MODEL CHECK AND GOODNESS-OF-FIT FOR NESTED CASE-CONTROL STUDIES

by

YING ZHANG

THESIS

for the degree of

MASTER OF SCIENCE

(Master i Modellering og dataanalyse)

Faculty of Mathematics and Natural Sciences
University of Oslo

May 2013

Det matematisk- naturvitenskapelige fakultet
Universitetet i Oslo

Acknowledgements

I would like to acknowledge that the thesis is submitted as part of my Master's degree in Modelling and Data Analysis at the University of Oslo. It is written during the period from Oct 2012 to May 2013 under the direction of Professor Ørnulf Borgan.

First and foremost, I would like to sincerely thank Professor Ørnulf Borgan for his patient guidance, invaluable advice and insightful comments regarding my thesis work, which have offered me the greatest depth of academic experience as I use what I have learned in the past. I feel that he is the best supervisor that I could ever ask for.

In addition, a special thanks is dedicated to my parents back home in Nanjing, who keep supporting me and inspiring me to stay the course. I really appreciate it.

Oslo, May 2013,
Ying Zhang

Contents

Chapter 1

Introduction

Survival data describe the duration of time from entering into the study to the occurrence of an event of interest. These days survival data are found in a wide range of applications in various fields. For instance, we use survival data to study the failure time of a system, the lifetime of cancer patients after surgery, or the time for finding new oil resources, etc. A feature of survival data is that they typically cannot be fully observed, thus they often contain the information of censoring as indication of the missing data. The most common reasons for censoring are that the event of interest has not occurred by the end of the study, or that one is unable to follow up the individuals. Special methods are needed to handle censored survival data. Indeed, by using the Kaplan-Meier estimator and the Nelson-Aalen estimator, we can manage to estimate the survival function and cumulative hazard rate of survival data.

In survival analysis, we are interested in estimating the hazard rate and survival function of individuals. By accessing survival data we are able to investigate the impact of covariates on the survival time of individuals. The effect of categorical covariates can be analyzed by grouping of the individuals according to their values of covariates. But in practice, it is quite common that we deal with multiple covariates which might contain numeric types. Hence there is a demand for regression models. The most widely used regression model in survival analysis is the Cox model, which assumes that the covariates of each individual are related to its hazard rate. The regression coefficients can be estimated by comparing the covariates of the individual that has experienced the event of interest with those who have not. Finally, by using the regression coefficients of the Cox model, we can easily assess the relative risk of each covariate.

Two model assumptions must be satisfied for Cox model. The first one is called log-linearity, that is to say that the hazard ratio must be a linear function of a numeric covariate on the log-scale. The other one is proportional

hazard, which implies that the hazard ratio of two individuals must be a constant and not depend on time. A number of methods have been developed for checking the fit of Cox regression models for cohort data; cf. Klein and Moeschberger (2003, chap 11). One option is to consider cumulative sums of martingale-based residuals along the lines of Lin et al. (1993). By specifying the cumulative residuals process properly, we can obtain the special case of both partial-sum process and score process, which can be applied for the checking of two model assumptions.

Sampling is a key step in the process of model checking. This is due to the fact that by only looking at the observed curve of cumulative residuals process, we cannot reach any conclusion about whether the model assumptions are satisfied or not. In other words, it is impossible to verify the randomness of the process. However, if we can sample a great amount of replicates of cumulative residuals, it will allow us to compare the observed curve with the sampled ones. In the end, we may acquire a formal test and visualize how many of the sampled processes are actually having a larger absolute supremum value than the observed one.

The Cox regression model is not only used in cohort studies, but it has also been extended to nested case-control studies. Generally, Cox regression model is based on the fact that we have obtained information of the covariates for all individuals. For large cohorts it may be extremely difficult and expensive to obtain such cohort data, especially considering that there might be only a small proportion of individuals having experienced the event of interest. To solve this problem, nested case-control studies turn out to be an efficient alternative to cohort studies. By using nested case-control methods, we only need to choose a small number of controls at each event time. This will reduce the workload of Cox regression greatly without missing much information compared to the cohort data.

As a matter of fact, the methodology for model checking is much less developed when Cox's model is used to analyze nested case-control data. The main reason for this is that the available data in a nested case-control study do not allow for an easy generalization of the common goodness-of-fit methods that are developed for the full cohort. Therefore, the aim of the thesis is to extend the cumulative residuals process of Lin et al. (1993) to nested case-control data and to study its performance on real data sets as well as on simulated data.

The outline of the thesis is as follow. In Chapter 2 we will first present the German breast cancer study (GBCS) data that will be used for illustration in Chapters 2 and 3. Then we will give some introductions about survival analysis, counting processes and the Cox regression model. In Chapter 3, we will discuss the model checking techniques for cohort studies. We will start by introducing the two model assumptions, martingale residual pro-

cesses and martingale residuals. Then we will use GBCS data to illustrate how the model checking using cumulative sums of martingale-based residuals along the lines of Lin et al. (1993) is done for cohort data. Further, nested case-control studies on a real data set will be discussed in Chapter 4, it is an extension of cohort studies that follows a quite similar structure to Chapter 3. We will start by giving a brief overview about the simple random sampling and the counter-matched sampling methods. All formulations regarding nested case-control data will be derived and referred to the cohort in Chapter 3 for comparison. Later on we will use the Radiation and breast cancer data for illustration, where both of two model assumptions will be checked using cohort, nested case-control with simple random sampling and counter-matched sampling respectively. In Chapter 5, we will run simulation and study the efficiency of model checking for nested case-control data. We will explain how to simulate data from both a correct model and a wrong model, followed by performing model checking for nested case-control data in comparison with cohort data. Finally, in Chapter 6, we will summarize our findings and have some further discussion about the results as well as the problems that we have encountered.

Chapter 2

Survival Analysis

The chapter is based on Sections 3.1, 3.2 and 4.1 in the book by Aalen, Borgan and Gjessing (2008). In Section 2.1 we present data from the German breast cancer study (GBCS), which is an example that we use for cohort data. In Section 2.2 we give an introduction about survival analysis, including some concepts, censoring and two important non-parametric estimators, followed by Section 2.3, where we have a brief discussion about counting processes. In Section 2.4 we focus on the Cox regression model. The model we obtain for the GBCS study will be later used to illustrate model checking in the next chapter. In the final section, we find an estimator for the cumulative baseline hazard and show that how it can be used.

2.1 German breast cancer study example

We start off by giving a brief overview about the data from German breast cancer study (GBCS) group, provided by Sauerbrei and Royston (1999). These data have been used to demonstrate methods for building prognostic models. From July 1984 to December 1989, there were in total 720 patients with primary node positive breast cancer recruited for this breast cancer study, in which 686 observations are accessible and valid. In the study, patients were followed from the date of breast cancer diagnosis until censoring or dying from breast cancer. The total number of events, or the number of deaths due to breast cancer, is 171. A summary of the data is shown in Table 2.1.

There are a total of 8 covariates on the list, with 5 of them being numeric, including Age at diagnosis, Tumor size, Number of positive lymph nodes, Number of progesterone receptors and Number of Estrogen Receptors, while the remaining are categorical covariates which include Menopausal Status, Hormone Therapy and Tumor Grade.

Table 2.1: Summary of German breast cancer study data

Covariate	mean(sd)	Number	Percentage
Age at Diagnosis	53.15(10.12)		
Tumor Size	29.33(14.30)		
Number of positive lymph nodes	5.01(5.48)		
Number of progesterone receptors	110.00(202.33)		
Number of Estrogen Receptors	96.25(153.08)		
Menopausal Status			
Yes		290	42.3%
No		396	57.7%
Hormone Therapy			
Yes		440	64.1%
No		246	35.9%
Tumor Grade			
Grade 1		81	11.8%
Grade 2		444	64.8%
Grade 3		161	23.5%

2.2 An overview of survival analysis

2.2.1 Concepts

Survival analysis is a collection of statistical methods for studying survival times. In present-day society, survival analysis finds a number of applications in many different fields, especially in biology, insurance, medicine, sociology, reliability engineering and economics. A survival time may be the lifetime of an individual. Some other examples of survival times are:

- Period from surgical resection to death.
- Period from disease remission to relapse.
- Period from smoking to lung cancer.
- Duration from graduation to employment.
- Time from driving to car accident.

To be more general, a survival time is the time from an initial event to an endpoint where the event of interest occurs. For illustration purpose, here we take the German breast cancer study (GBCS) as an example. In this context, survival times are from the very times when individuals were diagnosed of their own breast cancer to they meet their deaths.

The survival function $S(t)$, which gives the probability that an individual has not experienced the event of interest by time t, can be written in the form of

$$S(t) = P(T > t),$$

where T is a continuous random variable which indicates the survival time. This implies that the survival function $S(t)$ has a starting value of 1 and then it will decrease over time, and finally approach zero or a positive value as t goes to infinity.

The hazard rate $\alpha(t)$, which is defined by a conditional probability, can be written as

$$\alpha(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t | T \geq t).$$

Note that $\alpha(t)dt$ indicates the probability that the event will happen before time $t + dt$, given that it has not happened before time t. The hazard rate function is a bit more complex than the survival function, since the curve can be rising, dropping or even fluctuating as time goes by. The cumulative hazard rate is given by

$$A(t) = \int_0^t \alpha(s)ds.$$

The survival function and the hazard rate have a relation as

$$A'(t) = \alpha(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} \log S(t). \tag{2.1}$$

Notice that by the definition of the survival function we have $S(0)$=1, so by integrating on both sides of (2.1) we obtain that

$$S(t) = \exp\left\{ - \int_0^t \alpha(s)ds \right\}.$$

2.2.2 Censoring

In practice, due to closure of study, we will not be able to observe all survival times. In some cases we will only know that the survival time of an individual exceeds the follow-up time at closure. Such survival times are said to be censored. Observations can be censored for various reasons, except the study termination as we have mentioned. It might also be caused by withdrawal of participants as well as observation ceases. But either way, the survival times of those who have not experienced the event of interest by the end of the study will be censored, which is called right-censoring.

By taking the German breast cancer study (GBCS) as an example, there are in total 686 patients involved in the study, out of which the survival times of 515 patients are censored. Generally speaking, being unable to observe the

event of interest is the top reason for censoring. In this case, we are mainly interested in the patients dying from breast cancer, thus for those who were still alive by the end of the study and those who died from other causes, their survival times will be censored. Beyond that, being unable to track down the status of the participants might be another reason for the censoring of observations, as it is quite often that for some reasons participants suddenly want to drop out of the study, or they somehow lose contact with the researchers during the course of the study.

We will now look more formally at censoring. We assume that we have uncensored, independent survival times $T_1^0, ...T_n^0$ for n individuals, and let $\alpha_i(t)$ be the hazard rate of the i-th individual. Then what we observe is a right-censored survival time $\widetilde{T}_i$ together with the indicator D_i, where $D_i = 1$ if $\widetilde{T}_i = T_i^0$ and $D_i = 0$ if $\widetilde{T}_i < T_i^0$. Thus $D_i = 1$ if we observe the real survival time and $D_i = 0$ if we observe the censored survival time. We assume that censoring is independent. This means that individuals who have not experienced the event of interest or been censored at time t should have the same probability of experiencing the event in a short time interval $[t, t + dt)$ as in the situation without censoring. Hence the independent censoring assumption can be written as

$$P(t \leq \widetilde{T}_i < t + dt, D_i = 1 | \widetilde{T}_i \geq t, past) = P(t \leq T_i^0 < t + dt | T_i^0 \geq t).$$

2.2.3 Kaplan-Meier and Nelson-Aalen estimator

We here assume that all the individuals have the same hazard, i.e. $\alpha_i(t) = \alpha(t)$ for all i. The Nelson-Aalen estimator is a non-parametric estimator for estimating the cumulative hazard rate function from censored data. It is a sum over the observed survival times $T_1 < T_2 < ...$ (i.e. the $\widetilde{T}_i$ with $D_i = 1$ in increasing order) given by

$$\widehat{A}(t) = \sum_{T_j \leq t} \frac{1}{Y(T_j)},$$

where $Y(t)$ is the number of individuals at risk just before time t. It can be shown that the variance of the Nelson-Aalen estimator can be estimated by

$$\widehat{\sigma}^2(t) = \sum_{T_j \leq t} \frac{1}{Y(T_j)^2}.$$

The Kaplan-Meier estimator is the non-parametric maximum likelihood estimate of the survival function from lifetime data. It takes a form of

$$\widehat{S}(t) = \prod_{T_j \leq t} \left\{ 1 - \frac{1}{Y(T_j)} \right\}.$$

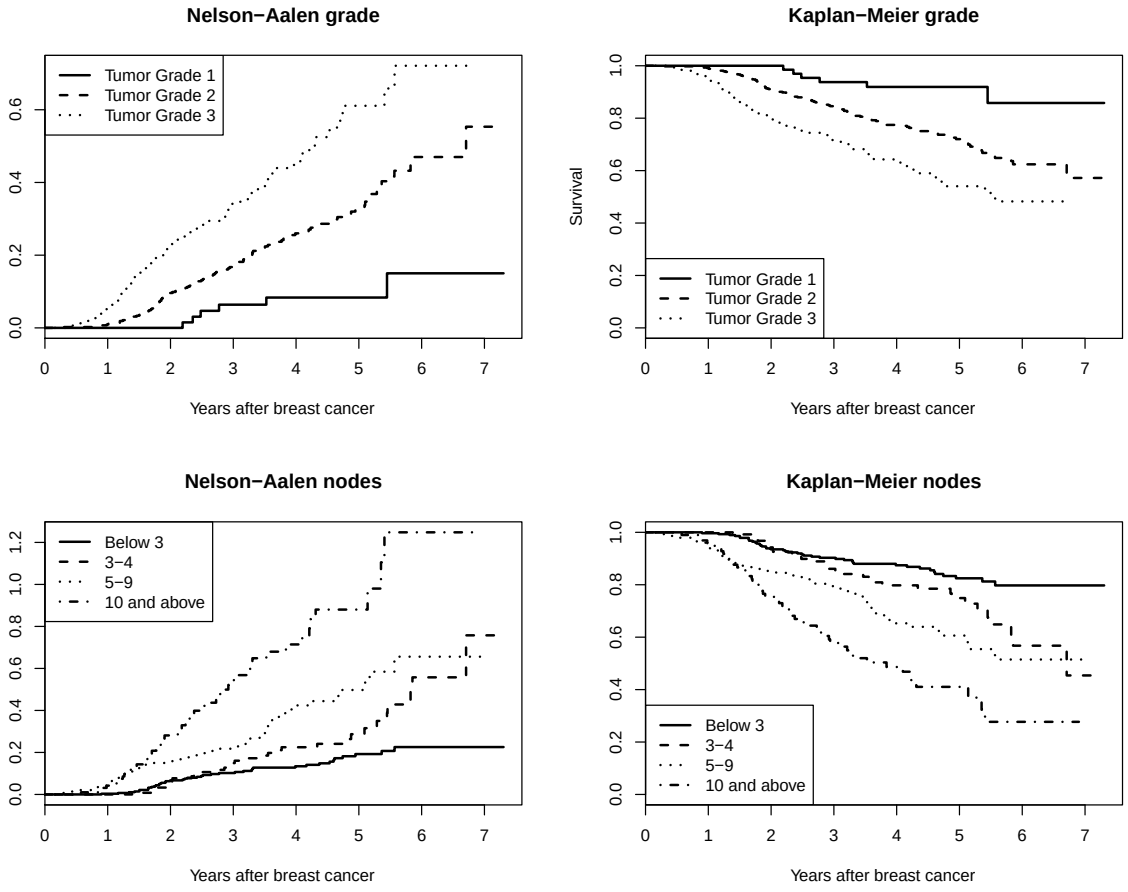

Figure 2.1: Nelson-Aalen and Kaplan-Meier plots for the breast cancer patients according to tumor grade (upper panel) and number of positive lymph nodes(lower panel)

The variance of the Kaplan-Meier estimator can by estimated by using the Greenwood's formula, which takes the form

$$\widehat{\tau}^2(t) = \widehat{S}(t)^2 \sum_{T_j \leq t} \frac{1}{Y(T_j)\left\{Y(T_j) - 1\right\}}.$$

By using the German breast cancer study (GBCS) data, Figure 2.1 shows the Nelson-Aalen and Kaplan-Meier plots for the time from breast cancer diagnosis to death. The first plot gives Nelson-Aalen estimates for cumulative hazard of death for patients in different tumor grades, with the upper curve being the highest grade and the lower curve being lowest. By looking at the slopes of the Nelson-Aalen plot we observe that the hazard rate for patients with high tumor grades is larger than those with low grades, which indicates that for high tumor grades patients, death from the breast cancer will takes place earlier and at a higher rate. We also see that the Nelson-Aalen plots are fairly linear, corresponding to constant hazards. From the

Kaplan-Meier plot, we can see that low tumor grade patients have a higher survival probability than high tumor grade patients. More specifically, the estimated survival probabilities for tumor grade 1 and 2 patients three years after breast cancer are 0.938 and 0.843 respectively. But when it comes to tumor grade 3 patients, however, the corresponding estimate is only 0.711.

To do a similar study of marginal effect of number of positive lymph nodes, we first use the Nelson-Aalen estimator. It can be clearly seen from Figure 2.1 that the hazard rate of patients with more than 10 nodes is stably at a high level. While for the patients with less than 3 nodes, they seem to be living fairly well in the first two years, after which their hazard rate is kept at a low level until the end of the study. In conclusion, it shows that patients with larger number of positive lymph nodes are much more likely to experience death from breast cancer. Finally, the Kaplan-Meier plot indicates that the lower number of nodes that the patients have, the higher survival probability they will have. Indeed, by the end of the study, the estimated survival probability for patients with less than 3 nodes, which is 0.797, is almost three times as large as that of patients group with 10 nodes and above (0.277).

