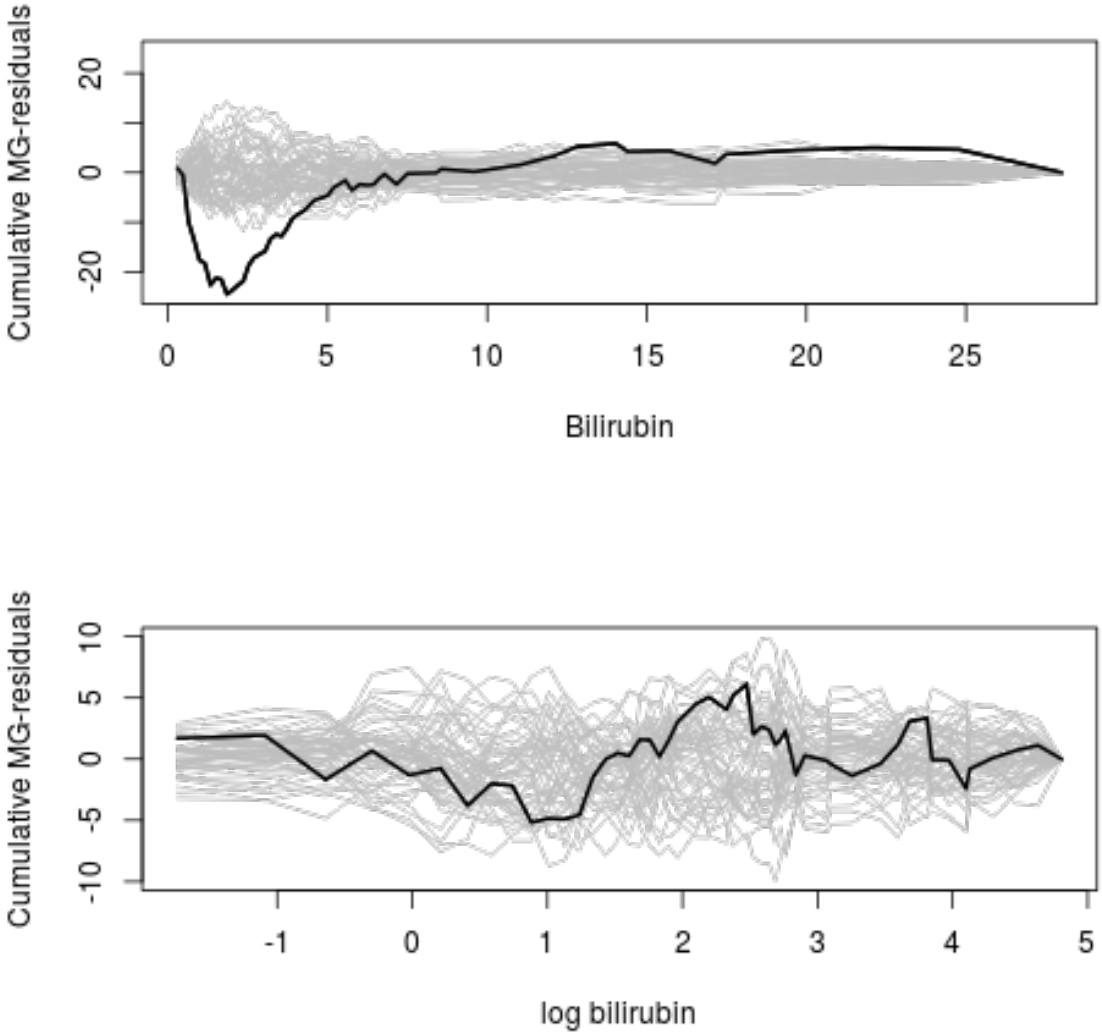

Figure 3.5: Plot of the covariates bilirubin and log bilirubin with their cumulative martinagle residuals.

Table 3.9: Supremum-test for the numeric covariates bilirubin, albumin and age from the pbc data.

Covariate	s_k value	p-value
Bilirubin	24.492	0.000
Albumin	6.706	0.403
Age	6.616	0.368

Table 3.10: Supremum-test for the numeric covariates log bilirubin, albumin and age from the pbc data.

Covariate	s_k value	p-value
Log bilirubin	6.135	0.511
Albumin	6.867	0.389
Age	8.268	0.130

erated untransformed bilirubin covariate which gives a very large peak. This may indicate that the functional form of the covariate is clearly inappropriate. The large $s_k = \sup_x |w_k(x)| = 24.5$ value (with 0 p-value) corresponding to bilirubin, given in table 3.9, supports our conclusion, namely that the functional form of the covariate bilirubian inappropriate or the assumption of the log-linear effect does not seem fine. The pattern suggests a logarithmic transformation. The lower plot is the logarithmic transformation of the covariate which shows much improvement in the functional form of the covariate. The value $s_k = \sup_x |w_k(x)| = 6.135$ with p-value 0.511, which clearly does not reject the null hypothesis. Table 3.10 summarizes the results of the logarithmic transformation of the covariate bilirubin in the Cox model with the covariates log bilirubin, albumin, age and sex.

Checking log-linearity for the remaining covariates (albumin and age), both in terms of plotting and formal test (supremum), indicates that the functional form of the covariates seem fine. The p-values for albumin and age are roughly 0.4, which does not reject the null hypothesis. See figure 3.5 and table 3.9. However, the p-values for the mentioned covariates are changed when we take the logarithm of bilirubin in the Cox model with albumin, age and sex as the remaining covariates, but still the null hypothesis is not rejected. In particular, the p-values corresponding to albumin and age are 0.39 and 0.13, respectively. Table 3.10 summarizes our finding that we just discussed.

3.6 Summary

In this chapter, we utilized different methods for checking the log-linearity assumption of the numeric covariates, in Cox model (3.1). In section 3.2.1, two

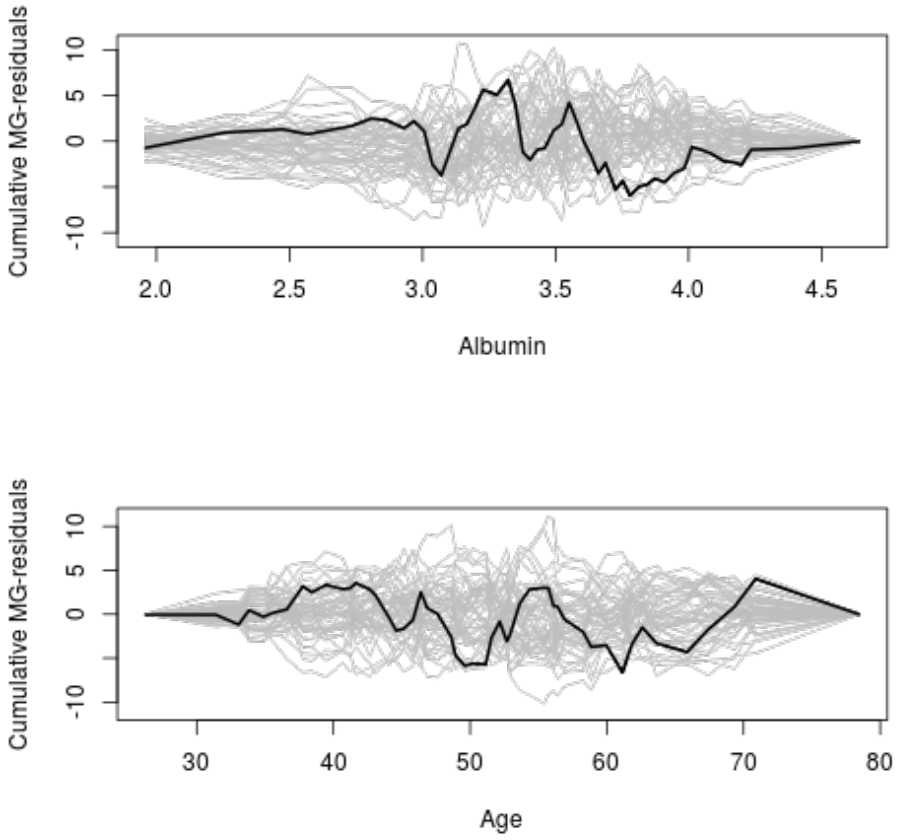

Figure 3.6: Plot of the covariates albumin and age versus their cumulative martingale residuals in the Cox model with covariates bilirubin, albumin and age.

simple models were introduced: (i) Increasing a numeric covariate by one unit, and then checking if the hazard ratio corresponding to x_j and $x_j + 1$ gives log-linear effect. We readily saw that this was in fact the case. (ii) Here we introduced a new covariate based on x_j. This means we made model (3.5) more complex by imposing a functional covariate, that is, $g(x) = x^2$, and then we saw if it gives log-linear effect. After a little calculation, we found out that this was not the case. We used the melanoma and the pbc datasets to illustrate the method. It turned out that the log-linearity assumption holds for the numeric tumor thickness from the melanoma data, and age and albumin from the pbc dataset. However, the log-linearity assumption is violated for the numeric covariates bilirubin from the pbc dataset.

