Universitetet
i Stavanger

FACULTY OF SCIENCE AND TECHNOLOGY

MASTER'S THESIS

Study programme/specialisation: **Lektor realfag 8.-13. trinn**	**Spring semester, 2020** **Open**
Author: **Dag Recep Eroglu Eilertsen**	**Dag Recep Eroglu Eilertsen** (signature of author)

Programme coordinator:

Supervisor:
Jan Terje Kvaløy

Title of master's thesis:
Survival Analysis using Cox Regression on Breast Cancer Data

Credits: **30**

Keywords: **Survival Analysis, Cox Regression, Breast Cancer**	Number of pages: **88** + supplemental material/other: **22** **Kopervik, 14.06.2020**

Summary

In this report, survival data from a german breast cancer study has been analysed using the programming software R. For the 686 female patients participating in the study, the value of eight explanatory variables were recorded at the start of the study. These variables were age, menopause status, whether the patient received tamoxifen or not, tumor grade, tumor size, number of positive lymph nodes, and amount of progesterone and estrogen bound to proteins in the cytosol of the primary tumor. Both time to recurrence of tumor and time to death were recorded for each patient. The focus has been to find out how important the explanatory variables are when it comes to time to recurrence and time to death. After creating *Kaplan-Meier curves* and doing *log-rank tests*, the data were analysed using *Cox regression.* The method of *purposeful selection* was used to choose which of the explanatory variables that should be included in the Cox regression model. *Schoenfeld residuals plots* were used to identify wheter or not the assumption of proportional hazards has been obeyed. *Martingale residuals plots* were used to detect the functional form that should be used for the explanatory variable values in the models. After performing purposeful selection, size, grade, nodes and progesterone were the variables that remained for time to death. For time to recurrence, tamoxifen, grade, nodes and progesterone were the ones that remained. An attempt to model recurrence as a time-dependent variable was made for time to death, and it was found that people experiencing recurrence has a much higher chance of death than those not experiencing recurrence.

Weibull distributed survival times were simulated by assuming the value of three explanatory variables (normally, exponentially and uniformly distributed) and their associated regression coefficients. A data frame of the simulated survival data were created, and Cox regression were runned on this data frame to check if the assumed regression coefficients were reproduced. The 95 % confidence intervals for the regression coefficients produced by the Cox regression machinery were found to include the assumed regression coefficient values. It was found that increasing the standard deviation of the normally distributed explanatory variable increased the accuracy of the regression coefficient estimates. Increasing the number of simulations was also found to increase the accuracy of the estimates.

Survival data which had non-proportional hazards were simulated by an inbuilt R-function called *sim.survdata.* These data were used to test whether or not the Schoenfeld residuals plot could detect the assumed functional form of a time-dependent regression coefficient. From the plot it was possible to detect that the assumed functional form had the graph of a parabola.

Preface

I would like to thank those who have been important to me during my work on the master's thesis.

First, I want to thank Jan Terje Kvaløy for supervising me during this master's thesis project. He has done a tremendous job guiding me during my work on this report. During our weekly meetings he has been giving me feedback on my report work, both in terms of content, layout, R-programming and how to use the Latex typesetting system.

I also want to thank my loved ones for their moral support. Both my mother, father, brothers, my partner and her family and my friends believed in me and motivated me to stay focused on the task of finishing this report.

Finally, I want to thank the University of Stavanger for the time I have been a student in their institution. I have spent the last eight years together with them and have learned a lot about mathematics, science and myself during this time. Now I'm looking forward to start working as a teacher to share the knowledge I've acquired myself during these eight years.

Enjoy reading.

Sincerely yours,
Dag Recep Eroglu Eilertsen.

Contents

1 Introduction 1

2 Intoduction to Survival Analysis 2

2.1 Survivor and hazard function 4

2.2 The Kaplan-Meier estimator 9

2.3 The log-rank test 15

3 Cox Regression 19

3.1 Introducing the model 19

3.2 Interactions 25

3.3 Including categorical variables in the model 25

3.4 Residuals 27

3.4.1 Schoenfeld residuals 27

3.4.2 Martingale residuals 29

3.5 Time-dependent explanatory variables 30

4 Application 32

4.1 Introducing the data set 32

4.2 Kaplan-Meier curves 36

4.2.1 Time to death 38

4.2.2 Time to recurrence 40

4.3 Log-rank test 41

4.3.1 Time to death 42

4.3.2 Time to recurrence 42

4.4 Determining the regression coefficients 43

4.4.1 Time to death 47

4.4.2 Time to recurrence 55

4.4.3 Recurrence as time-dependent variable 65

5 Simulations 71

5.1 The Weibull distribution 71

5.2 The Inverse Transform Method for generating random variables 72

5.2.1 Applying The Inverse Transform Method on a Weibull distributed random variable 73

5.3 Simulation of Weibull distributed survival times 74

5.3.1 Trying to predict the regression coefficients' values . . . 75

5.4 Survival data for which the proportional hazards assumtion is not satisfied 79

6 Conclusion 81
6.1 Chapter 4: Application 81
6.1.1 Time to death 81
6.1.2 Time to recurrence 82
6.1.3 Recurrence as time-dependent variable 83
6.2 Chapter 5: Simulations 83
6.2.1 Simulation of Weibull distributed survival times 83
6.2.2 Survival data for which the proportional hazards assumtion is not satisfied 84
6.3 Further work 84

Bibliography 87

Appendix A 89
A.1 Programming script from the R software 89

Chapter 1

Introduction

This report is about Survival Analysis and how to use Cox Regression to analyse survival data. Before presenting any details, a short explanation of the content in the different chapters in this report will be given.

In chapter 2 an introduction to basic survival analysis is given. Basic terminology and principles are presented. The survivor and hazard function, the Kaplan-Meier estimator and the log-rank test are the important topics in this chapter.

In chapter 3, the basics of Cox Regression is presented. The main focus is the proportional hazard model. Schoenfeld and Martingale residuals are also important topics in this chapter.

In chapter 4, the theory established in the two first chapters is applied on survival data from a german breast cancer study. Analysis is done for both time to death and time to recurrence.

In chapter 5, it is shown how to simulate survival data where the survival times are Weibull distributed. Cox regression is then applied to the simulated survival data to see if the values of the regression coefficients that was assumed during simulation are reproduced.

In chapter 6 the main focus is to present conclusions of the results from chapter 4 and 5. Suggestions to further work will also be given.

To do the estimates and simulations in chapter 4 and 5, the programming software R was used. The scripts showing the code that was used are found in Appendix A that follows right after chapter 6.

The report is ended with a bibliography list, showing the literary works used in this report. In this report the IEEE reference style was used.

Chapter 2

Intoduction to Survival Analysis

In this chapter, there will be given an introduction to a branch of Statistics called "***Survival Analysis***". The source of information is Chapter 1 and 2 in David Collett's book "*Modelling Survival Data in Medical Research*" [1].

In survival analysis, the data being analysed is *time*. The times are measured "*from a well-defined origin until the occurence of some particular event or end-point.*", [1, p. 1]. Later in this report, time data from medical research will be analysed. In medical research, the *time origin* t_0 is set to be the time a patient enters the medical study[1]. If the *end-point* is the death of the patient, the time data is called "*survival data*" [1]. If the end-point is not the death of the patient (but some other event), then the time data is called "*time to event data*"[1]. *Study time*[1] is the time from start of study t_{start} until the end of study t_{end}. During the study time patients are recruited and followed up. When the study time is over, the analysis of the survival times starts. *Patient time* is the time a patient spends in a study[1].

Now, consider a medical research study that looks at survival data for a group of patients during the study time in the interval $[t_{\text{start}}, t_{\text{end}}]$. Since the study is interested in survival data, the information of interest is when the patients in the study dies. A particular patient enters the study at time t_0. Now one of three events can happen to the patient during the study time[1]:
1) The patient dies
2) The patient is lost to follow-up
3) The patient is alive when the study time is over

In case 1, the patient stays in the study from t_0 to $t_0 +\ t$, where t is called the patient's *survival time.* That is, the patient dies t time units after it entered the study. Since the patient is not confirmed "dead" in case 2 and 3, these cases will not give rise to actual survival times. They will give rise to *censored* survival times. When a patient's survival time is censored, it simply means that an actual death/ end point is not observed for that particular patient. In case 2, for some reason the patient is lost to follow-up (the patient could have moved to another city or country, which would make it difficult to show up to participate in the study). The only thing known is that the patient

was alive during the last show-up at the research center, which was at time $t_0 + c_2$. This makes c_2 a censored survival time.
At time $t_0 + c_3$ the study time is over, and the patient from case 3 is still alive. c_3 is also a censored survival time.
Note that, each patient has its own t_0, the time the patient enters the study. This time is not used when the survival data analysis is made. Then, only the censored and actual survival times c and t are being used. A patient's survival time (actual or censored) starts at *zero* (this is usually when the patient is diagnosed with a disease), and ends when one of the three events mentioned above occurs.

There exist different types of *censoring*. Three types are mentioned in Chapter 1 of Collett's book [1]:

1) *Left censoring*
This is when we only know that the actual survival time is smaller than the censored one. That is, $t < c$, where t and c are actual and censored survival times, respectively. An example could be recurrence of a cancer tumor. Imagine you have a patient who's cancer tumor has been removed by surgery. One is interested in finding how long it goes before the tumor recurs. The survival time of the patient starts after the tumor is removed, and ends when the tumor's recurrence is observed. At the time incident the tumor is observed to recur (this will be the censored survival time), one knows that the actual time of recurrence must have been some time before the observed time (tumors do not grow in a second). If we had drawn a time line, the actual time would lie to the left of the censored time. This is the reason this kind of censoring is called "left censoring".

2) *Right censoring*
This is when we observe that the actual survival times is larger than the censored one. That is, when $c < t$. An example of this could be if a patient is alive when the study time is over. Then the only information we have is that the patient will die at some later time. If we again had drawn a time line, this actual time would lie to the right of the censored time.

3) *Interval censoring*
Interval censoring is when the actual survival time is known to lie inside a specific time interval.

Another aspect of censoring is whether the censoring is informative or not.

The three type of censoring mentioned above can all be subject to *informative censoring.* Assume a medical research study is investigating whether a given treatment is increasing the survival time of the patients participating in the study. It could happen that, the treatment has such a negative effect on the health of some of the patients that the treatment must be withdrawn from the study for those patients. For the sake of the study, continuing observing patients who is not receiving treatment anymore makes no sense. For this reason, the study must be terminated for those patients it concerns, and the patients' survival times must be censored. This type of censoring is called "*informative.* As a matter of fact, in this example the survival time is also right censored, since it is known that the patient is alive after the censored survival time. Collett emphasizes in his book that, for the methods presented in his book to be valid, the censored survival times must be of *non-informative* character. Non-informative censoring means that the censoring is "*not related to any factors associated with the actual survival time*"[1, p. 318]. In other words, the censored survival time doesn't carry any information about the patient's risk of death. Non-informative censoring will be assumed in the remainder of this report.

2.1 Survivor and hazard function

In the following, the *survivor function* and the *hazard function* will be presented similar to how Collett presents them in chapter 1.3 of his book [1, p. 11-13]. To understand these functions and the relationship between them, it is necessary to establish an understanding of the following terminology:
Continuous random variable, *Probability distribution* and
Cumulative density function. Løvås presents these terms in his book in Statistics [2]:

1) *Continuous random variable*
The survival time of persons in medical research studies is an example of a continuous random variable. The randomness is manifested in that it is not possible to predict the survival time of a patient (he/ she could die at any instant), and the continuity arises from the fact that the survival time of a patient constitutes a time interval from time origin to end-point (and since there is infinitely many discrete time values inside this time interval, the variable is continuous, opposed to discrete).

2) *Probability distribution*
A *probability distribution* presents how the probability is distributed among the different possible values of a random variable. For a continuous random variable the probability distribution is represented by a function f(x) that is called the "*probability density function*" (often denoted by "pdf"). A pdf has the following properties [2, p. 133]:

a) The total area under the function's curve is equal to 1

b) The probability for the random variable to have a value between a and b (denoted by $P(a \leq X \leq B)$) is equal to the area under the curve of the function $f(x)$ from a to b. With integral notation, this is written as

$$P(a \leq X \leq b) = \int_a^b f(x)dx \tag{2.1}$$

c) The curve is never negative, that is $f(x) \geq 0$.

3) *Cumulative distribution function* (cdf)
The *cumulative distribution function* F(x) (often abbrevated as "cdf") is defined such that its first order derivative yields the corresponding *probability density function* f(x). That is, F'(x)=f(x), which means $F(x) = \int_{-\infty}^{x} f(x)dx$. Its properties is such that it converges to zero when the argument approaches minus infinity from above, and it converges to one when the argument approches plus infinity from below. For any two arbitrary constants a and b (assuming $a < b$), the cdf obeys the following three properties [2, p. 125,135]:

a)

$$P(a \leq X \leq b) = F(b) - F(a) \tag{2.2}$$

b)

$$P(X > a) = 1 - F(a) = 1 - \int_{-\infty}^{a} f(x)dx \tag{2.3}$$

c)

$$P(X \leq b) = F(b) = \int_{-\infty}^{b} f(x)dx \tag{2.4}$$

Now we have established the terminology needed to define the survivor and hazard function.

Let T be a continuous random variable representing the survival time t of a patient. *f(t)* is the pdf representing the probability distribution of T. The cdf of T is "*given by*

$$F(t) = P(T < t) = \int_0^t f(u)du, \tag{2.5}$$

and represents the probability that the survival time is less than some value t." [1, p. 11]. The *survivor function* is then defined as

$$S(t) = P(T \geq t) = 1 - F(t), \tag{2.6}$$

which is *"the probability that the survival time is greater than or equal to t.* [1, p. 11].

The *hazard function*, sometimes denoted by "*hazard rate, the instantaneous death rate, the intensity rate, or the force of morality.*"[1, p. 11], is given by

$$h(t) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t}. \tag{2.7}$$

$P(t \leq T < t + \delta t | T \geq t)$ is the probability that T has a value in the interval $[t, t + \delta t]$, given that T is greater than or equal to t.

It is of interest to establish a relationship between the hazard and survivor function. A first step in doing so is to make use of *Bayes' Theorem*. By using Bayes' Theorem, it is possible to obtain an expression for the probability in the numerator of the hazard function. Given two events A and B, *Bayes' Theorem* is given by [1, p. 86]

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)}, \tag{2.8}$$

both $P(B|A)$ and $P(A|B)$ being conditional probabilities, where $P(B|A)$ is the probability of B given that A has occured, and $P(A|B)$ is the probability of A given that B has occured. $P(A)$ and $P(B)$ are the individual probabilities of A and B, respectively.

Now let A be the event "$T \geq t$" and B the event "$t \leq T < t + \delta t$". Inserting this into Bayes' Theorem yields:

$$P(t \leq T < t + \delta t | T \geq t) = \frac{P(t \leq T < t + \delta t) P(T \geq t | t \leq T < t + \delta t)}{P(T \geq t)} \tag{2.9}$$

Observing that $P(T \geq t | t \leq T < t + \delta t) = 1$ simplifies the equation to

$$P(t \leq T < t + \delta t | T \geq t) = \frac{P(t \leq T < t + \delta t)}{P(T \geq t)} \tag{2.10}$$

Usıng *property a* of the cdf given by (2.2), the numerator of (2.10) can be written as:

$$P(t \leq T < t + \delta t) = F(t + \delta t) - F(t) \tag{2.11}$$

The denominator of (2.10) is simply the survivor function defined in (2.6). Substituting the results from (2.11) and (2.6) into (2.10) gives:

$$P(t \leq T < t + \delta t | T \geq t) = \frac{F(t + \delta t) - F(t)}{S(t)} \tag{2.12}$$

Replacing the numerator of the the hazard function (2.7) by the result from (2.12) yields:

$$h(t) = \lim_{\delta t \to 0} \frac{F(t + \delta t) - F(t)}{\delta t} \frac{1}{S(t)} \tag{2.13}$$

Note that

$$\lim_{\delta t \to 0} \frac{F(t + \delta t) - F(t)}{\delta t} = F'(x) = f(x).$$

Thus, the hazard function can be rewritten as

$$h(t) = \frac{f(t)}{S(t)} \tag{2.14}$$

By further manipulating the expression of the hazard function it is possible to end up with a function called the "integrated *or* cumulative hazard" [1, p. 12]. Let's do this manipulation. First step is to take the natural logaritm of the survivor function:

$$ln(S(t)) = ln(1 - F(t)) \tag{2.15}$$

Taking the first order derivative of this gives:

$$\frac{d}{dt} ln(S(t)) = \frac{1}{1 - F(t)} \frac{d}{dt}(-F(t)) \tag{2.16}$$

Recognizing that $\frac{d}{dt}(-F(t))$ is equal to $-f(t)$ results in:

$$\frac{d}{dt}ln(S(t)) = \frac{-f(t)}{1-F(t)} \tag{2.17}$$

Multiplying both sides by -1 and identifying the denominator of (2.17) as the survivor function gives:

$$-\frac{d}{dt}ln(S(t)) = \frac{f(t)}{S(t)} \tag{2.18}$$

With a closer look, one can see that the right hand side of (2.18) is the hazard function. Thus,

$$-\frac{d}{dt}ln(S(t)) = h(t) \tag{2.19}$$

Integrating both sides of (2.19) from zero to t gives:

$$\int_0^t -\frac{d}{dt}ln(S(u))du = \int_0^t h(u)du \tag{2.20}$$

$$-\left[ln(S(u))\right]_0^t = \int_0^t h(u)du \tag{2.21}$$

$$ln(S(t)) - ln(S(0)) = -\int_0^t h(u)du \tag{2.22}$$

Noting that $ln(S(0)) = P(T \geq 0) = 1$ yields $ln(S(0)) = ln(1) = 0$. Thus

$$ln(S(t)) = -\int_0^t h(u)du \tag{2.23}$$

$$-ln(S(t)) = \int_0^t h(u)du \tag{2.24}$$

The right-hand side of (2.24) is the *cumulative hazard function* $H(t)$.

$$H(t) = \int_0^t h(u)du \tag{2.25}$$

Note that the survivor function can be expressed through the cumulative hazard function:

$$S(t) = e^{-H(t)} \tag{2.26}$$

2.2 The Kaplan-Meier estimator

In the following, methods for estimating the survivor, hazard and cumulative hazard functions associated with a group of survival data will be presented. The *Log Rank test* (method for comparing two or more groups of survival times) will also be considered. The source of information is Chapter 2 i Collett's book [1].

When none of the survival times in a survival data set is censored, the survivor function "*can be estimated by the empirical survivor function, given by*"[1, p. 15]

$$\hat{S}(t) = \frac{\text{Numer of individuals with survival times} \geq t}{\text{Number of individuals in the data set}} = 1 - \hat{F}(t), \quad (2.27)$$

where $\hat{F}(t)$ is the *empirical cumulative distribution function* defined by

$$\hat{F}(t) = \frac{\text{Numer of individuals alive at time t}}{\text{Number of individuals in the data set}}. \quad (2.28)$$

It is possible to plot the empirical survivor function. Collett mentions three features of this plot [1, p. 15-16]:

1) For survival times less than the lowest survival time in the data set, the empirical survivor function is equal to 1

2) For survival times larger than the largest survival time in the data set, it is equal to 0

3) Its value between two adjacent survival times are constant

Now, imagine that there exist 11 patients with the following survival times (in months), where no censoring is assumed:

12 14 14 14 14 14 15 15 16 16 18

From this data set it is possible to create a plot of the estimated survivor function. To make such a plot, I've used the in-built function "StepGraph" belonging to the software "GeoGebra". I did the calculation of the values of the estimated survivor function using Microsoft Excel. See the plot in figure 2.1 below.

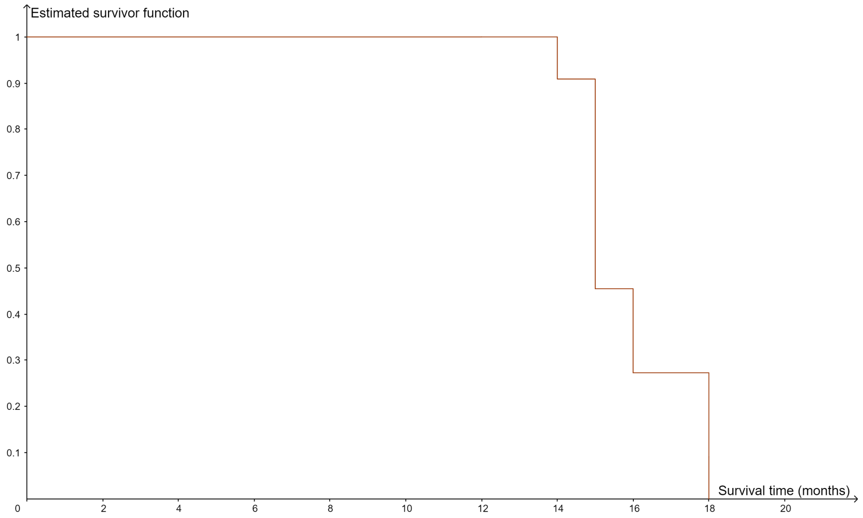


Figure 2.1: Plot of estimated survivor function for 11 imaginary survival times, no censoring.

Along the vertical axis of figure 2.1 you have the values of the estimated survivor function calculated from (2.27) using Microsoft Excel. Along the horisontal axis you have the corresponding survival times in months. From the plot, you can see that the survivor function has a step-wise nature.

When there exist censored survival times in a data set, other methods than the empirical survivor function must be used to estimate the survivor function. The *Kaplan-Meier estimate* of the survivor function (”*a generalisation of the empirical survivor function*” [1, p. 22]) is one such method. The method can be summarized in the following four steps [1]:

1) Start with a data set consisting of the survival times from n different patients, that is, the survival times $t_1, t_2, t_3, ..., t_n$. Allowing different patients to have the same survival time, find out at which times there are death events (actual survival times). Write down the death event times and count how many they are. Label the counted number as ”r”. This will give rise to the

iteration variable $j = 1, 2, 3, ..., r$.

2) Now that you know at which times there are death events, order these death event times in ascending order:
$t_1 < t_2 < t_3 < \cdots < t_r$

3) Now you have to use the survival times from step 2 to create intervals. The idea is that, inside each interval, there is only one death event time. In Figure 2.2 below, a time line is drawn to illustrate how the intervals are constructed.

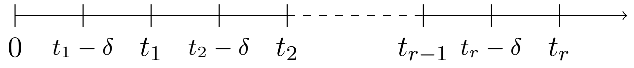


Figure 2.2: Illustration of Kaplan-Meier estimate interval

From Figure 2.2 you see that the first time value is 0. This is the time of diagnosis of the different patients, which is set as the time origin. Then you have $t_1 - \delta$, which is the time just before the first death time, which is at t_1. δ is infinitesimal. In other words, in the interval stretching from 0 to $t_1 - \delta$, there are no deaths. Now we allow ourselves to let the first death time t_1 lie inside the interval stretching from $t_1 - \delta$ to t_1. The next time on the time line is $t_2 - \delta$. This is the time right before the second death time, which is at t_2. Similarly, there are no deaths in the interval stretching from t_1 to $t_2 - \delta$, and we let the second death time fall inside the interval stretching from $t_2 - \delta$ to t_2. If generalising by using the iteration variable j, we can say that: The interval stretching from $t_j - \delta$ to t_j is the time interval where the deaths occurs, and the interval stretching from t_j to $t_{j+1} - \delta$ is the interval without any deaths. In addition to this, we denote n_j to be the number of patients alive at the beginning of the time interval from $t_j - \delta$ to t_j (including deaths at t_j). Also, we denote d_j to be the number of patients who's deaths lie inside the interval from $t_j - \delta$ to t_j, keeping in mind that we allow several patients to die at the same time. The censored survival times are constructed to lie inside those time intervals without any deaths. What is meant by that, is: If it happened that a censored survival time was exactly the same as a death time t_j, then this censored survival time is placed in the interval from t_j to $t_{j+1} - \delta$.