2.3 Counting processes

A counting process records events that are occurring over time. Some examples of such events contain:

- Catching a cold.
- Going to the gym.
- Increasing of food price.
- Running into an old friend on the street.
- Replacing the batteries of the TV remote control.

Let $N_i(t)$ be a counting process which counts the observed occurrences of the event of interest for individual i in the time interval $[0, t]$. From censored survival data, the counting process for individual i can be given as

$$N_i(t) = I(\widetilde{T}_i \leq t, D_i = 1), \tag{2.2}$$

where $\widetilde{T}_i$ indicates the right-censored survival time for individual i, while $D_i = 1$ is an indicator of observing the real survival time. Now we consider

the aggregated counting process, which is the sum of the individual counting processes. Thus it takes the form of

$$N_{\cdot}(t) = \sum_{i=1}^{n} N_i(t) = \sum_{i=1}^{n} I(\widetilde{T}_i \leq t, D_i = 1).$$

The intensity process of $N_i(t)$ is denoted by $\lambda_i(t)$, and can be written as

$$\lambda_i(t)dt = P(dN_i(t) = 1|past),$$

where $dN_i(t)$ is the increment of $N_i(t)$ in $[t, t + dt)$. Based on survival data, $\lambda_i(t)$ for the counting process takes the form

$$\lambda_i(t) = \alpha_i(t)Y_i(t),$$

where

$$Y_i(t) = I\left\{\widetilde{T}_i \geq t\right\}.$$

We introduce the process

$$M_i(t) = N_i(t) - \int_0^t \lambda_i(s)ds. \tag{2.3}$$

Then we have that

$$\begin{aligned} E(dM_i(t)|past) &= E(dN_i(t) - \lambda_i(t)dt|past) \\ &= P(dN_i(t) = 1|past) - \lambda_i(t)dt = 0 \end{aligned} \tag{2.4}$$

This shows that $M_i(t)$ is a martingale.

2.4 Cox regression

2.4.1 Model methods

In survival analysis, a covariate is an explanatory variable which possibly influences the hazard rate of an individual. Covariates can be either of numeric or categorical type, and in most circumstances there is more than one covariate. For instance, recall the German breast cancer study (GBCS) from Table 2.1, there are many covariates which turn out to be having a huge impact on the hazard rate of individuals. In order to assess the effect of covariates, we want to relate the hazard of an individual to its covariates.

Assume that the vector of covariates $\mathbf{x}_i = (x_{i1}, ...x_{ip})^T$ for individual i is related to the hazard rate $\alpha(t|\mathbf{x}_i)$ of the individual, and it is in the form of

$$\alpha(t|\mathbf{x}_i) = \alpha_0(t)r(\boldsymbol{\beta}, \mathbf{x}_i), \tag{2.5}$$

where $r(\boldsymbol{\beta}, \mathbf{x}_i)$ is a relative risk function with $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_p)^T$ being the vector of regression coefficients and $\alpha_0(t)$ denotes a baseline hazard rate. Notice that when $\mathbf{x}_i = \mathbf{0} = (0, ..., 0)^T$, we have $r(\boldsymbol{\beta}, \mathbf{0}) = 1$. We will throughout consider the relative risk function

$$r(\boldsymbol{\beta}, \mathbf{x}_i) = \exp\left\{\boldsymbol{\beta}^T \mathbf{x}_i\right\}$$

corresponding to Cox's regression model. Then the hazard ratio of two individuals becomes

$$\frac{\alpha(t|\mathbf{x}_2)}{\alpha(t|\mathbf{x}_1)} = \frac{\alpha_0(t) \exp\left\{\boldsymbol{\beta}^T \mathbf{x}_2\right\}}{\alpha_0(t) \exp\left\{\boldsymbol{\beta}^T \mathbf{x}_1\right\}} = \frac{\exp\left\{\boldsymbol{\beta}^T \mathbf{x}_2\right\}}{\exp\left\{\boldsymbol{\beta}^T \mathbf{x}_1\right\}}.$$

If $\mathbf{x}_1$ and $\mathbf{x}_2$ are the same except that $x_{2j} = x_{1j} + 1$, then the hazard ratio becomes

$$\frac{\alpha(t|\mathbf{x}_2)}{\alpha(t|\mathbf{x}_1)} = \exp\left\{\boldsymbol{\beta}^T (\mathbf{x}_2 - \mathbf{x}_1)\right\} = e^{\beta_j},$$

where e^{β_j} denotes the hazard rate ratio, or the relative risk of the jth covariate. Define $Y_l(t)$ as an indicator which takes the value of 1 if individual l is at risk just before time t and otherwise takes the value 0. Then $R_j = \{l|Y_l(T_j) = 1\}$ is the risk set at T_j. Let i_j further be the individual who experiences an event at time T_j. Then the partial likelihood for $\boldsymbol{\beta}$ is given as

$$L(\boldsymbol{\beta}) = \prod_{T_j} \frac{r(\boldsymbol{\beta}, \mathbf{x}_{i_j})}{\sum_{l \in R_j} r(\boldsymbol{\beta}, \mathbf{x}_l)}. \tag{2.6}$$

The maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\beta}$ that maximizes $L(\boldsymbol{\beta})$. Notice that $\hat{\boldsymbol{\beta}}$ is approximately multivariate normally distributed around $\boldsymbol{\beta}$, and the covariance matrix can be estimated by $I(\hat{\boldsymbol{\beta}})^{-1}$, where $I(\boldsymbol{\beta}) = \left\{-\frac{\partial^2}{\partial \beta_h \partial \beta_j} \log L(\boldsymbol{\beta})\right\}$.

For testing the null hypothesis $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, where typically $\boldsymbol{\beta}_0 = \mathbf{0}$, there are three different types of test statistics that we can use:

- The likelihood ratio test statistic:

$$\chi^2_{LR} = 2\left\{\log L(\hat{\boldsymbol{\beta}}) - \log L(\boldsymbol{\beta}_0)\right\}$$

- The Wald test statistic:

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

- The score test statistic:

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

  where $U(\boldsymbol{\beta}) = \frac{\partial}{\partial \beta} \log L(\boldsymbol{\beta})$.

Notice that all three tests are asymptotically equivalent and approximately χ^2_p distributed under H_0.

Table 2.2: Cox regression analysis for the marginal effect of the tumor grade as categorical covariates

Covariate	$\hat{\beta}_j$	Hazard ratio	se($\hat{\beta}_j$)	z	Pr(>|z|)
Tumor grade 2	1.24	3.46	0.42	2.96	0.003
Tumor grade 3	1.86	6.44	0.43	4.34	1.43e-05

2.4.2 The GBCS study

We are going to study the effect of some covariates using Cox regression for the German breast cancer study (GBCS). As we mentioned before, there are quite a lot of covariates. We start out by focusing on two of them. The tumor grades are grouped into three different levels, while the number of positive lymph nodes is a numeric covariate. A Cox model with grade as the only covariate has the breast cancer death rate for individual i of the form

$$\alpha(t|\mathbf{x}_i) = \alpha_0(t) \exp\left\{\beta_1 x_{i1} + \beta_2 x_{i2}\right\},$$

where

$$x_{i1} = \begin{cases} 1 & \text{if individual } i \text{ has tumor grade 2} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if individual } i \text{ has tumor grade 3} \\ 0 & \text{otherwise} \end{cases}$$

From Table 2.2 we see that the estimated relative risk for tumor grade 2 is exp(1.24)=3.46, therefore the hazard rate of this group is 246% larger than Tumor grade 1 group, corresponding to a P-value of 0.3%. As what we have expected, there is a clear difference between these two groups. Similarly we observe that the hazard rate of Tumor grade 3 group is over 6 times larger than Tumor grade 1 group, indicating a significant difference.

Then we study the marginal effect of the number of positive lymph nodes. If we group the lymph nodes into four groups, the breast cancer death rate for individual i is taking the form

$$\alpha(t|\mathbf{x}_i) = \alpha_0(t) \exp\left\{\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}\right\},$$

where

$$x_{i1} = \begin{cases} 1 & \text{if individual } i \text{ has 3-4 positive nodes} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if individual } i \text{ has 5-9 positive nodes} \\ 0 & \text{otherwise} \end{cases}$$

Table 2.3: Cox regression analysis for the marginal effect of positive lymph nodes as categorical covariates

Covariate	$\hat{\beta}_j$	Hazard ratio	se($\hat{\beta}_j$)	z	Pr(>|z|)
Nodes number 3-4	0.58	1.78	0.24	2.44	0.015
Nodes number 5-9	1.04	2.84	0.21	4.88	1.08e-06
Nodes number 10 and above	1.70	5.50	0.21	8.02	9.99e-16

Table 2.4: Cox regression analysis for the marginal effect of positive lymph nodes as numeric covariates

Covariate	$\hat{\beta}_j$	Hazard ratio	se($\hat{\beta}_j$)	z	Pr(>|z|)
Positive lymph nodes	0.068	1.07	0.008	8.20	2.22e-16

$$x_{i3} = \begin{cases} 1 & \text{if individual } i \text{ has 10 or more positive nodes} \\ 0 & \text{otherwise} \end{cases}$$

According to Table 2.3, when it comes to the number of positive lymph nodes, the estimated relative risks are 1.78, 2.84 and 5.50, indicating that the hazard rate of 3-4 nodes, 5-9 nodes and 10 nodes and above groups exceed the below 2 nodes group by 78%, 184% and 450% respectively. Besides, the P-values for all of the three groups are quite small, which also proves this significant difference. In conclusion, the tumor grade and number of positive nodes are both highly effective covariates for this study. Moreover, we find that the Cox regression analysis results are fairly consistent with the previous analysis that we did using the Kaplan-Meier and Nelson-Aalen estimators.

To illustrate the difference between categorical and numeric covariates, we now do a Cox regression analysis with nodes as a numeric covariate. According to the result given in Table 2.4, we can clearly see that positive lymph nodes will cause the hazard rate to rise. A Cox regression fit with log base 2-transformed positive lymph nodes as numeric covariates has also been done, as in Table 2.5. It turns out we obtain a quite similar result, corresponding to a hazard ratio of 1.57 and fairly significant p-value, that positive lymph nodes will trigger a higher hazard rate.

Finally, we will fit a multivariate Cox regression model with all covariates taken into account and without grouping the numeric covariates. We first

Table 2.5: Cox regression analysis for the marginal effect of log base 2-transformed positive lymph nodes as numeric covariates

Covariate	$\hat{\beta}_j$	Hazard ratio	se($\hat{\beta}_j$)	z	Pr(>|z|)
log(Positive lymph nodes)	0.45	1.57	0.057	7.88	3.22e-15

Table 2.6: Cox regression analysis for all covariates

Covariate	$\hat{\beta}_j$	Hazard ratio	se($\hat{\beta}_j$)	z	Pr(>|z|)
Age at diagnosis	0.01	1.01	0.01	0.56	0.57
No menopause	0.09	1.10	0.25	0.36	0.72
No hormone	-0.27	0.76	0.17	-1.59	0.11
Tumor size	0.01	1.01	0.00	2.73	0.006
Tumor grade 2	0.78	2.17	0.43	1.82	0.069
Tumor grade 3	1.13	3.09	0.44	2.55	0.011
Positive lymph nodes	0.05	1.05	0.01	5.50	3.84e-08
Progesterone receptors	-0.01	0.99	0.00	-4.49	7.00e-06
Estrogen Receptors	0.00	1.00	0.00	-0.47	0.64

consider the case when no transformation is applied. The result is given in Table 2.6. We can see that in the multivariate Cox regression model there are many insignificant covariates which need to be removed to get the best model. At this point, we will reduce those non-significant covariates one by one, starting from the covariate "Menopause", as it has a P-value of 0.72 which indicates a high insignificance. Then we fit the model with the remaining covariates, now "Estrogen Receptors" becomes the most non-significant covariate, corresponding to a P-value of 0.64. So we remove this covariate and repeat this process until all the covariates irrelevant to the Cox model are removed. The final Cox regression model is shown in Table 2.7, with only three effective covariates: Tumor size, Positive lymph nodes, and Progesterone receptors, where the first two will increase the hazard rate.

Then we do Cox regression fit with log-transformation on some covariates. The numeric covariates are Tumor size, Positive lymph nodes, Progesterone receptors and Estrogen Receptors. These have all been log base 2-transformed. Notice that one has been added to both Progesterone receptors and Estrogen Receptors such that their values are non-zero before applying the log transformation. According to Table 2.8 and 2.9, the result is slightly different compared with non-transformation. In this case, the final model ends up in three significant covariates: Hormone, Positive lymph nodes and Progesterone receptor. It implies that both hormone therapy experience and high number of positive lymph nodes will cause the hazard rate to rise. On the contrary, the higher number of progesterone receptors that patients have, the more likely they will be able to survive. Besides, it is necessary to make clear that we do not check model assumptions at this stage, as this will be done in later sections.

Table 2.7: Cox regression analysis for the best model

Covariate	$\hat{\beta}_j$	Hazard ratio	se($\hat{\beta}_j$)	z	Pr(>|z|)
Tumor size	0.01	1.01	0.005	2.94	0.003
Positive lymph nodes	0.05	1.06	0.009	5.65	1.57e-08
Progesterone receptors	-0.006	0.99	0.001	-5.39	6.99e-08

Table 2.8: Cox regression analysis for all covariates with log-transformation

Covariate	$\hat{\beta}_j$	Hazard ratio	se($\hat{\beta}_j$)	z	Pr(>|z|)
Age at diagnosis	0.01	1.01	0.01	0.53	0.60
No menopause	0.07	1.07	0.25	0.29	0.77
No hormone	-0.36	0.70	0.17	-2.1	0.035
Log(Tumor size)	0.20	1.22	0.12	1.62	0.11
Tumor grade 2	0.66	1.93	0.43	1.54	0.12
Tumor grade 3	0.88	2.42	0.45	1.97	0.049
Log(Positive lymph nodes)	0.37	1.45	0.06	6.18	6.29e-10
Log(Progesterone receptors)	-0.19	0.83	0.04	-5.09	3.60e-07
Log(Estrogen Receptors)	0.00	1.00	0.04	0.01	0.99

Table 2.9: Cox regression analysis for the best model with log-transformation

Covariate	$\hat{\beta}_j$	Hazard ratio	se($\hat{\beta}_j$)	z	Pr(>|z|)
No hormone	-0.36	0.69	0.16	-2.22	0.027
Log(Positive lymph nodes)	0.41	1.51	0.06	7.36	1.84e-13
Log(Progesterone receptors)	-0.21	0.81	0.03	-7.78	7.11e-15

2.5 Estimation of cumulative baseline hazard

In order to obtain an estimator for the cumulative baseline hazard $A_0(t) = \int_0^t \alpha_0(u)du$, we introduce the aggregated counting process

$$N_{\cdot}(t) = \sum_{l=1}^{n} N_l(t).$$

This has intensity process given by

$$\lambda_{\cdot}(t) = \left(\sum_{l=1}^{n} Y_l(t) r(\boldsymbol{\beta}, \mathbf{x}_l) \right) \alpha_0(t).$$

By using a Nelson-Aalen type estimator, we find that the cumulative baseline hazard can be estimated by

$$\hat{A}_0(t) = \int_0^t \frac{dN_{\cdot}(u)}{\sum_{l=1}^{n} Y_l(u) r(\widehat{\boldsymbol{\beta}}, \mathbf{x}_l)} = \sum_{T_j \leq t} \frac{1}{\sum_{l \in R_j} r(\widehat{\boldsymbol{\beta}}, \mathbf{x}_l)}. \tag{2.7}$$

This estimator is often denoted as the Breslow estimator. The cumulative hazard for an individual with a given covariate vector $\mathbf{x}_0$ can be written as

$$A(t|\mathbf{x}_0) = \int_0^t \alpha(u|\mathbf{x}_0)du = r(\boldsymbol{\beta}, \mathbf{x}_0) A_0(u),$$

which can be estimated by

$$\widehat{A}(t|\mathbf{x}_0) = r(\widehat{\boldsymbol{\beta}}, \mathbf{x}_0) \widehat{A}(t).$$

The corresponding survival function takes the form

$$S(t|x_0) = \prod_{u \leq t} \left\{ 1 - dA(u|\mathbf{x}_0) \right\}.$$

The estimator is given by

$$\widehat{S}(t|x_0) = \prod_{u \leq t} \left\{ 1 - d\widehat{A}(u|\mathbf{x}_0) \right\} = \prod_{T_j \leq t} \left\{ 1 - \triangle \widehat{A}(T_j|\mathbf{x}_0) \right\},$$

which is nearly equivalent to the following estimator

$$\widehat{S}(t|\mathbf{x}_0) = \exp\left\{ -\widehat{A}(t|\mathbf{x}_0) \right\}.$$

For illustration we will apply the above estimators for data from the German breast cancer study (GBCS). We then consider the model with two categorical covariates: Hormone Therapy status and Tumor Grade. The plots of the