In subsection 3.9, we tried another simple method which is based on working with quartiles of a numeric covariate, i.e. to make a categorical variable for the numeric covariates tumor thickness, bilirubin, albumin and age, then estimate a separate effect for each numeric group and then see if we get a fairly linear trend for the estimates. We then illustrated the method on the given datasets which turned to not rejecting the null hypothesis for the case tumor thickness, albumin and age. However, for the covariate bilirubin, the null hypothesis was clearly rejected.

In section 3.3, we attempted to check log-linearity of a numeric covariate with a more advanced method - Fractional polynomials (FP). To make model (3.4) more complex, polynomial terms, possibly fractional consisting of both positive and negative powers, can be imposed, as described earlier. We have then illustrated this method for checking log-linearity of the covariates (numeric) tumor thickness and bilirubin, age and albumin. The results of our finding turned to be as follows: the effects of tumor thickness, albumin and age are log-linear, however, the effect of bilirubin turned not to be log-linear (which is in accordance with the other methods).

A more flexible approach than FP for checking log-linearity of a numeric covariate is penalized smoothing splines. This method consist of piecewise polynomials connected by the number of knots. In particular, to apply this method for checking log-linearity of a numeric covariate, say x_1, we fitted a penalized smoothing spline $s(x_1)$. Similarly as for the other methods, we illustrated this method on the two datasets for the numeric covariates tumor thickness, bilirubin, albumin and age. The log-linearity assumption holds for the covariate albumin, but age is on the borderline for being significantly non-log-linear. However, the log-linearity assumption for the covariates tumor thickness and bilirubin is clearly violated.

The final method we have utilized for checking log-linearity of a numeric covariate is the so-called Martingale-based residuals. This is the difference between counting processes and the estimated intensity processes, when evaluating the processes at end time point. Lin et al. suggested to work with the partial-sum of the martingale residuals, rather working directly with the raw residuals. If the model is correctly specified, the process $W_k(.)$ turns to fluctuate randomly around zero. When we illustrated this method similarly as for

the other methods, it turned out that the functional form of the tumor thickness, albumin and age seems to be fine, i.e. there are log-linear effects of the covariates. However, the functional form of the billirubin covariate does not give log-linear effect which is in accordance with the other methods except for the P-spline method.

Chapter 4

Simulation studies

4.1 Introduction

Monte Carlo methods play an important role in modern statistical analysis. They can be utilized to examine the performance of the statistical methods where analytical methods are not very feasible. In survival analysis, to generate appropriate survival data for the underlying statistical models, we must generate the data through the Cox model, which is based on the hazard ratio $\exp\{\boldsymbol{\beta}^T \mathbf{x}_i\}$ and the baseline hazard $h_0(t)$. Parametric distributions are utilized to model survival times. The Exponential, the Weibull, and the Gompertz distributions are common for this purpose. In this chapter, Bender et al. [2005] is used as reference.

In this chapter, we first start by formulating survival time modeling, and then we describe how to generate experimental survival data on the computer, and further we will perform the various test procedures from Chapter 3 to analyse the generated data. Eventually, we will provide a few examples where we look at different scenarios to examine the performance of the test procedures for checking log-linearity of numeric covariates.

4.2 Simulating survival times

The cumulative hazard function corresponding to model (3.1), for individual i, is given by

$$H(t|\mathbf{x}_i) = H_0(t) \exp\{\boldsymbol{\beta}^T \mathbf{x}_i\}, \tag{4.1}$$

where t is the time, $\boldsymbol{\beta}$ is the coefficient vector corresponding to $\mathbf{x}$ as before, and $H_0(t)$ is the cumulative of the baseline hazard $h_0(t)$.

The survival function corresponding to the Cox model (2.15) is achieved through (2.4) and (4.1). That is,

$$\begin{aligned} S(t|\mathbf{x}_i) &= \exp\{-H(t|\mathbf{x}_i)\} \\ &= \exp\{-H_0(t) \exp(\boldsymbol{\beta}^T \mathbf{x}_i)\} \end{aligned} \tag{4.2}$$

If we denote Y to be a random variable with distribution function F, then $U = F(Y)$ is uniformly distributed in the interval $[0,1]$. In short, U is written as $U \sim U[0,1]$. Moreover, if $U \sim U[0,1]$, then $(1-U) \sim U[0,1]$. In the same manner, if we denote S be the the survival function, then $S(U) \sim U[0,1]$. Now, we denote T to be the survival time corresponding to (4.1), then

$$U = \exp\{-H_0(T)\exp(\boldsymbol{\beta}^T\mathbf{x}_i)\} \sim U[0,1] \tag{4.3}$$

Now we may find an expression for the survival time T, that is achieved after a simple manipulation of (4.2). That is,

$$T = H_0^{-1}\{-\frac{\log(U)}{\exp(\boldsymbol{\beta}^T\mathbf{x}_i)}\} \tag{4.4}$$

where U is random with $U \sim U[0,1]$. Thus we may conclude that we can generate T on the computer by using equation (4.4).

Below we give an example on how to incorporate the inverse of the Weibull cumulative hazard within formula (4.4).

4.2.1 Generating survival data with Weibull baseline hazard

Let $h_0(t) = \lambda\rho t^{\rho-1}$ with shape $\rho > 0$ and scale $\lambda > 0$. Then its cumulative hazard and its inverse functions are given by $H_0(t) = \lambda t^\rho$ and $H_0^{-1}(t) = (\frac{t}{\lambda})^{\rho^{-1}}$, respectively. Following the inversion method, T is obtained by computing

$$T = \{-\frac{\log U}{\lambda\exp(\boldsymbol{\beta}^T\mathbf{x}_i))}\}^{\frac{1}{\rho}} \tag{4.5}$$

Equation (4.5) can be used to generate survival data without censoring. To generate survival data with censoring, first we may sample censoring time C which we let to be exponentially distributed, then the censored survival time is obtained as $\tilde{T} = min(T, C, \tau)$, and the censoring indicator as $D = I(T = \tilde{T})$.

4.3 Simulation studies in practice

In this section, we will describe how to generate the survival time T and the censored survival time $\tilde{T}$ by means of software, and how to utilize the procedures from Chapter 3 to analyze the simulated data. We will explain this in details.

We define, for convenience, a Cox regression model similar to the Cox model (3.1). For simplicity we define only two covariates; x_1 as numeric covariate whereas x_2 as a binary covariate. Then the Cox's regression model yields

$$h(t|\mathbf{x}) = h_0(t)\exp\{f(\mathbf{x})\} \tag{4.6}$$

where $f(\mathbf{x}) = \beta_1 x_1 + \beta_2 x_2$ and $\mathbf{x} = (x_1, x_2)$. We let x_1 be Gamma distributed, i.e. $x_1 \sim Gamma(\alpha, s)$ with the shape parameter α and the scale parameter s. In addition, the expected and variance values are, respectively,

$$E(x_1) = \alpha s, \quad Var(x_1) = \alpha s^2 \tag{4.7}$$

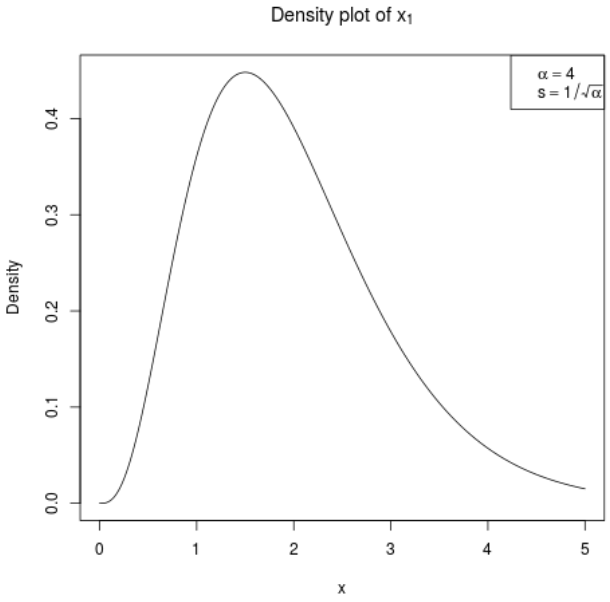

Figure 4.1: Density plot of x_1 which is sampled from the Gamma distribution with shape parameter $\alpha = 4$, and scale parameter $s = 1/\sqrt{\alpha}$.

To choose appropriate values for the Gamma parameters, we have that $Var(x) = \alpha s^2 = 1$, which implies that $s = 1/\sqrt{\alpha}$. Figure 4.1 illustrates the distribution of x_1, which is a Gamma distribution.

Further, we let $x_2 \sim binom(1, p)$, where p is the probability of success. If we define 1 to be the success and the 0 to be the failure, then we may use the binomial distribution function to generate success/failure data on the computer. The point is to make the simulated data similar to real data such as the melanoma data, where for example female was coded as 1 whereas male was coded as 2.