4) Now that the intervals are constructed, it is possible to start calculating with probabilities that will eventually lead to an estimate of the survivor function. The question of interest is: What is the probability of surviving through the different intervals constructed in step 3 above? Since there are no deaths occuring in the intervals stretching from t_j to $t_{j+1} - \delta$, the estimated probability of surviving through each of these intervals is 1. What about the intervals in which deaths occur, the intervals from $t_j - \delta$ to t_j? For each of these intervals, the probability of dying is estimated by

$$P(death) = \frac{d_j}{n_j}. \tag{2.29}$$

Using the *Complement rule* [2, p. 77] of Statistics, the probability of surviving through one of these intervals is estimated by:

$$P(survival) = 1 - \frac{d_j}{n_j} = \frac{n_j}{n_j} - \frac{d_j}{n_j} = \frac{n_j - d_j}{n_j}. \tag{2.30}$$

Recall from (2.6) that, the survivor function $S(t)$ gives the probability that the survival time is greater than or equal to an arbitrary survival time t. If you were to choose a survival time in the interval $[0, t_1 - \delta]$, then $\hat{S}(t) = 1$, because no one has died yet. On the other hand, if you choose t inside the interval $[t_r, \infty]$, then one of two things can happen (depending on whether the largest survival time is censored or not):
a) If the largest survival time is an actual death event time, then $\hat{S}(t) = 0$, because everyone is dead after t_r.
b) If there exists a censored event time c that is larger than the largest death time t_r, then $\hat{S}(t)$ will not tend to zero, but will remain constant. The reason for this is that, when the largest survival time in the data set is censored, the last observed event is not a death. Thus, $\hat{S}(t)$ will never become zero, because one or several individuals are not observed to be dead during the course of the study (as a matter of fact, these individuals' survival times are right censored).
What if you were to choose t to lie inside any of the intervals between t_1 and t_r (that is, $t_1 \leq t < t_r$)? What would $\hat{S}(t)$ be then? To calculate this, we must make use of the multiplication rule of independent events [2, p. 94]:

$$P(A_1 \cap A_2 \cap A_3 \cap \cdots \cap A_n) = P(A_1) \cdot P(A_2) \cdot P(A_3) \cdot \ldots \cdot P(A_n) \tag{2.31}$$

To exemplify, let's assume we have a data set with 20 death times, where we look at the four first of them. Also, define the following events:

A_1 = Survival through the interval $[0, t_1 - \delta]$
A_2 = Survival through the interval $[t_1 - \delta, t_1]$, given survival up to $t_1 - \delta$
A_3 = Survival through the interval $[t_1, t_2 - \delta]$, given survival up to t_1
A_4 = Survival through the interval $[t_2 - \delta, t_2]$, given survival up to $t_2 - \delta$
A_5 = Survival through the interval $[t_2, t_3 - \delta]$, given survival up to t_2
A_6 = Survival through the interval $[t_3 - \delta, t_3]$, given survival up to $t_3 - \delta$
A_7 = Survival through the interval $[t_3, t_4 - \delta]$, given survival up to t_3
A_8 = Survival through the interval $[t_4 - \delta, t_4]$, given survival up to $t_4 - \delta$

If we choose t to lie in the interval $[t_4 - \delta, t_4]$, then the Kaplan-Meier estimate of $S(4)$ would be:

$$\hat{S}(t) = P(A_1) \cdot P(A_2) \cdot P(A_3) \cdot P(A_4) \cdot P(A_5) \cdot P(A_6) \cdot P(A_7) \cdot P(A_8) \tag{2.32}$$

Setting the probability to be 1 for those intervals where there are no deaths gives:

$$\hat{S}(t) = 1 \cdot P(A_2) \cdot 1 \cdot P(A_4) \cdot 1 \cdot P(A_6) \cdot 1 \cdot P(A_8) \tag{2.33}$$

With a closer examination of (2.33), one can see that each of the remaining probabilities in the product are given by (2.30). Thus:

$$\begin{aligned}\hat{S}(t) &= P(A_2) \cdot P(A_4) \cdot P(A_6) \cdot P(A_8) \\ &= \frac{n_1 - d_1}{n_1} \cdot \frac{n_2 - d_2}{n_2} \cdot \frac{n_3 - d_3}{n_3} \cdot \frac{n_4 - d_4}{n_4}\end{aligned} \tag{2.34}$$

It is possible to generalise (2.34) for any t bounded by
$t_k \leq t < t_{k+1}$, where $k = 1, 2, 3, \ldots, r$. The generalized formula is known as the *Kaplan-Meier estimate of the survivor function*, and is given by:

$$\hat{S}(t) = \prod_{j=1}^{k} \frac{n_j - d_j}{n_j}. \tag{2.35}$$

Earlier, it was mentioned that the Kaplan-Meier estimate of the survivor function is a generalisation of the empirical survivor function. Let's show how this is the case.
If we assume no censoring, then the numerator in (2.35) becomes n_{j+1}. If we insert this into the equation and espand the product, we get:

$$\hat{S}(t) \quad = \quad \prod_{j=1}^{k} \frac{n_{j+1}}{n_j} \quad = \quad \frac{n_2}{n_1} \cdot \frac{n_3}{n_2} \cdot \frac{n_4}{n_3} \cdot \ldots \cdot \frac{n_{k-1}}{n_{k-2}} \cdot \frac{n_k}{n_{k-1}} \cdot \frac{n_{k+1}}{n_k} \tag{2.36}$$

If we cancel common factors, we are left with:

$$\hat{S}(t) = \frac{n_{k+1}}{n_1} \tag{2.37}$$

The empirical survivor function from (2.27) was given by

$$\hat{S}(t) = \frac{\text{Numer of individuals with survival times} \geq t}{\text{Number of individuals in the data set}} = 1 - \hat{F}(t).$$

The numerator of (2.27) corresponds to n_{k+1} (”*the number of individuals with survival times greater than or equal to* t_{k+1}” [1, p. 21]), and the denominator corresponds to n_1 (”*the number of individuals in the sample*”).

Recall that, (2.29) gave us the estimated probability of dying through the jth interval from t_j to t_{j+1}:

$$P(death) = \frac{d_j}{n_j}.$$

If we divide (2.29) by $\tau_j = t_{j+1} - t_j$, we get the Kaplan-Meier estimate of the hazard function:

$$\hat{h}(t) = \frac{\frac{d_j}{n_j}}{\tau_j} \tag{2.38}$$

The Kaplan-Meier estimate of the hazard function ”*is an estimate of the risk of death per unit time in the jth interval*”[1, p. 31].

By combining (2.24) and (2.25) we get the following expression for the cumulative hazard function:

$$H(t) = -ln(S(t)). \tag{2.39}$$

So, an estimate of the cumulative hazard function would be

$$\hat{H}(t) = -ln(\hat{S}(t)). \tag{2.40}$$

If we insert the expression for $\hat{S}(t)$ from (2.35) into (2.40) we get

$$\hat{H}(t) = -ln\left(\prod_{j=1}^{k} \frac{n_j - d_j}{n_j}\right). \tag{2.41}$$

Recall that, for two numbers a and b,

$$ln(a \cdot b) = ln(a) + ln(b). \tag{2.42}$$

If we apply the result from (2.42) to (2.41) we get:

$$\hat{H}(t) = -\sum_{j=1}^{k} ln\left(\frac{n_j - d_j}{n_j}\right). \tag{2.43}$$

Another well-known estimate of the cumulative hazard function $\hat{H}(t)$ is the Nelson-Aalen estimate [1, p. 33]:

$$\widetilde{H}(t) = \sum_{j=1}^{k} \frac{d_j}{n_j}. \tag{2.44}$$

2.3 The log-rank test

Now the log-rank test for comparing two groups of survival data will be presented. The method "*can be extended to enable three or more groups of survival data to be compared*'[1, p. 48], but this will not be shown for the moment being. Before continuing, it is recommended to recap what was presented under the section concerning the Kaplan-Meier estimate of the survivor function, because the content covered there founds the basis of what is to be presented.

The first step in constructing the log-rank test is to start with two separate groups of patients (Group 1 and Group 2) with each their set of survival times. Then, looking at the survival times from both groups simultaneously, note down the death event times. Thereafter, exactly the same as we did when establishing the Kaplan-Meier estimate of the survivor function, we order the death event times in ascending order and note down how many deaths that occur at each death time, and how many that is alive right before a given death time (see section about Kaplan-Meier estimate of survivor function for further details about how this is done). Let now d_{1j} and d_{2j} represent the deaths occuring at time t_j in group 1 and 2, respectively. Further, let n_{1j} and n_{2j} be the number of patients alive right before t_j (including those dying at t_j) in group 1 and 2, respectively. The difference $n_{1j} - d_{1j}$ and $n_{2j} - d_{2j}$ are the number of patients surviving beyond the time t_j in group 1 and 2, respectively. Also define the following quantities: $d_j = d_{1j} + d_{2j}$, $n_j = n_{1j} + n_{2j}$ and $n_j - d_j = (n_{1j} - d_{1j}) + (n_{2j} - d_{2j})$. The different quantities are summarized in Table 2.1 below (reconstruction of table 2.7 in Collett [1, p. 42]).

Group number	Number of deaths at t_j	Number surviving beyond t_j	Number at risk just before t_j
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

Table 2.1: Log-rank test table for group I and II

The parameters in the table are needed to construct the test statistic used in the hypothesis testing procedure associated with the log-rank test. For a detailed explanation about what is meant with hypothesis testing and its related terminology, take a look in Løvås [2] or Collett [1] (both give a comprehensive introduction to the topic). The test statistic is a quantity used to determine whether or not to reject the null hypothesis of interest. In the hypothesis test associated with the log-rank test, the null hypothesis is the following event [1, p. 42]:

$$H_0 : \text{No difference in survival experience between the two patient groups.}$$

A possible scenario would be that patients in group 1 and 2 have the same disease, but they receive different treatment. The question one wishes to find the answer to is then if one of the treatments are better (gives rise to longer survival time) than the other. In case of difference between the two treatments, H_0 must be rejected. Assuming the null hypothesis is true, the test statistic is given by [1, p. 44]:

$$W_L = \frac{(U_L)^2}{V_L}, \tag{2.45}$$

where W_L has an asssociated probability distribution known as *chi-squared distribution* with one degree of freedom. For those interested in knowing how to derive this test statistic, see Collett[1]. U_L, V_L and their related quantities are given by [1, p. 42-43]:

$$U_L = \sum_{j=1}^{r} (d_{1j} - e_{1j}). \tag{2.46}$$

$$e_{1j} = \frac{n_{1j}}{n_j} \cdot d_j. \tag{2.47}$$

$$V_L = \sum_{j=1}^{r} v_{1j}. \tag{2.48}$$

$$v_{1j} = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{(n_j)^2 (n_j - 1)}. \tag{2.49}$$

Here (2.47) is the theoretical (expected) number of patients who dies in group 1 at time t_j, under the assumption that there are no differences between the two patient groups. $(d_{1j} - e_{1j})$ gives the difference between the observed and expected number of patients who dies in group 1 at time t_j, under the assumption of no difference. The smaller the sum of these differences U_L are (given by (2.46)), the stronger indication exists of there being no difference between the two groups. V_L is an estimator of the variance of U_L. Thus, large values of W_L indicate that H_0 does not hold.

The next step is to calculate the probability value (p-value) associated with W_L, which is given by

$$P - value = P(W_L \geq w_L) = 1 - P(W_L \leq w_L) = 1 - F(w), \tag{2.50}$$

where $F(w)$ is the cumulative distribution function of the chi-squared distribution with one degree of freedom (which values' are calculated by numerical integration performed by computers [2]).

When do we reject the null hypothesis and conclude that one of the group of patients tend to live longer than the other? Collett suggests the following approach [1, p. 40]:

If we denote the p-value by "P", then:

1) If $P > 0.1$, there is no evidence to reject the null hypothesis

2) If $0.05 < P \leq 0.1$, there is slight evidence against the null hypothesis

3) If $0.01 < P \leq 0.05$, there is moderate evidence against the null hypothesis

4) If $0.001 < P \leq 0.01$, there is strong evidence against the null hypothesis

5) If $P \leq 0.001$, the evidence against the null hypothesis is overwhelming.

Chapter 3

Cox Regression

In chapter 2, methods for analysing survival times were presented. In addition to survival time data, information like gender, age, heart rate, size of tumour, life style (smoking, physical activity, dietary etc.) and other factors that may play a role on the patient's survival time will be recorded in a medical research study. Such factors are called *explanatory variables*. Collett states that, there are two types of explanatory variables [1]: *variates* and *factors*. Variates are explanatory variables that take numerical values, like age, amount of hemoglobin in your blood stream, size of tumor etc. Factors on the other hand is not numerical, but categorical. Exaples of factors are sex (male or female), degree of burns (first, second or third), if you have been pregnant or not (yes/ no) etc. By making use of statistical modeling it is possible to establish a relationship between the survival time and the explanatory variables associated with a patient [1].
The main emphasis in this chapter will be to establish a statistical regression model for survival data, and this model will be based on the hazard function. The model that will be used is called *Cox regression model* (also known by "*proportional hazard model*"). It is worth mentioning that no particular probability distribution is assumed for the Cox regression model [1]. The main source of information will be chapter 3 of Collett [1].

3.1 Introducing the model

Now, consider two groups of patients. All the patients are diagnosed with the same disease, but each group receives different treatment: One of the groups receives a standard treatment and the other one a new treatment. Each of the groups have their associated hazard function at time t, denoted by $h_S(t)$ and $h_N(t)$ for the patients on the standard and the new treatment, respectively. The proportional hazard model is then given by [1, 56]:

$$h_N(t) = \psi h_S(t), \tag{3.1}$$

where ψ is a constant called the *relative hazard* or *hazard ratio*. The reason for adopting these two names to ψ can be understood by observing that

$$\psi = \frac{h_N(t)}{h_S(t)}. \tag{3.2}$$

Depending on whether $\psi < 1$ or $\psi > 1$, there is less or greater risk of death associated with the new treatment compared to the standard treatment, respectively.
It is possible to express (3.1) by

$$h_i(t) = e^{\beta x_i} h_0(t). \tag{3.3}$$

The terms in (3.3) are as follows:

- $h_i(t)$ is the hazard function associated with patient i, where i goes from 1 to n.
- x_i is either *zero* or 1, depending on whether the patient is on the standard or the new treatment, respectively.
- $\beta = ln(\psi)$.
- $h_0(t)$ is the hazard function for a patient on the standard treatment.

Now, assume you have p explanatory variables $X_1, X_2, X_3, \ldots, X_p$ who's values are $x_1, x_2, x_3, \ldots, x_p$. Assume for the time being that these variables are given specific values at the start of the study. Further, define

$$\boldsymbol{x} = [x_1, x_2, x_3, \ldots, x_p]^T = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_p \end{bmatrix} \tag{3.4}$$

as the vector containing the explanatory variable values, where T denotes the transpose of the row vector. In addition, we define $h_0(t)$ to be *the baseline hazard function*[1, p. 57], the hazard function corresponding to $\boldsymbol{x} = 0$ (all the explanatory variable values are zero). The hazard function associated with patient i is then given by [1, 58]:

$$h_i(t) = \psi(\boldsymbol{x_i}) h_0(t), \tag{3.5}$$

where $\psi(\boldsymbol{x_i})$ is the relative hazard function associated with the explanatory variable values of patient i. $\boldsymbol{x_i}$ is the column vector given by

$$\boldsymbol{x_i} = [x_{1i}, x_{2i}, x_{3i}, \dots, x_{pi}]^T. \tag{3.6}$$

The relation given by (3.5) is also known by the name *general proportional hazard model.* We then rewrite $\psi(\boldsymbol{x_i})$ in the following form:

$$\psi(\boldsymbol{x_i}) = e^{\boldsymbol{\beta}^T \boldsymbol{x_i}}, \tag{3.7}$$

where $\boldsymbol{\beta}^T \boldsymbol{x_i}$ is defined as

$$\boldsymbol{\beta}^T \boldsymbol{x_i} = \sum_{j=1}^{p} \beta_j x_{ji} = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_p x_{pi}. \tag{3.8}$$

The β_i's are "*the coefficients of the explanatory variables $x_1, x_2, x_3, \dots, x_p$ in the model*"[1, p. 58], and "*$\boldsymbol{\beta}^T \boldsymbol{x_i}$ is called the linear component of the model*"[1, 58].
$\boldsymbol{\beta}^T$ is the transposed of a column vector containing the coefficients of the p explanatory variables:

$$\boldsymbol{\beta}^T = [\beta_1, \beta_2, \beta_3, \dots, \beta_p]. \tag{3.9}$$

With the terminology defined above, the expression for the *Cox proportional hazard model* is given by [1, p. 63]:

$$h_i(t) = e^{\boldsymbol{\beta}^T \boldsymbol{x_i}} h_0(t) \tag{3.10}$$

Let's now look at the interpretation of hazard ratio by looking at a simple special case. Assume you have a proportional hazard model on the form given by (3.10), where you only have one continuous explanatory variable $\boldsymbol{X}$ [1]. Then, the proportional hazard model becomes

$$h_i(t) = e^{\beta x_i} h_0(t), \tag{3.11}$$

where β is the coefficient of the explanatory variable value x_i corresponding to patient i. Now, assume x_i takes the value x. This gives rise to the hazard function

$$h_1(t) = e^{\beta x} h_0(t). \tag{3.12}$$

Further, increase x_i by 1. This gives rise to

$$h_2(t) = e^{\beta(x+1)} h_0(t). \tag{3.13}$$

Then, dividing $h_2(t)$ by $h_1(t)$ gives:

$$\frac{h_2(t)}{h_1(t)} = \frac{e^{\beta(x+1)}h_0(t)}{e^{\beta x}h_0(t)} = \frac{e^{\beta(x+1)}}{e^{\beta x}} = e^{\beta(x-x+1)} = e^{\beta}. \tag{3.14}$$

Thus, e^{β} can be interpreted as the hazard ratio associated with a one unit increase in the explanatory variable.
Taking the natural logaritm of (3.14) gives an expression for β:

$$\beta = ln\left(\frac{h_2(t)}{h_1(t)}\right). \tag{3.15}$$

In other words, β can be interpreted as the natural logaritm of the relative change in the hazard function when the explanatory variable value "*is increased by one unit.*"[1, p. 90].
In the proportional hazard model it is assumed that the β's of the model (often called *regression coefficients*) doesn't vary with time. By applying numerical methods together with "*the method of maximum likelihood*"[1, p. 63], it is possible to estimate these coefficients. After estimates of these coefficients have been obtained, it is possible to estimate the baseline hazard function $h_0(t)$ [1]. In our case, a script runned by the computer software R will estimate the coefficients and the baseline hazard function. The likelihood function used to estimate the coefficients will now be stated and explained to some degree, but for further details, see Collett [1]. It is worth mentioning that, the likelihood function that we now present assumes no ties (no patients have the same survival time), but with some modification it can easily be adjusted to be valid for situations with ties. Ties can occur in a data set if for example the patient's survival time is rounded to the nearest whole integer. Say for example that one patient has survival time 14.4 and the other patient 14.3. Then both of them could be rounded to 14, which would result in a tie (even if they to start with didn't have the same survival time).
It is worth mentioning that the accuracy of the proportional hazard model can be further increased by allowing the explanatory variable values to depend on time [1].
Start with assuming you have n patients, among which r will result in death and $n-r$ will be right-censored. Assume further that, all the death times are different from one another. You will then get the following ascending order of death times: $t_1 < t_2 < t_3 < \cdots < t_r$
Further, let t$_j$ denote the jth death time. Then, define the *risk set* $R(t_j)$ to

be the group of patients "*who are alive and uncensored at a time just prior to t_j*"[1, p. 63]. Then, the likelihood function is given by [1, p. 63]:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{h_0(t_j)e^{\boldsymbol{\beta}^T\boldsymbol{x_j}}}{\sum_{l\in R(\mathrm{t}_j)} h_0(t_j)e^{\boldsymbol{\beta}^T\boldsymbol{x_l}}} = \prod_{j=1}^{r} \frac{e^{\boldsymbol{\beta}^T\boldsymbol{x_j}}}{\sum_{l\in R(\mathrm{t}_j)} e^{\boldsymbol{\beta}^T\boldsymbol{x_l}}}. \tag{3.16}$$

The product in (3.16) is an expression for the probability of observing that the patient who dies among the people in the risk set $R(t_j)$ is patient j. The sum $\sum_{l\in R(\mathrm{t}_j)} e^{\beta^T x_l}$ is the sum of the exponential terms corresponding to all the patients in the risk set $R(\mathrm{t}_j)$. The idea is to find those β-coefficients that will maximize the value of the likelihood function defined by (3.16). The likelihood procedure can also give rise to standard deviation, confidence intervals and p-values associated with the β-coefficients. Harrell gives a description of how this is done [3]. A summary of the results from Harrell will now be given.

Start with defining $\ell(\boldsymbol{\beta})$ as the natural logaritm of (3.16). This gives:

$$\ell(\boldsymbol{\beta}) = ln(L(\boldsymbol{\beta})) = \sum_{j=1}^{r}\left[\boldsymbol{\beta}^T\boldsymbol{x_j} - ln\left(\sum_{l\in R(\mathrm{t}_j)} e^{\boldsymbol{\beta}^T\boldsymbol{x_l}}\right)\right]. \tag{3.17}$$

Then, take the first order derivative of (3.17) and set the result equal to zero:

$$\ell'(\boldsymbol{\beta})) = 0 \tag{3.18}$$

The equation in (3.18) gives rise to a set of equations that must be solved numerically.

After some calculations, (3.18) will give rise to the column vector $\hat{\boldsymbol{\beta}}$ containing the different regression coefficient estimates. Then, define

$$I(\boldsymbol{\beta}) = -\ell''(\boldsymbol{\beta})) \tag{3.19}$$

to be the negative of the matrix of second order derivative of (3.17), and define

$$J(\boldsymbol{\beta}) = I^{-1}(\boldsymbol{\beta}) \tag{3.20}$$

to be the inverse of (3.19). It can then be shown that, the regression coefficient estimates $\hat{\boldsymbol{\beta}}$ are approximately normally distributed with expectation $\boldsymbol{\beta}$ and covariance matrix $J(\hat{\boldsymbol{\beta}})$:

$$\hat{\boldsymbol{\beta}} \approx N(\boldsymbol{\beta}, J(\hat{\boldsymbol{\beta}})) \tag{3.21}$$

In particular, this gives that the regression coefficient estimate corresponding to individual i is also approximately normally distributed:

$$\hat{\beta_i} \approx N(\beta_i, J_{ii}), \tag{3.22}$$

where J_{ii} is element ii in $J(\boldsymbol{\beta})$ and is the estimated variance of $\hat{\beta_i}$. Now, define the test statistic

$$\text{Ƶ} = \frac{\hat{\beta_i} - \beta_i}{\sqrt{J_{ii}}} \approx N(0,1), \tag{3.23}$$

which is approximately standard normally distributed (expectation is zero, variance is 1). Then, perform a two-sided hypothesis test called the *Wald test* to check the null hypothesis

$$H_0 : \beta_i = 0$$

against the hypothesis

$$H_1 : \beta_i \neq 0.$$

Depending on which p-values the Wald test gives you, either reject or accept the null hypothesis. Rejecting the null hypothesis will in this situation mean that we conclude that $\beta_i \neq 0$ and thus that the survival times depends on the explanatory variable β_i. The p-value is calculated by

$$p - value = 2 \cdot P(\text{Ƶ} \geq |z|) = 2(1 - P(\text{Ƶ} \leq |z|)) = 2(1 - G(|z|)), \tag{3.24}$$

where z is the observed value of Z in 3.23, and $G(z)$ is the cumulative distribution function of the standard normal distribution given by [2, p. 178])

$$G(z) = P(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{\frac{-t^2}{2}} dt. \tag{3.25}$$

The value of $G(z)$ is calculated by numerical integration executed by a computer software. The associated confidence interval becomes:

$$\left[\hat{\beta_i} - \text{ƶ}_{\alpha/2} \cdot \sqrt{J_{ii}}, \hat{\beta_i} + \text{ƶ}_{\alpha/2} \cdot \sqrt{J_{ii}}\right], \tag{3.26}$$

where $\text{ƶ}_{\alpha/2}$ is the $\alpha/2$-quantile of Z defined by

$$P(-\text{ƶ}_{\alpha/2} < Z < \text{ƶ}_{\alpha/2}) = 1 - \alpha. \tag{3.27}$$

The associated confidence interval for the hazard ratio becomes:

$$\left[e^{\hat{\beta}_i - \text{ƶ}_{\alpha/2}\sqrt{J_{ii}}}, e^{\hat{\beta}_i + \text{ƶ}_{\alpha/2}\sqrt{J_{ii}}} \right]. \tag{3.28}$$

3.2 Interactions

The idea of interactions will now be illustrated by the help of an example.

Given a patient group with two associated explanatory variables X_1 and X_2, representing for instance age and size of tumor, respectively. Their hazard function becomes

$$h_i(t) = e^{\beta_1 x_1 + \beta_2 x_2} h_0(t) \tag{3.29}$$

What now, if the effect of the size of the tumor depended on the age of the patient? One way to model this would be to include a so-called interaction term, being the product of the values of the two variables:

$$h_i(t) = e^{\beta_1 x_1 + \beta_2 x_2 + \beta_{12} \cdot x_1 \cdot x_2} h_0(t), \tag{3.30}$$

where $\boldsymbol{\beta}_{12}$ is the regression coefficient corresponding to the interaction between X_1 and X_2. When doing Cox regression, it is of interest to check whether there exists any interaction between some of the explanatory variables involved in the data set.

3.3 Including categorical variables in the model

In his book, Collett says that the hazard function can depend on two types of variables: *variates* and *factors*, where variates are continuous (for example age, height, blood pressure) and factors are catergoric (for example gender)[1]. Including a variate in the hazard function is straight forward: You include it in the linear component of the model by multiplying the regression coefficient with its associated variate value, as shown in (3.8). When the variable is categorical, the story is a little bit different. To illustrate how it is done, start imagining you have a group of patients with a specific disease. This disease have three stages; Stage 1, 2 and 3. Each stage has its own regression coefficient, but they are all part of the same explanatory variable, which we here denote X_{stage}. Then, we define stage 1 to be the reference state, in which the regression coefficient value is assigned the value zero. The regression coefficients of stage 2 and 3 will take values with respect to the reference state stage 1. The three possible values of X_{stage} is summarized in equation 3.31 below.

$$X_{stage} = \begin{cases} 1 \\ 2 \\ 3 \end{cases} \tag{3.31}$$

For the patient group, the hazard function takes the following form:

$$h_i(t) = e^{\beta_1 \cdot I_2 + \beta_2 \cdot I_3} h_0(t), \tag{3.32}$$

where β_1 and β_2 are the regression coefficients of stage 2 and 3, respectively, and I_2 and I_3 are defined by

$$I_2 = \begin{cases} 1, if\ X_{stage} = 2 \\ 0, else \end{cases} \tag{3.33}$$

$$I_3 = \begin{cases} 1, if\ X_{stage} = 3 \\ 0, else \end{cases} \tag{3.34}$$

Depending on the stage of the particular patient, the hazard function becomes:

$$X_{stage} = 1 \Rightarrow h(t) = h_0(t)$$
$$X_{stage} = 2 \Rightarrow h(t) = h_0(t) \cdot e^{\beta_1}$$
$$X_{stage} = 3 \Rightarrow h(t) = h_0(t) \cdot e^{\beta_2}$$

3.4 Residuals

How can we know if the proposed proportional hazard model is a good model for our target patient group? One way to explore this is by looking at so-called *residuals*. Residuals are values one calculates in such a way that they may reveal whether the model assumptions are fulfilled. The estimated proportional hazard model for the ith patient is given by

$$\hat{h}_i(t) = e^{\hat{\beta}^T \mathbf{x}_i} \hat{h}_0(t), \tag{3.35}$$

where $\hat{h}_i(t)$ is the estimated hazard function for patient i and $\hat{h}_0(t)$ is the estimated baseline hazard function. $\hat{\beta}^T \mathbf{x}_i$ is given by

$$\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \cdots + \hat{\beta}_p x_{pi}, \tag{3.36}$$

which corresponds to (3.8) defined earlier. $\hat{\boldsymbol{\beta}}^T$ is the estimate of the regression coefficient vector given by (3.9).