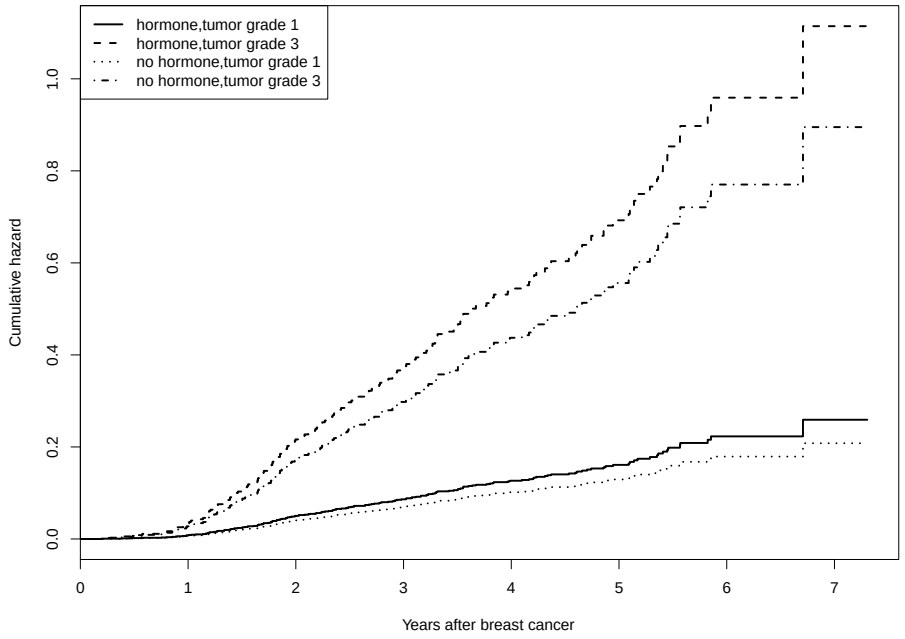

Figure 2.2: Estimated cumulative hazards curves for the breast cancer patients according to Hormone Therapy status and tumor grades

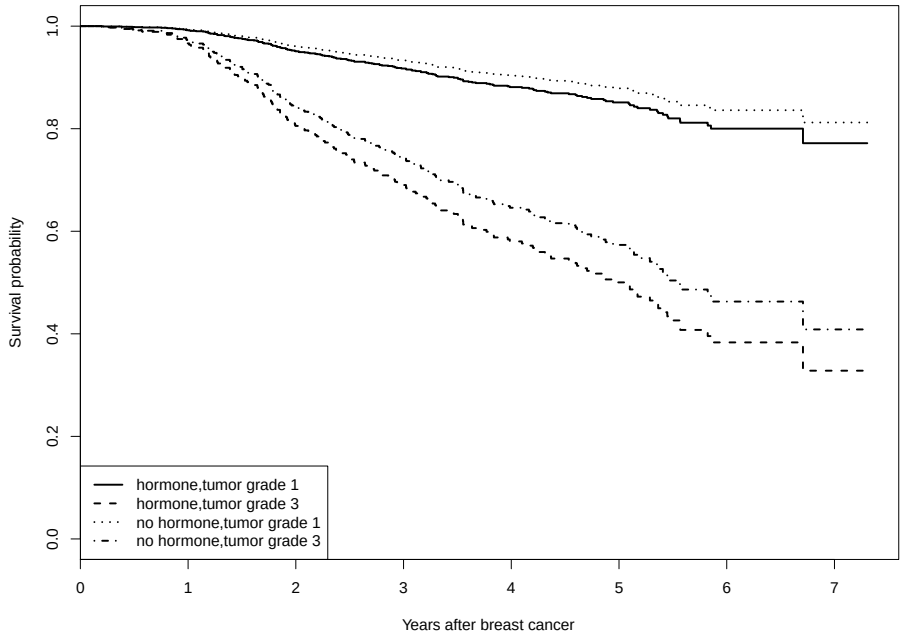

Figure 2.3: Estimated survival curves for the breast cancer patients according to Hormone Therapy status and tumor grades

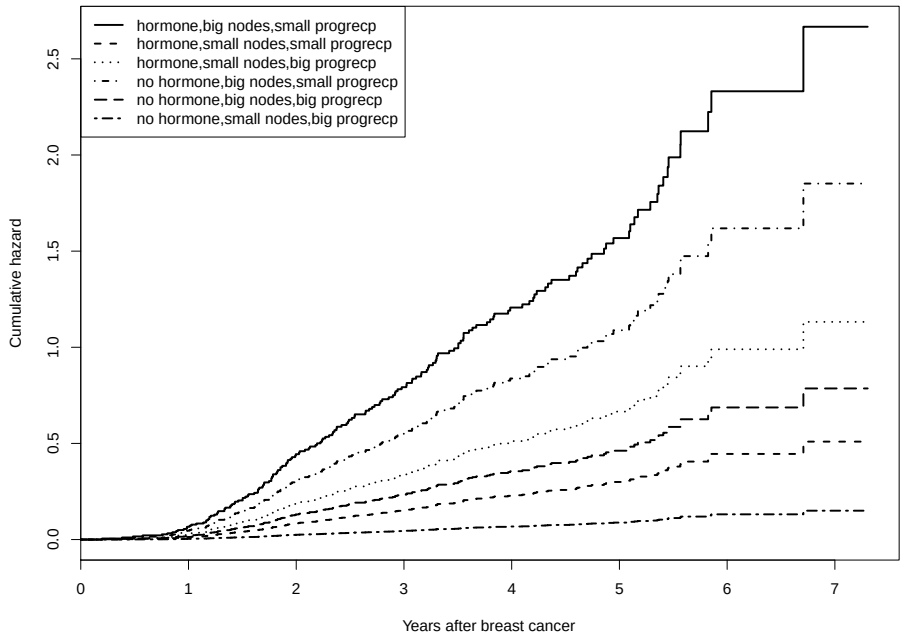

Figure 2.4: Estimated cumulative hazards curves for the breast cancer patients according to Hormone Therapy status, Positive lymph nodes and Progesterone receptors with log-transformation

cumulative hazards and survival curves for these covariates combinations are shown in Figure 2.2 and Figure 2.3.

From the survival curves in Figure 2.3 we can see that the probability of breast cancer patients with no Hormone Therapy and tumor grade 1 to survive 7 years after diagnosis is 81.2%, which is the highest one among these four combinations. When it comes to those patients with Hormone Therapy and tumor grade 3, however, the survival probability is only 32.8%. Finally, the survival probabilities of the other two combinations of covariates are 77.2% and 40.9% respectively.

The cumulative hazards plots in Figure 2.2 indicates a similar result. Moreover, when covariates Hormone Therapy and Tumor grade are fitted together for Cox regression model, the Tumor grade is having a more decisive role on survival rate. Beyond that, we can observe that the Hormone Therapy status shows a more significant impact on survival rate in larger tumor grade categories.

It is important to point out that we have already obtained a final model with log base 2-transformed numbers. In the continuation of this section we will use this model for further analysis. Recall that it is fitted by three covariates: Hormone, Positive lymph nodes and Progesterone receptors. The plots are

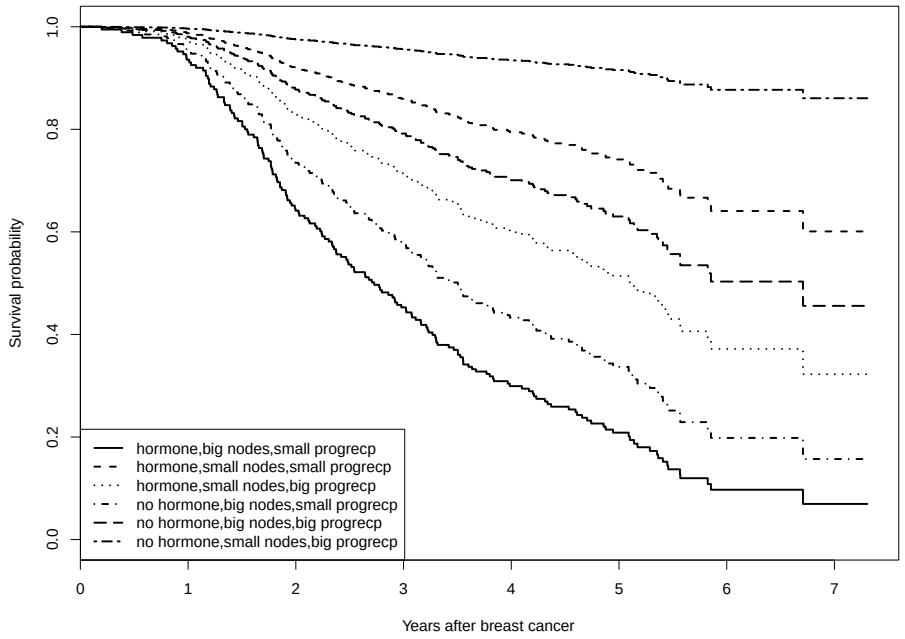

Figure 2.5: Estimated survival curves for the breast cancer patients according to Hormone Therapy status, Positive lymph nodes and Progesterone receptors with log-transformation

shown in Figure 2.4 and Figure 2.5, where original value 32 and 2 are used for representing large and small number of positive lymph nodes, while 255 and 15 are for large and small number of progesterone receptors respectively. Figure 2.5 indicates that patients with Hormone Therapy, a large number of Positive lymph nodes and a small number of Progesterone receptors will have the lowest survival rate which is even less than 7%, while the opposite situation will increase the survival rate up to over 86%.

To summarize, so far we have covered some basic concepts in survival analysis, counting process and Cox model, combined with detailed illustrations using cohort data. Consequently we are able to obtain the best Cox model and study the effect of the covariates on the survival probabilities. Further, it is important to check if both of the two assumptions of Cox model are satisfied. So in the next chapter we will discuss the model checking techniques and demonstrate how they can be used for cohort studies.

Chapter 3

Model checking for cohort studies

The chapter is based on Section 4.1 in the book by Aalen, Borgan and Gjessing (2008) and the paper by Lin et al (1993). In Section 3.1 we give a brief overview about the two basic model assumptions in the Cox model. Then we continue with Section 3.2, where the martingale residual processes and martingale residuals are introduced. In Section 3.3 we discuss cumulative sums based on the martingale residual processes, which is a very useful model checking technique for cohort studies. Further we divide the process that we obtain in Section 3.3 into two special cases: the partial-sum process and the score process. They are discussed separately in Sections 3.4 and 3.5. In the meantime, examples of model checking using GBCS data are shown for illustration.

3.1 Model assumptions

In a Cox model, there are two basic assumptions which must be satisfied, namely log-linearity of numeric covariates and proportional hazards. Many methods can be used for checking whether these assumptions are violated when fitting a Cox model. In the following sections, we will consider some such methods. To start with, we introduce the first key point of the Cox model assumption, which is the log-linearity for the hazard. That is to say that the hazard ratio must be a linear function of a numeric covariate on the log-scale. Hence we have that

$$\log\{\alpha(t|\mathbf{x})\} = \log\{\alpha_0(t)\} + \boldsymbol{\beta}^T\mathbf{x}. \tag{3.1}$$

Another model assumption is called proportional hazard, which suggests that the hazard rates for two individuals must be proportional. Therefore they must satisfy that

$$\frac{\alpha(t|\mathbf{x}_2)}{\alpha(t|\mathbf{x}_1)} = \exp\left\{\boldsymbol{\beta}^T(\mathbf{x}_2 - \mathbf{x}_1)\right\}, \tag{3.2}$$

where the hazard rate ratio is a constant that depends on the covariates, but not on time.

3.2 Martingale residual processes and martingale residuals

Martingale residuals are of great importance for checking the fit of the Cox model. In order to describe the martingale residuals processes, we write the cumulative intensity process of the ith individual at time t as

$$\Lambda_i(t) = \int_0^t \lambda_i(u)du = \int_0^t Y_i(u)r(\boldsymbol{\beta}, \mathbf{x}_i)\alpha_0(u)du. \tag{3.3}$$

From (2.7) we have that

$$\widehat{A}_0(t) = \int_0^t \frac{dN_{\cdot}(u)}{\sum_{l=1}^n Y_l(u)r(\widehat{\boldsymbol{\beta}}, \mathbf{x}_l)} = \sum_{T_j \leq t} \frac{1}{\sum_{l \in R_j} r(\widehat{\boldsymbol{\beta}}, \mathbf{x}_l)}$$

By plugging the maximum partial likelihood estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ and $d\widehat{A}_0(u)$ for $\alpha_0(u)du$ into (3.3), it follows that

$$\widehat{\Lambda}_i(t) = \int_0^t Y_i(u)r(\widehat{\boldsymbol{\beta}}, \mathbf{x}_i)d\widehat{A}_0(u) = \sum_{T_j \leq t} \frac{Y_i(T_j)r(\widehat{\boldsymbol{\beta}}, \mathbf{x}_i)}{\sum_{l \in R_j} r(\widehat{\boldsymbol{\beta}}, \mathbf{x}_l)}$$

By (2.3), the martingale residual processes can be written as

$$\widehat{M}_i(t) = N_i(t) - \widehat{\Lambda}_i(t),$$

where $N_i(t)$ is the observed number of events for the ith individual while $\widehat{\Lambda}_i(t)$ represents the expected ones. We define the upper time limit for the study as τ. Then we can obtain the martingale residuals as

$$\widehat{M}_i = \widehat{M}_i(\tau) = N_i(\tau) - \widehat{\Lambda}_i(\tau).$$

3.3 Cumulative sums of martingale based residuals

We are now starting to get to grips with the model checking techniques. The idea is to group the martingale based residuals cumulatively according to time or covariate components. In the paper by Lin et al (1993), it has been

shown that the general formulation of the cumulative sums of the martingale based residuals is given by the stochastic process

$$W(t,\mathbf{x}) = \sum_{i=1}^{n} f(\mathbf{x}_i) I(\mathbf{x}_i \leq \mathbf{x}) \widehat{M}_i(t). \tag{3.4}$$

Here $f(\mathbf{x}_i)$ is a smooth function, $\mathbf{x} = (x_1, ...x_p)^T$ is a p-component upper limit, and $\mathbf{x}_i \leq \mathbf{x}$ is an indication that all the components of the covariate vector $\mathbf{x}_i$ are less than or equal to the corresponding components of $\mathbf{x}$.

The process (3.4) is a smooth function of the maximum partial likelihood estimator $\widehat{\boldsymbol{\beta}}$. According to the studies in Lin et al. (1993), formula (3.4), we have that

$$\begin{aligned} W(t,\mathbf{x}) \approx & \sum_{i=1}^{n} \int_0^t \left\{ f(\mathbf{x}_i) I(\mathbf{x}_i \leq \mathbf{x}) - g(\boldsymbol{\beta}_0, u, \mathbf{x}) \right\} dM_i(u) \\ & - \sum_{i=1}^{n} \int_0^t Y_i(u) \exp\left\{ {\boldsymbol{\beta}_0}^T \mathbf{x}_i \right\} f(\mathbf{x}_i) I(\mathbf{x}_i \leq \mathbf{x}) \left\{ \mathbf{x}_i - \frac{S^{(1)}(\boldsymbol{\beta}_0, u)}{S^{(0)}(\boldsymbol{\beta}_0, u)} \right\}^T \alpha_0(u) du \\ & \times I(\boldsymbol{\beta}_0)^{-1} \sum_{i=1}^{n} \int_0^\tau \left\{ \mathbf{x}_i - \frac{S^{(1)}(\boldsymbol{\beta}_0, u)}{S^{(0)}(\boldsymbol{\beta}_0, u)} \right\} dM_i(u). \end{aligned} \tag{3.5}$$

Here $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$, and

$$g(\boldsymbol{\beta}, u, \mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i(u) \exp\left\{ \boldsymbol{\beta}^T \mathbf{x}_i \right\} f(\mathbf{x}_i) I(\mathbf{x}_i \leq \mathbf{x})}{S^{(0)}(\boldsymbol{\beta}, u)},$$

$$S^{(0)}(\boldsymbol{\beta}, u) = \sum_{l=1}^{n} Y_l(u) \exp\left\{ \boldsymbol{\beta}^T \mathbf{x}_l \right\},$$

and

$$S^{(1)}(\boldsymbol{\beta}, u) = \sum_{l=1}^{n} Y_l(u) \mathbf{x}_l \exp\left\{ \boldsymbol{\beta}^T \mathbf{x}_l \right\}.$$

The point now is that when normalized by $n^{-\frac{1}{2}}$, the process $W(t,\mathbf{x})$ and the right-hand side of (3.5) asymptotically have the same distribution when the Cox model is correctly specified. However, the problem is that we are unable to track the distribution. Therefore Lin et al (1993) suggest to find the distribution by simulation, i.e. one may simulate a certain number of realizations of a process with similar large sample properties as the right-hand side of (3.5). To this end, on the right-hand side of (3.5), we replace $\boldsymbol{\beta}_0$ by $\widehat{\boldsymbol{\beta}}$ and $\alpha_0(u)du$ by $d\widehat{A}_0(u)$. Further we replace $dM_i(u)$ by $G_i dN_i(u)$, where the G_i's are independent and standard normally distributed. Thus we

obtain the process

$$\begin{aligned}\widehat{W}(t,\mathbf{x}) =& \sum_{i=1}^{n}\int_0^t \left\{f(\mathbf{x}_i)I(\mathbf{x}_i\le\mathbf{x}) - g(\widehat{\boldsymbol{\beta}},u,\mathbf{x})\right\} G_i dN_i(u)\\ &-\sum_{i=1}^{n}\int_0^t Y_i(u)\exp\left\{\widehat{\boldsymbol{\beta}}^T\mathbf{x}_i\right\} f(\mathbf{x}_i)I(\mathbf{x}_i\le\mathbf{x})\left\{\mathbf{x}_i - \frac{S^{(1)}(\widehat{\boldsymbol{\beta}},u)}{S^{(0)}(\widehat{\boldsymbol{\beta}},u)}\right\}^T d\widehat{A}_0(u)\\ &\times I(\widehat{\boldsymbol{\beta}})^{-1}\sum_{i=1}^{n}\int_0^\tau \left\{\mathbf{x}_i - \frac{S^{(1)}(\widehat{\boldsymbol{\beta}},u)}{S^{(0)}(\widehat{\boldsymbol{\beta}},u)}\right\} G_i dN_i(u).\end{aligned} \tag{3.6}$$

To simulate replicates from process (3.6), we should keep the observations fixed and only sample the $G_i's$ from $N(0,1)$.