When we have x_1 and x_2 in order, next we may choose appropriate values of β_1 and β_2. The way we have defined our Gamma distribution, its variance is 1. Hence it may be reasonable to let $\beta_1 = 0.5$ be fixed at the first part of the analysis (in Example 1 below). In Example 2, we will examine β_1 for different values. Since x_2 has either value 1 or 0, so we may let $\beta_2 = 1$.

Algorithm 1 summarizes the simulation discussion throughout this chapter. In particular, the algorithm summarizes how we generate simulated data by means of the software, and how we perform the various test procedures from Chapter 3 for checking log-linearity of a numeric covariate, say x_1, and then draw a conclusion by computing the power of test, which is the proportion of rejection at 5% level.

Algorithm 1: How to perform the various test procedures from Chapter 3

Input:$\lambda, \gamma, \rho, \beta_1, \beta_2, \tau, \alpha$
Sample $x_1 \sim Gamma(\alpha, s)$
Sample $x_2 \sim Binomial(1, p)$
Return $f(\mathbf{x}) \leftarrow \beta_1 x_1 + \gamma x_1^2 + \beta_2 x_2$
Sample $U \sim$ uniform
Return hazard ratio $HR \leftarrow \exp\{f(\mathbf{x})\}$
Return survival time $T \leftarrow \{\frac{-\log(U)}{\lambda \times HR}\}^{1/\rho}$, i.e. formula (4.5).
Sample censoring time $C \sim \exp(n, rate)$
Return censored survival time $\tilde{T} \leftarrow min(T, C, \tau)$
for $k = 1, 2, \ldots, B$ **do**
    generate simulated data
    Perform the simple quadratic test from section 3.2.1
    Perform the simple quartile test from section 3.2.2
    Perform the Fractional Polynomial (FP) test from section 3.3
    Perform the P-spline test from section 3.4
    Perform the Martingale-residuals test from section 3.5
end for
Calculate the power of test for each test statistic, i.e. $\sum(\text{p-values} < 0.05)/B$

Example 1: Extending the Cox model by a function

We know that the Cox model (3.1) assumes log-linearity in covariates. Similarly, the Cox model (4.6) assumes log-linearity in covariates. This implies that,

$$\log \frac{h(t|x)}{h_0(t)} = \beta_1 x_1 + \beta_2 x_2 \tag{4.8}$$

is log-linear in covariates x_1 and x_2 with β_1 and β_2 as their associated coefficients. By this assumption, for instance, if we plot (4.8), we may get a fairly log-linear trend (straight line). We notice that when one wants to plot (4.8), only the first term, $\beta_1 x_1$, is of interest, since x_1 is a numeric covariate.

To check log-linearity of the numeric covariate x_1, we know that log-linearity is fine for x_2 since it is binary. Now we may impose an additional term, a quadratic function $g(x_1) = x_1^2$, based on x_1. That is, $\beta_1 x_1 + \gamma x_1^2 + \beta_2 x_2$. This equation is not linear anymore for different choice of γ, except for $\gamma = 0$. The next task is to choose appropriate γ values. This coefficient plays a vital role, either we get the plot of $\beta_1 x_1 + \gamma x_1^2$ as linear curve-shape, convex or concave curve-shape. It may therefore be reasonable to choose symmetrical values for γ such as $-0.5, -0.2, 0, 0.2, 0.5$.

Next we can utilize the test procedures from Chapter 3 to see if we can detect the violation of the log-linearity assumption, and also to figure out which procedures detect the violation best. This is actually the main objective of this chapter.

Since we have x_1 at our disposal, we are now able to illustrate the effect of different choice of γ, by plotting the equation $\beta_1 x_1 + \gamma x_1^2$. Notice that when

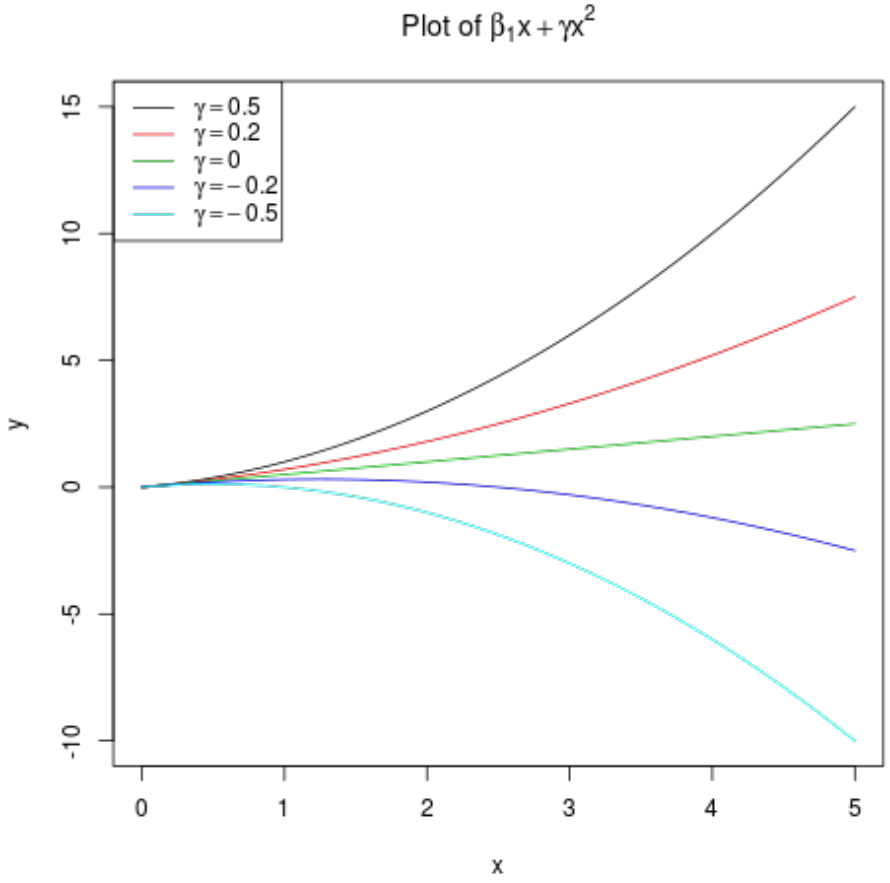

Figure 4.2: Plot of $\beta_1 x + \gamma x^2$ for fixed $\beta_1 = 0.5$, and when vary γ values symmetrical, i.e. when $\gamma = 0.5, 0.2, 0, -0.2, 0.5$.

plotting this equation, the term $\beta_2 x_2$ is already disregarded since x_2 is binary. Figure 4.2 displays plots of this equation for this purpose. We notice that when $\gamma = 0$, the second term of the equation vanishes, so we remain only with the first term. This is indeed linear, and the curve is thus a straight line. Further, as γ increases to 0.2 and 0.5, the curves get more like convex shapes. However, when the equation evaluates for negative γ, the curve has more like concave shapes, as we see in figure 4.2.

Now we have $\beta_1 x_1 + \gamma x_1^2$ at our disposal. We shall use this to generate survival data by utilizing (4.4), and also to generate censored survival data which is explained in section 4.2.1 and in Algorithm 1. Now we have the simulated dataset, the next step is to analyze the survival data by performing the test procedures from Chapter 3.

Analysis of the generated data

Next we will generate 1000 datasets for the number of observations $N = 200, 400$. For each generated data, all the test procedures from Chapter 3 are performed simultaneously, and the results of the computed p-values are to be used for computing the power of test. For $N = 400$ means we have more observations and hence more experiences to rely on than, for example, for $N = 200$. This means in turn that the results of the analysis may be significantly more accurate.

The hypothesis testing procedure is similar to Chapter 3, where the local-test statistics are utilized. For example, under the first simple method, the null hypothesis is: $H_0 : \gamma = 0$ vs. the alternative hypothesis $H_A : \gamma \neq 0$. We expect to reject H_0 when γ is different from zero. The larger γ is, the larger power of test we may achieve, i.e. when γ goes either towards 0.5 or -0.5, we achieve a larger power of test. One thing we must be aware of, under simulation, is that we have tried to ensure that we get approximately 50% events. To this end, the choice of λ, from formula (4.5), is crucial. Larger γ corresponds to smaller λ.

Table 4.1 summarizes the power of test for the various test procedures such as imposing a quadratic function on $\beta_1 x_1$ from section 3.2.1 (for simplicity we call it the quadratic test), imposing an indicator function according to its quartiles of x_1 on the term $\beta_1 x_1$ from section 3.2.2 (for simplicity we call it the quartile test), the Fractional polynomial (FP) test from section 3.3, the P-splines test from section 3.4 and the Martingale residuals (MR) test from section 3.5. In this table, the results yield for 1000 simulations ($B = 1000$) with $N = 200$. When $\gamma = 0$, we examine the test procedures against a log-linear term, so we expect that all the test procedures give a power of test about 5%. As we see from the table, this is the case.

For $\gamma = 0.2$, the test procedures give significantly larger power of test than compared to $\gamma = 0$. That is, the power of tests for the quadratic test is about 70%, for the quartile test is about 30%, for the FP test is about 80%, for the P-spline test is about 50%, last but not least, for the MR test it is about 40%. Looking at figure 4.2 on the preceding page, we notice that the "red curve" which corresponds to $\gamma = 0.2$, deviates from the "green-line" (straight line), therefore we get larger power of test compared to when $\gamma = 0.2$.