3.4.1 Schoenfeld residuals

Recall that, for the Cox regression (proportional hazard model) to be valid, the regression coefficients must be assumed to be independent of time. This assumption is called the *proportional hazards assumption*. If the proportional hazards assumption is not satisfied, Cox regression in its basic form cannot be used. One type of residuals called the *Schoenfeld residuals* can be used to check whether the assumption of time-independent regression coefficients is reasonable. Schoenfeld residuals are about assigning a residual for each explanatory variable. Begin with defining "*an event indicator*"[1, p. 114] δ_i, who's value is zero or unity depending on whether the patient's survival time is censored or not, respectively. The Schoenfeld residual for the ith patient associated with the jth explanatory variable X_j is given by [1, p. 117]

$$r_{P_{ji}} = \delta_i (x_{ji} - \hat{a}_{ji}), \tag{3.37}$$

where x_{ji} is the value of the jth explanatory variable and a_{ji} is given by [1, p. 117]

$$a_{ji} = \frac{\sum_{l \in R(\mathrm{t}_j)} x_{jl} e^{\hat{\boldsymbol{\beta}}^T x_l}}{\sum_{l \in R(\mathrm{t}_j)} e^{\hat{\boldsymbol{\beta}}^T x_l}}. \tag{3.38}$$

$R(t_j)$ is the risk set defined in the previous section.

Now, create a row vector consisting of the Schoenfeld residuals for the ith patient:

$$\mathbf{r}^T{}_{Pi} = [r_{P1i}, r_{P2i}, r_{P3i}, \ldots, r_{Ppi}]. \tag{3.39}$$

Further, denote the *scaled* Schoenfeld residual associated with the jth explanatory variable X_j for the ith patient by $r^*{}_{P_{ji}}$. The $r^*{}_{P_{ji}}$'s are the components of the row vector defined by [1, p. 118]

$$\mathbf{r}^T{}_{Pi}{}^* = f \cdot var(\hat{\beta}) \cdot \mathbf{r}^T{}_{Pi}, \tag{3.40}$$

where f is number of deaths and $var(\hat{\beta})$ is the variance-covariance matrix associated with the β-coefficients estimates.

It can be shown that [1, p. 144], the expected value of the $r^*{}_{P_{ji}}$'s is approximately given by

$$E(r^*{}_{P_{ji}}) \approx \beta_j(t_i) - \hat{\beta}_j, \tag{3.41}$$

where $\beta_j(t_i)$ is set to be the β-coefficient associated with explanatory variable X_j at death time t_i. If the calculated value of $r^*{}_{P_{ji}}$ is close to its expected value, we would get that

$$r^*{}_{P_{ji}} \approx \beta_j(t_i) - \hat{\beta}_j. \tag{3.42}$$

Adding $\hat{\beta}_j$ to both sides of (3.42) gives

$$r^*{}_{P_{ji}} + \hat{\beta}_j \approx \beta_j(t_i). \tag{3.43}$$

If you plot $r^*{}_{P_{ji}} + \hat{\beta}_j$ as a function of survival time, the appearance of the points in the plot would resemble the functional form of the regression coefficient as a function of survival time. Thus, if the points fits a horizontal straight line, it would mean that $\beta_j(t_i)$ is constant. Thus, a horizontal straight line would indicate no change in the β-coefficients with time, thus indicating the validity of using the proportional hazard model. Anything other

than a straight horizontal line would indicate that the regression coefficients are time-dependent. To handle time-dependent regression coefficients, an extension must be done to the Cox model.
In addition to making a plot of (3.43) as a function of survival time, a hypothesis test with the following test statistic can be performed:

$$\beta_j(t) = \beta j + \theta_j(g_j(t) - \bar{g}_j), j = 1, \ldots, p, \tag{3.44}$$

where p is the number of explanatory variables. It can be shown that, if the null hypothesis of $\theta = 0$ is not rejected, there is evidence of that the assumption of proportional hazards holds. Thus, the lower the p-value, the less evidence there is for proportional hazards. For further details about this hypothesis test, see [4].

3.4.2 Martingale residuals

The Martingale residual for the i*th* individual is defined as [1, p. 115]

$$r_{Mi} = \delta_i - r_{Ci}, \tag{3.45}$$

where δ_i is defined as in section 3.4.1, and r_{Ci} is the Cox-Snell residual for the i*th* individual given by [1, p. 112]

$$r_{Ci} = e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i} \hat{H}_0(t_i), \tag{3.46}$$

where $\hat{H}_0(t_i)$ is the Nelson-Aalsen/ Breslow estimate of the baseline cumulative hazard function given by [1, p. 101]

$$\widetilde{H}_0(t) = \sum_{j=1}^{k} \frac{d_j}{\sum_{l \in R(\mathrm{t}_j)} e^{\hat{\boldsymbol{\beta}}^T x_l}} \tag{3.47}$$

If you are unsure about the functional form of the explanatory variable that should be used, you can start with creating a scatter plot by plotting the martingale residuals for the null model against the value of the explanatory variable. The null model is the model without explanatory variables. It is possible to show that, this scatter plot reveales the functional form of the explanatory variable [4]. But there is no guarantee that it will be crystal clear from this scatter plot what the functional form should be.
It is also possible to suggest a functional form that should be used for the explanatory variable in the model. In section 4.2.3 Collett presents how this

is done [1]. What is done is to create a scatter plot by plotting the martingale residuals for the null model against the suggested functional form of the explanatory variable. Because such scatter plots can become quite noisy and therefore difficult to interpret, Collett suggests using the "*LOESS smoother*" [1, p. 127], an algoritm used to create a smooth curve to the data points in the plot. If the LOESS smoother resembles a straight line, this would indicate that he correct functional form is used. If not, it is suggested to try another functional form to see if this would make the LOESS smoother more straight.
It is also possible to create the martingale scatter plot described above by using a non-empty model (that is, the explanatory variables are included). Then, in addition to being straight, the LOESS smoother should be horizontal if the correct functional form of the explanatory variable has been used.

Note that, making plots of martingale residuals gives meaning for continuous variables but not for categorical. It makes sense to talk about the functional form of a continuous variable, but not of a categorical. For this reason, plots of martingale residuals will not be created for explanatory variables that are categorical.

3.5 Time-dependent explanatory variables

Recall that, the assumtion of proportional hazards is that the regression coefficients are independent of time. It is possible to let the values of the explanatory variables to be time-dependent, at the same time assuming that the regression coefficients stay constant. Assume now that the hazard function only contains one explanatory variable X_1 with regression coefficient β_1 and explanatory variable value x_1. If the explanatory variable value is time dependent, the hazard function would take the form of [5, p. 15]

$$h(t) = h_0(t)e^{\beta_1 x_1(t)} \tag{3.48}$$

where $x_1(t)$ is the time-dependent explanatory variable value. On the other hand: If the proportional hazard assumtion is not satisfied, the regression coefficients are time-dependent, and the hazard function will then take the form

$$h(t) = h_0(t)e^{\beta_1(t)x_1}, \tag{3.49}$$

where $\beta_1(t)$ is the time-dependent regression coefficient. If it was the case that both the explanatory variable value and the regression coefficients were time-dependent, the hazard function would become

$$h(t) = h_0(t)e^{\beta_1(t)x_1(t)}, \tag{3.50}$$

where both $\beta_1(t)$ and $x_1(t)$ are functions depending on time.

Chapter 4

Application

In this chapter, the theory presented in the two previous chapters will be applied on a survival data set available in the programming software R. Before showing applications, the data set will be presented in detail.

4.1 Introducing the data set

The name of the data set is "German Breast Cancer Study Data", and contains survival data from 686 female patients, all diagnosed with breast cancer. The data set is accessed from the statistical software R by executing the command "*data("gbcsCS")*" after installing and loading the *condSURV*-library. If the data is inspected in R, 16 columns will appear. In the following, each of these columns will be explained. The source of information is a book [6] and a research paper [7].
Column 1 is "*id*" and is simply an integer number from 1 to 686 to make it possible to distinguish between the different patients.
In column 2, 3 and 4 the three dates "*diagdateb*", "*recdate*" and "*deathdate*" are registered. These are the date of diagnosis, recurrence and death, respectively. Diagnosis date is the date the patient is diagnosed with breast cancer. Soon after this date the breast cancer cells are removed by surgery. It is of interest to find out how long it goes before the breast cancer cells recurs. When the breast cancer cells recurs, the recurrence date is registered. Column 13 is called "*rectime*" and is the number of days from diagnosis to recurrence, also known as the *recurrence time*. In column 14 you have "*censrec*", which is a variable that takes the value zero if censoring occurs and the value "1" if a true recurrence date is registered. Note that, if "*censrec*" is zero, censoring has occured, and the recurrence time is equal to the patient's survival time "*survtime*" found in column 15. The survival times is in days. Also note that, if "*censrec*" is zero, so is "*censdead*" in colum 16, which is a variable that takes the value zero if censoring occurs and a value "1" if an actual death is observed. See table 4.1 below for an interpretation of the different combinations of the censoring variables "*censrec*" and "*censdead*" that are possible.

censrec	censdead	Interpretation
0	0	Censoring occurs before any recurrence is registered. Thus, no death.
1	0	Censoring occurs after the recurrence date is registered. No death
1	1	No censoring occurs. Both actual recurrence and death are registered.

Table 4.1: Censoring in German Breast Cancer Study Data

In addition to recurrence and survival time, each patient has different explanatory variables associated with it. Let's inspect these explanatory variables one by one.

First, we have a look at the *age* of the patients, which is found in column 5. In the data set, the age of a patient is given as an integer number. The youngest of the patients was 21 years old and the oldest 80. The average age was 53. By using R a histogram is drawn to visualize the age distribution of the patients (see figure 4.1 below). Along the horisontal axis you have age, which is split into intervals of five years. Along the vertical axis you have the number of patients in each age group. The histogram tells us that there are most patients in the age group from 45 to 50. The second most crowded patient group is from 60 to 65.

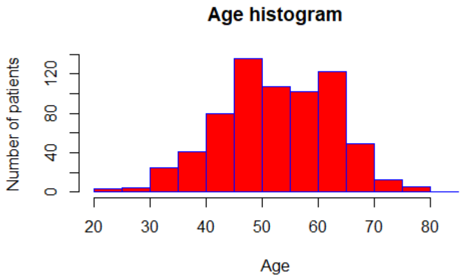


Figure 4.1: Histogram showing the age distribution of the patients.

The next variable of interest is *menopause*. A woman who has reached menopause doesn't have any menstrual bleeding anymore. For patients who has reached menopause the menopause variables is given the value "2", and for patients who hasn't reached menopause the variable is given the value "1". 396 of the women (58%) had reached menopause while 290 (42%) had not.

Another variable of interest is *hormone*. This variable has either the value "1" or "2", depending on whether the patient is receiving a daily dose of 30 mg tamoxifen or not, respectively. The growth of breast cancer cells are stimulated if the female sex hormone estrogen binds to the estrogen receptors in breast cancer cells [8]. Tamoxifen is a drug that binds to the estrogen receptors in breast cancer cells and thus prevents estrogen from binding to these receptors and thus inhibits stimulation of breast cancer cell growth [8]. 64 % were receiving tamoxifen and 36 % were not.

The next variable is the size of the primary tumor in millimeter, denoted as "*size*".

Tumor grade is the next variable, with three possible values: "1", "2" or "3". These values are scores *"that tells you how different the cancer cells' appearance and growth patterns are from those of normal, healthy breast cells."* [9]. The variable is denoted as "*grade*".

Next on the list is number of *positive lymph nodes*, denotes as "*nodes*". A lymph node is *positive* if it contains cancer cells [10].

The last two explanatory variables are the amount of *progesterone* and *estrogen* bound to proteins in the cytosol of the primary tumor (measured in 10^{-15} moles per milligram cytosol protein). These variables are denoted as "*prog_recp*" and "*estrg_recp*", respectively.

Here follows a figure showing histograms and pie charts of the explanatory variables.

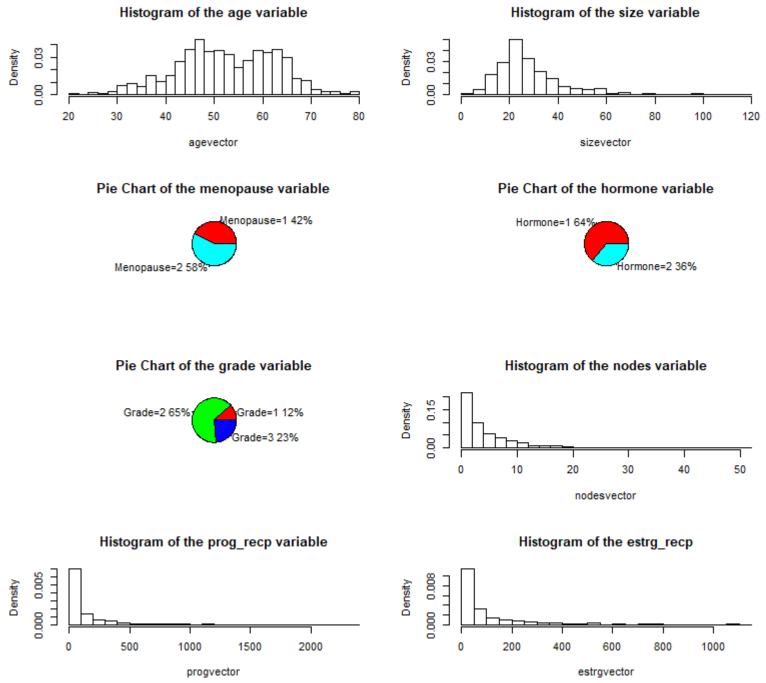


Figure 4.2: Figure showing histograms and pie charts illustrating the distribution of the explanatory variables

From figure 4.2 it seems like the age and size variable have a fairly symmetric distribution. The menopause, hormone and grade variable are all categorical. The nodes, *estrg_recp* and *prog_recp* seem to have a right skewed distribution. Knowledge about how the explanatory variables are distributed can become useful in the chapter about simulation, where the values of the explanatory variables will be simulated. Also, have in mind that, there is a correlation between the age and menopause variable. The reason for this is that women reach menopause in a specific period of their life, usually around their fifthies. It was also checked for correlation between the other variables (by plotting their scatter plots in the same figure and looking for patterns), but none were found.

Let's now present a table that summarizes the descriptive statistics (mean, standard deviation, percentage) of the explanatory variables.

Explanatory variable	**Descriptive statistics**
Age	$\bar{x}$=53.1, $\sigma_{\bar{x}}$=0.4

Menopause	Yes:(42%)
	No:(58%)

Hormone	Yes:(64%)
	No:(36%)

Size	$\bar{x}$=29.3, $\sigma_{\bar{x}}$=0.5

Grade	1:(12%) 2:(65%)
	3:(23%)

Nodes	$\bar{x}$=5.0, $\sigma_{\bar{x}}$=0.2
Progesteron	$\bar{x}$=110, $\sigma_{\bar{x}}$=8
Estrogen	$\bar{x}$=96.3, $\sigma_{\bar{x}}$=6.0

Table 4.2: Descriptive statistics of explanatory variables associated with German Breast Cancer Study Data. $\bar{x}$ denotes the mean, $\sigma_{\bar{x}}$ denotes the standard deviation of the mean and % denotes percentage.

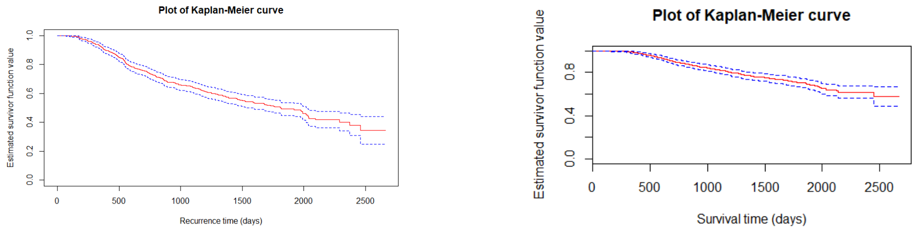
In table 4.2, the standard deviation of the mean is given by

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}}, \tag{4.1}$$

where σ_x is the sample standard deviation.

4.2 Kaplan-Meier curves

It is possible to use the statistical software R to create Kaplan-Meier curves of the time to recurrence and the time to death. See the appendix for the R-code used in this report.



(a) Kaplan-Meier curve of the time to recurrence

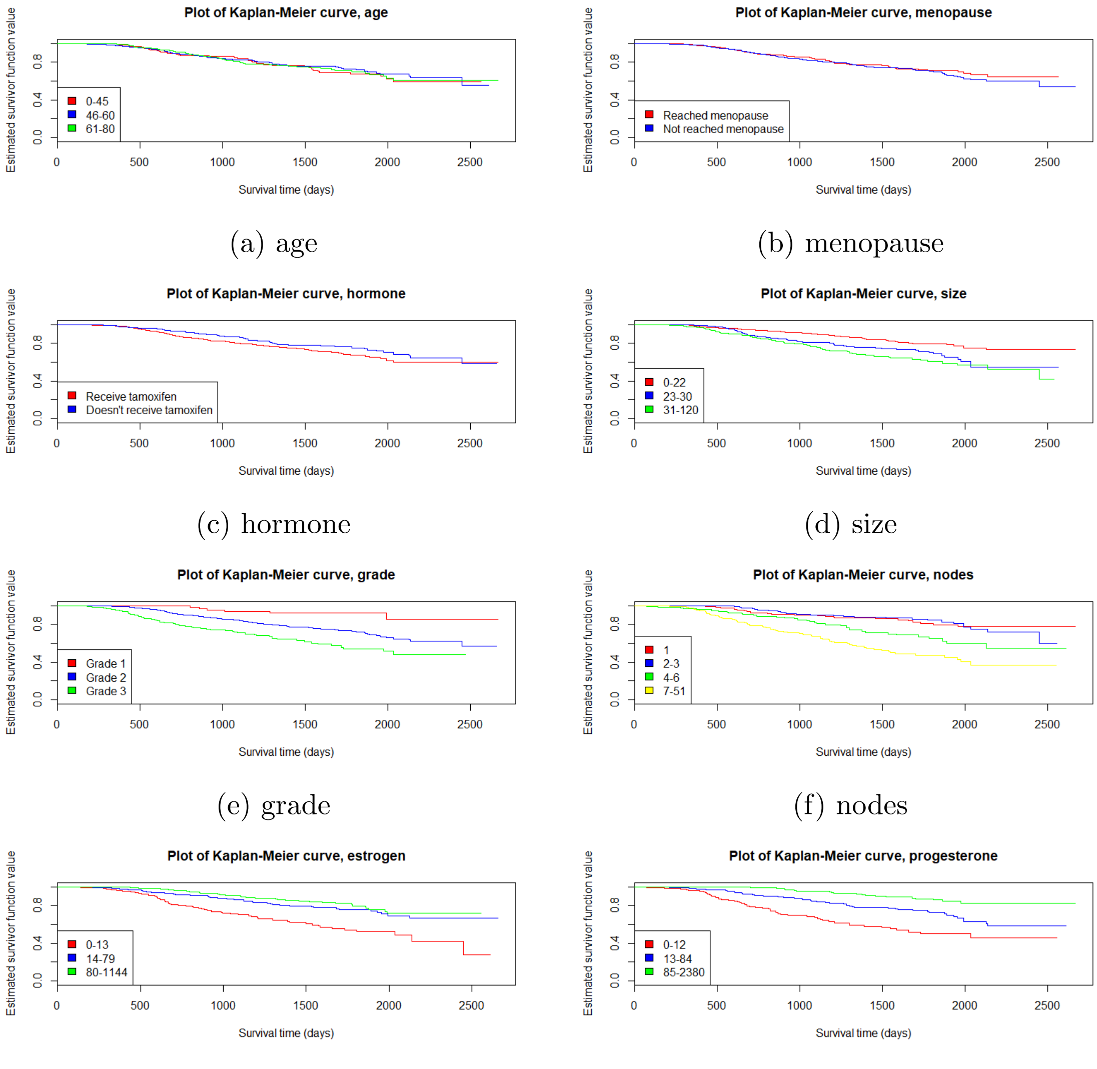
(b) Kaplan-Meier curve of the time to death

Figure 4.3: Kaplan-Meier curves for time to death and time to recurrence.

The red solid curve in figure 4.3a above is the Kaplan-Meier curve for the time to recurrence. The dashed blue curves above and below the red solid curve are the upper and lower confidence interval curves, respectively. Equivalently, figure 4.3b shows the Kaplan-Meier curve (red) and the associated confidence interval curves (blue) for time to death.

The Kaplan-Meier curves in figure 4.3a and 4.3b includes all the patients in the data set, regardless of the values of their explanatory variables. It is also possible to create Kaplan-Meier curves for specific categories of patients. One way to do this is to group together those patients who have the same or similar values of a specific explanatory variable. For example, you could make two Kaplan-Meier curves in the same plot, where one of the curves is of those who has reached menopause and the other of those who hasn't. In the following, such categoric Kaplan-Meier curves for time to death will be presented. Based on the appearance of such plots it is possible to get an understanding of how the values of the explanatory variables affect the survival time of the patients. Each of the Kaplan-Meier curve plots below have a legend in the bottom left corner which describes which curve is associated with the given value of the explanatory variable. Menopause, hormone and grade were already categorized from the start of, but each of the explanatory variables age, size, nodes, progesterone and estrogen I had to categorize myself. These variables were grouped into intervals, each interval containing approximately the same amount of patients. This resulted in several Kaplan-Meier curves for a given explanatory variable, one curve for each possible categorical value. First, let's do this for time to death, then for time to recurrence.

4.2.1 Time to death



(a) age

(b) menopause

(c) hormone

(d) size

(e) grade

(f) nodes

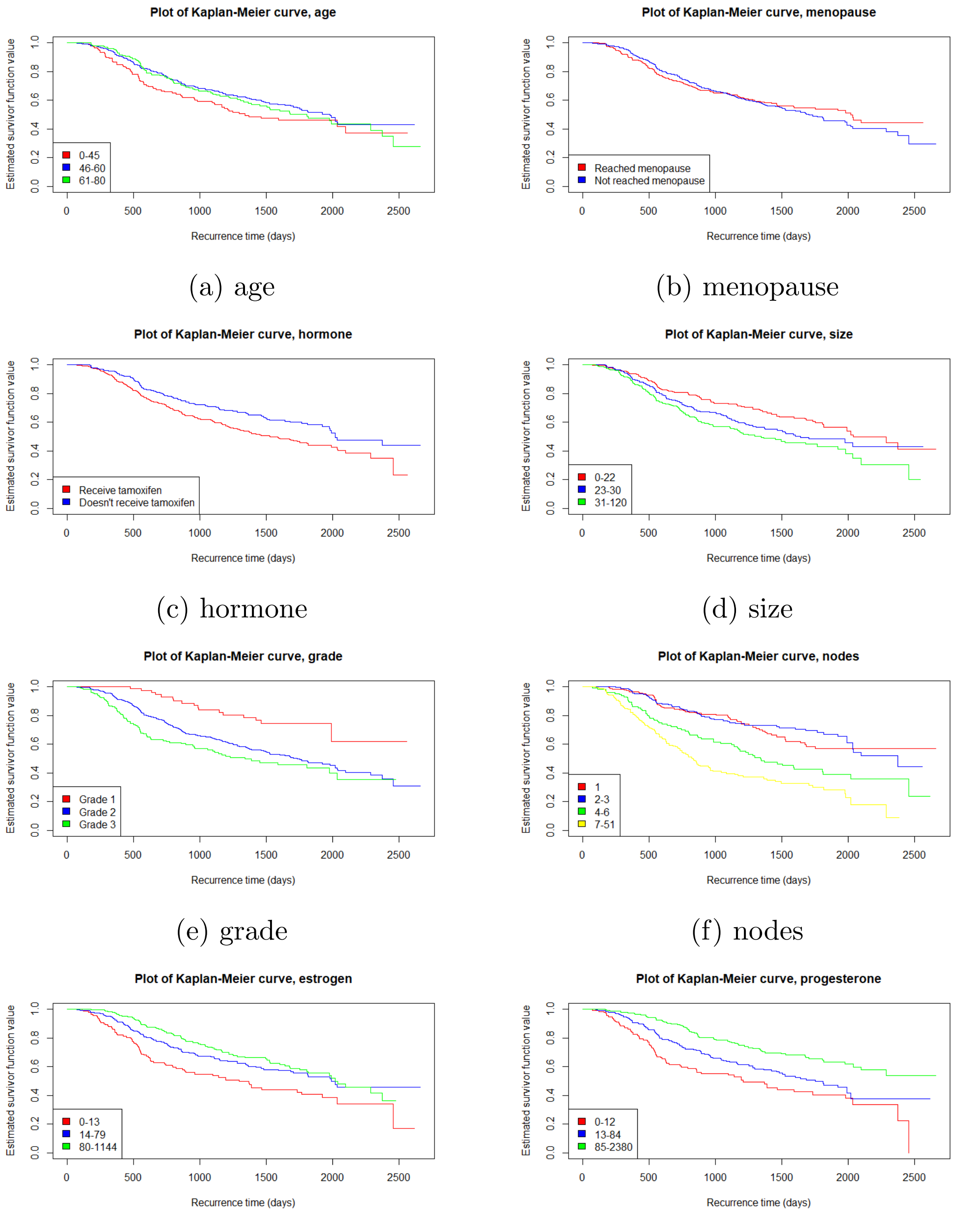
(g) estrogen

(h) progesterone

Figure 4.4: Kaplan-Meier curves for time to death for the different explanatory variables.