3.4 Special case of partial-sum process

3.4.1 Observation of partial-sum process

We will look at two special cases of the process (3.4). They are selected for checking the two assumptions of Section 3.1, namely (3.1) and (3.2), where both have distributions that can be approximated simply through the simulation of Gaussian processes given as (3.6). We start by introducing the special case of partial-sum process. By inserting $f(\mathbf{x}_i) = 1$ in (3.4), when $t = \tau$ denotes the maximum time limit for the study, $x_k = \infty$ for all $k \neq j$, and x_{ji} denotes the jth component of covariate vector $\mathbf{x}_i$ corresponding to the ith individual, it follows that

$$W_j(x) = \sum_{i=1}^{n} I(x_{ji}\le x)\widehat{M}_i. \tag{3.7}$$

We will use the German breast cancer study (GBCS) data to illustrate the use of the cumulative martingale residuals (3.7). Note that for each numeric covariate we get one such process. In reference to the Cox model without log-transformation that we obtained in Section 2.4.2, it contains three covariates: tumor size, number of lymph nodes and number of positive progesterone receptors. The cumulative martingale residual plots are shown in Figure 3.1. To start with, Figure 3.1 shows that the cumulative martingale residuals for tumor size are fluctuating around 0, indicating that the estimated hazard is quite close to the observation. When it comes to the number of positive lymph nodes which are smaller than 8, the estimated hazards are exceeding the observations to a great extent. In terms of the hazards for the number of progesterone receptors lower than 60, however, they are seen to be underestimated. Therefore, it suggests that a log-transformation on some covariates

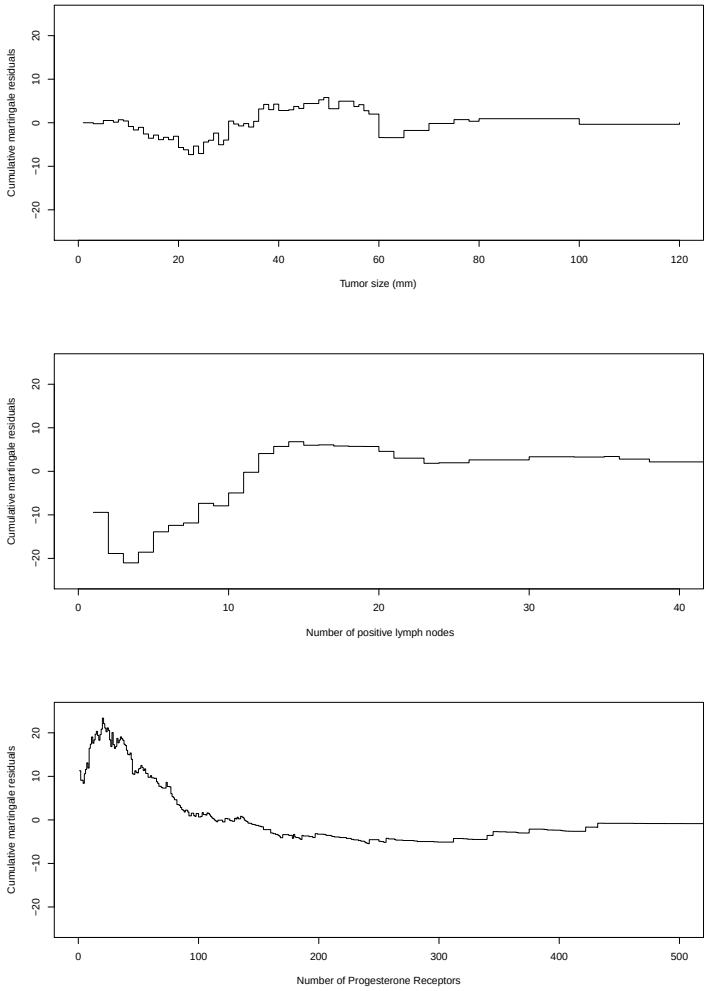

Figure 3.1: Cumulative martingale residuals against Tumor size, number of lymph nodes and number of positive progesterone receptors in the Cox model with Tumor size, Positive lymph nodes, and Progesterone receptors as covariates

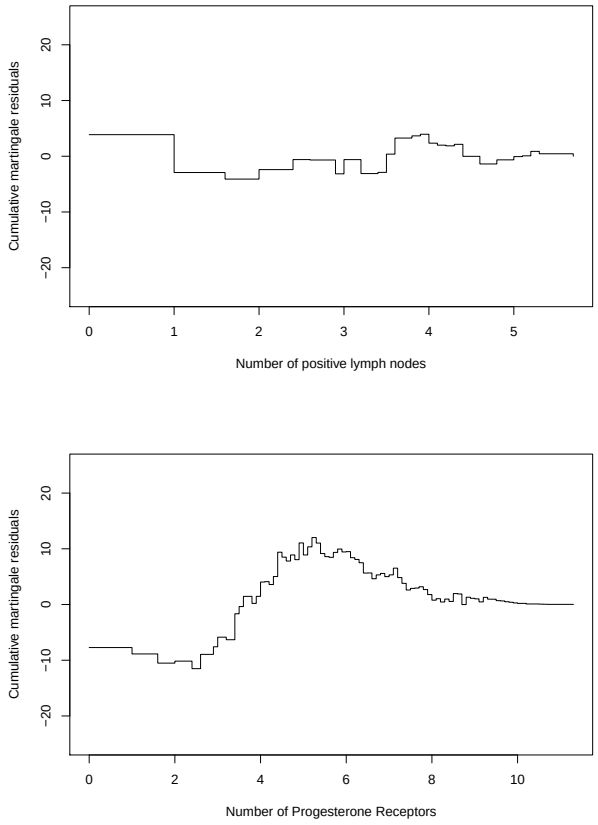

Figure 3.2: Cumulative martingale residuals against log base 2-transformed number of positive lymph nodes and number of progesterone receptors in the Cox model with Hormone Therapy, Log(Positive lymph nodes), and Log(Progesterone receptors) as covariates

is necessary, which leads to the cumulative martingale residual plot in Figure 3.2. As we have obtained earlier in Section 2.4.2, there are three covariates involved in this second Cox model, including hormone therapy, positive lymph nodes, and progesterone receptors. We can see that now the cumulative martingale residuals, which correspond to the number of positive lymph nodes and the number of progesterone receptors, are both floating around 0 without huge deviations. Obviously when log-transformation is applied, the model fit has been much improved. But we still have no knowledge about how much improvement we have obtained. Thus at this stage it requires that we compare the plots of (3.7) with plots that reflect the randomness when the model is correctly specified.

3.4.2 Simulation of partial-sum process

In order to obtain the simulated partial-sum process, we specialize the process (3.6) to the case in (3.7). For each individual in the survival data set we have the observations $\widetilde{T}_i$, D_i and $\mathbf{x}_i$ fixed. We first specialize (3.6) to situation with $f(\mathbf{x}_i) = 1$, $t = \tau$ and $\mathbf{x} = (\infty, \infty, ..., x_j, \infty, ...\infty)^T$, then specialize to censored survival data. We then have that

$$\begin{aligned}\widehat{W}_j(x) =& \sum_{i=1}^{n} \left\{ I(x_{ij} \leq x) - g(\widehat{\boldsymbol{\beta}}, \widetilde{T}_i, x) \right\} G_i D_i \\ & - \sum_{i=1}^{n} Y_i(\widetilde{T}_i) \exp\left\{ \widehat{\boldsymbol{\beta}}^T \mathbf{x}_i \right\} I(x_{ij} \leq x) \left\{ \mathbf{x}_i - \frac{S^{(1)}(\widehat{\boldsymbol{\beta}}, \widetilde{T}_i)}{S^{(0)}(\widehat{\boldsymbol{\beta}}, \widetilde{T}_i)} \right\}^T \widehat{A}_0(\widetilde{T}_i) \\ & \times I(\widehat{\boldsymbol{\beta}})^{-1} \sum_{i=1}^{n} \left\{ \mathbf{x}_i - \frac{S^{(1)}(\widehat{\boldsymbol{\beta}}, \widetilde{T}_i)}{S^{(0)}(\widehat{\boldsymbol{\beta}}, \widetilde{T}_i)} \right\} G_i D_i. \end{aligned} \tag{3.8}$$

In the continuation of the discussion above, we will in the following example demonstrate how to use (3.7) and (3.8) for checking log-linearity. Recall the Cox model we obtained from the German breast cancer study (GBCS) data. We start by checking the non-transformation case, the plot is shown in Figure 3.3, which gives (3.7) together with replicates of (3.8). The three dark lines shown in Figure 3.3 are the observed cumulative martingale residuals, which are the same as Figure 3.1. Those in grey color are sets of simulations of cumulative martingale residuals using (3.8). Looking closely, we find that the observation curve with respect to tumor size is quite satisfactory as it falls within the area representing the simulations. When it comes to the other two covariates, we see that they both have some deviations in the far left.

To acquire a formal test, we look at the supremum over x of the absolute value of (3.7) and (3.8). To obtain the P-value of the test, we simulate N

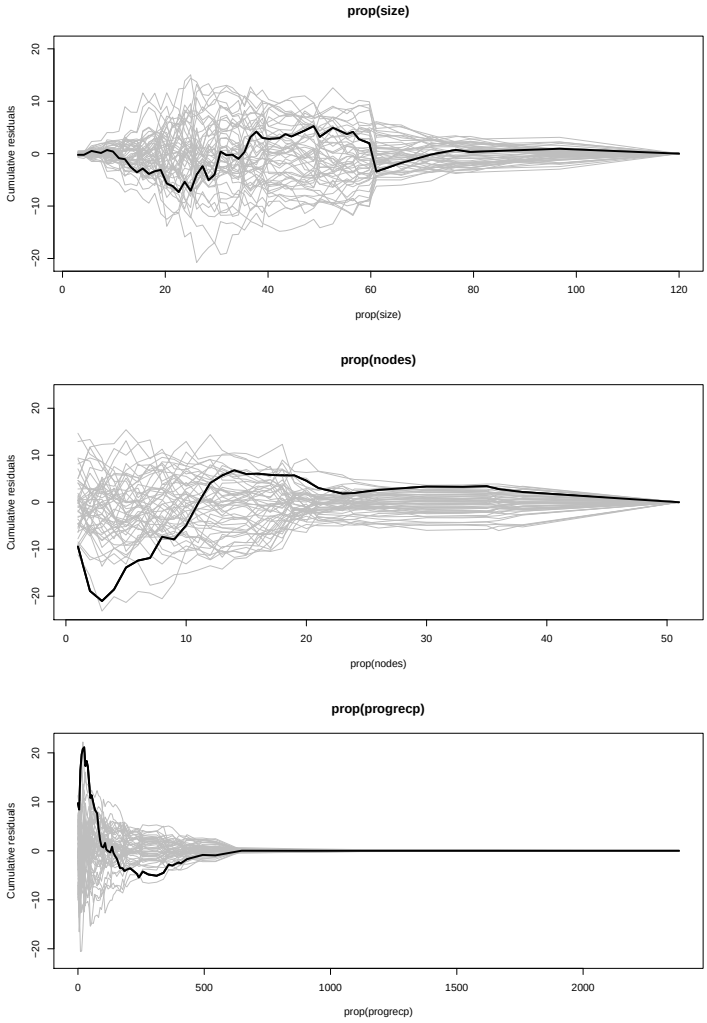

Figure 3.3: Simulation of cumulative martingale residuals against tumor size, number of lymph nodes and number of positive progesterone receptors in the Cox model with Tumor size, Positive lymph nodes, and Progesterone receptors as covariates

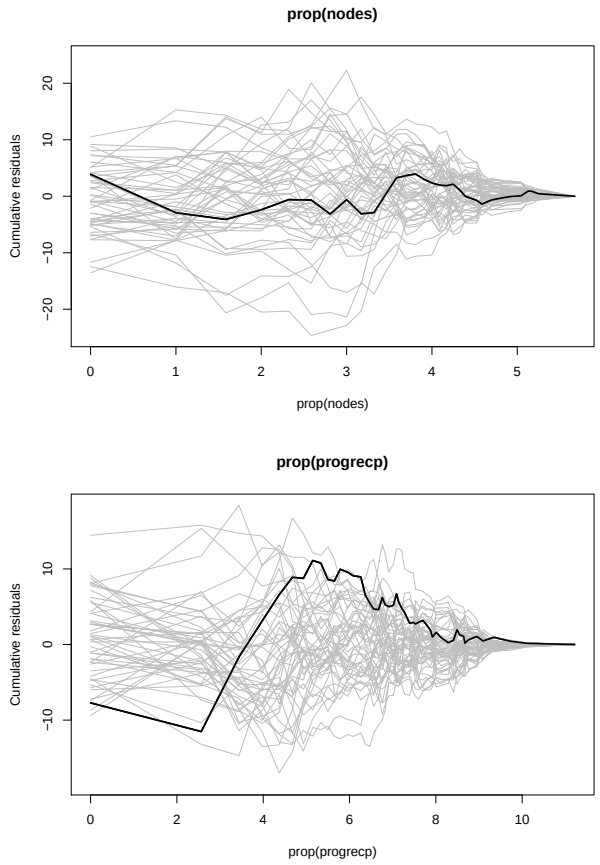

Figure 3.4: Simulation of cumulative martingale residuals against log base 2-transformed number of positive lymph nodes and number of progesterone receptors in the Cox model with Hormone Therapy, Log(Positive lymph nodes), and Log(Progesterone receptors) as covariates

replicates of (3.8), and compute the P-value as

$$P = \frac{1}{N}\sum_{n=1}^{N} I\left\{\sup_x |\widehat{W}_j^n(x)| \geq \sup_x |W_j(x)|\right\},$$

where $\widehat{W}_j^n(x)$ denotes the nth replicate of (3.8) from the simulation, and I takes value 1 if the expression in the bracket is true and otherwise 0. In fact, the P-value calculates the percentage of simulated cumulative residuals having a larger absolute supremum value than the observed one. For instance, if the observation plot is floating around 0 with a fairly small supremum, and there are 98 out of 100 simulated replicates higher than that, then P-value should be 0.98. This is a high P-value which indicates log-linearity.

We then go back to the GBCS example. To obtain an accurate computation for P-value, here we simulate using $N = 1000$. The test statistic considered is the supremum over x of $|W_j(x)|$. According to the summary statistics, the P-value with respect to tumor size is 0.772, while for the two covariates: number of lymph nodes and number of positive progesterone receptors, they are fairly small (0.007 and 0.005 respectively), indicating that log-linearity is not satisfied and hence log transformation is quite necessary. So we repeat the same process by using log base 2-transformation on the number of positive lymph nodes and number of progesterone receptors, the plot is shown in Figure 3.4. We can see that the cumulative martingale residuals are fluctuating around 0 along the y-axis, corresponding to the P-values of 0.992 and 0.293 respectively. In this log-transformation case the log-linearity is satisfied, thus we conclude that the two covariates should not be kept at the original scale and that log-transformation improves the model.

Figure 3.3 and 3.4 can be obtained in R using the *timereg* package, which was developed by Thomas.H.Scheike, see pages 202-204 in Martinussen and Scheike (2006). Note that when we use the package, the Cox model should be fitted by the command *cox.aalen* with all covariates as proportional effects, and then simply use the *cum.residuals* command to simulate replicates of cumulative residuals.

3.5 Special case of score process

3.5.1 Observation of score process

In this part we will look at the second special case of (3.4). Considering the situation when $f(\mathbf{x}_i) = \mathbf{x}_i$ and $\mathbf{x} = \infty$, it leads to the special case which can be written as

$$U(\widehat{\boldsymbol{\beta}}, t) = \sum_{i=1}^{n} \mathbf{x}_i \widehat{M}_i(t). \tag{3.9}$$

When specifying (3.9) to the jth covariate, we have that

$$U_j(\widehat{\boldsymbol{\beta}}, t) = \sum_{i=1}^{n} x_{ji} \widehat{M}_i(t). \tag{3.10}$$

Now we will explain how (3.9) is related to the score based on the partial likelihood (2.6). For Cox's regression model, the score function can be written as

$$U(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ \mathbf{x}_i - \frac{S^{(1)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)} \right\} dN_i(t), \tag{3.11}$$

where $L(\boldsymbol{\beta})$ is the partial likelihood in (2.6). By replacing τ with t in the expression (3.11), we obtain the score process

$$U(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} \int_0^{t} \left\{ \mathbf{x}_i - \frac{S^{(1)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)} \right\} dN_i(t).$$

Now the expression (3.9) can be written as

$$\begin{aligned}
\sum_{i=1}^{n} \mathbf{x}_i \widehat{M}_i(t) &= \sum_{i=1}^{n} \mathbf{x}_i \left\{ N_i(t) - \widehat{\Lambda}_i(t) \right\} \\
&= \sum_{i=1}^{n} \left\{ \int_0^t \mathbf{x}_i dN_i(u) - \int_0^t \mathbf{x}_i \frac{Y_i(u) \exp\left\{ \widehat{\boldsymbol{\beta}}^T \mathbf{x}_i \right\}}{S^{(0)}(\widehat{\boldsymbol{\beta}}, u)} dN_{\cdot}(u) \right\} \\
&= \sum_{i=1}^{n} \int_0^t \mathbf{x}_i dN_i(u) - \int_0^t \frac{S^{(1)}(\widehat{\boldsymbol{\beta}}, u)}{S^{(0)}(\widehat{\boldsymbol{\beta}}, u)} dN_{\cdot}(u) \\
&= \sum_{i=1}^{n} \int_0^t \left\{ \mathbf{x}_i - \frac{S^{(1)}(\widehat{\boldsymbol{\beta}}, u)}{S^{(0)}(\widehat{\boldsymbol{\beta}}, u)} \right\} dN_i(u).
\end{aligned} \tag{3.12}$$

Thus we have (3.9).