When $\gamma = 0.5$, the same test procedures give even larger power of test, i.e. we reject the null hypothesis by approximately 95%, 50%, 85%, 85% and 60%, by the same test procedures. This is in accordance with figure 4.2 on the previous page, where the the "black-curve" corresponds to $\gamma = 0.5$ which deviates much from the green-curve (straight line). However, there are significant differences when it comes to the performance of the various tests. Although other tests perform well, yet the quartile test and MR test perform very poorly (particularly the quartile test).

We reject the null hypothesis similarly when performing the test procedures for negative γ. More concrete, for $\gamma = -0.2, -0.5$, it turns out that the power of test for rejecting of the null hypothesis is not as large as it was the case when we

Table 4.1: The power of test of the various test procedures for checking log-linearity of the numeric covariate x_1. The test procedures are tested on the function $f(\mathbf{x}) = \beta_1 x_1 + \gamma x_1^2 + \beta_2 x_2$, where x_1 is a numeric covariate, and x_2 is a binary covariate in the Cox model. In the simple methods (quadratic and quartiles), the p-values for the Likelihood-ratio, the Wald and the Score tests are given in parenthesis, respectively. The test procedures are simulated 1000 times, and the number of observation is $N = 200$.

λ	γ	Quadratic g(x_1)	Quartiles	FP	P-splines	MR
0.00004	0.5	$(0.94, 0.94, 0.94)$	$(0.49, 0.48, 0.49)$	0.84	0.84	0.57
0.0001	0.2	$(0.71, 0.72, 0.73)$	$(0.28, 0.28, 0.30)$	0.80	0.51	0.38
0.0003	0	$(0.05, 0.05, 0.06)$	$(0.06, 0.06, 0.06)$	0.03	0.05	0.05
0.0007	-0.2	$(0.53, 0.43, 0.41)$	$(0.14, 0.15, 0.16)$	0.28	0.21	0.27
0.003	-0.5	$(0.82, 0.76, 0.73)$	$(0.31, 0.31, 0.32)$	0.33	0.51	0.47

performed the same test procedures for $\gamma = 0.2, 0.5$. For instance, for $\gamma = -0.5$, the power of tests give respectively $82\%, 76\%, 73\%$ for the Likelihood-ratio, the Wald and the Score tests under the quadratic test, about $30\%, 33\%, 50\%$ and 47% under the quartile, the FP, the P-spline and MR tests, respectively. As we discussed above, the corresponding numbers are much larger when performing the same test procedures for $\gamma = 0.5$.

To rank the performance of the various test procedures, the results clearly indicate that the quadratic test performs the best, and the FP test performs the second best, the P-splines test performs the third best, the MR method performs the fourth best, and finally the quartile test performs the fifth best. The reason that the quadratic test performs so outstanding is because we have generated data using a quadratic function which the method itself is based on. In contrast to the quadratic test, the quartile test seems to perform the poorest of all the test procedures.

Table 4.2 displays the power of test for the test procedures when simulated 1000 times, and the number of observation is $N = 400$. The results indicate that for $\gamma = 0$, the power of test, for the various test procedures, are in accordance with table 4.1. However, when γ increases positively or negatively, it gives much larger power of tests in comparison to table 4.1. For instance, for $\gamma = 0.5$ it gives 100% power of test in the quadratic, the FP and the P-splines tests, while for $N = 200$ as indicated in table 4.1, it gives $94\%, 84\%$ and 84%, respectively, the power of test under the same test procedures. Ranking the performance of the various test procedures is in accordance with the ranking order of the results in table 4.1.

Figure 4.3 displays plots of one simulated p-spline against x_1 (the upper plot), and one simulated cumulative sum Martingale residuals against x_1 (the lower plot) both for $\gamma = 0.5$, in the Cox's regression model with the numeric covariate x_1 and binary covariate x_2. The upper plot clearly indicates deviation

Table 4.2: The power of test for the various test procedures for checking log-linearity of the numeric covariate x_1. The test procedures are tested on the function $f(\mathbf{x}) = \beta_1 x_1 + \gamma x_1^2 + \beta_2 x_2$, where x_1 is the numeric covariate, and x_2 is the binary covariate. In the simple methods (quadratic and quartiles), the p-values for the Likelihood-ratio, the Wald and the Score test procedures are given in parenthesis, respectively. The number of simulation is $B = 1000$, and the number of observation is $N = 400$.

λ	γ	Quadratic	Quartiles	FP	P-splines	MR
0.00004	0.5	(1,1,1)	(0.81,0.80,0.81)	1.0	1.0	0.93
0.0001	0.2	$(0.96, 0.96, 0.96)$	$(0.48, 0.48, 0.48)$	0.85	0.89	0.75
0.0003	0	$(0.06, 0.07, 0.07)$	$(0.05, 0.05, 0.05)$	0.04	0.06	0.06
0.0007	-0.2	(0.80,0.75,0.74)	(0.23,0.23,0.24)	0.58	0.49	0.44
0.003	-0.5	(0.98,0.98,0.98)	(0.57,0.57,0.58)	0.83	0.92	0.76

from the straight line. Since this is one simulated plot, other simulated plots may vary from this. However, according to the power of test (100%) which we discussed above, most of the curves deviate much from the straight line. The lower plot also indicates that the functional form of the numeric covariate x_1 clearly inappropriate (due to the large peak), i.e. for $\gamma = 0.5$, the numeric covariate x_1 does not give log-linear effect.

For more illustrations, we have also displayed two more plots for when $\gamma = 0$ and $\gamma = -0.5$, as in figure C.1 and figure C.2, respectively, which are attached and interpreted in appendix C.1.

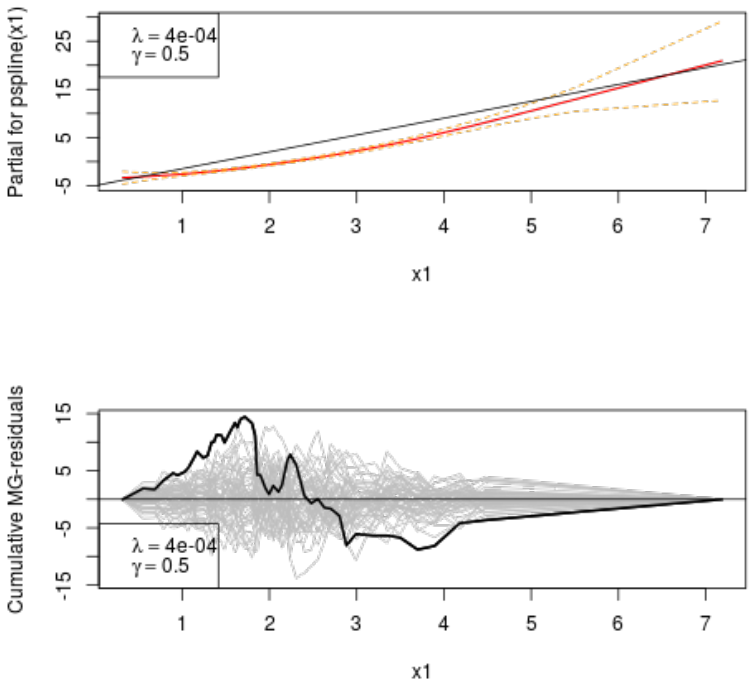

Figure 4.3: The upper plot is one simulated Cox regression partial P-splines against x_1. And the lower plot is the one simulated Martingale residuals against x_1. In both plots, $N = 400$, $B = 1000$, $\lambda = 0.00004$ and $\gamma = 0.5$.

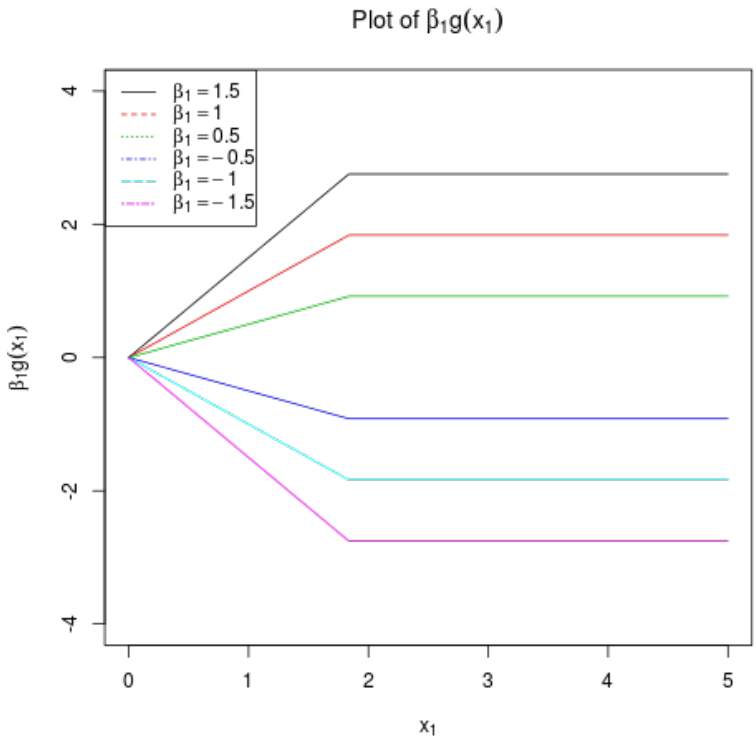

Figure 4.4: Plot of $\beta_1 g(x_1)$ when β_1 varies for $\beta_1 = 1.5, 1, 0.5, -0.5, -1, -1.5$ and $g(x_1)$ is an indicator function according to expression (4.9).