From figure 4.4a it doesn't seem that age plays a difference in the survival experience of the patients. The Kaplan-Meier curves lie approximately upon each other. As seen from figure 4.4b, the Kaplan-Meier curve doesn't change much for patient who has and hasn't reached menopause. This indicates that menopause seem to not play a crusial role in the survival experience of the patients. In figure 4.4c the red curve (receiving tamoxifen) is for the most under the blue curve (not receiving tamoxifen). This indicates that those patients receiving tamoxifen seem to have a higher mortality than those not receiving tamoxifen. In figure 4.4d you can se that, the Kaplan-Meier curves of those patients with smaller tumors is above those with larger tumors. This indicates those with smaller tumors seem to have better odds surviving compared to those with larger tumors. In figure 4.4e it is observed that, the Kaplan-Meier curve of grade 1 is above grade 2, and the Kaplan-Meier curve of grade 2 is above grade 3. This indicates that the lower the grade, the better odds there are for survival. From figure 4.4f the trend is that, the patients with few positive nodes have Kaplan-Meier curves above those with many positive nodes. This suggest that people with few positive nodes have better odds of survival. The appearance of the curves in plot 4.4g suggests that, the larger amount of estrogen that is bound to proteins in the cytosol of the primary tumor, the better survival experience. Thus, individuals with large values of the explanatory variable "*estrg_recp*" seem to have better odds of surviving compared with individual with low values of this variable. Similar to plot 4.4g, the appearance of the curves in plot 4.4h suggests that, the larger amount of progesterone that is bound to proteins in the cytosol of the primary tumor, the better the survival experience of the individual. Thus, individuals with large values of the explanatory variable "*prog_recp*" seem to have better odds of surviving compared with individual with low values of this variable.

4.2.2 Time to recurrence



(a) age

(b) menopause

(c) hormone

(d) size

(e) grade

(f) nodes

(g) estrogen

(h) progesterone

Figure 4.5: Kaplan-Meier curves for time to recurrence for the different explanatory variables.

Figure 4.5a have a small indication of that, the younger the patients, the higher the risk of recurrence. From 4.5b it doesn't seem that menopause status plays a role when it comes to time to recurrence. Figure 4.5c indicates that, those patients receiving tamoxifen have a higher chance of recurrence. Confirming our intuition, figure 4.5d indicates that, the bigger the larger the primary tumor, the higher risk of recurrence. Figure 4.5e indicate that, the chance of recurrence increase with increasing grade. Figure 4.5f indicate that, the more positive nodes, the larger the chance for recurrence. Figure 4.5g indicate that the odds of recurrence increase with decreasing value of the *estrg_recp* variable. From figure 4.5h it is indicated that the chance of recurrence is increasing with decreasing value of the *prog_recp* variable.

When comparing the results from the Kaplan-Meier plots in section 4.2.1 and 4.2.2, one see that, for most of the time, the same variables important for time to recurrence is also important for time to death. Later on (in section 4.4.3), recurrence will be included in the model as a time-dependent variable for time to death. Then recurrence's importance when it comes to time to death will become more clear.

4.3 Log-rank test

By just looking at the Kaplan-Meier curves in section 4.2, we could get an idea of how the survival experience of the patients depended on the different explanatory variables. Now, the Log-rank test presented in section 2.3 will be applied to compare the different groups of patients presented in the Kaplan-Meier curve plots found in section 4.2. Each Kaplan-Meier curve in a given plot corresponds to a group, and the log-rank test was executed to compare all the groups associated with a given plot. If the p-value from the log-rank test was less than 0.05, it was concluded that there were evidence for rejecting the null hypothesis "*No difference in survival experience between the patient groups*". The results from the log-rank test was found by applying the function "survdiff" in R to the groups of interest. The p-values associated with the different explanatory variables for time to death and time to recurrence are listed in table 4.3 and 4.4 below, respectively.

4.3.1 Time to death

Explanatory variable	**p-value**
Age	0.9
Menopause	0.5
Hormone	0.1
Grade	$3 \cdot 10^{-7}$
Size	$1 \cdot 10^{-4}$
Nodes	$9 \cdot 10^{-15}$
Progesterone	$6 \cdot 10^{-13}$
Estrogen	$2 \cdot 10^{-7}$

Table 4.3: p-values from log-rank test. Time to death

The p-values in table 4.3 confirms the observations made by inspecting the Kaplan-Meier curves in section 4.2.1. Both the age and hormone variables have p-values indicating that they do not play an important role when it comes to the survival of the patient. The hormone variable has a p-value of 0.1, which is close to 0.05. For this reason, combined with the observation done from the Kaplan-Meier curve in figure 4.4c, there is a slight indication that the value of the hormone variable might have an influence on the survival experience of the patients. The remaining variables' p-value are less than 0.05, thus they indicate significant difference in survival experience between the patient groups.

4.3.2 Time to recurrence

Explanatory variable	**p-value**
Age	0.1
Menopause	0.6
Hormone	0.003
Grade	$3 \cdot 10^{-5}$
Size	0.001
Nodes	$2 \cdot 10^{-16}$
Progesterone	$1 \cdot 10^{-7}$
Estrogen	$4 \cdot 10^{-4}$

Table 4.4: p-values from log-rank test. Time to recurrence

The p-values in table 4.4 confirms observations done in section 4.2.2. The p-value of the age variable is 0.1, not far from 0.05. This, combined with the appearance of figure 4.5a, there is indication of that the age variable plays a role when it comes to recurrence, with indications of younger patients being more exposed to recurrence. As observed from figure 4.5a, the p-value of 0.6 of the menopause variable gives strong indications of that the menopause variable doesn't play a role when it comes to recurrence. The remaining variables have p-values suggesting that they play a significant role when it comes to recurrence.

4.4 Determining the regression coefficients

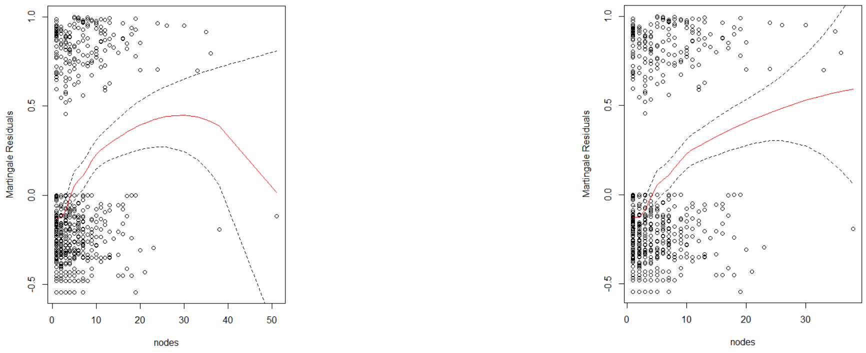
It is of interest to find out at which extent the survival times of a patient group depends on the different explanatory variables. The different explanatory variables contribute differently and thus will have different regression coefficient values in the hazard function. Based on knowledge about contribution you can choose to include in the model those explanatory variables that would make the model more close to reality. In chapter 3.6 Collett emphasizes that there could be several equally good models. For this reason, he suggest to not restrict your focus on finding one superior model, but rather allowing the possibility to find several models that individually gives a good representation of the dependence between survival time and explanatory variables [1]. There exist several methods for selecting which explanatory variables should be included in your survival analysis model. One such method is *purposeful selection*, and Hosmer, Lemeshow and May gives a step-by-step explanation of this method [6]. In the following these steps will be applied with the aim of finding a suitable combination of explanatory variables to be included in our model.

Depending on whether you consider the explanatory variables together (multivariable analysis) or seperately (univariable analysis) you will obtain different values for the regression coefficients for each explanatory variable. The choice of looking at uni - or multivariable model is made by specifying in the programming script how many variables your model should include and thus for how many variables regression coefficients should be estimated for. For more information on how this is done, look up the documentation for the function "Coxph" in R. Also, see appendix.

Let's start with univariable analysis, that is, considering each explanatory variable separately, one at a time, and investigate their survival time contribution (in practice, this means letting the hazard function only depend on one variable and thus omitting the existance of the other explanatory variables). For each of them, we use R to calculate the regression coefficient, hazard ratio confidence interval, and p-value from Wald test. Then, we present the results in a table. We start with time to death. Then, we repeat the same procedure for time to reccurence. After that is done, we repeat the procedure once again for time to death, but by treating the recurrence time as a time-dependent explanatory variable.

Before presenting the results from the analysis part, two remarks will be given.

The nodes variable of patient with ID variable value 684 had 51 positive nodes, a much larger value compared with the rest of the patient group. After running the analysis it was found that the large nodes value dominated the result compared with the more common smaller observations, in particular for the univariable martingale plot for time to death . For this reason, it was decided to remove this patient from the data set and then run the analysis. This turned out to be successful. Figure 4.6 below shows the martingale residuals plots before and after removal of ID 684, respectively.



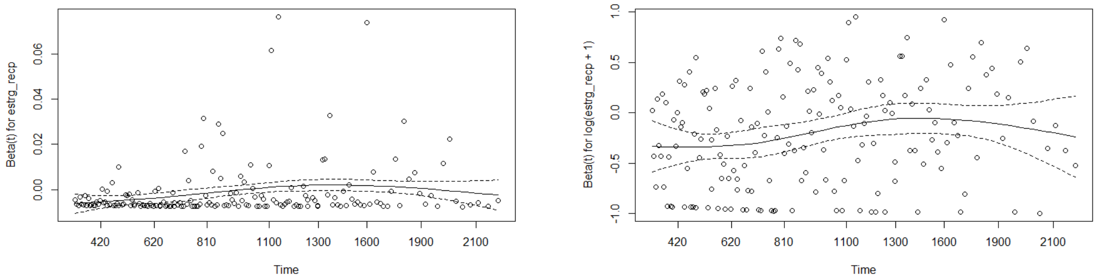
(a) Before removal of patient with ID 684

(b) After removal of patient with ID 684

Figure 4.6: Univariable martingale residuals plots for time to death associated with the nodes variable. Figure 4.6a is before and figure 4.6b is after removal of the patient from the data se, respectively.

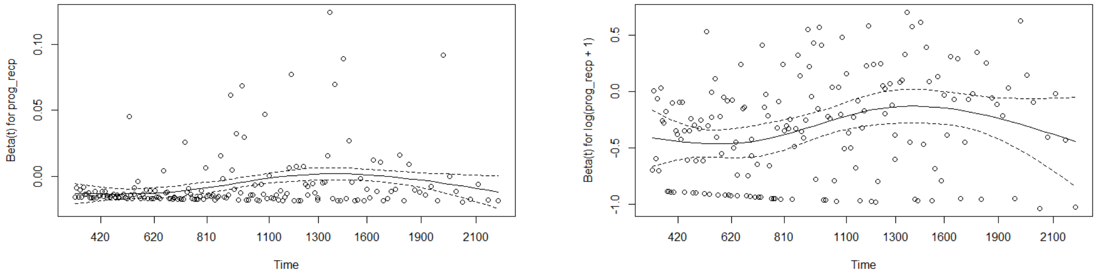
In the Martingale residuals plots in figure 4.6 above, the red line is the LOESS smoother and the dashed lines are the accompanied confidence interval boundaries (95% confidence interval). See appendix for details on how the Martingale residuals plots are created.
It was also found to be an issue that, large values of the progesterone and estrogen receptor variables were suppressing their associated smaller values. The largest values of the variables *prog_recp* and *estrg_recp* were 2380 and

1144, respectively. These values are a magnitude larger than their mean values of 110 and 96.3, respectively. This big difference resulted in suppression of the smaller values. The skewed distribution of these two variables can be seen from their histograms in figure 4.2. Both the Schoenfeld and Martingale residuals plots were affected by the suppression. To reduce the effect of suppression, a *log transformation* [11] was performed to these two variables. The Schoenfeld residuals plot and its associated p-value were compared before and after log transformation, and it was found that the requirements were better met for the log transformation case (the p-values from the hypothesis test became larger and it was easier to fit a straight horizontal line in the plots). Also, the Martingale residuals plots were compared before and after log transformation, and the plot requirements were better met after log transformation. The comparison of the situation before and after log transformation is shown in figure 4.7 for the Schoenfeld residuals and in figure 4.8 for the Martingale residuals. Since the values of both the *prog_recp* and *estrg_recp* variables could take the value zero, the number "1" were added to all values corresponding to these variables before performing the log transformation. If not else stated, log transformation will be performed to these two variables during Cox regression calculations.



(a) Estrogen before log transformation. Associated p-value from hypothesis test: 0.00006.

(b) Estrogen after log transformation. Associated p-value from hypothesis test: 0.021.

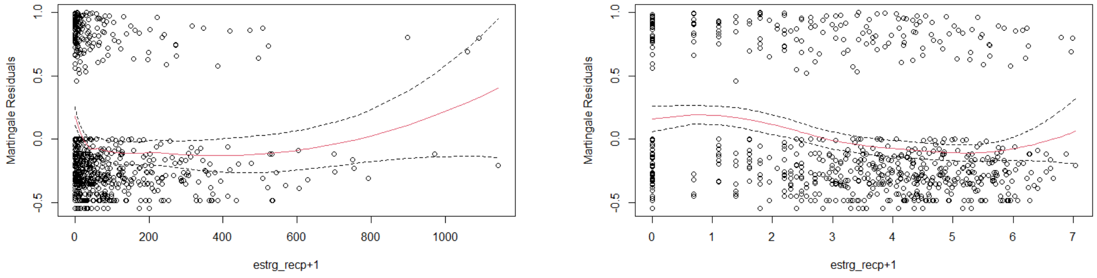


(c) Progesterone before log transformation. Associated p-value from hypothesis test: 0.00031.

(d) Progesterone after log transformation. Associated p-value from hypothesis test: 0.018.

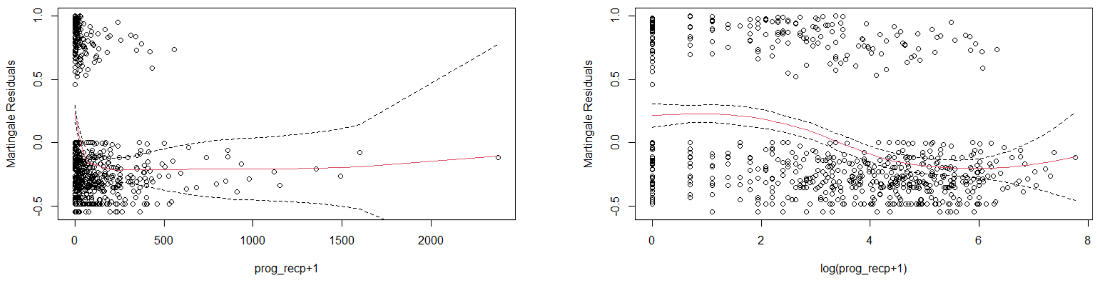
Figure 4.7: Univariable Schoenfeld residuals plots for time to death before and after log transform of the estrogen and progesterone variable. Figure 4.7a and 4.7b show the Schoenfeld residuals plots for the estrogen variable before and after log transform, respectively. Similarly, figure 4.7c and 4.7d show the Schoenfeld residuals plots for the progesterone variable before and after log transform, respectively.

In the Schoenfeld residuals plots shown above, the two dashed curves illustate upper and lower confidence interval boundaries. The solid curve visible in betwen the two dotted curves is a curve fitted to the points in the plot and should work as an aid to detect any trends in the plot. A rough interpretation of these plots is that, if it is possible to fit a straight horizontal line in between the two dashed curves, the assumption of proportional hazards holds. Details on how to create the Schoenfeld residuals plot can be found in appendix.



(a) Estrogen before log transformation.

(b) Estrogen after log transformation.



(c) Progesterone before log transformation. (d) Progesterone after log transformation.

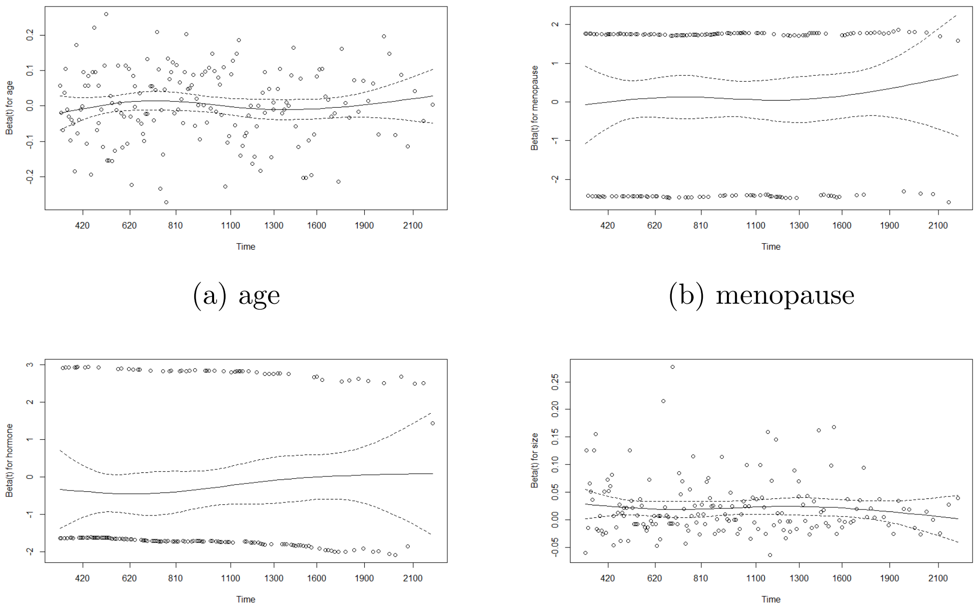
Figure 4.8: Univariable Martingale residuals plots for time to death before and after log transform of the estrogen and progesterone variable. Figure 4.8a and 4.8b show the Martingale residuals plots for the estrogen variable before and after log transform, respectively. Similarly, figure 4.8c and 4.8d show the Martingale residuals plots for the progesterone variable before and after log transform, respectively.

4.4.1 Time to death

Explanatory variable	$\hat{\beta}$	$e^{\hat{\beta}}$	Confidence interval for e^{β}	p-value
age	0.0016	1.0016	[0.99, 1.0]	0.83
menopause = 1	ref			
menopause = 2	0.11	1.1	[0.82, 1.5]	0.48
hormone = 1	ref			
hormone = 2	-0.26	0.77	[0.56, 1.1]	0.11
size	0.021	1.0	[1.01, 1.03]	$8.54 \cdot 10^{-7}$
grade 1	ref			
grade 2	1.2	3.5	[1.5, 7.9]	0.0030
grade 3	1.9	6.4	[2.8, 15]	$1.4 \cdot 10^{-5}$
nodes	0.078	1.1	[1.06, 1.10]	$2 \cdot 10^{-16}$
prog_recp	-0.32	0.72	[0.67, 0.78]	$2 \cdot 10^{-16}$
estrg_recp	-0.21	0.81	[0.75, 0.88]	$1.22 \cdot 10^{-7}$

Table 4.5: Regression coefficient table from univariable analysis. Time to death. "*ref*" means reference group for categorical variables.

The first column in table 4.5 tells you which explanatory variable is considered. The remaining columns from left to right are quantities associated with the explanatory variable in column 1. The quantities are as follows: Estimated regression coefficient $\hat{\beta}$, estimate of hazard ratio $e^{\hat{\beta}}$, 95 percent confidence interval for the hazard ratio e^{β} and p-value from Wald test. Note that menopause, hormone and grade are all categoric. For these variables the quantities in column 2 to 5 is calculated with respect to a reference state. For menopause the reference state is "*menopause is reached*", for hormone it is "*is receiving tamoxifen*" and for grade it is "*grade 1*". The regression coefficient of the reference state is zero, and thus the corresponding hazard ratio is 1. To illustrate this principle, consider the grade variable. This variable has three categories: Grade 1, Grade 2 and Grade 3. When reporting a hazard ratio for the grade variable, Grade 1 is put as the reference state (that is, $\beta = 0$ and thus $e^0 = 1$ for Grade 1). If $e^{\hat{\beta}} = 3.5$ for Grade 2, this means that the hazard rate of Grade 2 is 3.5 times bigger that of Grade 1. Similarly, if $e^{\hat{\beta}} = 6.4$ for Grade 3, it means that the hazard rate of Grade 3 is 6.4 times that of Grade 1. Note that, the p-values in figure 4.5 are almost the same as those in figure 4.3. Some differences exist, though. The differences occur because, to calculate the p-values in figure 4.5, regression coefficient estimates calculated from Cox regression are required. To calculate the p-values in table 4.3 on the other hand, regression coefficient estimates and thus Cox regression is not required.

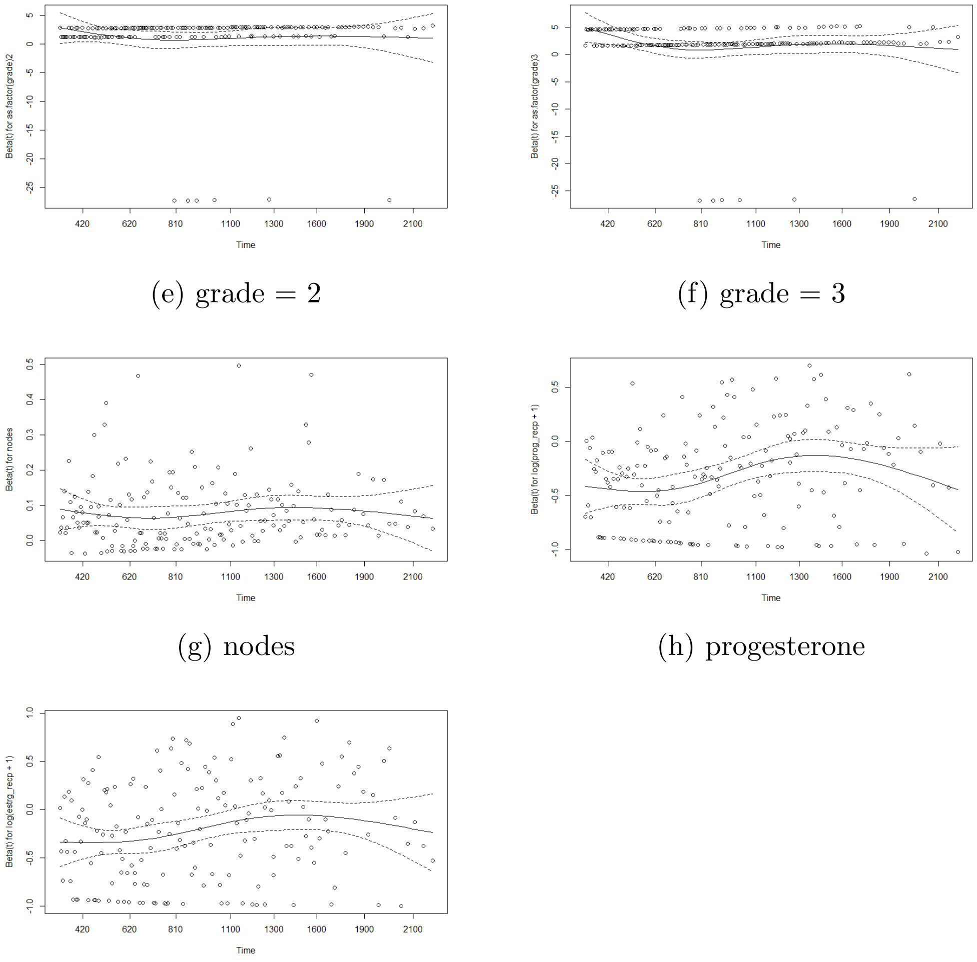


(a) age

(b) menopause

(c) hormone

(d) size



(e) grade = 2

(f) grade = 3

(g) nodes

(h) progesterone

(i) estrogen

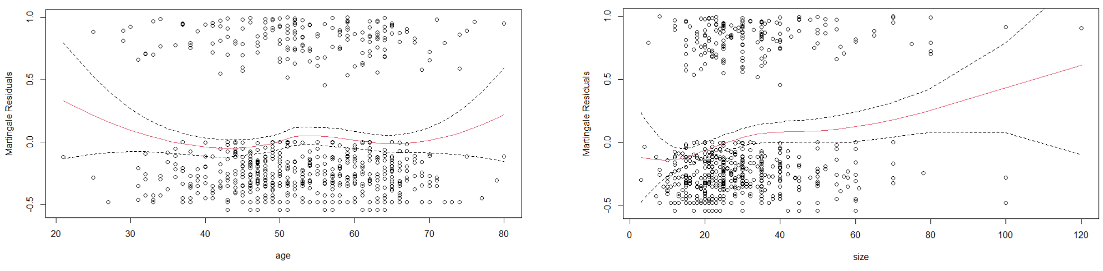
Figure 4.9: Schoenfeld residuals plots associated with the different explanatory variables to check assumption of proportional hazards. Univariable analysis. Time to death.

In all the plots in figure 4.9 the solid black curve fitting the points was found to be quite horizontal. It was possible to fit a straight horizontal line between the dashed curves for all variables, except for the estrogen and progesterone variable. Even if the requirement was not quite satisfied for these two, the solid fitted curve had a non-monotonic nature (neither monotonically increasing nor monotonically decreasing). Because of the non-monotonic nature of the plots and the deviations not being large, the deviations from proportional hazards was concluded to not be severe/ critical, and therefore a decision was made to continue the analysis without introducing methods for handling non-proportional hazards.