We make a score process plot of (3.9) to check for proportionality. For illustration purpose we use the German breast cancer study (GBCS) data. What we expect to see is that if the proportionality assumption is satisfied, the curve corresponding to each covariate should fluctuate around 0 over the entire course of the study. Moreover, we should notice that by definition of $\widehat{\boldsymbol{\beta}}$, (3.10) will take the value zero when $t = \tau$. We start with the non-transformation case. As we can see in Figure 3.5 three covariates: Tumor size, Positive lymph nodes and Progesterone receptors have been used for the fit of the Cox model. Although the plots for the first two covariates look fine, when it comes to the Progesterone receptors, the score is declining rapidly from the very start to the middle of the study period before going back to around zero again in the 4th year. Then we check the score process for the

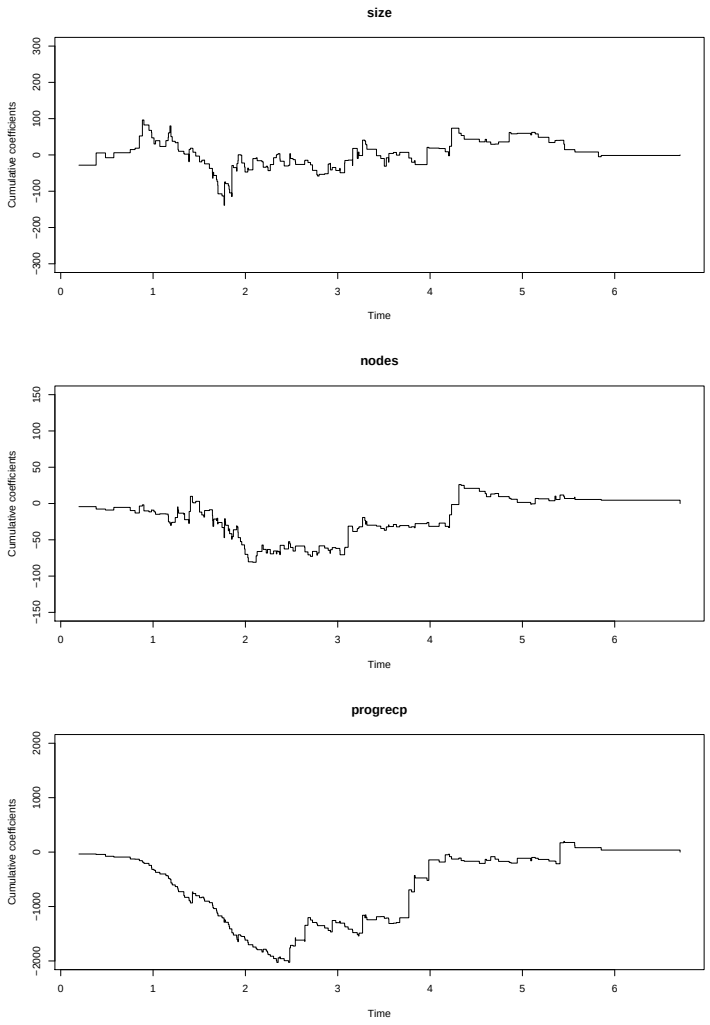

Figure 3.5: Score process against time for tumor size, number of lymph nodes and number of positive progesterone in the Cox model with Tumor size, Positive lymph nodes, and Progesterone receptors as covariates

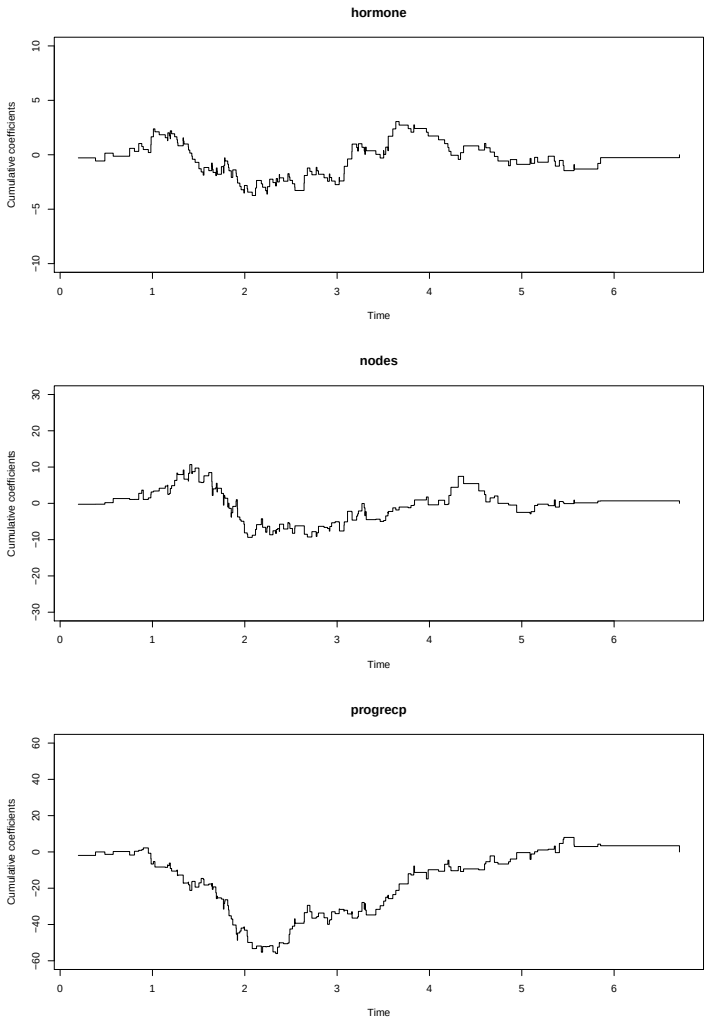

Figure 3.6: Score process against time for hormone therapy, log base 2-transformed number of positive lymph nodes and number of progesterone receptors in the Cox model with Hormone Therapy, Log(Positive lymph nodes), and Log(Progesterone receptors) as covariates

Cox model using Hormone therapy, Positive lymph nodes and Progesterone receptors as covariates, where log based 2-transformation has been applied on the number of positive lymph nodes and the number of progesterone receptors. As shown in Figure 3.6, we can see that the plots corresponding to hormone therapy and positive lymph nodes are quite smooth but the negative score trend for progesterone receptors still exist. We will later use simulated score process to have a further check of the non-proportional effect. This is needed to know how the plots may look like when proportionality is satisfied.

It is worth mentioning that the commands for obtaining Figures 3.5 and 3.6 are based on the Schoenfeld residuals, see page 360 in Klen and Moeschberger (2003). The Schoenfeld residuals are the expression in curly brackets in the last line of (3.12). When given for all $\widetilde{T}_i$ with $D_i = 1$, by the command $residuals(coxfit, type = "schoenfeld")$, where $coxfit$ represents the Cox model fitted earlier, we can obtain the Schoenfeld residuals matrix. It is a three-dimensional vector, with one component for each covariate in the Cox model. Then by using the *cumsum* command we can sum up the Schoenfeld residuals for $\widetilde{T}_i \leq t$, and hence calculate the expression (3.12) with respect to the jth covariate.

3.5.2 Simulation of score process

We will at this stage specialize the process (3.6) to the case in (3.9), so as to obtain the simulated score process. For each individual in the survival data set we have the observations $\widetilde{T}_i$, D_i and $\mathbf{x}_i$ fixed. We start by specializing (3.6) to the case when $f(\mathbf{x}_i) = \mathbf{x}_i$ and $\mathbf{x} = (\infty, \infty, ..., \infty, ...\infty)^T$, then specialize to censored survival data. Then we obtain that

$$\begin{aligned}\widehat{U}(\widehat{\boldsymbol{\beta}}, t) = &\sum_{i=1}^{n} \left\{ \mathbf{x}_i - \frac{S^{(1)}(\widehat{\boldsymbol{\beta}}, \widetilde{T}_i)}{S^{(0)}(\widehat{\boldsymbol{\beta}}, \widetilde{T}_i)} \right\} G_i D_i \\ &- \sum_{i=1}^{n} Y_i(\widetilde{T}_i) \exp\left\{ \widehat{\boldsymbol{\beta}}^T \mathbf{x}_i \right\} \mathbf{x}_i \left\{ \mathbf{x}_i - \frac{S^{(1)}(\widehat{\boldsymbol{\beta}}, \widetilde{T}_i)}{S^{(0)}(\widehat{\boldsymbol{\beta}}, \widetilde{T}_i)} \right\}^T \widehat{A}_0(\widetilde{T}_i) \qquad (3.13) \\ &\times I(\widehat{\boldsymbol{\beta}})^{-1} \sum_{i=1}^{n} \left\{ \mathbf{x}_i - \frac{S^{(1)}(\widehat{\boldsymbol{\beta}}, \widetilde{T}_i)}{S^{(0)}(\widehat{\boldsymbol{\beta}}, \widetilde{T}_i)} \right\} G_i D_i.\end{aligned}$$

Now we will have a check of proportionality using (3.13). Considering the German breast cancer study (GBCS) data, when using the original data without transformation, we have the plot shown in Figure 3.7, where both the experimental and simulated processes are given. The dark lines are the same as in Figure 3.5. The plot for tumor size shows that the observed score is falling within the simulation process quite nicely. Moreover, the score process

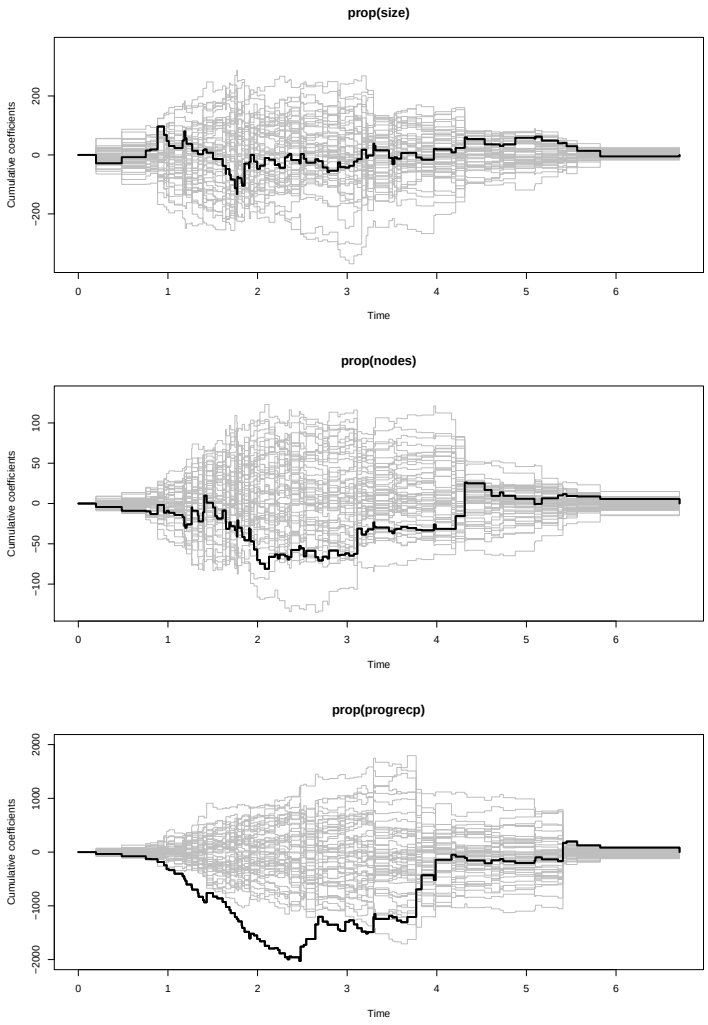

Figure 3.7: Simulation of score process against time for tumor size, number of lymph nodes and number of positive progesterone in the Cox model with Tumor size, Positive lymph nodes, and Progesterone receptors as covariates

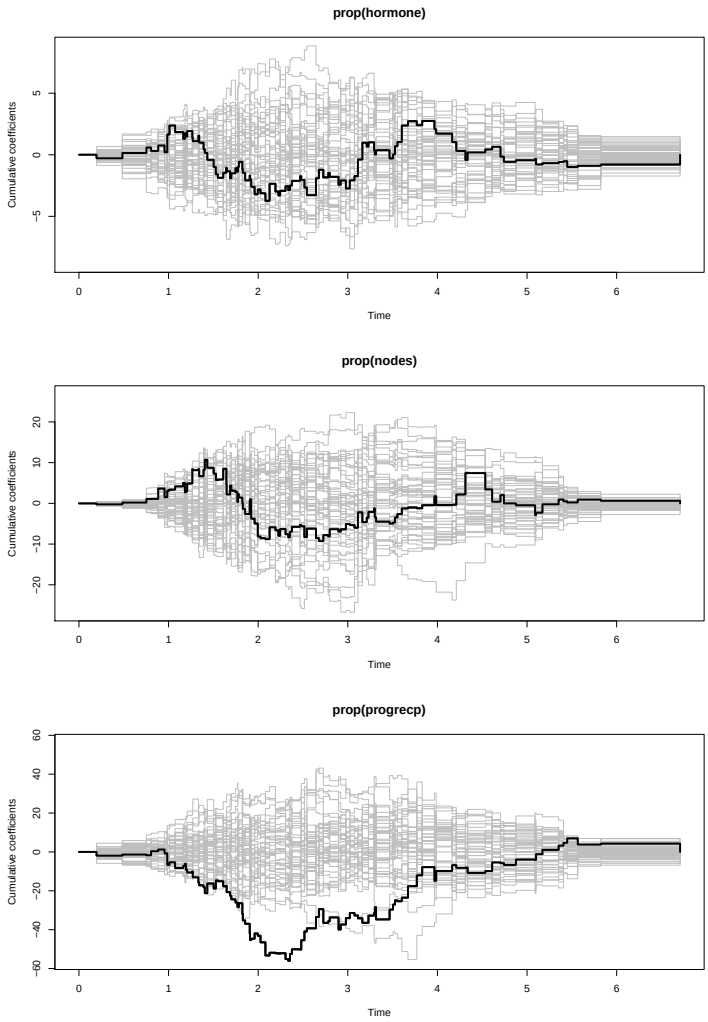

Figure 3.8: Simulation of score process against time for hormone therapy, log base 2-transformed number of positive lymph nodes and number of progesterone receptors in the Cox model with Hormone Therapy, Log(Positive lymph nodes), and Log(Progesterone receptors) as covariates

for positive lymph nodes is at the lower edge of the simulated area. By looking at the plot corresponding to the progesterone receptors, it seems like the score is going extremely negative when comparing with the simulated score process, especially during the study period between the 2nd year and the 3rd year, indicating that the fit of the Cox model in this case is not so impressive. Note that this is also seen by the summary statistics, where the P-values of the three covariates: Tumor size, Positive lymph nodes and Progesterone receptors, calculated by a simulation of 1000 replicates, are 0.762, 0.388 and 0.003 respectively. When applying log based 2-transformation on the number of positive lymph nodes and the number of progesterone receptors, we have the results given in Figure 3.8, where the dark lines are the same as in Figure 3.6. Specifically, the hormone therapy has a quite proportional effect as it falls right in the middle area of the simulated data. The plot with respect to positive lymph nodes is improved after log-transformation and is continually fluctuating around 0. But clearly there is no improvement regarding the covariate progesterone receptors since it still departs from the simulated score process. The same result is reflected from the summary statistics. For the three covariates: Hormone Therapy, Positive lymph nodes and Progesterone receptors, the corresponding P-values are 0.754, 0.781 and 0.007. Thus there is an obvious non-proportional effect of progesterone receptors.

To conclude, we have illustrated how model checking is performed for cohort studies by using cumulative sums of martingale-based residuals. Regarding the GBCS data example, it turns out that the best model that we obtained in Section 2.4.2 using the original data does not satisfy both the log-linearity assumption and the proportionality assumption. As a matter of fact, the model is improved after log based 2-transformation on the covariates, nonetheless, through model checking it reveals that there is still one covariate that violates the proportionality assumption. It needs to make clear that this thesis work only focuses on model checking, therefore we will not work on finding the final Cox model for the GBCS data. So far we have seen that cohort studies are always required to process all the information of individuals, which can be quite costly. In the continuation of our studies on model checking, we will extend the cumulative sums of martingale-based residuals to nested case-control data with two different sampling methods. Then we will verify if the model checking for nested case-control data is able to give a similar result with that for cohort data.

Chapter 4

Nested case-control studies

The chapter is based on the papers by Borgan and Samuelsen (2013) and Borgan and Langholz (2007). In Section 4.1 we introduce two methods for sampling of controls, followed by Section 4.2 where the counting process formulation for nested case-control data is given. In Section 4.3 we obtain the partial likelihood of regression coefficients in Cox model for nested case-control data. After that we present the Radiation and breast cancer data in Section 4.4, which is an example for nested case-control data. Then in Section 4.5, the martingale residual processes are derived in a similar manner as in Chapter 3. Section 4.6 is the material for showing how the program for cohort can be tricked to compute for nested case-control simulation. Finally, in Sections 4.7 and 4.8, we specialize the processes in order to check for the two model assumptions using nested case-control data.

In previous chapters we have covered the methods appropriate for cohort studies. Cohort studies are taking all the information of individuals at risk into consideration, regardless of how many of them have actually experienced the event of interest, which gives a complete and detailed analysis. But the drawbacks are also obvious, cohort studies can be quite costly and time-consuming to perform.

Nested case-control studies is an useful alternative to cohort studies. The idea is that for each case when we observe an event of interest, instead of using data for all the individuals at risk as we did in cohort studies, now we only select a small number of controls. Still most of the information in the cohort is captured. This way it saves us a lot of time and effort in terms of data collection and checking, thereby making the studies much more cost efficient.

4.1 Sampling of controls

There are two important sampling designs which can be used in nested case-control studies. We start off by introducing the simple random sampling. Assume that we observe an event of interest at time t. Then by simple random sampling we pick $m-1$ controls from the remaining $Y(t)-1$ individuals in the risk set $R(t)$, where m is a number that needs to be defined. It is typically so that m is chosen to take a small value like 2 or 4. Now we have obtained a sampled risk set which is denoted by $\widetilde{R}(t)$. It contains the observed case together with the $m-1$ controls. Notice that controls are selected independently and do not contain any information of each other at the event times.