Example 2: Extending the Cox model by indicator function

Now we consider another situation where we impose an indicator function $g(x_1)$ on the Cox model (4.6). That is,

$$h(t|\mathbf{x}) = h_0(t) \exp\{f(\mathbf{x})\} \tag{4.9}$$

where $f(\mathbf{x}) = \beta_1 g(x_1) + \beta_2 x_2$ and

$$g(x_1) = \begin{cases} x_1 & \text{if} \quad x_1 \leq x^\star \\ x^\star & \text{if} \quad x_1 > x^\star \end{cases}$$

when $x^\star$ is in turn the median in the distribution of x_1.

We follow the same strategy as in section 4.3. The only change we make in Algorithm 1 is $f(\mathbf{x})$, as is described above. In figure 4.4, we have plotted the function $f(\mathbf{x})$ in equation (4.9) to illustrate the behavior of this function for symmetrical choices of β_1. The curves are linear (straight line) until it encounters the median of the distribution of x, and then suddenly the curves lose their linearity behavior. From there on the curves are constant throughout. Next we examine the performance of the various test procedures for checking

log-linearity of x_1 (numeric), in such a situation for both $N = 200$ and $N = 400$.

Analysis of the generated data for N=200

Table 4.3 displays the power of test for the various test procedures given symmetrical β_1. For $\beta_1 = 0.5$, the quadratic test performs the best (especially the Likelihood ratio test (LRT) gives 19% as the power of test), and the MR test performs the second best (14%). The quartile test and P-spline test perform almost at the same level (11%). However, the FP test performs the poorest (10%). For $\beta_1 = 1$, the quadratic test still performs the best (0.43%, 0.36%, 0.36%), and then the FP, the P-spline and the MR tests perform at the same level (i.e. the power of test is 31%). Now, the quartile test performs the poorest (26%). Increasing β_1 further to 1.5, the quadratic test and P-spline test perform very well with the power of test about 85%. The FP test and the MR test perform almost at the same level, but worse than the quadratic test and P-spline test with the power of test about 81%. The quartile test again perform very poorly with the power of test around 70%.

Similarly, when decreasing β_1, the power of test increases accordingly. For example, for $\beta_1 = -0.5$, the power of tests give $(20\%, 26\%, 25\%)$ for the quadratic test, about $(15\%, 15\%, 15\%)$ for the quartile test, 10% for the FP test, 11% for the P-spline test, and 16% for the MR test. The FP test performs the worst as it was also the case for $\beta_1 = 0.5$. While decreasing β_1 further to -1 and -1.5, the power of tests show further increase in all the tests. In particular, the power of test shows much increase in the quadratic test (58%, 65%, 65%) and the P-spline test (57%), when $\beta_1 = -0.5$. Thus, the MR test performs the third best with the power of test 45%. The FP test and quartile test again perform poorly, but the FP test shows a little improvement now compared to when $\beta_1 = -0.5$. For $\beta_1 = -1.5$, the quadratic test and P-spline test give the largest power of test (about 92%), while the FP test produces the second largest (84%), the MR test produces the third largest (79%). Not surprisingly, the quartile test performs the poorest with the power of test about 67%.

Thus, our findings indicate that the quadratic test and the P-spline test perform very well, and are the best methods among all the methods we have considered, while the FP test performs the poorest for small β_1 values. However, increasing the β_1 values either positively or negatively, the FP test performs the third best. The MR test comes at fourth place, last but not least, the quartile test comes, again, last because of its poor performance.

Analysis of the generated data when N=400

Table 4.4 displays the power of tests for the various test procedures for checking log-linearity of numeric covariate x_1 when $N = 400$.

For $\beta = 0.5$, the quartile test, the FP test and the P-spline test perform almost at the same level (with the power of test around 20%). On the other hand, the quadratic test and the MR test perform at the same level (in particular the

Table 4.3: The power of test of the various test procedures for checking log-linearity of the numeric covariate x_1. The procedures are examined on the extended Cox model by the function $f(\mathbf{x}) = \beta_1 g(x_1) + \beta_2 x_2$, where x_1 is a numeric covariate, and x_2 is a binary covariate. In the simple methods (the quadratic and the quartiles tests), the p-values of the test procedures for the Likelihood-ratio, the Wald and the Score tests are given in parenthesis, respectively. The number of simulations are $B = 1000$ and the number of observations are $N = 200$.

β_1	λ	Quadratic	Quartiles	FP	P-splines	MR
1.5	0.0001	(0.87,0.84,0.83)	(0.69,0.69,0.70)	0.82	0.86	0.81
1	0.0001	(0.43,0.36,0.36)	(0.25,0.26,0.27)	0.31	0.31	0.31
0.5	0.0003	(0.19,0.15,0.15)	(0.11,0.12,0.12)	0.10	0.11	0.14
-0.5	0.0017	(0.20,0.26,0.25)	(0.14,0.14,0.15)	0.12	0.19	0.16
-1	0.0033	(0.58,0.65,0.65)	(0.37,0.36,0.38)	0.39	0.57	0.45
-1.5	0.008	(0.92,0.95,0.95)	(0.67,0.66,0.67)	0.84	0.92	0.79

Likelihood test under the quadratic method performs well), yet the latter mentioned tests perform better than the former mentioned tests. Increasing β_1 to 1 and 2, we achieve accordingly larger power of test. As we notice from the table, increasing β_1 from 0.5 to 1, the power of test increases roughly by about 60%, in almost all the tests. Increasing β_1 further to 1.5 gives about $98 - 99\%$ in the power of test under all the tests - except for the quartile test which gives 93% as the power of test.

However, performing the tests for $\beta_1 = -0.5$, we achieve larger power of test compared to when $\beta_1 = 0.5$. In particular, we notice substantial increase in the performance of the quadratic test (especially under the Wald-test and the Score test) and P-spline test. Decreasing β_1 further to -1, and -2, we achieve accordingly further increase in the power of test, for all the tests. More specifically, for $\beta_1 = -1$ the quadratic test and P-spline test perform the best among all the tests. The FP test and the MR test perform the second best. Again the quartile test performs the poorest of all the tests.

To rank the performance of the test procedures, under the discussed scenario, the ranking results are not in accordance with earlier where the quadratic test performed the best. Despite that, the quadratic test still performs very well (surely better than quartile test) or at the same level as the FP, the P-spline and the MR tests when increasing β_1 positively. On the other hand, when decreasing β_1 further to -1, the quadratic test and p-spline test perform better than the MR, the FP and quartile tests. We see that when decreasing β_1 to -0.5 and -1, we achieve similar result under the FP test as the other tests. Decreasing β_1 further to -1.5, the quadratic test and P-spline test achieve the largest power of test, while the FP test and the MR test achieve the second largest, namely with the power of test almost 100%. The quartile method again performs the poorest with the power of test 96%.

Table 4.4: The power of test of the various test procedures for checking log-linearity of the numeric covariate x_1. The test procedures are tested on the non-linear function $f(\mathbf{x}) = \beta_1 g(x_1) + \beta_2 x_2$, where x_1 is numeric covariate, and x_2 is binary covariate. In the simple methods (the quadratic and quartile test), the p-values for the Likelihood-ratio, the Wald and the Score test procedures are given in parenthesis, respectively. The number of simulation are $B = 1000$ and the number of observation are $N = 400$.

β_1	λ	Quadratic	Quartiles	FP	P-splines	MR
1.5	0.00008	(0.98,0.97,0.97)	(0.93,0.93,0.93)	0.98	0.99	0.98
1	0.0002	(0.85,0.80,0.80)	(0.66,0.67,0.67)	0.78	0.82	0.81
0.5	0.00031	(0.30,0.25,0.25)	(0.19,0.19,0.19)	0.20	0.22	0.26
-0.5	0.0017	(0.36,0.41,0.41)	(0.22,0.22,0.23)	0.24	0.35	0.27
-1	0.0033	(0.87,0.90,0.90)	(0.67,0.67,0.67)	0.81	0.89	0.80
-1.5	0.009	(1.00,1.00,1.00)	(0.96,0.96,0.96)	0.99	1.00	0.99

We summarize our findings as follows. Decreasing β_1 we achieve larger power of test for all the tests compared to when increasing. The quadratic test and P-spline test perform the best, the MR method and the FP method are the second best, while the quartile test performs the poorest of all the tests. One remark regarding the FP test is in order. Large values of β_1, cause convergence problems for this test, and thus the power of test may be very unstable, and hence not reliable.