Explanatory variable	**p-value**
Age	0.96
Menopause = 2	0.55
Hormone = 2	0.31
Size	0.59
Grade = 2	0.090
Grade = 3	0.043
Nodes	0.55
Progesterone	0.018
Estrogen	0.021

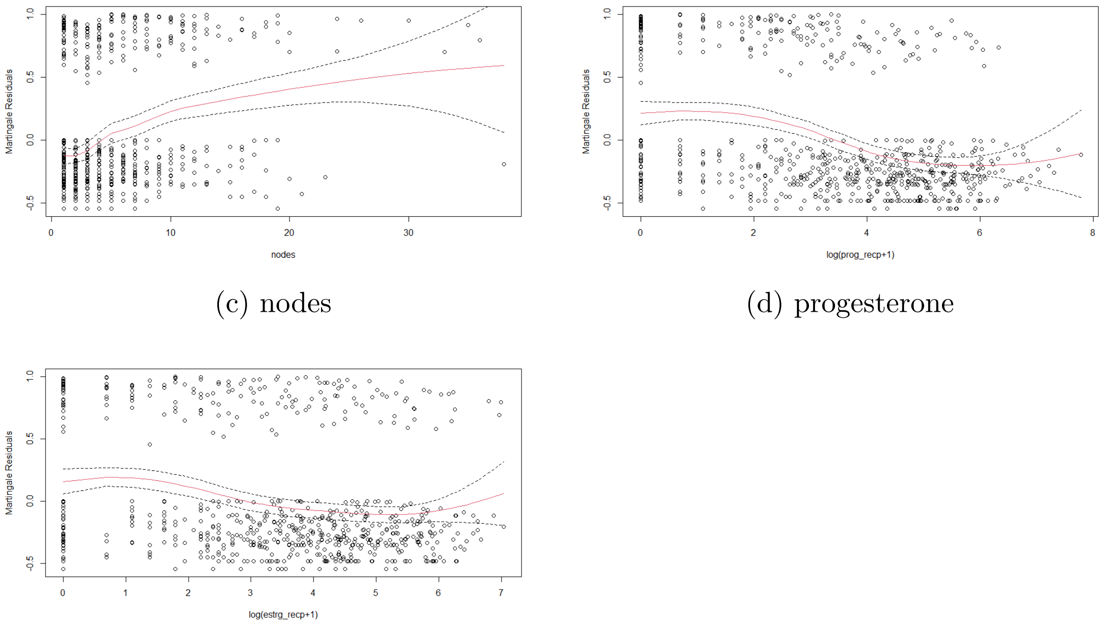
Table 4.6: p-values for hypothesis test of proportional hazards. Univariable analysis. Time to death

It is of interest to check whether the functional form of the variables included in the model are on a preferable form. This is done by looking at the plot resulting from plotting the martingale residuals against a given functional form of the explanatory variable value of interest. The plots are shown below. Along the vertical axis is the martingale residuals, and along the horizontal axis is the functional form of the explanatory variable value. As mentioned in the introduction of this section, the red line is the LOESS smoother and the dashed lines are the boundaries of a 95% confidence interval. As explained in section 3.4.2, if the LOESS smoother resembles a straight line, this indicates that the functional form that is used along the horisontal axis is adequate. If it was possible to fit a straight line in between the dashed lines, it was concluded that the suggested functional form was adequate. See R-code in appendix to see how the plots were created.



(a) age

(b) size



(c) nodes

(d) progesterone

(e) estrogen

Figure 4.10: Martingale residuals plots associated with the different explanatory variables. Univariable analysis. Time to death.

None of the variables has a LOESS smoother that perfectly fits a straight line, but they all meet the requirement of fitting a straight line inside the confidence interval boundaries.

Now that we have performed our univariable analysis, it is time to start the multivariable analysis. The first step is to decide which of the explanatory variables the hazard function should depend on. To start with, we will include all those explanatory variables whose p-value in table 4.5 is lower than 0.2. Then, we will use R to calculate the new regression coefficient estimates for the explanatory variables included in the model. One by one, we will remove from the model those explanatory variables that have a Wald test p-value larger than 0.05. If several variables have p-values larger than 0.05, the one with the largest p-value will be removed, and then regression coefficient estimates will be calculated again for the remaining variables. This process will be repeated until none of the explanatory variables' p-values are larger than 0.05. For the grade variable, it was decided to include it if either one or both of $grade = 2$ or $grade = 3$ had a p-value smaller than 0.05. In table 4.7 below you can see which explanatory variables remained after the elimination process was finished.

Explanatory variable	$\hat{\beta}$	$e^{\hat{\beta}}$	Confidence interval for e^{β}	p-value
size	0.011	1.0	[1.00, 1.02]	0.02
grade 1	ref			
grade 2	0.71	2.0	[0.88, 4.7]	0.10
grade 3	0.94	2.6	[1.1, 6.1]	0.037
nodes	0.061	1.1	[1.04, 1.08]	$5.0 \cdot 10^{-9}$
prog_recp	-0.27	0.77	[0.71, 0.83]	$1.87 \cdot 10^{-10}$

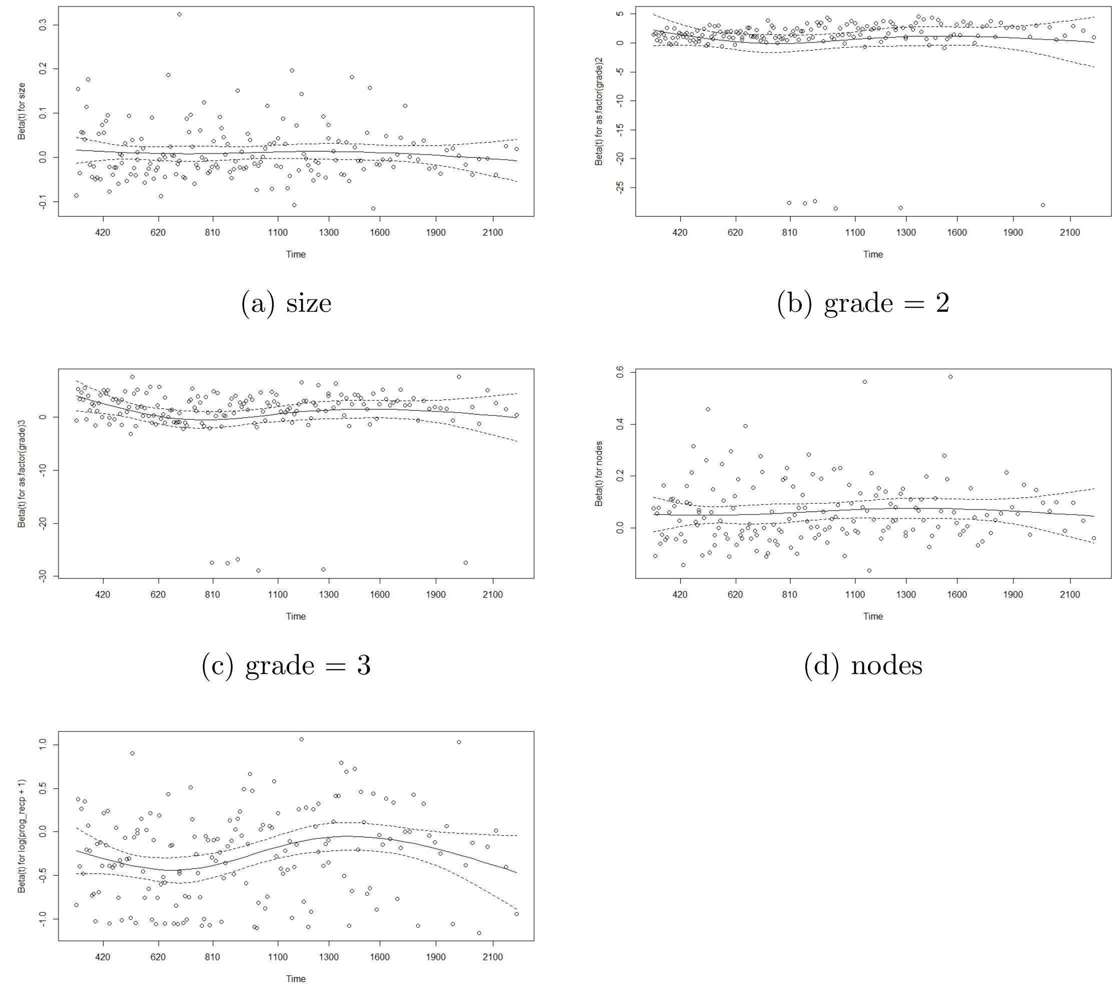
Table 4.7: Regression coefficient table from multivariable analysis. Time to death. "*ref*" means reference group for categorical variables.

Based on the regression coefficient estimate values in table 4.7 the assumption of proportional hazard for each of the explanatory variables will now be checked.

Explanatory variable	**p-value**
Size	0.68
Grade = 2	0.15
Grade = 3	0.081
Nodes	0.79
Progesteron	0.024

Table 4.8: p-values for hypothesis test of proportional hazards. Multivariable analysis. $prog\_recp$ is transformed to $log(prog\_recp + 1)$. Time to death

We choose to assume that the assumption of proportional hazards is met if the p-value in table 4.8 is larger than 0.05, combined with checking if it is possible to fit a horizontal straight line inside the confidence interval boundaries in the residual plots. Under these requirements, we see that the progesterone variable is not accepted. The associated residual plots are shown below.



(a) size

(b) grade = 2

(c) grade = 3

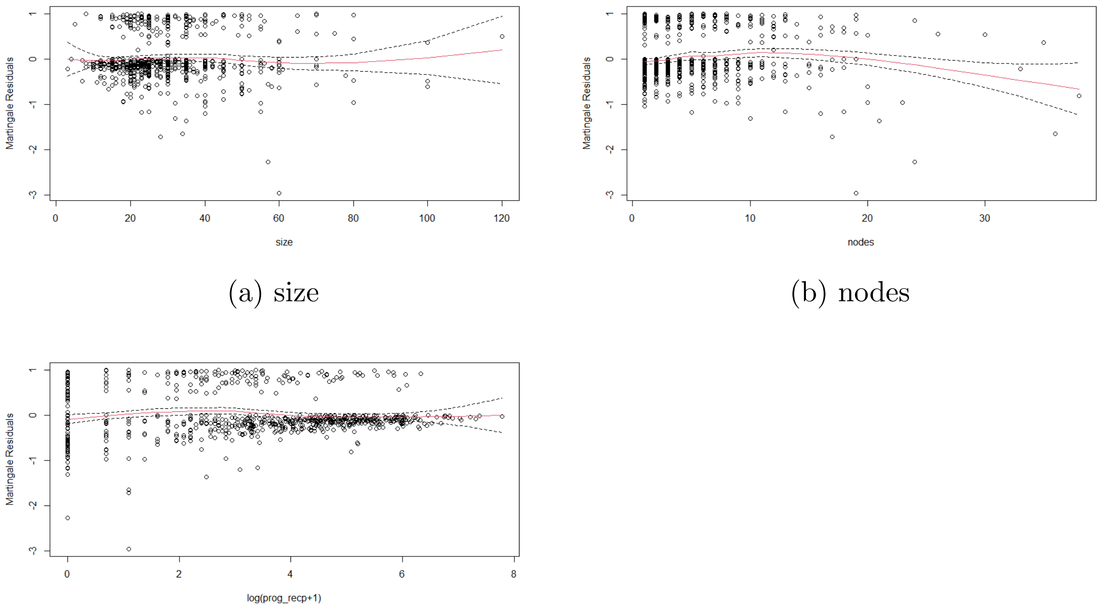
(d) nodes

(e) progesterone

Figure 4.11: Schoenfeld residuals plots associated with the different explanatory variables to check assumption of proportional hazards. Multivariable analysis. Time to death.

All of the variables fulfill the requirement of fitting a straight horizontal line inside the confidence interval boundaries, except for the progesterone variable. But since the variation is small and the smoother doesn't show any monotone behavior, the deviance from horizontal linear behavior is concluded to not be critical.

The martingale residuals plots corresponding to the model including the explanatory variables are as follows:



(a) size

(b) nodes

(c) progesterone

Figure 4.12: Martingale plots associated with the different explanatory variables. Univariable analysis. Time to death.

Plot 4.12b suggests that the nodes variable has some issues for large values of nodes. Then the functional form suggested does not hold the requirement of fitting a straight horizontal line. Further investigation can be done to see if any transformation of the nodes variable will meet the requirement better. Note also that the requirement of fitting a straight horizontal line is not fully met for the progesterone variable in plot 4.12c, but since there is only a minor non-monotonic behavior of the smoother, the deviance is not considered to be serious.

Next in the multivariable analysis, we check for interactions. The number of possible interactions are $\binom{4}{2} = 6$, and they are:

1) size iteracted with nodes
2) size interacted with grade
3) size interacted with *prog_recp*
4) grade interacted with nodes
5) grade interacted with *prog_recp*
6) *prog_recp* interacted with nodes

To check for intractions, the programming script in R was runned six times, each time with one of the interactions listen above together with the explanatory variables in table 4.7. For each run of the script it was checked if the interaction's p-value was smaller than 0.05. After having identified which of the interactions had a p-value smaller than 0.05, the script was runned again with these interactions together with the explanatory variables in table 4.7. Then, one at a time, those interactions who's p-values were larger than 0.05 were eliminated (if several of the interactions had p-value larger than 0.05, the one with the largest was eliminated first). After each elimination, the script was runned again. The elimination process was executed until none of the interactions had a p-value larger than 0.05.
After executing the procedure explained above, it turned out that none of the interactions listen above survived, and it was concluded that there were no interactions between the variables in 4.7 for time to death.

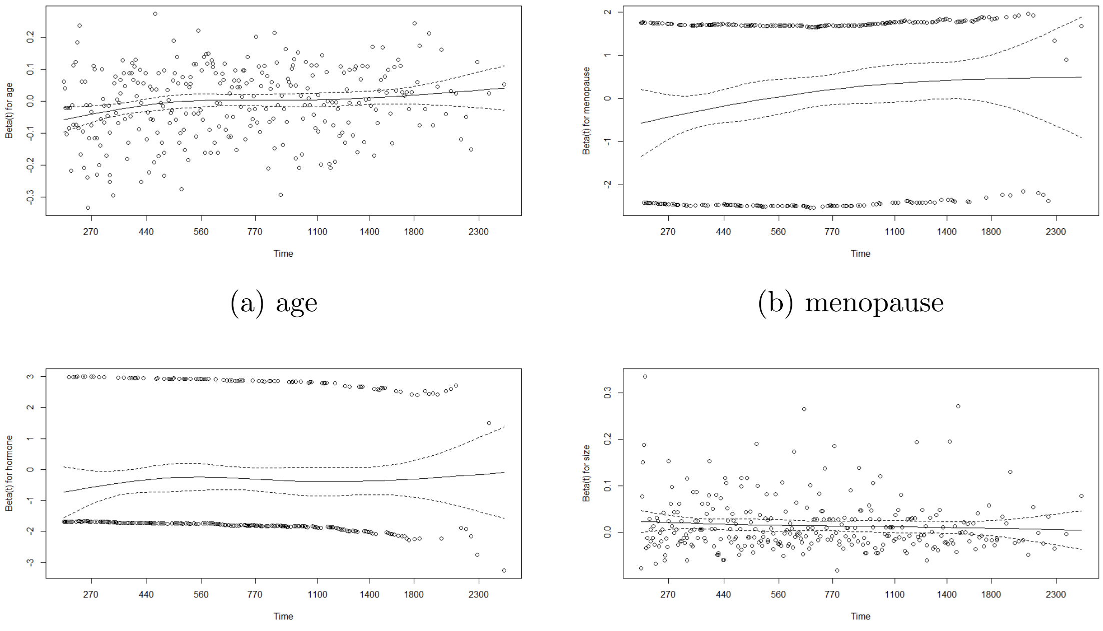
4.4.2 Time to recurrence

Explanatory variable	$\hat{\beta}$	$e^{\hat{\beta}}$	Confidence interval for e^{β}	p-value
age	-0.0045	1.0	[0.98, 1.0]	0.46
menopause = 1	ref			
menopause = 2	0.064	1.1	[0.85, 1.3]	0.59
hormone = 1	ref			
hormone = 2	-0.37	0.69	[0.54, 0.89]	0.0035
size	0.015	1.02	[1.008, 1.020]	$1.84 \cdot 10^{-5}$
grade 1	ref			
grade 2	0.87	2.4	[1.5, 3.9]	0.00039
grade 3	1.2	3.2	[1.9, 5.3]	$1.0 \cdot 10^{-5}$
nodes	0.071	1.1	[1.06, 1.09]	$2 \cdot 10^{-16}$
prog_recp	-0.21	0.81	[0.76, 0.86]	$5.0 \cdot 10^{-13}$
estrg_recp	-0.14	0.87	[0.82, 0.93]	$9.4 \cdot 10^{-6}$

Table 4.9: Regression coefficient table from univariable analysis. Time to reccurence. "*ref*" means reference group for categorical variables.

Explanatory variable	**p-value**
Age	0.0025
Menopause = 2	0.011
Hormone = 2	0.63
Size	0.27
Grade = 2	0.062
Grade = 3	0.0014
Nodes	0.66
Progesteron	0.012
Estrogen	0.0007

Table 4.10: p-values for hypothesis test of proportional hazards. Univariable analysis. Time to recurrence

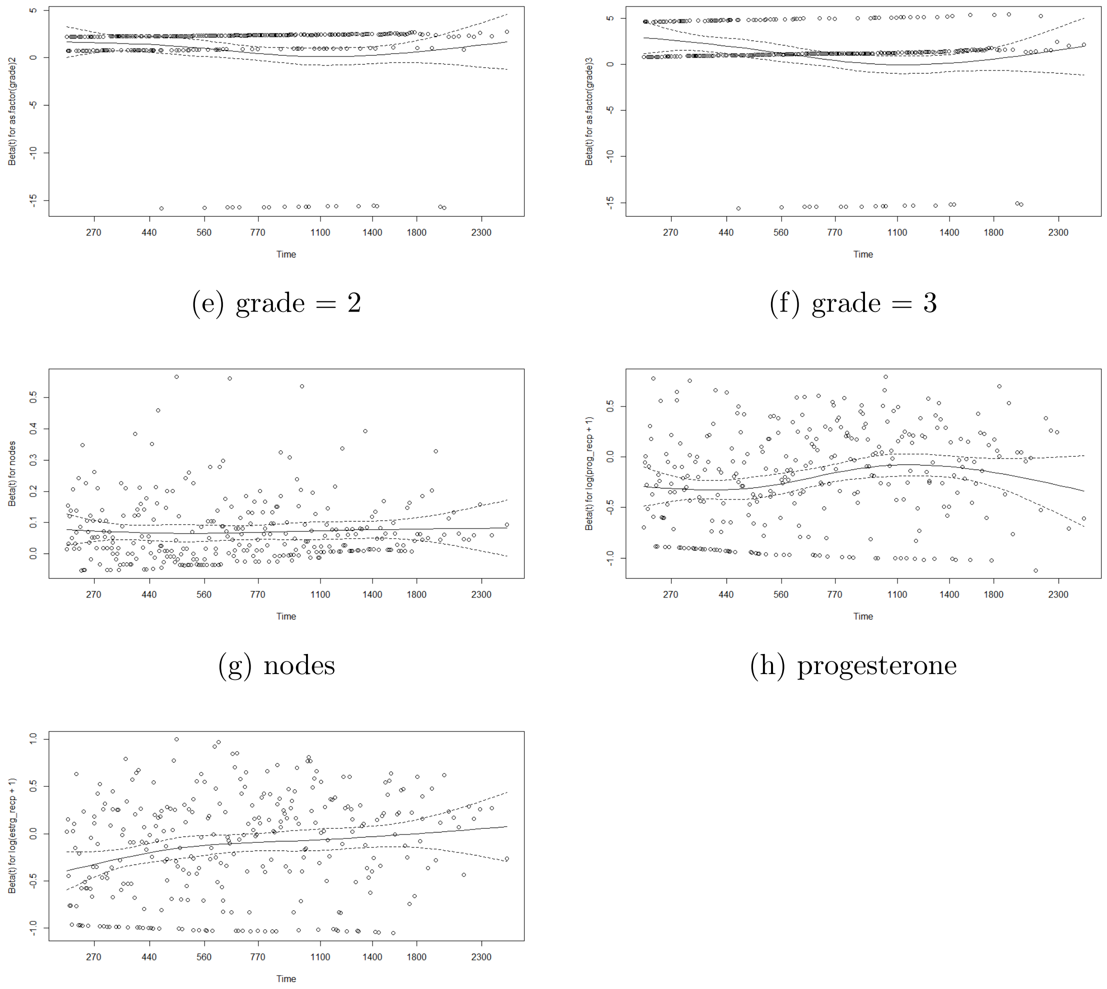


(a) age

(b) menopause

(c) hormone

(d) size



(e) grade = 2

(f) grade = 3

(g) nodes

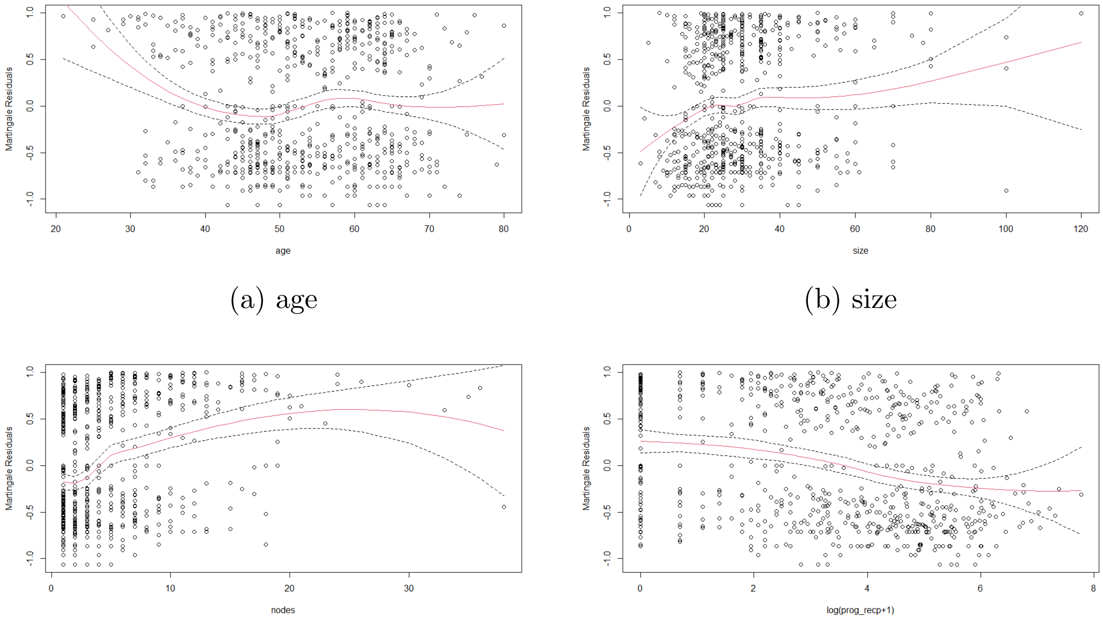
(h) progesterone

(i) estrogen

Figure 4.13: Schoenfeld plots associated with the different explanatory variables to check assumption of proportional hazards. Univariable analysis. Time to recurrence.

The p-values in table 4.10 show that the hormone, size, $grade = 2$ and nodes variables all have p-values above the chosen significance level of 0.05. Having a glance at their associated residuals plots in figure 4.13 shows that it is possible to fit a straight horizontal line inside the confidence interval boundaries for all of them, without any of them showing any serious sign of monotonic trend. The age, menopause, $grade = 3$, progesterone and estrogen variables on the other hand all have p-values below 0.05. Of these, the age, menopause and estrogen variables show a monotonic increasing behavior, indicating deviation from proportional hazards. Thus, investigation should be made to see of it is possible to modify the Cox model to take their time-dependent coefficients into consideration. For the $grade = 3$ and progesterone variables it is possible to fit a straight line inside the confidence interval boundaries.

The martingale residuals plots associated with the continuous variables in table 4.7 are shown in figure 4.10 below.

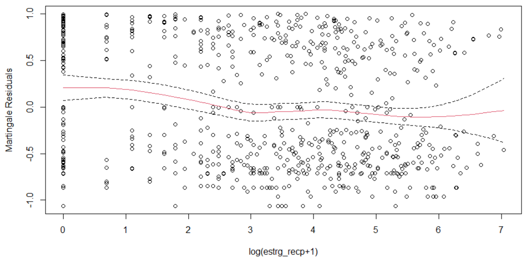


(a) age

(b) size

(c) nodes

(d) progesterone



(e) estrogen

Figure 4.14: Martingale residuals plots associated with the different explanatory variables. Univariable analysis. Time to recurrence.

It is clear from plot 4.14c that, the nodes variable doesn't meet the requirement of fitting a straight line inside the dahsed boundary lines. Therefore, it must be done some investigation to find a transformation of this variable to make it meet the requirement. The same applies for the age variable. One possibility is to find a transformation that would meet the requirement. An alternative approach would be to categorize these continuous variables into subgroups, such that each subgroup corresponds to a specific interval containing a portion of the continuous variable's values. The subgroups' intervals are chosen in a such way that the requirement for the martingale residuals are met for each group. It was found useful to compare the Kaplan-Meier curves with the martingale plots. It turned out that, for both the age and the nodes variable, the martingale requirement was met if the variables were grouped into the same intervals used in the Kaplan-meier curve plots.

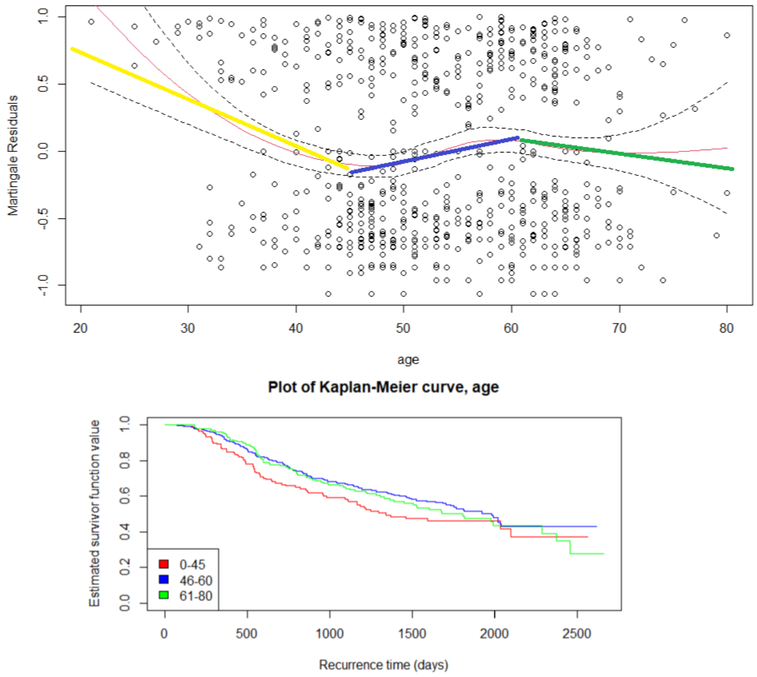

Plot of Kaplan-Meier curve, age

Figure 4.15: Comparison of Martingale residuals plot and Kaplan-Meier curve of the age variable. Time to recurrence.