The other method for sampling of controls is called counter-matched sampling, also known as stratified sampling. The idea is to use the information available for everyone in the cohort to classify each individual at risk into one of S distinct strata. We define $R_s(t)$ as the subset of $R(t)$ that corresponds to stratum s, where the number at risk just before time t is $Y_s(t) = |R_s(t)|$. If an event of interest is observed at time t in stratum $s(i)$, we will sample m_s different controls from $R_s(t)$ where $s \neq s(i)$. However, we only sample $m_{s(i)} - 1$ controls from $R_{s(i)}(t)$ because the sampled risk set already contains the observed case. Note that we obtain a total number of $m = \sum_{s=1}^{S} m_s$ individuals in the sampled risk set. In addition, it needs to be pointed out that the classification into strata may depend on time, and the information associated with the stratification should be known before time t.

4.2 Counting process formulation

We will introduce the counting process formulation for nested case-control data. Recall the counting process in (2.2) as we have described previously. When it comes to nested case-control studies, we define a potential sampled risk set as $\mathbf{r}$, let $t_1 < t_2 < ... < t_d$ be the observed event times and i_j be the corresponding cases. Then the counting process formulation can be written as

$$N_{i,\mathbf{r}}(t) = \sum_{j \geq 1} I(t_j \leq t, i_j = i, \widetilde{R}(t_j) = \mathbf{r}). \tag{4.1}$$

Note that $N_{i,\mathbf{r}}(t)$ in (4.1) counts how many times in $[0,t]$ that individual i has an event and $\mathbf{r}$ is the sampled risk set. Referring to the Cox model in (2.5), the intensity processes $\lambda_{i,\mathbf{r}}(t)$ of the counting process (4.1) can be written as

$$\lambda_{i,\mathbf{r}}(t) = Y_i(t)\alpha_i(t)\pi(\mathbf{r}|t,i) = Y_i(t)\alpha_0(t)\exp\left\{\boldsymbol{\beta}^T\mathbf{x}_i\right\}\pi(\mathbf{r}|t,i), \tag{4.2}$$

where $\pi(\mathbf{r}|t,i)$ defines the conditional probability of selecting $\mathbf{r}$ as the sampled risk set, given all the information in the past and also that individual i has

experienced the event at time t. For a set $\mathbf{r} \subset R(t)$ of size m with $i \in \mathbf{r}$ we may for simple random sampling write

$$\pi(\mathbf{r}|t,i) = \frac{1}{\binom{Y(t)-1}{m-1}} = \frac{Y(t)}{m}\frac{1}{\binom{Y(t)}{m}}. \tag{4.3}$$

In terms of counter-matched sampling, for the sampled risk set $\mathbf{r} \subset R(t)$ with $i \in \mathbf{r}$, and when the size of $\mathbf{r} \cap R_s(t)$ is m_s for $s = 1, ..., S$, the probability of selecting $\mathbf{r}$ as the sampled risk set is given as

$$\pi(\mathbf{r}|t,i) = \left\{ \binom{Y_{s(i)}(t)-1}{m_{s(i)}-1} \prod_{s \neq s(i)} \binom{Y_s(t)}{m_s} \right\}^{-1} = \frac{Y_{s(i)}(t)}{m_{s(i)}} \left\{ \prod_{s=1}^{S} \binom{Y_s(t)}{m_s} \right\}^{-1}. \tag{4.4}$$

Note that we for both sampling designs may write

$$\pi(\mathbf{r}|t,i) = w_i(t)\pi(\mathbf{r}|t), \tag{4.5}$$

where the weights $w_i(t)$ are the leading factor on the right-hand side of (4.3) and (4.4). The second factor $\pi(\mathbf{r}|t)$ is a probability distribution over all possible sampled risk sets $\mathbf{r}$. In other words, for the simple random sampling design it is defined that $w_i(t) = \frac{Y(t)}{m}$, $\pi(\mathbf{r}|t) = \frac{1}{\binom{Y(t)}{m}}$, while for the counter-matched sampling design $w_i(t) = \frac{Y_{s(i)}(t)}{m_{s(i)}}$, $\pi(\mathbf{r}|t) = \left\{ \prod_{s=1}^{S} \binom{Y_s(t)}{m_s} \right\}^{-1}$.

4.3 Partial likelihood

We will derive the partial likelihood for $\boldsymbol{\beta}$ when it comes to nested case-control data. Let $\pi(i|t,\mathbf{r})$ denote the conditional probability of individual i experiencing the event at time t, given the past information and that an event is observed in the sampled risk set $\mathbf{r}$. Thus it takes the form

$$\pi(i|t,\mathbf{r}) = \frac{\lambda_{i,\mathbf{r}}(t)}{\sum_{k \in \mathbf{r}} \lambda_{k,\mathbf{r}}(t)}. \tag{4.6}$$

We first look at the simple random sampling case. With reference to (4.2) and (4.3) we further obtain that

$$\begin{aligned} \pi(i|t,\mathbf{r}) =& \frac{Y_i(t)\exp\left\{\boldsymbol{\beta}^T\mathbf{x}_i\right\}\pi(\mathbf{r}|t,i)}{\sum_{k\in\mathbf{r}} Y_k(t)\exp\left\{\boldsymbol{\beta}^T\mathbf{x}_k\right\}\pi(\mathbf{r}|t,k)} \\ =& \frac{Y_i(t)\exp\left\{\boldsymbol{\beta}^T\mathbf{x}_i\right\}}{\sum_{k\in\mathbf{r}} Y_k(t)\exp\left\{\boldsymbol{\beta}^T\mathbf{x}_k\right\}}, \end{aligned} \tag{4.7}$$

where the last line is derived since the sampling probabilities $\pi(\mathbf{r}|t,i)$ and $\pi(\mathbf{r}|t,k)$ are equal and hence can be cancelled. By multiplying (4.7) over all event times, cases and sampled risk sets, we can obtain the partial likelihood

$$L_{ncc}(\boldsymbol{\beta}) = \prod_{j=1}^{d} \pi\left(i_j|t_j, \widetilde{R}(t_j)\right) = \prod_{j=1}^{d} \frac{\exp\left\{\boldsymbol{\beta}^T \mathbf{x}_{i_j}\right\}}{\sum_{k\in\widetilde{R}(t_j)} \exp\left\{\boldsymbol{\beta}^T \mathbf{x}_k\right\}}. \tag{4.8}$$

Notice that the partial likelihood in (4.8) is similar to the one for cohort data in (2.6). But we sum over individuals in the sampled risk set instead of everyone of risk. Now we will obtain the partial likelihood for $\boldsymbol{\beta}$ when controls are selected by counter-matched sampling. Let $w_i(t) = \frac{Y_{s(i)}(t)}{m_{s(i)}}$, by plugging (4.2) and (4.4) into (4.6), we have that

$$\begin{aligned} \pi\left(i|t,\mathbf{r}\right) =& \frac{Y_i(t)\exp\left\{\boldsymbol{\beta}^T\mathbf{x}_i\right\}\pi(\mathbf{r}|t,i)}{\sum_{k\in\mathbf{r}} Y_k(t)\exp\left\{\boldsymbol{\beta}^T\mathbf{x}_k\right\}\pi(\mathbf{r}|t,k)} \\ =& \frac{Y_i(t)\exp\left\{\boldsymbol{\beta}^T\mathbf{x}_i\right\}w_i(t)}{\sum_{k\in\mathbf{r}} Y_k(t)\exp\left\{\boldsymbol{\beta}^T\mathbf{x}_k\right\}w_k(t)}, \end{aligned} \tag{4.9}$$

where the last line is obtained by cancelling the common factor $\left\{\prod_{s=1}^{S}\binom{Y_s(t)}{m_s}\right\}^{-1}$ of the sampling probabilities $\pi(\mathbf{r}|t,i)$ and $\pi(\mathbf{r}|t,k)$. Then it follows that the partial likelihood is given as

$$L_{cm}(\boldsymbol{\beta}) = \prod_{j=1}^{d} \pi\left(i_j|t_j, \widetilde{R}(t_j)\right) = \prod_{j=1}^{d} \frac{\exp\left\{\boldsymbol{\beta}^T \mathbf{x}_{i_j}\right\} w_{i_j}(t_j)}{\sum_{k\in\widetilde{R}(t_j)} \exp\left\{\boldsymbol{\beta}^T \mathbf{x}_k\right\} w_k(t_j)} \tag{4.10}$$

At this point, we can see that the partial likelihood expressions for simple random sampling (4.8) differ from counter-matched sampling (4.10) in terms of the weights. To be more specific, counter-matched sampling gives a weighted partial likelihood. But for simple random sampling, weights have been cancelled since they are equal for all individuals.

4.4 Radiation and breast cancer

Here we will introduce a data set that will be used to illustrate nested case-control sampling. Data are obtained from two hospitals in Massachusetts, where a total of 1720 female patients with tuberculosis were receiving medical treatment during the years 1930-1956 (Hrubec et al., 1989). We will focus on the breast cancer risk due to radiation exposure. During the course of tuberculosis therapy, patients were examined by different medical tests with a majority that involved X-ray fluoroscopies, which was a cause of the increase in radiation doses. Indeed, 1022 patients had taken lung examinations through X-ray fluoroscopies for 101 times in average, while the remaining 698

Table 4.1: Cox regression analysis for the effect of the covariate Dose in rad for cohort design, simple random sampling design, and counter-matched sampling design with m-1=2 controls

Covariate	$\hat{\beta}_j$	se($\hat{\beta}_j$)	z	Wald test	Pr(>|z|)
cohort	0.296	0.103	2.881	8.3	0.004
simple random	0.348	0.165	2.115	4.47	0.034
counter-matching	0.274	0.124	2.204	4.86	0.028

patients were examined by other non-fluoroscopic means that resulted in no radiation exposure. By the end of the study in 1980, there were 75 breast cancer cases being observed. As a matter of fact, 54 of them were identified with radiation dose. The data reveal that breast cancer risk is related to the dose in radiation, and more than expected cases of breast cancer occurred among those who had experienced radiation exposure in their lives.

Even though all data in this example are available for a cohort study, we would like to sample for illustration. The reason is that we observe a small number of events versus a high number of individuals, which is a sign that nested case-control sampling could have been a good design. At this point we will select samples by applying both the simple random sampling method and the counter-matched sampling method. Then by comparing them with the Cox model fitted with cohort data, we will be able to see how good results we may obtain from nested case-control studies. We first look at the simple random sampling design. To fit the Cox model to nested case-control data, we use Dose in radiation (mean=0.569) as the covariate. It is suggested that $m-1 = 2$ controls per case should be chosen. Therefore, at each event time we sample two controls from the remaining individuals in the risk set. Further, we will describe how the counter-matching is performed. For obtaining strata we will consider the covariate: Number of fluoroscopy examinations, which has a mean 60.15. Three strata for the data are then created according to the number of examinations. Specifically, the first stratum contains all with no fluoroscopy examinations, followed by the second and third stratum having 1-149, 150 and more of examinations respectively. Also, we select $m - 1 = 2$ controls per case to fit the Cox model, which implies $m_1 = m_2 = m_3 = 1$ for the three strata. In other words, we are not sampling any controls from the stratum which the case belongs to, but only selecting one control from each of the other two strata. Finally, Table 4.1 shows the Cox regression analysis for the effect of the covariate Dose in rad for the cohort design, the simple random sampling design, and the counter-matched sampling design. By looking at the P-values we see that all three designs give a significant effect of dose. The estimated parameters obtained from the nested case-control models are quite close to that from the cohort. Apparently, counter-matched sampling shows an even better result than simple random sampling. The standard errors of the two nested case-control estimates are about 60% and

20% larger compared with using the cohort design. In conclusion, we find that nested case-control designs can be fairly effective approaches.

4.5 Martingale residual processes

In this section we also for nested case-control studies would like to check the assumptions of the Cox model. Furthermore, we will study how the methods for the cohort based on martingale residual processes may be adapted to nested case-control data. To start with, we will derive the martingale residual processes for nested case-control data. Recall the counting process (4.1) and intensity process (4.2) for nested case-control data. Further, we can write the local square integrable martingales as

$$M_{i,\mathbf{r}}(t) = N_{i,\mathbf{r}}(t) - \Lambda_{i,\mathbf{r}}(t), \tag{4.11}$$

where

$$\Lambda_{i,\mathbf{r}}(t) = \int_0^t \lambda_{i,\mathbf{r}}(u)du = \int_0^t Y_i(u)\alpha_0(u)\exp\left\{\boldsymbol{\beta}^T\mathbf{x}_i\right\}\pi(\mathbf{r}|u,i)du. \tag{4.12}$$

With reference to (2.7) we introduce

$$\widehat{A}_{0\mathbf{r}}(t) = \sum_{t_j\le t,\widetilde{R}(t_j)=\mathbf{r}} \frac{1}{\sum_{l\in\mathbf{r}}\exp\left\{\widehat{\boldsymbol{\beta}}^T\mathbf{x}_l\right\}\pi(\mathbf{r}|t_j)w_l(t_j)}. \tag{4.13}$$

Here the weights $w_l(t_j)$ and the $\pi(\mathbf{r}|t_j)$ are given by the decomposition (4.5). Similar with what we did earlier for cohort data, by plugging the maximum partial likelihood estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ and $d\widehat{A}_{0\mathbf{r}}(t)$ for $\alpha_0(u)du$ into (4.12), it follows that

$$\begin{aligned}
\widehat{\Lambda}_{i,\mathbf{r}}(t) &= \int_0^t Y_i(u)\exp\left\{\widehat{\boldsymbol{\beta}}^T\mathbf{x}_i\right\}\pi(\mathbf{r}|u)w_i(u)d\widehat{A}_{0\mathbf{r}}(u) \\
&= \int_0^t \frac{Y_i(u)\exp\left\{\widehat{\boldsymbol{\beta}}^T\mathbf{x}_i\right\}\pi(\mathbf{r}|u)w_i(u)}{\sum_{l\in\mathbf{r}}\exp\left\{\widehat{\boldsymbol{\beta}}^T\mathbf{x}_l\right\}\pi(\mathbf{r}|u)w_l(u)}dN_{\mathbf{r}}(u) \\
&= \sum_{t_j\le t,\widetilde{R}(t_j)=\mathbf{r}} \frac{Y_i(t_j)\exp\left\{\widehat{\boldsymbol{\beta}}^T\mathbf{x}_i\right\}w_i(t_j)}{\sum_{l\in\mathbf{r}}\exp\left\{\widehat{\boldsymbol{\beta}}^T\mathbf{x}_l\right\}w_l(t_j)}.
\end{aligned} \tag{4.14}$$

Note that $N_{\mathbf{r}}(t)$ is the aggregation of the counting process (4.1) over all individuals, which can be written as

$$N_{\mathbf{r}}(t) = \sum_{i\in\mathbf{r}} N_{i,\mathbf{r}}(t) = \sum_{j\ge1} I(t_j\le t, \widetilde{R}(t_j)=\mathbf{r}).$$

According to (4.11), the martingale residual processes can be written as

$$\widehat{M}_{i,\mathbf{r}}(t) = N_{i,\mathbf{r}}(t) - \widehat{\Lambda}_{i,\mathbf{r}}(t). \tag{4.15}$$

It needs to be mentioned that since the majority of the martingale residual processes (4.15) will end up in zero, in practice it might not be working so impressive. However, it is an important building brick of the grouped martingale residual processes where we aggregate over individuals and sampled risk sets.

We define P as the set that contains all subsets of $\{1, 2, ..., n\}$, and let $P_i = \{\mathbf{r} : \mathbf{r} \in P, i \in \mathbf{r}\}$ be the sets $\mathbf{r}$ of P that contain individual i. Referring to the process (3.4), where we have discussed the cumulative sums of the martingale-based residuals for cohort data, our objective now is to generalize it to nested case-control data. Thus we introduce the process

$$\begin{aligned}
\widetilde{W}(t, \mathbf{x}) =& \sum_{i=1}^{n} \sum_{\mathbf{r} \in P_i} f(\mathbf{x}_i) I(\mathbf{x}_i \leq \mathbf{x}) \widehat{M}_{i,\mathbf{r}}(t) \\
=& \sum_{i=1}^{n} \sum_{\mathbf{r} \in P_i} f(\mathbf{x}_i) I(\mathbf{x}_i \leq \mathbf{x}) \left\{ N_{i,\mathbf{r}}(t) - \widehat{\Lambda}_{i,\mathbf{r}}(t) \right\} \\
=& \sum_{i=1}^{n} \sum_{\mathbf{r} \in P_i} f(\mathbf{x}_i) I(\mathbf{x}_i \leq \mathbf{x}) N_{i,\mathbf{r}}(t) - \sum_{i=1}^{n} \sum_{\mathbf{r} \in P_i} f(\mathbf{x}_i) I(\mathbf{x}_i \leq \mathbf{x}) \widehat{\Lambda}_{i,\mathbf{r}}(t) \\
=& \sum_{i=1}^{n} f(\mathbf{x}_i) I(\mathbf{x}_i \leq \mathbf{x}) N_i(t) - \sum_{\mathbf{r} \in P} \sum_{i \in \mathbf{r}} f(\mathbf{x}_i) I(\mathbf{x}_i \leq \mathbf{x}) \widehat{\Lambda}_{i,\mathbf{r}}(t) \\
=& \sum_{i=1}^{n} f(\mathbf{x}_i) I(\mathbf{x}_i \leq \mathbf{x}) N_i(t) \\
& - \sum_{t_j \leq t} \frac{\sum_{i \in \widetilde{R}(t_j)} f(\mathbf{x}_i) I(\mathbf{x}_i \leq \mathbf{x}) \exp\left\{ \widehat{\boldsymbol{\beta}}^T \mathbf{x}_i \right\} w_i(t_j)}{\sum_{l \in \widetilde{R}(t_j)} \exp\left\{ \widehat{\boldsymbol{\beta}}^T \mathbf{x}_l \right\} w_l(t_j)}.
\end{aligned} \tag{4.16}$$

It is worth noting that the process (3.4) for cohort data may also be written as

$$\begin{aligned}
W(t, \mathbf{x}) =& \sum_{i=1}^{n} f(\mathbf{x}_i) I(\mathbf{x}_i \leq \mathbf{x}) \left\{ N_i(t) - \widehat{\Lambda}_i(t) \right\} \\
=& \sum_{i=1}^{n} f(\mathbf{x}_i) I(\mathbf{x}_i \leq \mathbf{x}) N_i(t) - \sum_{i=1}^{n} f(\mathbf{x}_i) I(\mathbf{x}_i \leq \mathbf{x}) \widehat{\Lambda}_i(t) \\
=& \sum_{i=1}^{n} f(\mathbf{x}_i) I(\mathbf{x}_i \leq \mathbf{x}) N_i(t) - \sum_{T_j \leq t} \frac{\sum_{l \in R_j} f(\mathbf{x}_i) I(\mathbf{x}_i \leq \mathbf{x}) \exp\left\{ \widehat{\boldsymbol{\beta}}, \mathbf{x}_i \right\}}{\sum_{l \in R_j} \exp\left\{ \widehat{\boldsymbol{\beta}}, \mathbf{x}_l \right\}}.
\end{aligned} \tag{4.17}$$

Therefore we see that the process (4.16) for nested case-control data is similar to the process (3.4) for cohort data. To be more specific, the first term of both processes which represent the observations are exactly the same, the only difference lies in the second term where the nested case-control process has a sampled risk set and weight.