4.4 Summary

We started chapter 4 by formulating simulating survival time modeling, where we utilized the Weilbull distribution function for the baseline hazard, $h_0(t)$, in the Cox model to generate survival data. Next we sampled two covariates; x_1 as numeric covariate and x_2 as binary covariate such that the covariates involved resembling real survival data (e.g. the melnoma and the pbc data). The Cox's regression model, thus, yields model (4.6). This model assumes log-linearity in covariates. Since our objective has been to reveal violation of the log-linearity assumption, thus we imposed additional term, to make $f(\mathbf{x})$ in (4.6) non-linear. In particular, we looked at two scenarios: (i) We imposed $g(x_1) = \gamma x_1^2$ on $f(\mathbf{x})$ such that $f(\mathbf{x}) = \beta_1 x_1 + \gamma g(x_1) + \beta_2 x_2$, (ii) We imposed $g(x_1)$ on $f(\mathbf{x})$ such that $f(\mathbf{x}) = \beta_1 g(x_1) + \beta_2 x_2$, where $g(x_1)$ is the indicator function defined as in (4.9). In (i) we worked with fixed β_1, but varied γ symmetrical, while in (ii) we varied β_1 symmetrical.

Next we performed the various test procedures from chapter 3 to reveal violation of the log-linearity assumption, and also to examine which test procedures perform the best or better than others. It turned out that in scenario (i) the quadratic test performs the best, the FP test performs the second best, the P-splines test performs the third best, the MR test performs the fourth best, and finally the quartile test performs the fifth best. In contrast to the quadratic test, the quartile test seems to perform the poorest of all the tests. In scenario (ii), we decreased β_1, then we achieved larger power of test for all the test compared to when increasing. The quadratic test and P-spline are the best, the MR test and the FP test are the second best, while the quartile test performs the poorest of all the tests. So in short, the quadratic test performs the best in both scenarios, while the P-spline test, in addition to the quadratic test, performs the best in scenario (ii). The quartile test seems to perform the worst in both scenarios.

One remark regarding the FP test is appropriate here. Large β_1 and γ, cause issues for the performance of the FP test. We found this test procedure quite unstable for large values of the mentioned parameters.

Chapter 5

Summary and Conclusion

Summary

The main objective of this thesis has been to check the log-linearity assumption of numeric covariates for the Cox regression, and also to examine how the test procedures involved for checking this assumption perform.

In chapter 3, we presented the test procedures for checking this assumption. In particular, we applied two simple methods, and several advanced methods to inspect the assumption. All the methods, except the Martingale-residuals method, are based on making the Cox's regression model non-log-linear by imposing one or more terms such that when applying the methods on the model, one may easily detect violation of the log-linearity assumption. The Martingale-residuals method, on the other hand, is based on working with the partial cumulative sums of the Martingale-based residuals, which is the difference between the observed and expected number of events for the *ith* individual, then checking if the functional form of numeric covariates appropriate; meaning that is an alternative way of checking if numeric covariates give log-linear effect. We applied the methods on the real datasets; the melanoma dataset and the pbc dataset. In chapter 4, we performed the test procedures corresponding to the various methods on generated datasets.

Table 5.1 summarizes the p-values associated with various test procedures from chapter 3 for checking the log-linearity assumption of the numeric covariates. Note that for simplicity we call the two simple methods (as in chapter 4), the quadratic method (from section 3.2.1 on page 23) and the quartile method (from section 3.2.2 on page 27), where both test procedures that corresponds to these methods are based on the Likelihood ratio test (LRT), the Wald test and the Score test. Thus the p-values associated with these test statistics are hence given respectively in parenthesis.

To assess the effect of tumor thickness covariate for the melanoma data, all the p-values of the test statistics, given in the table, reveal that we fail to reject the null hypothesis, except for the p-values from P-spline test statistic which rejects the null hypothesis. However, some test procedures give larger p-values

Table 5.1: The p-values associated with the five different test procedures for checking log-linearity of the numeric covariates. The p-values that correspond to likelihood ratio test (LRT), the Wald test and the Score test all are, respectively, given in parenthesis. Note that FP, P-S and MR are abbreviations of fractional polynomial, P-spline and martingale residuals, respectively.

Covariate	Quadratic	Quartile	FP	P-S	MR
Thickness	(0.639,0.646,0.646)	(0.190.0.177,0.167)	0.192	0.038	0.161
Bilirubin	(5.6E-6,2.5E-5,1.6E-5)	(0.000,0.000,9.4E-5)	6.6E-9	3.8E-7	0.000
Albumin	(0.729,0.729,0.729)	(0.307,0.290,0.287)	0.760	0.690	0.403
Age	(0.308,0.292,0.290)	(0.820,0.820,0.820)	0.625	0.055	0.368

than others. In particular, the quadratic test gives the largest p-values which is about 65%. The quartile test, the FP test and the MR test give almost the same level of p-values (about 18%). The p-values of the test statistics indicate that assessing the log-linearity assumption of tumor thickness, the P-spline test may be the best test procedure for this purpose, where the other test procedures are not capable of revealing the non-log-linearity. The size of the p-values indicates that the quadratic test is, on the other hand, the least good test procedure to reveal such an non-linearity effect. The other three tests are at the same level and therefore also do not reveal non-log-linear effect of tumor thickness.

The p-values of the test statistics corresponding to the covariates bilirubin and albumin are not of interest to assess since these p-values either all reject (bilirubin) or all fail to reject (albumin) the null hypothesis. So there is not much information to gain by looking at the p-values in order to assess how these test procedures perform. The p-values associated with the age covariate, under the P-spline test statistic, is 5.5%, which is on the borderline for being significantly non-log-linear. The p-values and graphical check (i.e. figure 3.2 on page 37) indicate that the P-spline test procedure is the most sensitive (i.e. the best) test procedure to reveal non-log-linearity effect of the numeric covariate age (as tumor thickness) where the other test procedures are in lack of such capability.

In chapter 4, we attempted to check the log-linearity assumption of numeric covariates through simulation studies. We performed the test procedures we had developed in chapter 3 on two examples where the main objective was to examine the performance of the test procedures corresponding to the methods. In Example 1, when testing $f(\mathbf{x}) = \beta_1 x_1 + \gamma g(x_1) + \beta_2 x_2$ where we imposed the term $g(x_1)$ based on x_1, it turned out that the quadratic test, not surprisingly, performed the best, and the FP test performed the second best, the P-splines test performed the third best, the MR performed the fourth best, and finally the quartile test performed the poorest of all the tests. In Example 2, when testing $f(\mathbf{x}) = \beta_1 g(x_1) + \beta_2 x_2$, where $g(x_1)$ is an indicator function according to (4.9), and especially when decreasing β_1, we achieved larger power of test for all the tests. The quadratic test and P-spline test performed the best, the MR test and the FP test performed the second best, while the quartile test

performed the poorest of all the tests. One remark regarding to the FP test is in order. We found the FP test procedure quite unstable for large β_1 and γ.

Conclusion

So what makes this thesis unique from other works? As indicated in the introductory chapter, this thesis has provided a systematic review of the various graphical procedures and the formal tests that have been proposed in the literatures, for checking log-linearity of numeric covariates. Thus this work sums up the proposed methods, and also hopefully helps analysts or students to choose methods that perform better than others for checking the log-linearity assumption of numeric covariates. (see below).

Based on the analysis of the real datasets (the melanoma and the pbc datasets), our findings indicate that the P-spline test (which includes both graphical procedure and formal test) is the best test procedure to reveal non-log-linearity effect of numeric covariates.

However, to assess the performance of the various procedures under the simulated data, we found the quadratic test to be the best among the existing tests when extending the Cox model by imposing a quadratic function. On the other hand, extending the Cox model by imposing an indicator function such as $g(x_1)$ in (4.9), the quadratic test in addition to the P-spline test indicate to be the best methods among the existing methods. In such a scenario, the FP test and MR test perform equally well, however, not as good as the first two mentioned methods. The quartile test is found to be the poorest of all the tests procedures. Based on our findings, the quadratic method and P-spline method may be preferred over the other methods. The last mentioned method is quite complex analytically, however, this method is, luckily, already implemented quite professionally in **R**, and thus it is quite straight forward to use it. The first mentioned method is, on the hand, a simple method. It is easy to implement by oneself in, for example **R**, and it also performs quite well. The FP method is adequate when performing a real dataset, however, this method is not properly implemented in the **R** software such that when performing the test procedure on simulated data, it gives unstable results. The quartile method is not recommended at all.

Further challenges

If time had permitted, we could have inspected the test procedures in some more examples. One such example would be to extend the Cox model by imposing even more complex functions than the two examples we presented in chapter 4. One such a extension would be to incorporate the function $f(\mathbf{x}) = \beta_1 x_1 + \gamma x_1^2 + \beta_2 g(x_2)$ within the Cox model, where both x_1 and x_2 are numeric covariates that can be analyzed simultaneously, and $g(x_2)$ is an indicator function defined in similar manner as $g(x_1)$ in (4.9). Few days before the thesis's deadline, we tried to perform the test procedures on such an example, however, the simulated program we had developed in chapter 4 introduced

bugs and unstable results that would require more time to debug.