The upper part of figure 4.15 is the martingale residuals plot from figure 4.14a and the lower part is the Kaplan-Meier curve plot from figure 4.5a. The yellow, blue and green straight lines drawn onto the martingale residuals plot is resembling the following age groups: Group 1 (0 to 45 years), groups 2 (46 to 60 years) and group 3 (61 to 80 years). By categorizing the age variable in this way it was found possible to meet the martingale residuals requirement.

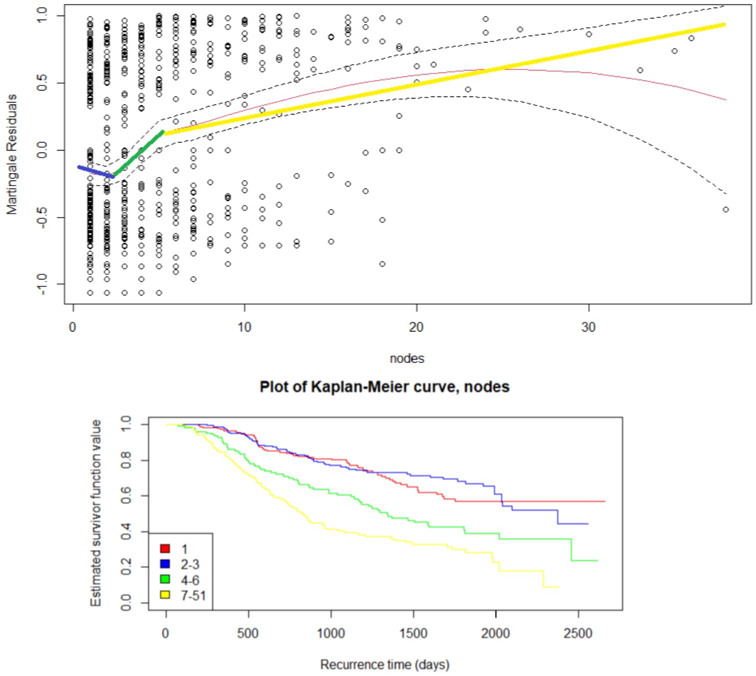


Figure 4.16: Comparison of Martingale residuals plot and Kaplan-Meier curve of the nodes variable. Time to recurrence.

The upper part of figure 4.16 is the martingale residuals plot from figure 4.14c and the lower part is the Kaplan-Meier curve plot from figure 4.5f. The blue, green and yellow straight lines drawn onto the martingale residuals plot is resembling the following nodes groups: Group 1 (1 to 3 nodes), groups 2 (4 to 6 nodes) and group 3 (7 to 51 nodes). By categorizing the nodes variable in this way it was found possible to meet the martingale residuals requirement.

After the elimination process (by only including those variables from table 4.9 that has a p-value below 0.05), we see that the variables that the hazard function depends on for time to recurrence is hormone, grade, nodes and *prog_recp*. Note that these are almost the same set of variables as for time to death. The difference is that hormone has taken the place of the size variable. The nodes variable was made a categorical variable before performing the multivariable analysis. The results of the multivariable Cox regression analysis are shown in table 4.11 below.

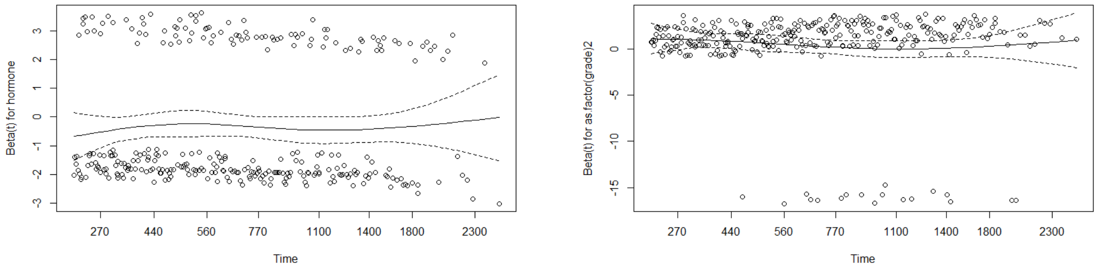
Explanatory variable	$\hat{\beta}$	$e^{\hat{\beta}}$	Confidence interval for e^{β}	p-value
hormone	-0.37	0.69	[0.54, 0.88]	0.0033
grade 1	ref			
grade 2	0.52	1.7	[1.0, 2.7]	0.039
grade 3	0.57	1.8	[1.0, 3.0]	0.039
nodescategorical = 1	ref			
nodescategorical = 2	0.74	2.1	[1.6, 2.8]	$1.4 \cdot 10^{-6}$
nodescategorical = 3	1.1	2.9	[2.2, 3.7]	$4.9 \cdot 10^{-15}$
prog_recp	-0.17	0.84	[0.79, 0.90]	$8.0 \cdot 10^{-8}$

Table 4.11: Regression coefficient table from multivariable analysis. Time to recurrence. ”*ref*” means reference group for categorical variables.

In table 4.11, ”$nodescategorical = 1$”, ”$nodescategorical = 2$” and ”$nodescategorical = 3$” are the categorical nodes groups earlier refered to as group 1 (1 to 3 nodes), group 2 (4 to 6 nodes) and group 3 (7 to 51 nodes), respectively.

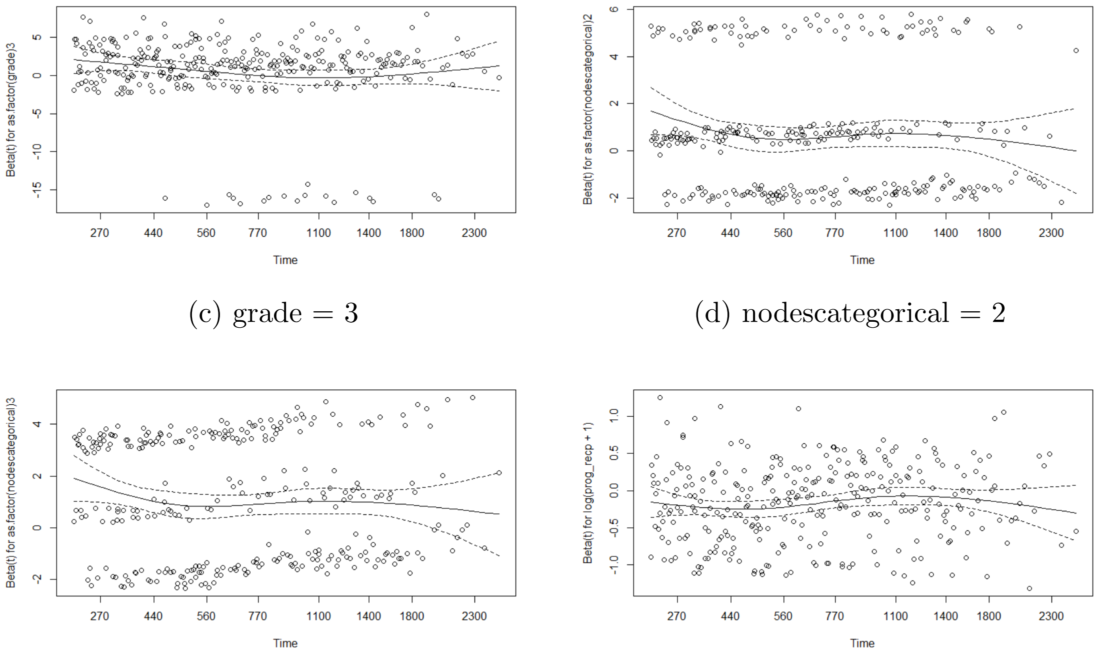
Explanatory variable	**p-value**
hormone	0.64
Grade = 2	0.10
Grade = 3	0.0039
nodescategorical = 2	0.46
nodescategorical = 3	0.23
Progesteron	0.025

Table 4.12: p-values for hypothesis test of proportional hazards. Multivariable analysis. *prog_recp* is transformed to $log(prog\_recp + 1)$. Time to recurrence



(a) hormone

(b) grade = 2



(c) grade = 3

(d) nodescategorical = 2

(e) nodescategorical = 3

(f) Progesterone

Figure 4.17: Schoenfeld plots associated with the different explanatory variables to check assumption of proportional hazards. Multivariable analysis. Time to recurrence.

It is possible to fit a straight horizontal line inside the confidence interval boundaries for all variables in figure 4.20. The p-values in table 4.12 is below 0.05 for the progesterone variable, but since figure 4.17f doesn't show any clear sign of monotonic increasing or monotonic decreasing behavior, the variable is accepted to follow proportional hazards.

Only one of the variables in table 4.11 are continuous, the progesterone variable. Thus, only one Martingale plot will result for the multivariable case for time to recurrence. The plot is shown in figure 4.18 below.

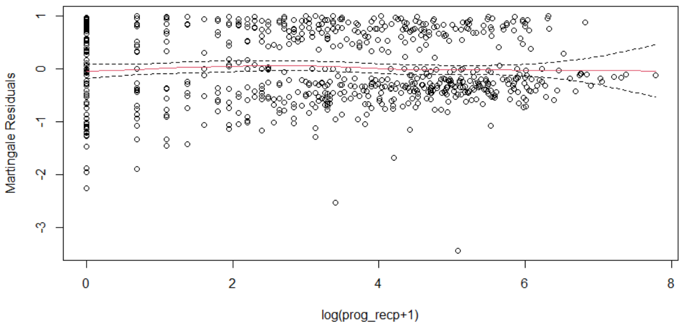


Figure 4.18: Martingale plot of the progesterone variable. Multivariable analysis. Time to recurrence.

For time to recurrence, the number of possible interactions are $\binom{4}{2} = 6$, and they are:

1) hormone iteracted with grade
2) hormone interacted with nodescategoric
3) hormone interacted with *prog_recp*
4) grade interacted with nodescategoric
5) grade interacted with *prog_recp*
6) *prog_recp* interacted with nodescategoric

After executing the elimination process, the interactions listed above that survived were 3 (hormone interacted with *prog_recp*
) and 4 (grade interacted with nodescategoric).

4.4.3 Recurrence as time-dependent variable

In section 3.5 the idea about time-dependent explanatory variables was introduced. In this section, the analysis procedure presented in section 4.4.1 and 4.4.2 will be executed on the same data set as before, but now looking at time to death and considering recurrence as a separate explanatory variable who's value is time-dependent but regression coefficient is constant. This will result in a separate regression coefficient for the recurrence. Details of how this is implemented in R is explained in the literature [5]. What was done was to assign a new categorical explanatory variable "*rec*" that either took the value zero or unity, depending on if a recurrence was observed or not, respectively. When recurrence occured, the variable "rec" would change value from 0 to 1 at the recurrence time. Intuition tells that the regression coefficient of the "rec" variable should be large, because recurrence is a bad sign when it comes to chance of fatality for cancer patients. The larger the regression coefficient, the more likely it is for the patient to die.

Explanatory variable	$\hat{\beta}$	$e^{\hat{\beta}}$	Confidence interval for e^{β}	p-value
rec	3.7	42	[26, 69]	$2 \cdot 10^{-16}$
age	0.0013	1	[0.99, 1.0]	0.87
menopause = 1	ref			
menopause = 2	0.11	1.1	[0.82, 1.5]	0.50
hormone = 1	ref			
hormone = 2	-0.26	0.77	[0.56, 1.0]	0.11
size	0.021	1.0	[1.01, 1.03]	$9.16 \cdot 10^{-7}$
grade 1	ref			
grade 2	1.2	3.5	[1.5, 7.9]	0.0031
grade 3	1.9	6.5	[2.8, 15]	$1.3 \cdot 10^{-5}$
nodes	0.068	1.1	[1.05, 1.09]	$2.79 \cdot 10^{-16}$
prog_recp	-0.32	0.72	[0.67, 0.78]	$2 \cdot 10^{-16}$
estrg_recp	-0.21	0.81	[0.75, 0.87]	$1.1 \cdot 10^{-7}$

Table 4.13: Regression coefficient table from univariable analysis. "*ref*" means reference group for categorical variables. Time to death. Recurrence as explanatory variable.

In principle, except of an additional row for the recurrence variable "*rec*", the values in table 4.13 should be the same as found in table 4.5. But since another script was used in the model with recurrence as time dependent variable (and the structure of this script was a little bit different), some minor differences were observed. The schoenfeld residuals plots and martingale residuals plots for the univariable analysis part should in principle also be the same as found in section 4.4.1. In the second column of the rec variable, note the value of $e^{\beta} = 42$. An interpreation of this number is that, a person with recurrence has 42 times higher chance of dying at a particular time incident compared with a person that didn't have recurrence.

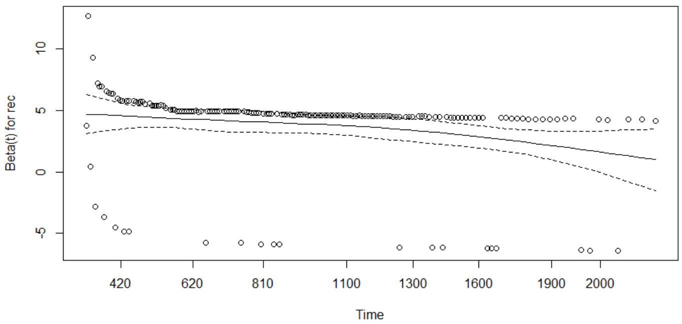


Figure 4.19: Plot to check assumption of proportional hazards for the recurrence variable. Univariable analysis. Time to death. Recurrence as explanatory variable.

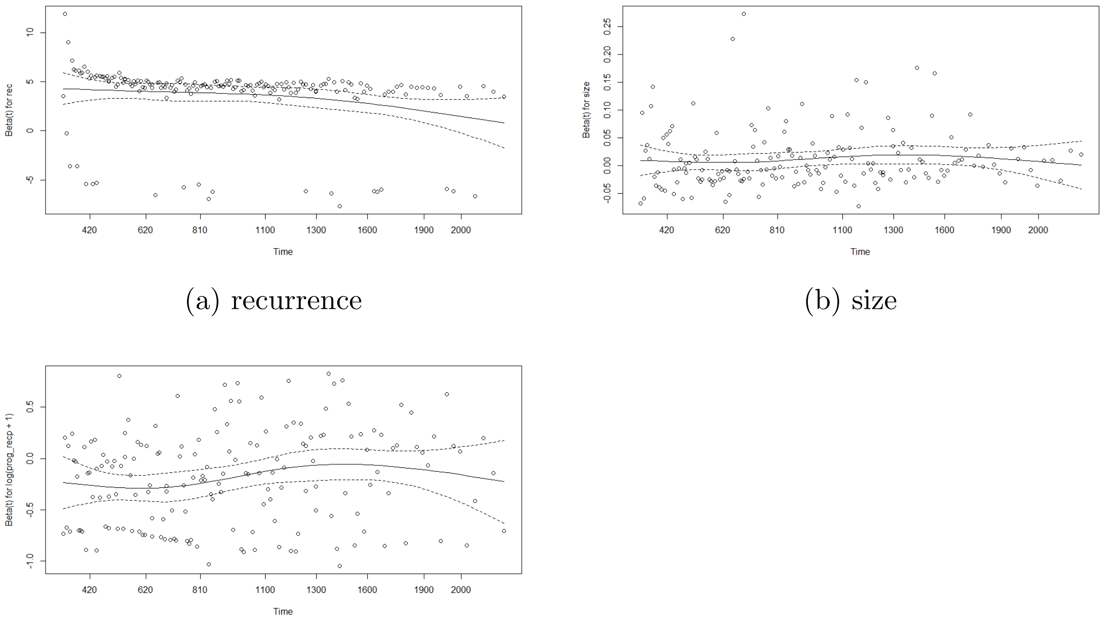
From figure 4.19 it is visible that a horizontal straight line is not possible to fit between the confidence boundaries, and the assumption of proportional hazards doesn't hold. The associated p-value was found to be 0.00079, which isn't adequate. Measures must be done to solve this issue.

Explanatory variable	$\hat{\beta}$	$e^{\hat{\beta}}$	Confidence interval for e^{β}	p-value
rec	3.5	35	[21, 57]	$2 \cdot 10^{-16}$
size	0.012	1.0	[1.0, 1.02]	0.0068
prog_recp	-0.19	0.83	[0.76, 0.89]	$2.2 \cdot 10^{-6}$

Table 4.14: Regression coefficient table from multivariable analysis. Time to death. ”*ref*” means reference group for categorical variables. Recurrence as explanatory variable

After executing the elimination procedure, the explanatory variables left are rec, size and *prog_recp*. Before including recurrence in the model, size and *prog_recp* were also significant, together with nodes and grade (see table 4.7). By including recurrence in the model nodes and grade turned out to not be significant anymore. But it is important to know that both nodes and grade are important variables when it comes to time to recurrence (see figure

4.5e and 4.5f for Kaplan-Meier curves and table 4.9 and 4.11 for regression coefficients). So, both of the nodes and grade variables are important in this model too, but through the recurrence variable.



(a) recurrence

(b) size

(c) progesterone

Figure 4.20: Schoenfeld residuals plots associated with the different explanatory variables to check assumption of proportional hazards. Multivariable analysis. Time to death. Recurrence included as explanatory variable.

Explanatory variable	**p-value**
rec	0.0016
size	0.68
prog_recp	0.041

Table 4.15: p-values for hypothesis test of proportional hazards. Multivariable analysis. Time to death. Recurrence included as explanatory variable.

From table 4.15 it can be seen that, the p-value for the size variable is very acceptable, but the progesterone variable's p-value is quite low and there is reason to ask the question of whether this variable has a time-dependent

regression coefficient or not. The Schoenfeld residuals plots of size and progesterone in figure 4.20b and 4.20c, respectively, show no sign of either monotonically increasing or decreasing pattern, and therefore it is assumed that both variables do not have time-dependent regression coefficients.
As seen from table 4.15 the recurrence variable has a low p-value, combined with difficulty of fitting a horizontal straight line inside the dashed confidence interval boundary curves in figure 4.20a. For most of the time span, its Schoenfeld residuals plot shows a montonically decreasing linear pattern, indicating a negative linear time-dependent regression coefficient. For survival times below 500 it seems like there is some non-linear behavior which will be neglected for the moment being (this because of few points follow this non-linear trend).
An attempt was made to resolve this issue by modifying the Cox model to allow time-dependent regression coefficients. This was done by including a so-called "tt"-term in the in-built coxph-function used in R to perform the Cox regression analysis. A linear term in time was used as the tt-term. The results are shown in table 4.16 below.

Explanatory variable	$\hat{\beta}$	$e^{\hat{\beta}}$	Confidence interval for e^{β}	p-value
rec	4.81	122	[44.2, 337]	$2 \cdot 10^{-16}$
size	0.0109	1.01	[1.00, 1.02]	0.0101
prog_recp	-0.190	0.827	[0.764, 0.895]	$2.39 \cdot 10^{-6}$
tt(rec)	-0.00135	0.999	[0.998, 1.00]	0.00170

Table 4.16: Regression coefficient table from multivariable analysis. Time to death. Recurrence as explanatory variable.

The total expression of the hazard function based on the variables in table 4.16 is given by

$$h(t) = h_0(t)e^{(\beta_0+\beta_1\cdot t)\cdot rec+\beta_2\cdot size+\beta_3\cdot prog\_recp}. \tag{4.2}$$

In (4.2), rec takes the value zero or unity depending on whether recurrence is absent or present, respectively. The term $\beta_0 + \beta_1 \cdot t$ make up the function of a straight line, where t is survival time, and β_0 and β_1 is the regression coefficients of rec and $tt(rec)$ in table 4.16, respectively. Both β_0 and β_1 in (4.2) can be associated with figure 4.20a, where β_0 can be thought of as the intercept of a straight line with the vertical axis and β_1 is the slope of the line. If we insert a straight line with intercept equal to $\beta_0 = 4.81$ and slope of $\beta_1 = -1.35 \cdot 10^3$ into the plot in figure 4.20a, we see the following:

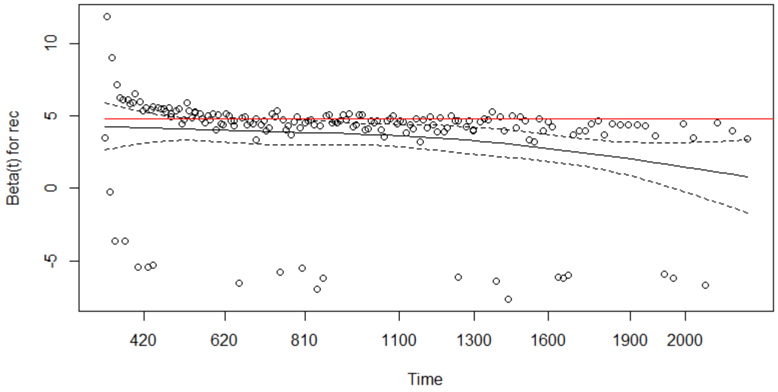
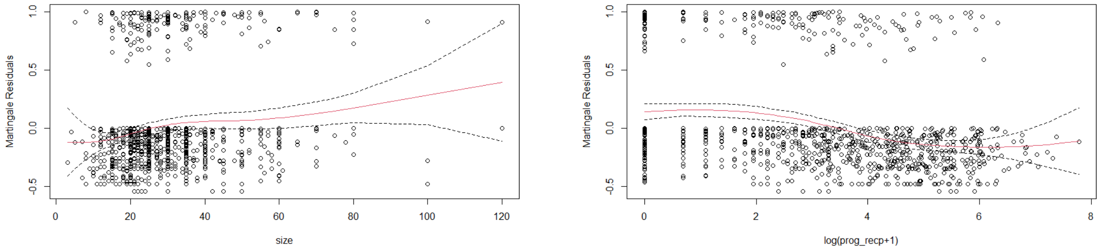


Figure 4.21: Figure 4.20a with inserted red straight line, showing the linear time-dependent recurrence variable.

The red line is the function $y = 4.81 - 0.00135t$. Over all, the line seem to represent the points in the plot quite well.

Let's now have a look at the Martingale residuals plots.



(a) size

(b) progesterone

Figure 4.22: Martingale plots associated with the different explanatory variables. Multivariable analysis. Time to death. Recurrence included as explanatory variable.

It is observed from figure 4.22a and figure 4.22b that, both of the size and progesterone variables have trouble meeting the requirement of fitting a straight horizontal line inside the confidence interval boundaries in the martingale residuals plots, respectively. It can be discussed whether the size variable's deviance from the requirement is so small that it should be accepted. For the progesterone variable the situation is more severe. It should be investigated more closely what can be done to these two variables for them to better meet the requirement.

Possible interactions are:

1) rec with size
2) rec with *prog_recp*
3) size with *prog_recp*

None of these interactions were found to be significant.

Chapter 5

Simulations

In this chapter there will be simulated survival times that will be analysed using Cox regression. First, survival data for which the assumption of proportional hazards is satisfied will be simulated, assuming that the survival times follow a probability distribution known as the *Weibull distribution*. Next, survival data for which the proportional hazards assumption is not satisfied will be simulated, with the aim of finding out how good are the methods for detecting non-proportional hazards used in chapter 4. Before starting with simulations and applying the Cox regression analysis procedure to the simulated data, a short presentation of the Weibull distribution and its relation to Cox regression will be presented.

5.1 The Weibull distribution

A Weibull distributed continuous random variable T with scale parameter λ and shape parameter γ is denoted by $T \sim W(\lambda, \gamma)$ [1, p. 155]. The hazard function of $T \sim W(\lambda, \gamma)$ is given by [1, p. 154]

$$h(t) = \lambda\gamma t^{\gamma-1}. \tag{5.1}$$

Recalling (2.23) given by

$$ln(S(t)) = -\int_0^t h(u)du$$

the survivor function becomes [1, p. 127]

$$S(t) = e^{-\int_0^t \lambda\gamma u^{\gamma-1}du} = e^{-\lambda t^{\gamma}}. \tag{5.2}$$

Using the results from (5.1) and (5.2) together with (2.14) given by

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

gives the following expression for the pdf of T:

$$f(t) = h(t)S(t) = \lambda\gamma t^{\gamma-1}e^{-\lambda t^{\gamma}}. \tag{5.3}$$

Further, it can be shown that, the expectation of T is given by [1, p. 155]

$$E(T) = \lambda^{\frac{-1}{\gamma}}\Gamma(\gamma^{-1} + 1), \tag{5.4}$$

where $\Gamma(x)$ is defined by [1, p. 155]

$$\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du = (x-1)!. \tag{5.5}$$

Recall the general proportional hazard model given by (3.10):

$$h_i(t) = e^{\boldsymbol{\beta}^T \boldsymbol{x_i}} h_0(t)$$

.

If the baseline hazard function $h_0(t)$ in (3.10) is replaced by the right-hand side of (5.1), we get [1, p. 176]

$$h_i(t) = e^{\boldsymbol{\beta}^T \boldsymbol{x_i}} \lambda\gamma t^{\gamma-1}. \tag{5.6}$$

The corresponding survivor function becomes [1, p. 176]

$$S_i(t) = e^{-\lambda t^\gamma e^{\boldsymbol{\beta}^T \boldsymbol{x_i}}}. \tag{5.7}$$

5.2 The Inverse Transform Method for generating random variables

How do we generate random numbers? The computer software R has a PRNG (pseudo random number generator) for uniformly distributed numbers. The in-built function in R functioning as a PRNG to generate uniformy distributed numbers is "*runif*". By combining this uniform PRNG with a method called *The Inverse Transform Method*, it is possible to generate random numbers from a probability distribution of interest. Rizzo summarizes the Inverse Transform Method in her book [12, p. 50]. First, start with finding the cumulative distribution function given by

$$F(x) = \int_{-\infty}^{x} f(x)dx, \tag{5.8}$$

where f(x) is the probability density function of your random variable X. Then, let

$$U = F(x). \tag{5.9}$$

Solve 5.9 for x, which gives

$$x = F^{-1}(U), \tag{5.10}$$

where F^{-1} denotes the inverse of the cumulative distribution function F. Then, generate a uniformly distributed number "U" between 0 and 1 (denoted by $U \sim Uniform(0, 1)$). Inserting this U into 5.10 yields a random variable value x.