The approximation (3.5) may also be generalized to nested case-control data. Following the arguments in Section 5.2 of Borgan and Langholz (2007), we may show that the process (4.16) may be approximated as follows (Borgan, personal communication)

$$\begin{aligned}\widetilde{W}(t,\mathbf{x}) \approx & \sum_{\mathbf{r}\in P}\sum_{i\in\mathbf{r}}\int_0^t \left\{f(\mathbf{x}_i)I(\mathbf{x}_i\le\mathbf{x}) - g_{\mathbf{r}}(\boldsymbol{\beta}_0,u,\mathbf{x})\right\} dM_{i,\mathbf{r}}(u) \\ & -\sum_{\mathbf{r}\in P}\sum_{i\in\mathbf{r}}\int_0^t Y_i(u)\exp\left\{\boldsymbol{\beta}_0{}^T\mathbf{x}_i\right\} w_i(u) f(\mathbf{x}_i)I(\mathbf{x}_i\le\mathbf{x}) \left\{\mathbf{x}_i - \frac{S_{\mathbf{r}}^{(1)}(\boldsymbol{\beta}_0,u)}{S_{\mathbf{r}}^{(0)}(\boldsymbol{\beta}_0,u)}\right\}^T \\ & \pi(\mathbf{r}|u)\alpha_0(u)du \times I(\boldsymbol{\beta}_0)^{-1}\sum_{\mathbf{r}\in P}\sum_{i\in\mathbf{r}}\int_0^\tau \left\{\mathbf{x}_i - \frac{S_{\mathbf{r}}^{(1)}(\boldsymbol{\beta}_0,u)}{S_{\mathbf{r}}^{(0)}(\boldsymbol{\beta}_0,u)}\right\} dM_{i,\mathbf{r}}(u).\end{aligned} \tag{4.18}$$

Note that here $I(\boldsymbol{\beta})$ is based on partial likelihood for nested case-control data, and we have

$$g_{\mathbf{r}}(\boldsymbol{\beta},u,\mathbf{x}) = \frac{\sum_{l\in\mathbf{r}} Y_i(u)\exp\left\{\boldsymbol{\beta}^T\mathbf{x}_i\right\} f(\mathbf{x}_i)I(\mathbf{x}_i\le\mathbf{x})w_i(u)}{S_{\mathbf{r}}^{(0)}(\boldsymbol{\beta},u)},$$

$$S_{\mathbf{r}}^{(0)}(\boldsymbol{\beta},u) = \sum_{l\in\mathbf{r}} Y_i(u)\exp\left\{\boldsymbol{\beta}^T\mathbf{x}_i\right\} w_i(u),$$

and

$$S_{\mathbf{r}}^{(1)}(\boldsymbol{\beta},u) = \sum_{l\in\mathbf{r}} Y_i(u)\mathbf{x}_i\exp\left\{\boldsymbol{\beta}^T\mathbf{x}_i\right\} w_i(u).$$

It should be noted that (4.18) is similar to (3.5) for cohort data. Further, in order to find the distribution of the process (4.16), similarly to cohort data, we need to simulate $\widetilde{W}(t,\mathbf{x})$. We do that by keeping all observations fixed, replacing $\boldsymbol{\beta}_0$ by $\widehat{\boldsymbol{\beta}}$, $\pi(\mathbf{r}|u)\alpha_0(u)du$ by $\pi(\mathbf{r}|u)d\widehat{A}_{0\mathbf{r}}(u)$ and $dM_{i,\mathbf{r}}(t)$ by $G_{i,\mathbf{r}}dN_{i,\mathbf{r}}(t)$, where the $G_{i,\mathbf{r}}$'s are independent $N(0,1)$. This gives us processes $\widehat{W}^*(t,\mathbf{x})$ that have the same distribution as $\widetilde{W}(t,\mathbf{x})$, if the model is correctly

specified. This gives

$$\begin{aligned}\widehat{W}^*(t,\mathbf{x}) =& \sum_{\mathbf{r}\in P}\sum_{i\in\mathbf{r}}\int_0^t \left\{f(\mathbf{x}_i)I(\mathbf{x}_i\le\mathbf{x}) - g_{\mathbf{r}}(\widehat{\boldsymbol{\beta}},u,\mathbf{x})\right\} G_{i,\mathbf{r}}dN_{i,\mathbf{r}}(u)\\ &-\sum_{\mathbf{r}\in P}\sum_{i\in\mathbf{r}}\int_0^t Y_i(u)\exp\left\{\widehat{\boldsymbol{\beta}}^T\mathbf{x}_i\right\} w_i(u)f(\mathbf{x}_i)I(\mathbf{x}_i\le\mathbf{x})\left\{\mathbf{x}_i - \frac{S_{\mathbf{r}}^{(1)}(\widehat{\boldsymbol{\beta}},u)}{S_{\mathbf{r}}^{(0)}(\widehat{\boldsymbol{\beta}},u)}\right\}^T\\ &\pi(\mathbf{r}|u)d\widehat{A}_{0\mathbf{r}}(u)\times I(\widehat{\boldsymbol{\beta}})^{-1}\sum_{\mathbf{r}\in P}\sum_{i\in\mathbf{r}}\int_0^\tau\left\{\mathbf{x}_i - \frac{S_{\mathbf{r}}^{(1)}(\widehat{\boldsymbol{\beta}},u)}{S_{\mathbf{r}}^{(0)}(\widehat{\boldsymbol{\beta}},u)}\right\} G_{i,\mathbf{r}}dN_{i,\mathbf{r}}(u).\end{aligned} \tag{4.19}$$

Now we are able to see that the process (4.19) is quite similar to (3.6) for cohort data.

4.6 Material on computing

It is generally acknowledged that Cox regression methodology does not make any difference between one individual over a time period and different individuals covering the same time period. Thus the thought is that the Cox regression program for cohort data may be tricked to do estimation from case-control set organized data. To achieve this we create a data set that contains all the observed cases together with the sampled controls. We also make it so that each case and its controls are "observed" over the same period, which is a tiny time interval just before the event time. In this way all individuals in the risk sets will only contribute at the risk set failure time. Due to the similarity of (4.16) and (4.17) as we mentioned earlier, by using case-control set organized data, we can plot the observations of cumulative martingale residuals for nested case-control data exactly the same way as for cohort. Further, we will run the simulation processes (4.23) by using *timereg* program as we did for cohort data. To make sure that the trick really works, we need to rewrite (3.6) along the lines of (4.17), denote $t_1,...,t_d$ as event times and $i_1,...i_j$ as cases, thus it gives that

$$\begin{aligned}\widehat{W}(t,\mathbf{x}) =& \sum_{t_j\le t}\left\{f(\mathbf{x}_{i_j})I(\mathbf{x}_{i_j}\le\mathbf{x}) - g(\widehat{\boldsymbol{\beta}},t_j,\mathbf{x})\right\}G_{i_j}\\ &-\sum_{t_j\le t}\exp\left\{\widehat{\boldsymbol{\beta}}^T\mathbf{x}_{i_j}\right\}f(\mathbf{x}_{i_j})I(\mathbf{x}_{i_j}\le\mathbf{x})\left\{\mathbf{x}_{i_j} - \frac{S^{(1)}(\widehat{\boldsymbol{\beta}},t_j)}{S^{(0)}(\widehat{\boldsymbol{\beta}},t_j)}\right\}^T\frac{1}{S^{(0)}(\widehat{\boldsymbol{\beta}},t_j)}\\ &\times I(\widehat{\boldsymbol{\beta}})^{-1}\sum_{t_j}\left\{\mathbf{x}_{i_j} - \frac{S^{(1)}(\widehat{\boldsymbol{\beta}},t_j)}{S^{(0)}(\widehat{\boldsymbol{\beta}},t_j)}\right\}G_{i_j}.\end{aligned} \tag{4.20}$$

Also rewrite (4.19) along the lines of (4.16), we obtain that

$$\begin{aligned}\widehat{W}^*(t,\mathbf{x}) =& \sum_{t_j\le t}\left\{f(\mathbf{x}_{i_j})I(\mathbf{x}_{i_j}\le\mathbf{x}) - g_{\widetilde{R}(t_j)}(\widehat{\boldsymbol{\beta}},t_j,\mathbf{x})\right\}G_{i_j,\widetilde{R}(t_j)}\\ &-\sum_{t_j\le t}\exp\left\{\widehat{\boldsymbol{\beta}}^T\mathbf{x}_{i_j}\right\}w_{i_j}(t_j)f(\mathbf{x}_{i_j})I(\mathbf{x}_{i_j}\le\mathbf{x})\left\{\mathbf{x}_{i_j}-\frac{S^{(1)}_{\widetilde{R}(t_j)}(\widehat{\boldsymbol{\beta}},t_j)}{S^{(0)}_{\widetilde{R}(t_j)}(\widehat{\boldsymbol{\beta}},t_j)}\right\}^T\\ &\frac{1}{S^{(0)}_{\widetilde{R}(t_j)}(\widehat{\boldsymbol{\beta}},t_j)}\times I(\widehat{\boldsymbol{\beta}})^{-1}\sum_{t_j}\left\{\mathbf{x}_{t_j}-\frac{S^{(1)}_{\widetilde{R}(t_j)}(\widehat{\boldsymbol{\beta}},t_j)}{S^{(0)}_{\widetilde{R}(t_j)}(\widehat{\boldsymbol{\beta}},t_j)}\right\}G_{i_j,\widetilde{R}(t_j)}.\end{aligned} \tag{4.21}$$

By comparing the expressions (4.20) with (4.21), we reach the conclusion that two processes are taking the same form. Therefore *timereg* program should be an effective approach for nested case-control simulation.

4.7 Specialize to partial-sum process

4.7.1 Partial-sum process for nested case-control data

In this part we will look at methods for checking log-linearity similar to those of Section 3.4. We will present a simplification to censored survival data and the simulation process. By inserting $f(\mathbf{x}_i)=1$ in (4.16), denoting $t=\tau$ as the maximum time limit for the study, $x_k=\infty$ for all $k\neq j$, and x_{ji} as the jth component of covariate vector $\mathbf{x}_i$ corresponding to the ith individual, we obtain the process

$$\begin{aligned}\widetilde{W}_j(\mathbf{x}) =& \sum_{i=1}^n\sum_{\mathbf{r}\in P_i} I(x_{ji}\le x)\widehat{M}_{i,\mathbf{r}}(t)\\ =& \sum_{i=1}^n I(x_{ji}\le x)N_i(t) - \sum_{t_j\le t}\frac{\sum_{i\in\widetilde{R}(t_j)} I(x_{ji}\le x)\exp\left\{\widehat{\boldsymbol{\beta}}^T\mathbf{x}_i\right\}w_i(t_j)}{\sum_{l\in\widetilde{R}(t_j)}\exp\left\{\widehat{\boldsymbol{\beta}}^T\mathbf{x}_l\right\}w_l(t_j)}.\end{aligned} \tag{4.22}$$

Further, we will specialize the process (4.19). For each individual in the nested case-control data set we have the observations t_i, D_i and $\mathbf{x}_i$ fixed. We first specialize (4.19) to situation with $f(\mathbf{x}_i)=1$, $t=\tau$ and $\mathbf{x}=(\infty,\infty,...,x_j,\infty,...\infty)^T$, then specialize to censored survival data. This gives that

$$\begin{aligned}\widehat{W}_j^*(x) = &\sum_{\mathbf{r}\in P}\sum_{i\in\mathbf{r}}\left\{I(x_{ij}\le x) - g_{\mathbf{r}}(\widehat{\boldsymbol{\beta}}, t_i, x)\right\} G_{i,\mathbf{r}} D_i \\ &- \sum_{\mathbf{r}\in P}\sum_{i\in\mathbf{r}} Y_i(t_i)\exp\left\{\widehat{\boldsymbol{\beta}}^T\mathbf{x}_i\right\} w_i(u) I(x_{ij}\le x)\left\{\mathbf{x}_i - \frac{S_{\mathbf{r}}^{(1)}(\widehat{\boldsymbol{\beta}}, t_i)}{S_{\mathbf{r}}^{(0)}(\widehat{\boldsymbol{\beta}}, t_i)}\right\}^T \\ &\pi(\mathbf{r}|t_i)\widehat{A}_{0\mathbf{r}}(t_i) \times I(\widehat{\boldsymbol{\beta}})^{-1}\sum_{\mathbf{r}\in P}\sum_{i\in\mathbf{r}}\left\{\mathbf{x}_i - \frac{S_{\mathbf{r}}^{(1)}(\widehat{\boldsymbol{\beta}}, t_i)}{S_{\mathbf{r}}^{(0)}(\widehat{\boldsymbol{\beta}}, t_i)}\right\} G_{i,\mathbf{r}} D_i.\end{aligned} \tag{4.23}$$

By comparison we can see that the expression in (4.23) resembles the one in (3.8) for full cohort data.

4.7.2 Check of log-linearity

In order to check the log-linearity of the Cox model for nested case-control data, we need to plot the cumulative martingale residuals processes using (4.22) and (4.23). To illustration how this works we will be using the Radiation and breast cancer data example. The Cox model is fitted by one covariate: Dose in radiation. The cumulative martingale residuals plots for cohort, nested case-control with simple random sampling and counter-matched sampling using $m-1=2$ controls are shown in Figure 4.1, it can be seen that the plots for both two nested case-control designs resemble the one for the cohort though slightly smoother. Also, in terms of dose in rad which is lower than 1, the estimated hazards are higher than the observations. To compute P-value, we simulate $N = 1000$ replicates, and the result is shown in Figure 4.2. The P-values corresponding to cohort, simple random design and counter-matching design are 0.251, 0.426 and 0.201 respectively, which is an indication that the log-linearity is satisfied. Besides, we find that both two nested case-control designs give a very close result to cohort design. Furthermore, by repeating the same process with $m-1=8$ and $m-1=14$ controls, we obtain the Figures 4.3 and 4.4, with the P-value of the log-linearity test given in Table 4.2, which indicates that in nested case-control design, the log-linearity is still satisfied. It is quite clear that the more controls that we choose for the nested case-control design, the closer they will approach the cohort, but in this case 2 controls are already effective enough.