The cumulative Martingale residuals test suggested by Lin et al., has the property that if the model is correctly specified, the processes $W_k(x)$ in (3.28) may be expected to obtain their largest absolute values for intermediate values of x. For this purpose, we looked at both plots and formal test procedures, which are implemented in **R**. Another maybe more relevant possibility for measuring the deviation of the observation processes from zero, is to integrate the area under the simulated curves which starts in zero and ends in zero, that is, $\int |w_k(x)|dx$, and then draw a conclusion based on how much the total area deviates from zero. An alternative approach would be to look at weighted supremum test statistics where one may put more emphasis on the start and end point of the processes. Another possibility would be to measure the relative standard deviation of the processes $W_k(x)$. That is, measuring $\frac{w_k(x)}{SD(w_k(x))}$. All of these alternative approaches would be appropriate a topic for a master thesis itself which is not possible for further investigation in such a short thesis.

Bibliography

Odd Aalen, Ørnulf Borgan, and Håkon Gjessing. *Survival and event history analysis: a process point of view*. Springer Science & Business Media, 2008.

Per Kragh Andersen, Ørnulf Borgan, Richard D. Gill, and Niels Keiding. *Statistical models based on counting processes*. Springer Ser. Statist., Springer, 1993.

Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics in medicine*, 24(11): 1713–1723, 2005.

Paul HC Eilers and Brian D Marx. Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102, 1996.

R. Thomas. Fleming and Harrington P. David. *Counting process and Survival analysis*. Willey series in probability and mathematical statistics, 1991.

Usha S Govindarajulu, Elizabeth J Malloy, Bhaswati Ganguli, Donna Spiegelman, and Ellen A Eisen. The comparison of alternative smoothing methods for fitting non-linear exposure-response relationships with cox models in a simulation study. *The international journal of biostatistics*, 5(1), 2009.

John P. Klein and Moeschberger Melvin L. *Survival Analysis techniques for censored and truncated Data*. Springer Ser. Statist., Springer, 1997.

Danyu Y Lin, Lee-Jen Wei, and Zhiliang Ying. Checking the cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80(3):557–572, 1993.

P. Royston and D. G. Altman. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied Statistics*, pages 429–467, 1994.

W. Sauerbrei, C. Meier-Hirmerb, A. Bennerc, and P. Royston. Multivariable regression model building by using fractional polynomials: Description of SAS, STATA and R programs. *Computational Statistics and Data Analysis 50 (2006) 3464 – 3485*, 2005.

Willi Sauerbrei and Patrick Royston. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(1):71–94, 1999.

Richard Van Noorden, Brendan Maher, and Regina Nuzzo. The top 100 papers. *Nature*, 514(7524):550–553, 2014.

Appendix A

Attchament to Chapter 2

The main objective of this appendix is to give readers opportunity to see how the main results are obtained. These calculations are, however, not required reading to understand the main content of the thesis.

A.1 Covariance and information matrix

The Wald test statistic is given by (2.23). For convinient, we recall it here:

$$\chi^2_W = (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})^T (I^{11})^{-1} (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \tag{A.1}$$

The term $(I^{11})^{-1}$ is calculating by the following calculation:

$$\begin{aligned} \boldsymbol{I} \times \boldsymbol{I^{-1}} &= \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix} \times \begin{pmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{pmatrix} \\ &= \begin{pmatrix} I_{11}I^{11} + I_{12}I^{21} & I_{11}I^{12} + I_{12}I^{22} \\ I_{21}I^{11} + I_{22}I^{21} & I_{21}I^{12} + I_{22}I^{22} \end{pmatrix} \qquad \text{(A.2)} \\ &= \begin{pmatrix} I_q & 0_{q,p-q} \\ 0_{p-q,q} & I_{q-p} \end{pmatrix}. \qquad \text{(A.3)} \end{aligned}$$

Further, based on this matrix we set up four equations:

$$I_{11}I^{11} + I_{12}I^{21} = I_q \tag{A.4}$$

$$I_{11}I^{12} + I_{12}I^{22} = 0_{q,p-q} \tag{A.5}$$

$$I_{21}I^{11} + I_{22}I^{21} = 0_{p-q,q} \tag{A.6}$$

$$I_{21}I^{12} + I_{22}I^{22} = I_{q-p} \tag{A.7}$$

From (A.5) and (A.6) it follows, respectively, that

$$I^{12} = -(I_{11})^{-1} I_{12} I^{22} \tag{A.8}$$

and

$$I^{21} = -(I_{22})^{-1} I_{21} I^{11} \tag{A.9}$$

Substituting (A.9) into (A.4), we get

$$(I_{11} - I_{12} I_{22}^{-1} I_{21}) I^{11} = I_q \tag{A.10}$$

This gives

$$I^{11} = (I_{11} - I_{12} I_{22}^{-1} I_{21})^{-1} \tag{A.11}$$

Inserting (A.8) into (A.7), it follows that

$$(I_{22} - I_{21} I_{11}^{-1} I_{12}) I^{22} = I_{q-p}, \Rightarrow I^{22} = (I_{22} - I_{21} I_{11}^{-1} I_{12})^{-1} \tag{A.12}$$

Substituting further (A.11) and (A.12) into (A.8) and (A.9), we arrive at

$$I^{12} = -(I_{11})^{-1} I_{12} (I_{22} - I_{21} I_{11}^{-1} I_{12})^{-1} \tag{A.13}$$

and

$$I^{21} = -(I_{22})^{-1} I_{21} (I_{11} - I_{12} I_{22}^{-1} I_{21})^{-1} \tag{A.14}$$

Finally, we achieve the following inverse information matrix:

$$\begin{aligned} \boldsymbol{I}^{-1} &= \begin{pmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{pmatrix} \\ &= \begin{pmatrix} (I_{11} - I_{12} I_{22}^{-1} I_{21})^{-1} & -(I_{11})^{-1} I_{12} (I_{22} - I_{21} I_{11}^{-1} I_{12})^{-1} \\ -(I_{22})^{-1} I_{21} (I_{11} - I_{12} I_{22}^{-1} I_{21})^{-1} & (I_{22} - I_{21} I_{11}^{-1} I_{12})^{-1} \end{pmatrix} \end{aligned}$$

At the end, we see that $(I^{11})^{-1} = I_{11} - I_{12} I_{22}^{-1} I_{21}$. This was what we aimed to find. For further information, we refer to Klein and L. [1997].

A.2 Calculating $\left(I^{11}\right)^{-1}$ in practise

To calculate $(I^{11})^{-1}$, we need information-, and covariance matrices. The information matrix obtained in R is given by:

$$I(\hat{\boldsymbol{\beta}}) = I(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = I(0.46, -1.67, 0.113) = \begin{pmatrix} 14.21 & -1.34 & 4.36 \\ -1.34 & 11.44 & -28.13 \\ 4.35 & -28.13 & 764.05 \end{pmatrix}$$

The covaraince matrix which is the inverse of the information matrix is given by

$$I(\hat{\boldsymbol{\beta}})^{-1} = I(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)^{-1} = \begin{pmatrix} 0.0711 & 0.0080 & -0.0001 \\ 0.0080 & 0.0970 & 0.0035 \\ -0.0001 & 0.0035 & 0.0014 \end{pmatrix} \tag{A.15}$$

Thus formula (A.11) is calcuated as $(I^{11})^{-1} = I_{11} - I_{12} I_{22}^{-1} I_{21} = 14.21 - (-1.34) \times 11.44^{-1} \times (-1.34) = 14.05$.

Appendix B

Attachment to Chapter 3

In this part we attach **R**-code that is used for analyzing the melanoma data. One can use similar **R**-code to analyse the pbc data.