5.2.1 Applying The Inverse Transform Method on a Weibull distributed random variable

Recall from (2.6) that, the survivor function can be expressed as

$$S(t) = P(T \geq t) = 1 - F(t),$$

where $F(t)$ is the cumulative distribution function. Solving for $S(t)$ yields

$$F(t) = 1 - S(t). \tag{5.11}$$

Let now $S(t)$ be expressed as in (5.7), but do the following substitution:

$$\lambda_i^* = \lambda \cdot e^{\boldsymbol{\beta}^T \boldsymbol{x_i}}. \tag{5.12}$$

Substituting (5.12) into (5.7) yields

$$S_i(t) = e^{-\lambda_i^* t^\gamma}. \tag{5.13}$$

By substituting (5.13) into (2.6) we get

$$F_i(t) = 1 - S_i(t) = 1 - e^{-\lambda_i^* t^\gamma}. \tag{5.14}$$

Doing as in (5.9) gives

$$U = 1 - e^{-\lambda_i^* t^\gamma}. \tag{5.15}$$

Solving 5.15 for t gives the following expression for the survival time of patient number i:

$$t_i = F_i^{-1}(U) = \left(- \frac{ln(1-U)}{\lambda_i^*} \right)^{\frac{1}{\gamma}}. \tag{5.16}$$

In the following, the expression in (5.16) will be used to generate Weibull distributed survival times. Values of λ, $\boldsymbol{\beta}^T$ and γ will be chosen, U will be simulated by the in-built R-function *runif* and $\boldsymbol{x_i}$ will be simulated from specific probability distributions chosen for the explanatory variables.

5.3 Simulation of Weibull distributed survival times

Now, the simulated survival data will be presented together with the Cox regression analysis that was performed.

In the simulations, it was to some extent made an attempt to mimic the German Breast Cancer Data presented in chapter 4. It was simulated 700 non-censored survival times with three associated explanatory variables. The scale parameter λ was set to 1.5 and the shape parameter γ was set to 0.75. The explanatory variables used were intended to present the size, progesterone and nodes variable. Now each explanatory variable will be presented.

1) Explanatory variable 1
This explanatory variable was to mimic the size variable, which was continuous and fairly symmetrically distributed. This variable was simulated from a normal distribution, using the in-built R-function called "*rnorm*". The input in this function were number of simulated values, mean and standard deviation. The mean was set to 30, and the standard deviation was to start with set to 0.5 but later altered to illustrate some observations about the Cox analysis machinery (to be presented later). Before running simulations, it was checked if some of the simulated values were negative (this could happen if a large enough standard deviation was chosen). If negative values appeared, then a smaller standard deviation was chosen such that negative values did not appear.

2) Explanatory variable 2
This explanatory variable was to mimic the progesterone variable, which was

continuous and non-symmetrically distributed. This variable was simulated from an exponential distribution, using the in-built R-function called "*rexp*". Input to this function were number of simulated values, including a rate parameter. The rate parameter was set to 0.4.

3) Explanatory variable 3
This explanatory variable was to mimic the nodes variable, which was a discrete variable. This variable was simulated from an uniform distribution rounded to the nearest integer number. The in-built R-function used to simulate from the uniform distribution is "*runif*". Input to this function were number of simulated values, minumum value and maximum value, which was set to 1 and 15, respectively.

To allow the reader of the report to reproduce the results presented in this report, the seed used was "*123*". See R-code in appendix for how to set the seed.
After simulating values of survival times and assiciated explanatory variables, it is possible to analyse these data with the Cox regression analysis scripts used in chapter 4. This will now be done to check if the scripts are able to reproduce the regression coefficients assumed when simulating the survival times.

5.3.1 Trying to predict the regression coefficients' values

Denote the three explanatory variables by X_1, X_2 and X_3. Further, assume their associated regression coefficients are time-independent and equal to $\beta_1 = 0.0110$, $\beta_2 = -0.0610$ and $\beta_3 = -0.270$, respectively. Given these parametric values, we can simulate survival times from (5.16). By organizing the survival times and associated explanatory variable values in a data frame in R, and then running the Cox regression analysis methods from chapter 4 on the data frame produce the following for the multivariable case:

Explanatory variable	$\hat{\beta}$	$e^{\hat{\beta}}$	Confidence interval for e^{β}	Confidence interval for β	p-value
X_1	0.0838	1.09	[0.929, 1.27]	[-0.0736, 0.239]	0.297
X_2	-0.0841	0.919	[0.892, 0.948]	[-0.114,-0.0534]	$4.45 \cdot 10^{-8}$
X_3	-0.278	0.757	[0.738, 0.777]	[-0.304, -0.252]	$2 \cdot 10^{-16}$

Table 5.1: Regression coefficient table from multivariable analysis. Time to death.

As seen from table 5.1, the regression coefficients are not estimated to be the exact value that was assumed in the simulations. Nonetheless, the assumed regression coefficient values of $\beta_1 = 0.0110$, $\beta_2 = -0.0610$ and $\beta_3 = -0.270$ lies inside the associated 95 percent confidence intervals of [-0.0736, 0.239], [-0.114,-0.0534] and [-0.304, -0.252], respectively.
It is of interest to check what can be done to increase the accuracy of the regression coefficient estimates. It was found that the accuracy of the estimates was increased by changing the value of some of the parametric variables contained in (5.16). The associated findings will now be presented.

5.3.1.1 Changing the standard deviation

Let us now try to estimate the regression coefficients again, but now by changing the standard deviation of X_1 from 0.5 to 5, and then observe if the estimates becomes more or less accurate.

Explanatory variable	$\hat{\beta}$	$e^{\hat{\beta}}$	Confidence interval for e^{β}	Confidence interval for β	p-value
X_1	0.0181	1.02	[1.00, 1.03]	[0, 0.0296]	0.0249
X_2	-0.0846	0.919	[0.892, 0.947]	[-0.114,-0.0545]	$3.62 \cdot 10^{-8}$
X_3	-0.279	0.757	[0.737, 0.776]	[-0.305, -0.254]	$2 \cdot 10^{-16}$

Table 5.2: Regression coefficient table from multivariable analysis. Time to death.

If comparing table 5.1 and 5.2, it can be seen that, when increasing the standard deviation of X_1 from 0.5 to 5, the accuracy of the estimates of the regression coefficients of X_1 and X_2 increases, but that of X_3 decreases (changes from 0.278 to 0.279, so the decrease in accuracy is not that large). Checking for standard deviation equal to 10:

Explanatory variable	$\hat{\beta}$	$e^{\hat{\beta}}$	Confidence interval for e^{β}	Confidence interval for β	p-value
X_1	0.0145	1.01	[1.01, 1.02]	[0, 0.0198]	0.000348
X_2	-0.0841	0.919	[0.892, 0.947]	[-0.0823,-0.0704]	$4.39 \cdot 10^{-8}$
X_3	-0.279	0.757	[0.737, 0.777]	[-0.677, -0.587]	$2 \cdot 10^{-16}$

Table 5.3: Regression coefficient table from multivariable analysis. Time to death.

Table 5.3 shows that increasing the standard deviation of X_1 from 5 to 10 increases the accuracy of the regression coefficient estimates of X_1 and X_2, but no increase for X_3. In the following simulation examples, the standard deviation of X_1 will be assumed to be 10. All other parameters will also be assumed to stay constant if else is not mentioned.

5.3.1.2 Changing the number of simulated values

In table 5.3, 700 survival times were simulated. What will happen to the regression coefficient estimates if the number of simulations is increased? Simulations were done for 700, 7000, 70 000 and 700 000 simulations, and the regression coefficient estimates were reported in table 5.4 below.

Number of simulations	700	7000	70 000	700 000
$\hat{\beta}_1$	0.01447	0.01118	0.01103	0.01092
$\hat{\beta}_2$	-0.08413	-0.06080	-0.06017	-0.06141
$\hat{\beta}_3$	-0.2787	-0.2725	-0.2704	-0.2697

Table 5.4: Regression coefficient table from multivariable analysis. Time to death.

From table 5.4 it can be seen that, when increasing the number of simulations from 700 to 7000, the accuracy of the regression coefficient estimates increases for all of the variables. When increasing the number of simulations from 7000 to 70 000, the accuracy of the regression coefficient estimates increases for X_1 and X_3, but not for X_2. When increasing the number of simulations from 70 000 to 700 000, the accuracy of the regression coefficient estimates increases for X_2 and X_3, but not for X_1. Summarized, increasing the number of simulations will increase the overall accuracy of the regression coefficient estimates.

5.3.1.3 Trying to detect a non-linear term

Now, it will be investigated if it will be possible to detect from the martingale residuals plot if the assumed functional form of one of the explanatory variables is not linear. Martingale residuals plot for non-empty model will be used (see section 3.4.2).

The following form of (5.12) is used in the simulation:

$$\lambda_i^* = \lambda \cdot e^{\beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \beta_4 \cdot (x_3 - \overline{x_3})^2}, \tag{5.17}$$

where $\overline{x_3}$ is the mean of x_3. Parameter values are as before. The only new parameter value is $\beta_4 = 0.1$.
When using the R-script to create the Martingale residuals plot, only the term $\beta_3 \cdot x_3$ is assumed for the X_3 variable. Then the resulting plot looks like the following:

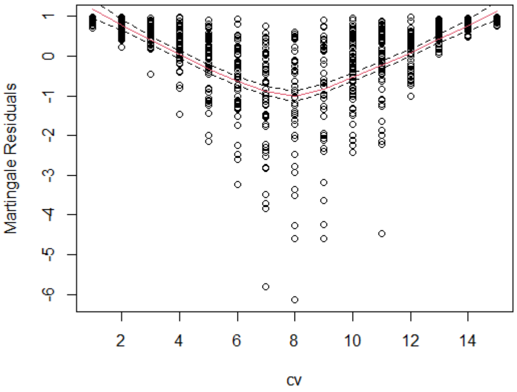


Figure 5.1: Martingale plot of the X_3 variable when only assuming linear term.

The curve in figure 5.1 resembles a parabola. If we now in the model add a quadric term in the model assumption (accomplished in R by adding the term "$I(x_3{}^2)$"), the plot changes to the following:

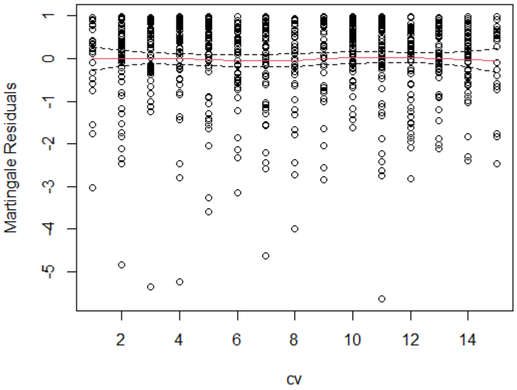


Figure 5.2: Martingale plot of the X_3 variable when including a quadric term in the model assumption.

As seen from figure 5.2, the LOESS smoother is now approximately straight and horizontal. It is fully possible to fit a straight horizontal line inside the confidence interval boundaries, which indicate that the correct functional form is assumed.

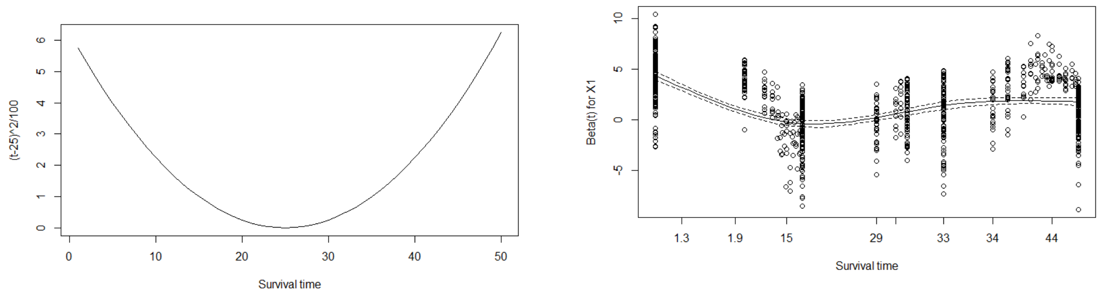
5.4 Survival data for which the proportional hazards assumtion is not satisfied

The library ”coxed” in the programming software R has an in-built function with the name ”sim.survdata”. This function can be used to simulate survival data. In the following, this function will be used to simulate survival data. Then, these data will be used to illustrate how good the Schoenfeld residuals are to detect non-proportional hazards.

What was done was to copy the example shown in senction 3.9 of the article ”How to simulate survival data with the sim.survdata function” [13, p. 50]. The only difference is that we defined β_1 to be

$$\beta_1 = \frac{(t-25)^2}{100}, \tag{5.18}$$

and we set T equal to 50 instead of 100 (have a look in the documentation of the sim.survdata-function for an understanding of the terminology). The Schoenfeld residuals plot corresponding to the simulated data is shown in 5.3b below. Figure 5.3a is a plot of (5.18).



(a) Plot of (5.18).

(b) Schoenfeld residuals plot of β_1 defined in 5.18.

Figure 5.3: Comparison of theoretical and observed functional form of regression coefficient from simulated survival data.

As seen from figure 5.3, the Schoenfeld residuals plot in 5.3b resembles a parabola, which is the shape of the theoretical form of the regression coefficient function shown in figure 5.3a. Thus, the Schoenfeld residuals plot seems to do its job in being an aid to identify the functional form of the regression coefficient.

Chapter 6

Conclusion

In this chapter, conclusions for the results from chapter 4 and 5 will be given. In addition, a suggestion to further work will be given.

6.1 Chapter 4: Application

6.1.1 Time to death

The Kaplan-Meier curves for time to death indicated that size, grade, nodes, estrogen and progesterone were the explanatory variables that had an influence on the survival experience of the patients. The Kaplan-Meier curves indicated that

1) The larger the size of the tumor, the bigger risk of death exists.

2) The risk of death increaes with increasing grade.

3) Increasing number of positive lymph nodes increases the risk of death in the patient.

4) The lower the amount of estrogen and progesterone bound to proteins in the cytosol of the primary tumor, the bigger the risk of death.

For the univariable Cox regression analysis, all except of the estrogen and progesterone variables were found to meet the Schoenfeld residuals plot requirement (fit a straight and horizontal line inside the confidence interval boundaries). Because of no critical deviance (no monotonically decreasing or increasing trend), multivariable analysis procedure was initiated without introducing methods for handling non-proportional hazards.

After performing Cox regression using the method of purposeful selection for the multivariable case, size, grade, nodes and progesterone were the variables remaining in the model. There were observed some deviance from the Schoenfeld and Martingale residuals plots for the progesterone variable, but the deviances were considered to not be critical. For the nodes variable the requirement for the Martingale residuals plot were not met for large nodes values. A possible fix of this could either be to find a suitable transformation or to categorize the variable.

6.1.2 Time to recurrence

For time to recurrence, the Kaplan-Meier curves indicated that all the variables except the menopause variable had an effect on the time to recurrence. More specifically,

1) There is a small indication of that younger patients have higher risk of recurrence.

2) Patients receiving tamoxifen have a higer risk of recurrence.

3) The larger the primary tumor, the higher the risk of recurrence.

4) Increasing grade increases the chance of recurrence.

5) The more positive lymph nodes, the higher the risk of recurrence.

6) The risk of recurrence increases with increasing amount of estrogen and progesterone bound to proteins in the cytosol of the primary tumor.

On the univariable level, monotonic increasing behavior of the age, menopause and estrogen variables were observed in the Schoenfeld residuals plots, indicating that these variables may have time-dependent regression coefficients. A decision was made to continue with the multivariable analysis without doing anything about these variable's possible deviance from proportional hazards. Also, Martingale residuals plot requirement was not met for the age and nodes variables on the univariable level. Before continuing on multivariable level, these variables were categorized to make sure these variables met the Maringale residuals plot requirement.

After performing the procedure for the multivariable case, hormone, grade, categorical nodes variable and the progesterone variable were the variables left in the model. All of them satisfied the Schoenfeld residuals plot requirement. It was found indication of that hormone may be interacted with the progesterone variable and also that the grade variable may be interacted with the categorical nodes variable.

6.1.3 Recurrence as time-dependent variable

Recurrence was modelled as a time-dependent variable for time to death. For the multivariable case, recurrence, size and progesterone turned out to be the important variables, with recurrence having a very large hazard rate. Thus, it was concluded that patients experiencing recurrence had a much higher risk of dying compared to patients not experiencing recurrence. It is important to keep in mind that, those variables important to recurrence are also important to time to death, but through the recurrence variable. For example are nodes and grade important for time to recurrence and therefore also important for time to death, even if they didn't survive the purposeful selection method for time to death when including recurrence as time-dependent variable.

Both the uni - and multivariable analysis showed that the Schoenfeld residuals plot had a monotonically decreasing linear behavior for the recurrence variable. The multivariable analysis was then runned again, now by modelling the recurrence variable with a time-dependent regression coefficient. The Martingale residuals plot requirement was not met for the size and progesterone variables.

6.2 Chapter 5: Simulations

6.2.1 Simulation of Weibull distributed survival times

When running Cox regression on the simulated Weibull distributed survival times, the programming script managed to estimate the regression coefficients inside its confidence interval boundaries. It was found that increasing the standard deviation of the normally distributed explanatory variable increased the accuracy of the estimated regression coefficient values. Increasing the number of simulated survival times also increased the accuracy of the estimates. When letting the value of one of the explanatory variables be expressed as a non-linear function (in our case a parabola) but assuming a linear term in the model, the Martingale residuals plot managed to detect this. After assuming the correct functional form in the model, the Martingale residuals plot requirement were met.

6.2.2 Survival data for which the proportional hazards assumtion is not satisfied

When using the *sim.survdata*-function from the *coxed*-library to simulate survival data for which the proportional hazards assumtion were not satisfied, the Schoenfeld residuals plot showed that one of the three explanatory variables had a time-dependent regression coefficient, resembling a parabola, in consistence with the quadric function that was chosen as input in the simulation script.

6.3 Further work

After working on this report, some possible extensions have been identified. They are:
1) Investigate more closely the cases were time-dependent regression coefficients may be present. For example have a look at the age, menopause and estrogen variables for time to recurrence (non-proportional hazards observed for these variables on the univariable level).
2) Have a closer look on the interactions that were found and see how they will affect the models. Examples for time to recurrence are the interaction between hormone and progesterone and between grade and categorical nodes variable.

3) Investigate if there exist some transformations of the explanatory variables from chapter 4 that would make the Martingale residuals for some of the variables fit better, or alternatively categorize them in suitable categories. Examples are the nodes variable for time to death, and the size and progesterone variables for time to death when including recurrence as time-dependent variable.

4) Further experiment with the sim.survdata function to investigate the goodness of the Schoenfeld residuals plot and its ability to detect time-dependent regression coefficients and their nature.

5) Investigate more closely the topic of non-proportional hazards. Information about this topic can be found in the literature [1, p. 313-318].

6) Have a look at multi-state models. In his book, Collett gives a short presentation of this topic [1, p. 323-326]. In a diagram, he presents a three-state model for analysing survival data from a patient group similar to that we have been working on in chapter 4 (patients with cancer, recurrence and survival time recorded) [1, p. 324]. In figure 6.1 below you see a reconstruction of this diagram.

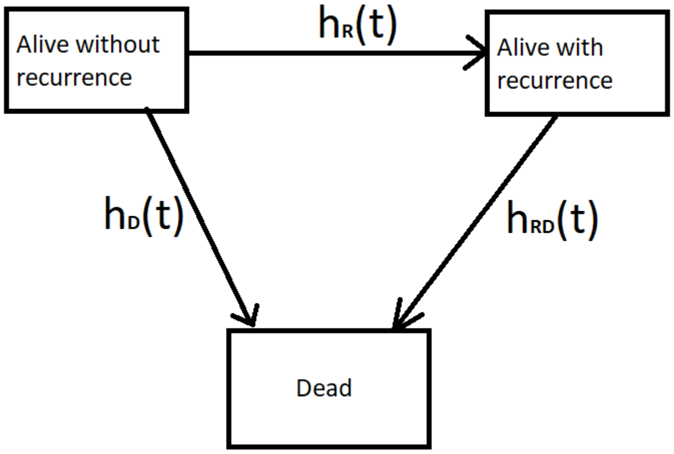


Figure 6.1: Diagram of three-state model that can be used in cancer studies.

In figure 6.1 you see three rectangles. These presents the state of the patient. Depending on whether the patient experiences recurrence or not, we can work with three different hazard functions:

1) $h_D(t)$. This is the hazard function the patient depends on if you look at time to death and ignore whether the patient has experienced recurrence or not.

2) $h_R(t)$. This is the hazard function the patient depends on if you look at time to recurrence.

3) $h_{RD}(t)$. This is the hazard function the patient depends on if you look at time to death, given that the patient has experienced recurrence.

In this report we have modelled case 1 ($h_D(t)$) and 2 ($h_R(t)$) above. The closest we have been to 3 ($h_{RD}(t)$) is when we modelled recurrence as a time-dependent variable. The main challenge with case 3 is that only patients who's recurrence time is known can be included in the study. But the reality is that people get recurrence at different times. Someone get recurrence after the study is finished, while others get recurrence after they have been censored. Such patients cannot be included in case 3, because their recurrence time will remain unknown. To investigate case 3, Collett suggests using a technique called *conditional logistic regression* [1, p. 325]. A possible progression of this report work could for example be to use conditional logistic regression to model case 3 above.

Bibliography

[1] D. Collett, *Modelling Survival Data in Medical Research, Second Edition.* Chapman and Hall/CRC, 2003.

[2] G. G. Løvås, *Statistikk for universitetet og høgskoler, 2nd Edition.* Universitetsforlaget, 2004.

[3] J. Frank E. Harrell, *Regression Modelling Strategies, 2nd Edition.* Springer, 2015.

[4] T. M. Therneau and P. M. Grambsch, *Statistics for Biology and Health, Modeling Survival Data, Extending th Cox Model.* Springer, 2000.

[5] T. Therneau, C. Crowson, and E. Atkinson, "Using time dependent covariates and time dependent coefficients in the cox model, url: https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf."

[6] D. W. H. Jr., S. Lemeshow, and S. May, *Applied Survival Analysis, Regression Modelling of Time-to-Event Data, Second Edition.* Wiley, 2008.

[7] C. Schmoor, W. Sauerbrei, G. Bastert, and M. S. for the German Breast Cancer Study Group, *Role of Isolated Locoregional Recurrence of Breast Cancer: Results of Four Prospective Studies.* American Society of Clinical Oncology, 2000.

[8] https://kreftlex.no/Brystkreft/ProsedyreFolder/BEHANDLING/Cellegift/med-bryst–Tamoxifen?lg=ks&CancerType=Bryst&containsFaq=False. accessed 04.06.2020.

[9] https://www.breastcancer.org/symptoms/diagnosis/cell_grade?gclid=EAIaIQobChMIwcTv9fj05wIVVamaCh0VSgyhEAAYASAAEgKHbfD_BwE. accessed 28.02.2020.

[10] https://www.breastcancer.org/symptoms/diagnosis/lymph_nodes. accessed 28.02.2020.

[11] L. Metcalf and W. Casey, *Cybersecurity and Applied Mathematics.* Syngress, 2016.

[12] M. L. Rizzo, *Statistical Computing with R.* Taylor & Francis Group, LLC, 2008.

[13] J. Kropko and J. J. Harden, “How to simulate survival data with the sim.survdata function, https://cran.r-project.org/web/packages/coxed/vignettes/simulating_survival_data.html# time-varying-coefficients, accessed 04.06.2020.”