B.1 R-code for analysing the melanoma data

```
# load the packages
library(survival)
library(mfp)
library(timereg)
#
# read the melanoma data
melanoma <- read.table("melanoma.txt",header=T)
#
#============================================================
#====================QUADRATIC METHOD=======================
#============================================================
#
# creating new variable x^2 for g(x1)=x1^2:
melanoma$thick2 <- (melanoma$thickn)**2
#
# fit a model with covariates thickness, ulceration and sex:
fit.mode0=coxph(Surv(lifetime,status==1)~thickn+factor(sex)+
factor(ulcer), data=melanoma)
summary(fit.mode0)
#
# specify an initial beta-vector with thick2=0:
coef0 <- c(0,fit.mode0$coef)
#
#  fit a model with covariates thickness, thickness^2,
# ulceration and sex:
#
fit.mode1 = coxph(Surv(lifetime,status==1)~ thick2 + thickn +
                  factor(sex)+factor(ulcer),init=coef0,
                  data=melanoma)
summary(fit.mode1)
fit.mode1$loglik
```

```
fit.mode0$loglik
#
#=============================================================
#====================QUARTILES METHOD=========================
#=============================================================
# a model based on thickness, sex and ulceration:
# ===============================================
fit.thickn=coxph(Surv(lifetime,status==1)~thickn+factor(sex)+
factor(ulcer), data=melanoma)
summary(fit.thickn)
#
# initial beta_vector
coef1 <- c(0,0,0,fit.thickn$coef)
#
# make categorical variables for thickness
b.thk=c(0,quantile(melanoma$thickn[melanoma$status==1])[2:4],
    max(melanoma$thickn))
melanoma$thick.group=cut(melanoma$thickn,b.tck,labels=1:4)
#
# create variables for quartiles (0.25%, 0.50%, 0.75%):
#=====================================================
#
melanoma$thick.q2 = melanoma$thickn*(melanoma$thick.group==2)
melanoma$thick.q3 = melanoma$thickn*(melanoma$thick.group==3)
melanoma$thick.q4 = melanoma$thickn*(melanoma$thick.group==4)
#
#=============================================================
#====model based on thickn, quartiles and other covariates===
# ============================================================
#
cox.thickq=coxph(Surv(lifetime,status==1)~thick.q2+thick.q3+
thick.q4+       thickn+factor(sex)+factor(ulcer),
data=melanoma,init=coef1)
summary(cox.thickq)
1-pchisq(4.76,3) #  0.190
1-pchisq(4.93,3) #  0.177
1-pchisq(5.06,3) #  0.167
#
# the formal test indicates non-log-linearity of the model
#
#=============================================================
#===============FRACTIONAL POLYNOMIAL METHOD ===============
#=============================================================
#
fp.fit=mfp(Surv(lifetime,status==1)~fp(thickn)+factor(sex)+
factor(ulcer),family=cox,
data=melanoma)
print(fp.fit)
summary(fp.fit)
# p-value
fp.fit$pvalues
#
#=============================================================
#====================P-SPLINE METHOD=========================
```

```
#=========================================================
#
fit.pstu=coxph(Surv(lifetime,status==1)~ pspline(thickn)+
factor(sex)+ factor(ulcer),data=melanoma)
#
print(fit.pstu)
par(mfrow=c(1,2))
termplot(fit.pstu,terms=1,se=T)
attributes(summary(fit.pstu))
summary(fit.pstu)
#
# Command to find the p-value corresponding to testing of
# log-linearity:
#
sum.fit=summary(fit.pstu)
sum.fit$coef
sum.fit$coef[2,6]
#
# We check then log-linearity of logthickness in a model
# with logthickness and ulceration and sex as covariates:
#
melanoma$log2thik=log2(melanoma$thickn)
fit.pslogtu=coxph(Surv(lifetime,status==1)~pspline(log2thck)
+ factor(ulcer) + sex,data=melanoma)
#
print(fit.pslogtu)
#
sum.fitl <- summary(fit.pslogtu)
sum.fitl$coef[2,6]
# term plot
par(mfrow=c(1,3))
termplot(fit.pslogtu,terms=1,se=T)
dev.copy(png,"pspline_thick_logthick.png")
dev.off()
#
#=========================================================
#================MARTINGALE-RESIDUALS METHOD =============
# ========================================================
#
# We first consider a model with ulceration and thickness
#       (not log-transformed) and sex:
#
fit.ut=cox.aalen(Surv(lifetime,status==1)~prop(ulcer)+
prop(thickn)+ prop(sex), data=melanoma, weighted.test=0,
residuals=1,rate.sim=0,n.sim=1000)
summary(fit.ut)
#
#============CHECKING OF LOG-LINEARITY=================
resids.ut=cum.residuals(fit.ut,data=melanoma,cum.resid=1,
n.sim=1000)
#
plot(resids.ut,score=2,xlab="Tumor thickness")
dev.copy(png,"lglin_thick.png")
dev.off()
```

```
summary(resids.ut)
#
# We then check log-linearity for a model with
# log-transformed thickness:
#
fit.ult=cox.aalen(Surv(lifetime,status==1)~prop(ulcer)+
prop(log2(thickn))+prop(sex),data=melanoma,weighted.test=0,
residuals=1,rate.sim=0,n.sim=1000
#
# checkin the functional form of tumor thickness
resids.ult=cum.residuals(fit.ult,data=melanoma,cum.resid=1,
n.sim=1000)
#
plot(resids.ult,score=2,xlab="log tumor thickness")
dev.copy(png,"lglin_log2thick.png")
dev.off()
summary(resids.ult)
```

Appendix C

Attachment to Chapter 4

In this part of appendix, we attach figures that are redundant being part of the main results. This part should be reading together with the results from chapter 4. **R**-code for this chapter 4 is not included since we have tried to explaine the simulation procedure with all the details, as also included an algorithm to summarize the steps. Therefore we thought it is not necessary to provide also the code.

C.1 Simulation studies

Figure C.1 is plotted similarly as figure 4.3, but for $\lambda = 0.0003$ and $\gamma = 0$. The curves seem fine - that's what we expect namely that the power-test is around 6%. Figure C.2 is also plotted in similar manner, it seems similarly to the other plots, p-splines does not detect violation from the log-linearity assumption good enough in terms of ploting in comparison to the Martingale residuals method.

Figure C.2 is plotted similarly as the two other plots, but for $\lambda = 0.003$ and $\gamma = -0.5$.

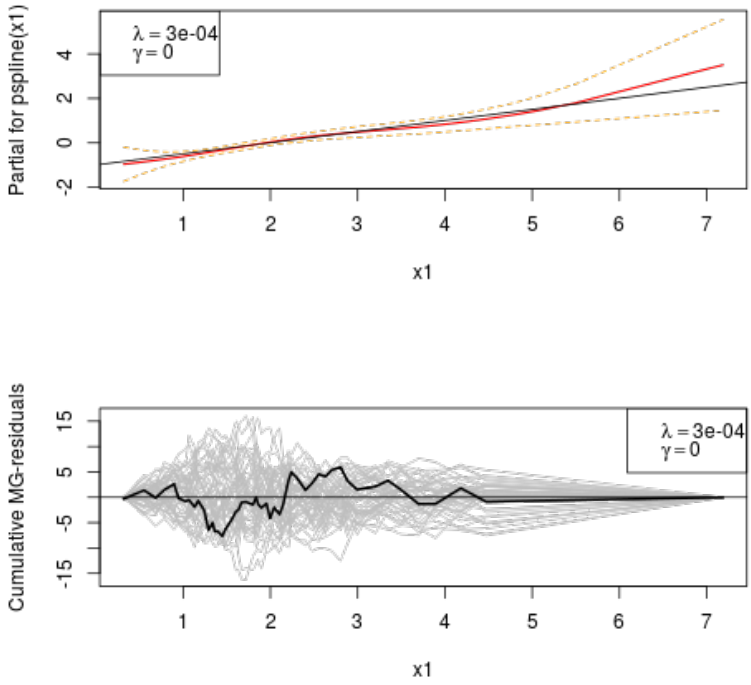

Figure C.1: The upper plot is one simulated Cox regression partial P-splines against x_1. And the lower plot is the one simulated Martingale residuals against x_1. In both plots, $N = 200, B = 1000, \lambda = 0.0003$ and $\gamma = 0$.

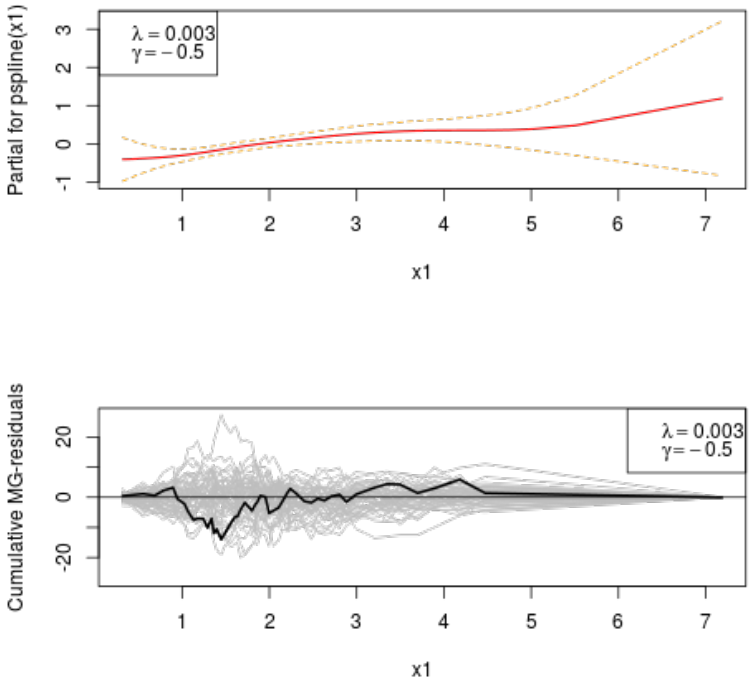

Figure C.2: The upper plot is one simulated Cox regression partial P-splines against x_1. And the lower plot is the one simulated Martingale residuals against x_1. In both plots, $N = 400$, $B = 1000$, $\lambda = 0.003$ and $\gamma = -0.5$