Appendix A

A.1 Programming script from the R software

```
rm(list=ls()) # Clearing the variables stored in the Global Environment



#### Libraries

# Below follows a list of the packages I have been using in my R scripts.
     Remember to first install these packages before calling them by the
    library function.

install.packages(”condSURV”)
library(condSURV) # Contains the German Breast Cancer study data
    frame

install.packages(”survival”)
library(survival) # Contains the coxph and cox.zph functions needed to
    perform Cox regression

install.packages(”coxed”)
library(coxed) # Includes the sim.survdata function. Used to simulate
    survival times for which the assumption of proportional hazards is
    not satisfied .



#### Data frame

# The German Breast Cancer study data frame’s is loaded as shown below
    .

data(”gbcsCS”) # Loading the data frame of the German Breast Cancer
    study
```

```
#### Drawing histograms and pie charts

# In the report histograms and pie charts were drawn.
# Below it is shown how they are drawn using R

agevector<-gbcsCS[,"age"] # Extracting the age column in the gbcsCS
    data frame

# Histogram of the age variable:
hist(agevector,breaks=c(20,25,30,35,40,45,50,55,60,65,70,75,80,85), freq=
    TRUE,
col="red", main="Age histogram", xlab="Age",ylab="Number of
    patients",border="blue")

# Figure showing the histograms and pie charts of the eight explanatory
    variables in the same figure:

par(mfrow=c(4,2)) # Making a figure consisting of eight subfigures,
    ordered with four rows and two columns.


# age

agevector<-gbcsCS[,"age"]
hist(agevector, prob = TRUE,nclass=sqrt(length(agevector)),
     main="Histogram of the age variable")


# size

sizevector<-gbcsCS[,"size"]
hist(sizevector, prob = TRUE,nclass=sqrt(length(sizevector)),
     main="Histogram of the size variable")


# menopause
```

```
menopausevector<-gbcsCS[,"menopause"]
meno1<-sum(menopausevector==1)
meno2<-sum(menopausevector==2)
# pie(c(meno1,meno2),labels=c("Menopause=1","Menopause=2"),edges =
      1000, radius=1, clockwise=FALSE, init.angle=0, density=NULL,
     angle=45, col=c(1,2), border=NULL, main="Pie chart of the
     menopause variable")
 slices  <- c(meno1,meno2)
 lbls  <- c("Menopause=1","Menopause=2")
pct <- round(slices/sum(slices)*100)
 lbls  <- paste(lbls, pct) # add percents to labels
 lbls  <- paste(lbls,"%",sep="") # ad % to labels
pie( slices ,labels = lbls, col=rainbow(length(lbls)),
     main="Pie Chart of the menopause variable")



# hormone

hormonevector<-gbcsCS[,"hormone"]
hormone1<-sum(hormonevector==1)
hormone2<-sum(hormonevector==2)
 slices  <- c(hormone1,hormone2)
 lbls  <- c("Hormone=1","Hormone=2")
pct <- round(slices/sum(slices)*100)
 lbls  <- paste(lbls, pct) # add percents to labels
 lbls  <- paste(lbls,"%",sep="") # ad % to labels
pie( slices ,labels = lbls, col=rainbow(length(lbls)),
     main="Pie Chart of the hormone variable")


# grade

gradevector<-gbcsCS[,"grade"]
grade1<-sum(gradevector==1)
grade2<-sum(gradevector==2)
grade3<-sum(gradevector==3)
```

```
slices  <− c(grade1,grade2,grade3)
lbls  <− c(”Grade=1”,”Grade=2”,”Grade=3”)
pct <− round(slices/sum(slices)∗100)
lbls  <− paste(lbls, pct) # add percents to labels
lbls  <− paste(lbls,”%”,sep=””) # ad % to labels
pie( slices ,labels = lbls,  col=rainbow(length(lbls)),
    main=”Pie Chart of the grade variable”)

# nodes

nodesvector<−gbcsCS[,”nodes”]
hist(nodesvector, prob = TRUE,nclass=sqrt(length(nodesvector)),
     main=”Histogram of the nodes variable”)

# prog_recp

progvector<−gbcsCS[,”prog_recp”]
hist(progvector, prob = TRUE,nclass=sqrt(length(progvector)),
     main=”Histogram of the prog_recp variable”)

# estrg_recp

estrgvector<−gbcsCS[,”estrg_recp”]
hist(estrgvector , prob = TRUE,nclass=sqrt(length(estrgvector)),
     main=”Histogram of the estrg_recp”)


par(mfrow=c(1,1))


#### Kaplan−Meier curves without categories


# Kaplan−Meier curves for time to death.
Kaplanmeier <− survfit(Surv(survtime,censdead) ˜ 1, data = gbcsCS,
    conf.type = ”plain”)
plot(Kaplanmeier, xlab=”Survival time (days)”,ylab=”Estimated survivor
    function value”, main=”Plot of Kaplan−Meier curve”,col=c(”red”,”
    blue”,”blue”))

# Kaplan−Meier curves for time to recurrence.
Kaplanmeier <− survfit(Surv(rectime,censrec) ˜ 1, data = gbcsCS, conf.
    type = ”plain”)
plot(Kaplanmeier, xlab=”Recurrence time (days)”,ylab=”Estimated
    survivor function value”, main=”Plot of Kaplan−Meier curve”,col=c(”
    red”,”blue”,”blue”))


#### Kaplan−Meier curves with categories, with associated log−rank
    test. Only shown for time to death. For time to recurrence, change
    survfit(Surv(survtime,censdead) in the code to survfit(Surv(rectime,
    censrec).


## Age

# Categorizing the age variable
agevector<−gbcsCS[, ”age”]
agecategory<−cut(agevector,c(0,45,60,Inf))
mygbcsCS<−data.frame(gbcsCS,agecategory)

# Kaplan−Meier curve with confidence interval
Kaplanmeierage <− survfit(Surv(survtime,censdead) ˜ agecategory, data
    = mygbcsCS, conf.type = ”plain”)
plot(Kaplanmeierage, xlab=”Survival time (days)”,ylab=”Estimated
    survivor function value”, main=”Plot of Kaplan−Meier curve, age”,
    col=c(”red”,”blue”,”green”))
legend(”bottomleft”,
       c(”0−45”,”46−60”,”61−80”),
        fill =c(”red”,”blue”,”green”))

# Log−rank test
survdiff (formula=Surv(survtime,censdead)˜ agecategory, data=
    mygbcsCS)
```

```
## Menopause

# Kaplan−Meier curve with confidence interval
Kaplanmeiermenopause <− survfit(Surv(survtime,censdead) ˜ menopause,
      data = mygbcsCS, conf.type = ”plain”)
plot(Kaplanmeiermenopause, xlab=”Survival time (days)”,ylab=”
     Estimated survivor function value”, main=”Plot of Kaplan−Meier
     curve, menopause”,col=c(”red”,”blue”))
legend(”bottomleft”,
         c(”Reached menopause”, ”Not reached menopause”),
          fill =c(”red”,”blue”))

# Log−rank test
survdiff (formula=Surv(survtime,censdead)˜ menopause, data=
     mygbcsCS)


## Hormone

# Kaplan−Meier curve with confidence interval
Kaplanmeierhormone <− survfit(Surv(survtime,censdead) ˜ hormone,
     data = mygbcsCS, conf.type = ”plain”)
plot(Kaplanmeierhormone, xlab=”Survival time (days)”,ylab=”Estimated
     survivor function value”, main=”Plot of Kaplan−Meier curve, hormone
     ”,col=c(”red”,”blue”))
legend(”bottomleft”,
         c(”Receive tamoxifen”, ”Doesn’t receive tamoxifen”),
          fill =c(”red”,”blue”))

# Log−rank test
survdiff (formula=Surv(survtime,censdead)˜ hormone, data=mygbcsCS)


## Grade

# Kaplan−Meier curve with confidence interval
Kaplanmeiergrade <− survfit(Surv(survtime,censdead) ˜ grade, data =
     mygbcsCS, conf.type = ”plain”)
```

```
plot(Kaplanmeiergrade, xlab=”Survival time (days)”,ylab=”Estimated
    survivor function value”, main=”Plot of Kaplan−Meier curve, grade”,
    col=c(”red”,”blue”,”green”))
legend(”bottomleft”,
       c(”Grade 1”, ”Grade 2”, ”Grade 3”),
        fill =c(”red”,”blue”, ”green”))

# Log−rank test
survdiff (formula=Surv(survtime,censdead)˜ grade, data=mygbcsCS)


## Size

# Categorizing the size  variable
sizevector<−gbcsCS[,”size”]
sizecategory<−cut(sizevector,c(0,22,30,Inf))
mygbcsCS<−data.frame(gbcsCS,sizecategory)

# Kaplan−Meier curve with confidence interval
Kaplanmeiersize <− survfit(Surv(survtime,censdead) ˜ sizecategory, data
     = mygbcsCS, conf.type = ”plain”)
plot(Kaplanmeiersize, xlab=”Survival time (days)”,ylab=”Estimated
    survivor function value”, main=”Plot of Kaplan−Meier curve, size”,
    col=c(”red”,”blue”,”green”))
legend(”bottomleft”,
       c(”0−22”, ”23−30”, ”31−120”),
        fill =c(”red”,”blue”, ”green”))

# Log−rank test
survdiff (formula=Surv(survtime,censdead)˜ sizecategory, data=
    mygbcsCS)


## Nodes

# Categorizing the nodes variable
nodesvector<−gbcsCS[,”nodes”]
nodescategory<−cut(nodevector,c(0,1,3,6,Inf))
```

```
mygbcsCS<−data.frame(gbcsCS,nodescategory)

# Kaplan−Meier curve with confidence interval
Kaplanmeiernode <− survfit(Surv(survtime,censdead) ˜ nodescategory,
    data = mygbcsCS, conf.type = ”plain”)
plot(Kaplanmeiernode, xlab=”Survival time (days)”,ylab=”Estimated
    survivor function value”, main=”Plot of Kaplan−Meier curve, nodes”,
    col=c(”red”,”blue”,”green”,”yellow”))
legend(”bottomleft”,
        c(”1”, ”2−3”, ”4−6”,”7−51”),
         fill =c(”red”,”blue”, ”green”,”yellow”))

# Log−rank test
survdiff (formula=Surv(survtime,censdead)˜ nodescategory, data=
    mygbcsCS)


## Prog_recp

# Categorizing the progesterone variable
progesteronevector<−gbcsCS[,”prog_recp”]
progesteronecategory<−cut(progesteronevector,c(0,12,84,Inf))
mygbcsCS<−data.frame(gbcsCS,progesteronecategory)

# Kaplan−Meier curve with confidence interval
Kaplanmeierprogesterone <− survfit(Surv(survtime,censdead) ˜
    progesteronecategory, data = mygbcsCS, conf.type = ”plain”)
plot(Kaplanmeierprogesterone, xlab=”Survival time (days)”,ylab=”
    Estimated survivor function value”, main=”Plot of Kaplan−Meier
    curve, progesterone”, col=c(”red”,”blue”,”green”,”yellow”))
legend(”bottomleft”,
        c(”0−12”, ”13−84”, ”85−2380”),
         fill =c(”red”,”blue”, ”green”))

# Log−rank test
survdiff (formula=Surv(survtime,censdead)˜ progesteronecategory, data=
    mygbcsCS)
```

```
## Estrg_recp

# Categorizing the estrogen variable
estrogenvector<-gbcsCS[,"estrg_recp"]
estrogencategory<-cut(estrogenvector,c(0,13,79,Inf))
mygbcsCS<-data.frame(gbcsCS,estrogencategory)

# Kaplan-Meier curve with confidence interval
Kaplanmeierestrogen <- survfit(Surv(survtime,censdead) ~
     estrogencategory, data = mygbcsCS, conf.type = "plain")
plot(Kaplanmeierestrogen, xlab="Survival time (days)",ylab="Estimated
     survivor function value", main="Plot of Kaplan-Meier curve, estrogen
     ", col=c("red","blue","green"))
legend("bottomleft",
          c("0-13", "14-79", "80-1144"),
           fill =c("red","blue", "green"))

# Log-rank test
 survdiff (formula=Surv(survtime,censdead)~ estrogencategory, data=
     mygbcsCS)


#### Using Cox regression to determine regression coeffisients. Only for
     time to death will be shown. To estimate regression  coefficients  for
     time to recurrence,  replace coxph(Surv(survtime,censdead) in the
     script with coxph(Surv(rectime,censrec).

mygbcsCS<-gbcsCS[-684,] # Deleting from the data set the patient that
      seem to make problems for the residuals plots.


## Univariable


# age variable
coxmod1age <- coxph(Surv(survtime,censdead) ~ age, data = mygbcsCS
     )
coxmod1age
```

```
summary(coxmod1age)

# menopause variable
coxmod1menopause <- coxph(Surv(survtime,censdead) ~ menopause,
    data = mygbcsCS)
coxmod1menopause
summary(coxmod1age)

# hormone variable
coxmod1hormone <- coxph(Surv(survtime,censdead) ~ hormone, data =
    mygbcsCS)
coxmod1hormone
summary(coxmod1hormone)

# size  variable
coxmod1size <- coxph(Surv(survtime,censdead) ~ size, data =
    mygbcsCS)
coxmod1size
summary(coxmod1size)

# grade variable
coxmod1grade <- coxph(Surv(survtime,censdead) ~ as.factor(grade),
    data = mygbcsCS)
coxmod1grade
summary(coxmod1grade)

# nodes variable
coxmod1nodes <- coxph(Surv(survtime,censdead) ~ nodes, data =
    mygbcsCS)
coxmod1nodes
summary(coxmod1nodes)

# prog_recp variable
coxmod1progesterone <- coxph(Surv(survtime,censdead) ~ log(prog_recp
    +1), data = mygbcsCS)
coxmod1progesterone
summary(coxmod1progesterone)
```

```
# estrg_recp variable
coxmod1estrogen <- coxph(Surv(survtime,censdead) ~ log(estrg_recp+1),
      data = mygbcsCS)
coxmod1estrogen
summary(coxmod1estrogen)


## Multivariable

# Showing one example of a multivariable model consisting of the
     variables  size ,  grade,  nodes and progesterone.
coxmod1multi <- coxph(Surv(survtime,censdead) ~ size + as.factor(
     grade) + nodes + log(prog_recp+1) , data = mygbcsCS)
coxmod1multi
summary(coxmod1multi)



#### Residuals


### Univariable


## Schoenfeld residuals

# Age
resmod1age <- cox.zph(coxmod1age)
resmod1age
plot(resmod1age)

# Menopause
resmod1menopause <- cox.zph(coxmod1menopause)
resmod1menopause
plot(resmod1menopause)


# Hormone
```

```
resmod1hormone <− cox.zph(coxmod1hormone)
resmod1age
plot(resmod1hormone)


# Grade
resmod1grade <− cox.zph(coxmod1grade , transform=”km”, terms=
    FALSE, singledf=FALSE, global=TRUE)
resmod1grade
plot(resmod1grade)


# Size
resmod1size <− cox.zph(coxmod1size)
resmod1size
plot(resmod1size)


# Nodes
resmod1nodes <− cox.zph(coxmod1nodes)
resmod1nodes
plot(resmod1nodes)


# Prog_recp
resmod1progesterone <− cox.zph(coxmod1progesterone)
resmod1progesterone
plot(resmod1progesterone)


# Estrg_recp
resmod1estrogen <− cox.zph(coxmod1estrogen)
resmod1estrogen
plot(resmod1estrogen)



## Martingale residuals
## Shown only for one of the variables. The method is similar for the
    other variables



# Prog_recp

progesteronevector<-gbcsCS[,"prog_recp"]
progesteronetransformation<-log(progesteronevector+1) # Make the
    transformation you wish for
mygbcsCS<-data.frame(gbcsCS,progesteronetransformation)
mygbcsCS<-mygbcsCS[-684,] # Deleting the observation that seem to
    make problems for the residuals plots
fit = coxph(Surv(survtime, censdead) ~ 1, data=mygbcsCS)
martingaleresiduals = residuals(fit, type="martingale")



# Ordne residualene etter stigende kovariatverdi
martingaleresidualsv2 <- martingaleresiduals[order(mygbcsCS$
    progesteronetransformation)]
progesteronetransformationv2 <- mygbcsCS$progesteronetransformation[
    order(mygbcsCS$progesteronetransformation)]

plot(progesteronetransformationv2, martingaleresidualsv2, xlab="log(prog
    _recp+1)", ylab="Martingale Residuals")

# Glattet kurve med 95% konfidensintervall
plx<-predict(loess(martingaleresidualsv2 ~
    progesteronetransformationv2), se=T)
lines(progesteronetransformationv2,plx$fit,col=2)
lines(progesteronetransformationv2,plx$fit - qt(0.975,plx$df)*plx$se, lty
    =2)
lines(progesteronetransformationv2,plx$fit + qt(0.975,plx$df)*plx$se, lty
    =2)
```

```
### Multivariable


## Schoenfeld residuals. Showing for model with size, grade, nodes and
    progesterone as variables.
coxmod1multiv2 <− coxph(Surv(survtime,censdead) ˜ size + as.factor(
    grade) + nodes + log(prog_recp+1) , data = mygbcsCS)
resmod1multiv2 <− cox.zph(coxmod1age, transform=”km”, terms=
    FALSE, singledf=FALSE, global=TRUE)
resmod1multiv2
plot(resmod1multiv2 )


## Martingale residuals. Shown only for one of the variables. The
    method is similar for the other variables.

#Prog_recp

 fit  = coxph(Surv(survtime, censdead) ˜ size + as.factor(grade) + nodes
    + log(prog_recp+1), data=mygbcsCS)
martingaleresiduals  = residuals(fit, type=”martingale”)

# Ordering the residuals in increasing variable value
martingaleresidualsv2 <− martingaleresiduals[order(mygbcsCS$
    progesteronetransformation)]
progesteronetransformationv2 <− mygbcsCS$progesteronetransformation[
    order(mygbcsCS$progesteronetransformation)]

plot(progesteronetransformationv2, martingaleresidualsv2, xlab=”log(prog
    _recp+1)”, ylab=”Martingale Residuals”)

# Smooth curves with 95% confidence interval
plx<−predict(loess(martingaleresidualsv2 ˜
    progesteronetransformationv2), se=T)
lines(progesteronetransformationv2,plx$fit,col=2)
lines(progesteronetransformationv2,plx$fit − qt(0.975,plx$df)∗plx$se, lty
    =2)
```

```
lines(progesteronetransformationv2,plx$fit + qt(0.975,plx$df)∗plx$se, lty
    =2)


#### Categorizing the age and nodes variable for time to recurrence


### Categorizing the age variable for time to recurrence

agevector<−gbcsCS[, ”age”]
agecategorical<− rep(NA, length(agevector))
for (i in 1:length(agevector)) {
    if (agevector[i] <= 45)
    {agecategorical[i] = 1} else if (agevector[i] >= 46 & agevector[i]
        <= 60)
    {agecategorical[i] = 2} else
    {agecategorical[i] = 3}
}

### Categorizing the nodes variable for time to recurrence

nodesvector<−gbcsCS[, ”nodes”]
nodescategorical<− rep(NA, length(nodesvector))
for (i in 1:length(nodesvector)) {
    if (nodesvector[i] <= 3)
    {nodescategorical[i] = 1} else if (nodesvector[i] >= 4 &
        nodesvector[i] <= 6)
    {nodescategorical[i] = 2} else
    {nodescategorical[i] = 3}
}


### Calculating regression coefficients for time to recurrence in
    multivariable model. For time to recurrence, the hormone, grade,
    nodes and progesterone variables turned out to be significant .

mygbcsCS<−data.frame(gbcsCS,nodescategorical)
```

```
coxmod1 <- coxph(Surv(rectime,censrec) ~ hormone + as.factor(grade)
    + as.factor(nodescategorical) + log(prog_recp+1) , data =
    mygbcsCS)
coxmod1
summary(coxmod1)



#### Recurrence as time-dependent variable

# Se https://cran.r-project.org/web/packages/survival/vignettes/
    timedep.pdf

# Creating a new data frame on the "start-stop" form and adding
    recurrence as time-dependent variable.
gbcsTD <- tmerge(gbcsCS[,c(1,5:12)], gbcsCS[,c(1,13:14,15:16)],
                  id=id, censdead=event(survtime, censdead),
                  rec=tdc(rectime))
head(gbcsTD,20)

mygbcsTD<-gbcsTD[-684,] # Deleting from the data set the patient
    that seem to make problems for the residuals plot


# Model with only recurrence
coxmoduni <- coxph(Surv(tstart,tstop,censdead) ~ rec , data =
    mygbcsTD)
summary(coxmoduni)

# Model with recurrence and other variables
coxmodmultiv3 <- coxph(Surv(tstart,tstop,censdead) ~ rec + size + log(
    prog_recp+1) , data = mygbcsTD)
summary(coxmodmultiv3)
#resmod1rec <- cox.zph(coxmod)


# Associated Schoenfeld residuals plot
```

```
resmod1rec <- cox.zph(coxmodmultiv3, transform="km", terms=
    FALSE, singledf=FALSE, global=TRUE)
plot(resmod1rec)



## Accounting for that recurrence has a time-dependent regression
    coefficient.

# Calculating regression  coefficients
coxmod1rec <- coxph(Surv(tstart,tstop,censdead) ~ rec + size + log(
    prog_recp+1) +tt(rec), data = mygbcsTD,
                  tt  = function(x, t,  ...)  x * t)
summary(coxmod1rec)

# Creating associated Schoenfeld residuals  plots.
resmod1rec <- cox.zph(coxmod1rec)
resmod1rec
plot(resmod1rec)



##### Simulation of Weibull distributed survival times using the inverse
     transform method

Nsim<-700000 # Number of simulations
scalelambda <- 1.5 # Scale parameter of Weibull distribution.
shapegamma <- 0.75 # Shape parameter of Weibull distribution.



## Covariates
set.seed(123) # Setting a particular  seed  to  ensure user  of  the  script
    can reproduce results
cv1<-rnorm(Nsim, mean = 30, sd= 10) # Covariate 1, similar to the
    size variable
```

```
cv2<-rexp(Nsim,rate=0.4) # Covariate 2, similar to the progesterone variable
cv3<-round(runif(Nsim,min=1, max=15)) # Covariate 3, similar to the nodes variable



## Setting the regression  coefficients
betacv1 <- 0.011 # Regression coefficient of covariate 1
betacv2 <- -0.061# Regression coefficient of covariate 2
betacv3 <- -0.27# Regression coefficient of covariate 3
betacv4 <- 0.1


#### Assuming linear effect

lambdastar <- scalelambda*exp(betacv1*cv1+betacv2*cv2+betacv3*cv3)
max(lambdastar)
min(lambdastar)
mean(lambdastar)
sd(lambdastar)

set.seed(123)
survtime <- (-log(1-runif(Nsim))/(lambdastar))^(1/shapegamma)
max(survtime)
min(survtime)
mean(survtime)
sd(survtime)


### Making data frame to use the Cox regression analysis on

id<-1:Nsim # Patient id number
censdead<-rep(1,Nsim) # No censoring for time to death
gbcsCSsimu<- data.frame(id,cv1,cv2,cv3,survtime,censdead)
```

```
### Multivariable Cox regression analysis to check if we can reproduce
    the regression  coefficients  assumed in the simulation

coxmod1 <- coxph(Surv(survtime,censdead) ~ cv1 + cv2 + cv3 , data =
     gbcsCSsimu)
coxmod1
summary(coxmod1)

### Producing Martingale plot associated with the cv3 variable.

cvcategory<-cv3 # Make the transformation you wish for
mygbcsCSsimu<-data.frame(gbcsCSsimu,cvcategory)


 fit  = coxph(Surv(survtime, censdead) ~ cvcategory, data=mygbcsCSsimu
    )
martingaleresiduals  = residuals(fit,  type="martingale")

# Ordering the residuals in  increasing  variable  value
omartingaleresiduals  <- martingaleresiduals[order(mygbcsCSsimu$
    cvcategory)]
ocvcategory <- mygbcsCSsimu$cvcategory[order(mygbcsCSsimu$
    cvcategory)]
plot(ocvcategory, omartingaleresiduals,  xlab="cv", ylab="Martingale
    Residuals")

# Smooth curves with 95% confidence interval
plx<-predict(loess(omartingaleresiduals ~ ocvcategory), se=T)
lines(ocvcategory,plx$fit ,col=2)
lines(ocvcategory,plx$fit  - qt(0.975,plx$df)*plx$se, lty=2)
lines(ocvcategory,plx$fit  + qt(0.975,plx$df)*plx$se, lty=2)



#### Repeating the procedure, but now by assuming non-linear effect:
```

```
lambdastar <- scalelambda*exp(betacv1*cv1+betacv2*cv2+betacv3*
    cv3
                              +betacv4*(cv3-mean(cv3))^2)
survtime <- (-log(1-runif(Nsim))/(lambdastar))^(1/shapegamma)
id<-1:Nsim # Patient id number
censdead<-rep(1,Nsim) # No censoring for time to death
gbcsCSsimu<- data.frame(id,cv1,cv2,cv3,survtime,censdead)
cvcategory<-cv3 # Make the transformation you wish for
mygbcsCSsimu<-data.frame(gbcsCSsimu,cvcategory)


### Cheking if the Martingale plot manages to identify the non-linear
    relation

 fit  = coxph(Surv(survtime, censdead) ~ cvcategory, data=mygbcsCSsimu
    )
martingaleresiduals  = residuals(fit, type="martingale")

# Ordering the residuals in  increasing  variable  value
omartingaleresiduals  <- martingaleresiduals[order(mygbcsCSsimu$
    cvcategory)]
ocvcategory <- mygbcsCSsimu$cvcategory[order(mygbcsCSsimu$
    cvcategory)]
plot(ocvcategory, omartingaleresiduals, xlab="cv", ylab="Martingale
    Residuals")

# Smooth curves with 95% confidence interval
plx<-predict(loess(omartingaleresiduals ~ ocvcategory), se=T)
lines(ocvcategory,plx$fit ,col=2)
lines(ocvcategory,plx$fit  - qt(0.975,plx$df)*plx$se, lty=2)
lines(ocvcategory,plx$fit  + qt(0.975,plx$df)*plx$se, lty=2)


### Checking if the Martingale residuals plot is  satisfied   after
    assuming the correct  functional  form of  the cv3  variable

 fit  = coxph(Surv(survtime, censdead) ~ cvcategory+I(cvcategory^2),
    data=mygbcsCSsimu)
```

```
martingaleresiduals  = residuals(fit,  type=”martingale”)

# Ordering the residuals in  increasing  variable  value
omartingaleresiduals  <− martingaleresiduals[order(mygbcsCSsimu$
     cvcategory)]
ocvcategory <− mygbcsCSsimu$cvcategory[order(mygbcsCSsimu$
     cvcategory)]
plot(ocvcategory, omartingaleresiduals,  xlab=”cv”, ylab=”Martingale
     Residuals”)

# Smooth curves with 95% confidence interval
plx<−predict(loess(omartingaleresiduals ~ ocvcategory), se=T)
lines(ocvcategory,plx$fit ,col=2)
lines(ocvcategory,plx$fit  − qt(0.975,plx$df)∗plx$se, lty=2)
lines(ocvcategory,plx$fit  + qt(0.975,plx$df)∗plx$se, lty=2)



##### Simulation of survival data for which the proportional hazards
     assumption is not satisfied

ttstop<−50
Nsim<−1000
fracnumb<−100

beta.mat <− data.frame(beta1 = (1:ttstop − 25)ˆ2/fracnumb,
                               beta2 = .5,
                               beta3 = −.25)
head(beta.mat)

simdata <− sim.survdata (N=Nsim,T=ttstop,type=”tvbeta”, num.data.
     frames = 1,beta=beta.mat)

head(simdata)

head(simdata$data, 10)

mycensoring<−rep(TRUE,times=Nsim)
```

```
mysimdata<-data.frame(simdata$data,mycensoring)


par(mfrow=c(1,2))

plot(1:ttstop ,(1: ttstop-25)^2/fracnumb,type="l",xlab="Survival time",
    ylab="(t-25)^2/100")
coxmodsimsurvdata <- coxph(Surv(y,failed) ~ X1, data = simdata$data
    ) # You can replace "failed" in coxmodsimsurvdata with "mycensoring
    " defined earlier if you want to remove censoring.
resmodsimsurvdata <- cox.zph(coxmodsimsurvdata)
resmodsimsurvdata
plot(resmodsimsurvdata,xlab="Survival time")
```