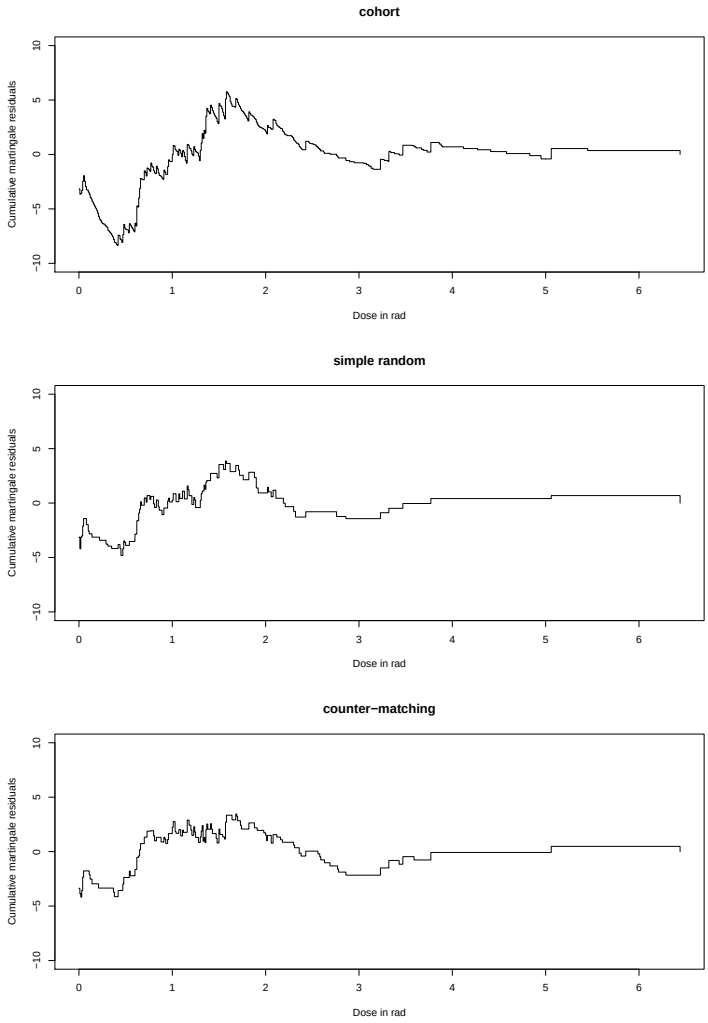

Figure 4.1: Cumulative martingale residuals against Dose in radiation in the Cox model for cohort, nested case-control with simple random sampling and counter-matched sampling using $m - 1 = 2$ controls

Table 4.2: The P-value of the log-linearity test of the Cox model fitted by Dose in radiation for cohort design, simple random sampling design, and counter-matched sampling design with different number of controls

cohort		0.251	
number of controls	$m-1=2$	$m-1=8$	$m-1=14$
simple random	0.426	0.226	0.284
counter-matching	0.201	0.371	0.310

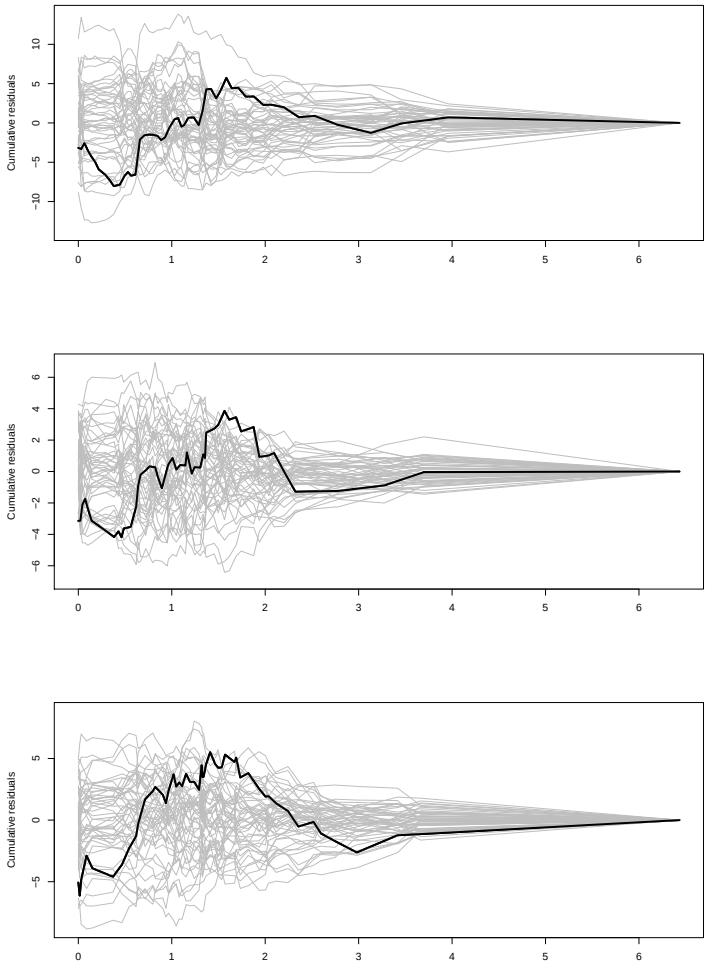

Figure 4.2: Simulation of cumulative martingale residuals against Dose in radiation in the Cox model for cohort, nested case-control with simple random sampling and counter-matched sampling using $m - 1 = 2$ controls

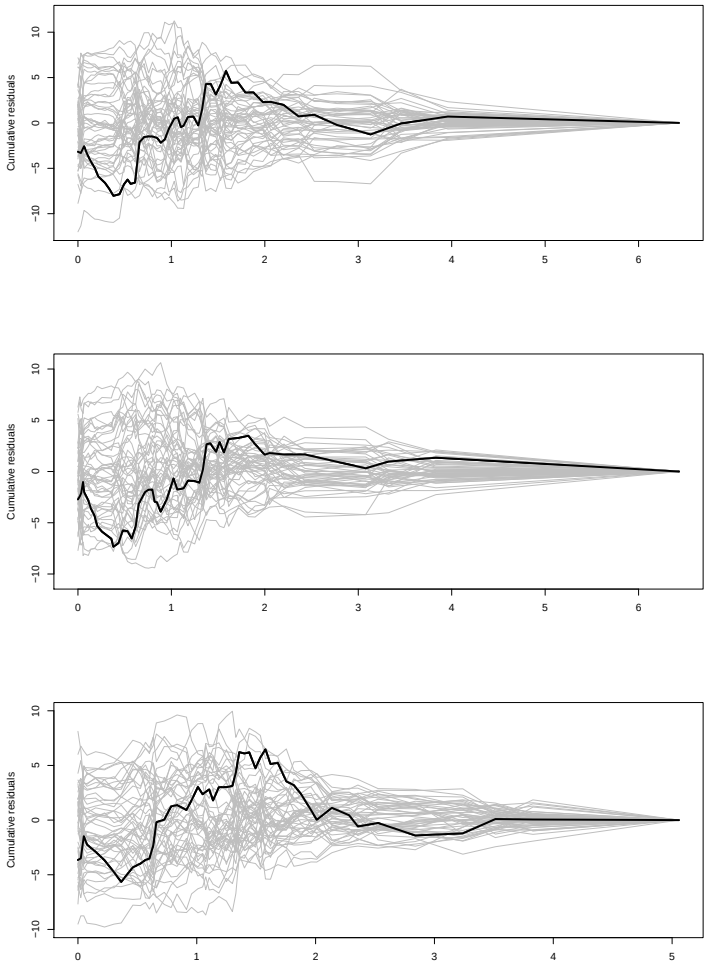

Figure 4.3: Simulation of cumulative martingale residuals against Dose in radiation in the Cox model for cohort, nested case-control with simple random sampling and counter-matched sampling using $m - 1 = 8$ controls

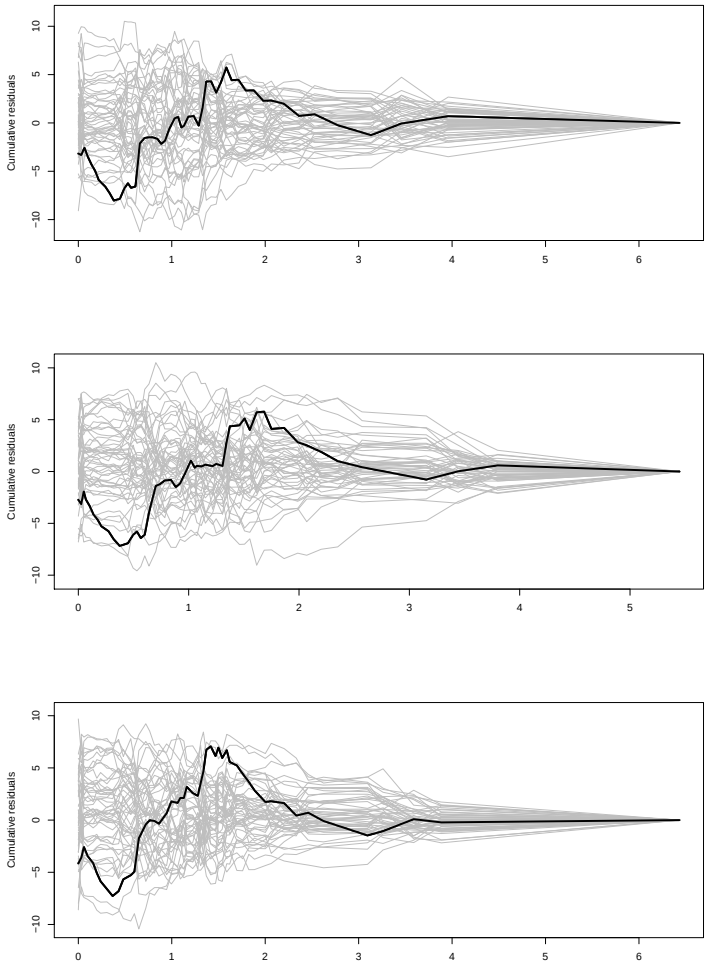

Figure 4.4: Simulation of cumulative martingale residuals against Dose in radiation in the Cox model for cohort, nested case-control with simple random sampling and counter-matched sampling using $m - 1 = 14$ controls

4.8 Specialize to score process

4.8.1 Score process for nested case-control data

In the continuation of the model checking for nested case-control designs, we will look at the specialization of (4.16) to the score process. Let $f(\mathbf{x}_i) = \mathbf{x}_i$ and $\mathbf{x} = \infty$, we can obtain that

$$\begin{aligned}\widetilde{U}(\widehat{\boldsymbol{\beta}}, t) =& \sum_{i=1}^{n} \sum_{\mathbf{r} \in P_i} \mathbf{x}_i \widehat{M}_{i,\mathbf{r}}(t) \\ =& \sum_{i=1}^{n} \mathbf{x}_i N_i(t) - \sum_{t_j \le t} \frac{\sum_{i \in \widetilde{R}(t_j)} \mathbf{x}_i \exp\left\{\widehat{\boldsymbol{\beta}}^T \mathbf{x}_i\right\} w_i(t_j)}{\sum_{l \in \widetilde{R}(t_j)} \exp\left\{\widehat{\boldsymbol{\beta}}^T \mathbf{x}_l\right\} w_l(t_j)}.\end{aligned} \tag{4.24}$$

Then we specify (4.24) to the jth covariate, it follows that

$$\widetilde{U}_j(\widehat{\boldsymbol{\beta}}, t) = \sum_{i=1}^{n} x_{ji} N_i(t) - \sum_{t_j \le t} \frac{\sum_{i \in \widetilde{R}(t_j)} x_{ji} \exp\left\{\widehat{\boldsymbol{\beta}}^T \mathbf{x}_i\right\} w_i(t_j)}{\sum_{l \in \widetilde{R}(t_j)} \exp\left\{\widehat{\boldsymbol{\beta}}^T \mathbf{x}_l\right\} w_l(t_j)}. \tag{4.25}$$

It needs to be pointed out that (4.24) is related to the score based on the partial likelihood of simple random sampling (4.8) and counter-matched sampling (4.10). In a Cox's regression model, the score function for nested case-control data can be expressed as

$$\widetilde{U}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log L_{ncc}(\boldsymbol{\beta}) = \sum_{\mathbf{r} \in P} \sum_{i \in \mathbf{r}} \int_0^{\tau} \left\{ \mathbf{x}_i - \frac{S_{\mathbf{r}}^{(1)}(\boldsymbol{\beta}, u)}{S_{\mathbf{r}}^{(0)}(\boldsymbol{\beta}, u)} \right\} dN_{i,\mathbf{r}}(u) \tag{4.26}$$

By replacing τ with t in the expression (4.26), we obtain the score process

$$\widetilde{U}(\boldsymbol{\beta}, t) = \sum_{\mathbf{r} \in P} \sum_{i \in \mathbf{r}} \int_0^{t} \left\{ \mathbf{x}_i - \frac{S_{\mathbf{r}}^{(1)}(\boldsymbol{\beta}, u)}{S_{\mathbf{r}}^{(0)}(\boldsymbol{\beta}, u)} \right\} dN_{i,\mathbf{r}}(u)$$

Also rewrite the expression (4.24) as

$$\begin{aligned}\sum_{i=1}^{n}\sum_{\mathbf{r}\in P_i}\mathbf{x}_i\widehat{M}_{i,\mathbf{r}}(t) &= \sum_{i=1}^{n}\sum_{\mathbf{r}\in P_i}\mathbf{x}_i\left\{N_{i,\mathbf{r}}(t)-\widehat{\Lambda}_{i,\mathbf{r}}(t)\right\}\\ &= \sum_{\mathbf{r}\in P}\sum_{i\in\mathbf{r}}\int_0^t \mathbf{x}_i dN_{i,\mathbf{r}}(u) - \int_0^t \mathbf{x}_i \frac{Y_i(u)\exp\left\{\widehat{\boldsymbol{\beta}}^T\mathbf{x}_i\right\}\pi(\mathbf{r}|u)w_i(u)}{\sum_{l\in\mathbf{r}}\exp\left\{\widehat{\boldsymbol{\beta}}^T\mathbf{x}_l\right\}\pi(\mathbf{r}|u)w_l(u)}dN_{\mathbf{r}}(u)\\ &= \sum_{\mathbf{r}\in P}\sum_{i\in\mathbf{r}}\int_0^t \mathbf{x}_i dN_{i,\mathbf{r}}(u) - \sum_{\mathbf{r}\in P}\int_0^t \frac{S_{\mathbf{r}}^{(1)}(\widehat{\boldsymbol{\beta}},u)}{S_{\mathbf{r}}^{(0)}(\widehat{\boldsymbol{\beta}},u)}dN_{\mathbf{r}}(u)\\ &= \sum_{\mathbf{r}\in P}\sum_{i\in\mathbf{r}}\int_0^t \left\{\mathbf{x}_i - \frac{S_{\mathbf{r}}^{(1)}(\widehat{\boldsymbol{\beta}},u)}{S_{\mathbf{r}}^{(0)}(\widehat{\boldsymbol{\beta}},u)}\right\}dN_{i,\mathbf{r}}(u).\end{aligned} \tag{4.27}$$

Thus it can be seen that the score process (4.24) for nested case-control data resembles the one for cohort data in (3.9).

4.8.2 Check of proportionality

In this following example we will illustrate how the score process plot of (4.24) can be used to check for proportionality. Here we have the nested case-control data obtained from the Radiation and breast cancer study, where only one covariate: Dose in radiation has been used for the fit of the Cox model. The observation of score process against time for cohort, nested case-control with simple random sampling and counter-matched sampling using $m-1=2$ controls is shown in Figure 4.5. We can see that the curves corresponding to three designs are fluctuating around 0 in similar shapes over the entire period, indicating that nested case-control designs have managed to capture most of the information.

Note that the trick for obtaining the nested case-control plot of score process is the same as we did earlier for the cumulative martingale residuals in the first special case. When the organized case-control data set is created, simply apply the commands for obtaining Figures 3.5 again to calculate the Schoenfeld residuals, which is a one-dimensional vector representing the only covariate of the Cox model.

According to the observation of score process plot, it seems like the results obtained from all three designs are quite similar. But in order to check for proportionality assumption, it is important to see the randomness of the score process. Thus, we are going to run simulation to calculate P-values and hence see if proportionality is satisfied for one cohort design and both nested case-control designs. To the end we specialize the process (4.19) to

Table 4.3: The P-value of the proportionality test of the Cox model fitted by Dose in radiation for cohort design, simple random sampling design, and counter-matched sampling design with different number of controls

cohort		0.588	
number of controls	$m-1=2$	$m-1=8$	$m-1=14$
simple random	0.466	0.426	0.482
counter-matching	0.518	0.576	0.502

the case in (4.24). For each individual in the nested case-control data set we have the observations t_i, D_i and $\mathbf{x}_i$ fixed. We first specialize (4.19) to the case when $f(\mathbf{x}_i) = \mathbf{x}_i$ and $\mathbf{x} = (\infty, \infty, ..., \infty, ...\infty)^T$, then to the censored survival data. This gives that

$$\begin{aligned}\widehat{U}_j^*(\boldsymbol{\beta}, t) = &\sum_{\mathbf{r}\in P}\sum_{i\in\mathbf{r}}\left\{\mathbf{x}_i - \frac{S_{\mathbf{r}}^{(1)}(\widehat{\boldsymbol{\beta}}, t_i)}{S_{\mathbf{r}}^{(0)}(\widehat{\boldsymbol{\beta}}, t_i)}\right\} G_{i,\mathbf{r}} D_i \\ &- \sum_{\mathbf{r}\in P}\sum_{i\in\mathbf{r}} Y_i(t_i)\exp\left\{\widehat{\boldsymbol{\beta}}^T\mathbf{x}_i\right\} w_i(u)\mathbf{x}_i\left\{\mathbf{x}_i - \frac{S_{\mathbf{r}}^{(1)}(\widehat{\boldsymbol{\beta}}, t_i)}{S_{\mathbf{r}}^{(0)}(\widehat{\boldsymbol{\beta}}, t_i)}\right\}^T \\ &\pi(\mathbf{r}|t_i)\widehat{A}_{0\mathbf{r}}(t_i) \times I(\widehat{\boldsymbol{\beta}})^{-1}\sum_{\mathbf{r}\in P}\sum_{i\in\mathbf{r}}\left\{\mathbf{x}_i - \frac{S_{\mathbf{r}}^{(1)}(\widehat{\boldsymbol{\beta}}, t_i)}{S_{\mathbf{r}}^{(0)}(\widehat{\boldsymbol{\beta}}, t_i)}\right\} G_{i,\mathbf{r}} D_i.\end{aligned} \tag{4.28}$$

The simulated score process plots are shown in Figure 4.6, where the upper panel corresponds to the full cohort while the lower two are for nested case-control data with simple random sampling and counter-matched sampling using 2 controls. The respective P-values for the plots are 0.588, 0.466 and 0.518. It is obvious that the result from both simple random sampling and counter-matched sampling resemble the cohort, where counter-matching gives closer result with cohort, thus the proportionality assumption is satisfied. We repeat the process by adding more controls, and the plots are shown in Figure 4.7 and 4.8. According to the plots and the P-values in Table 4.3, we find that by selecting at least $m-1=2$ controls, the proportionality tests for nested case-control data are able to give a fairly close result with cohort.

To this end, we have managed to extend the cumulative sums of martingale-based residuals from cohort to nested case-control studies, and we have examined the result of model checking for nested case-control data with simple sampling and counter-matched sampling. As expected they are fairly close to cohort studies even by using a small number of controls. So far we have only used one real data set for illustration. To make it more generalized, in the following chapter we would like to observe the performance of model checking for simulated nested case-control data, where the Cox models are defined in different forms as either good or wrong. Through various simulation processes we will be able to have an extensive understanding of model checking for nested case-control data.

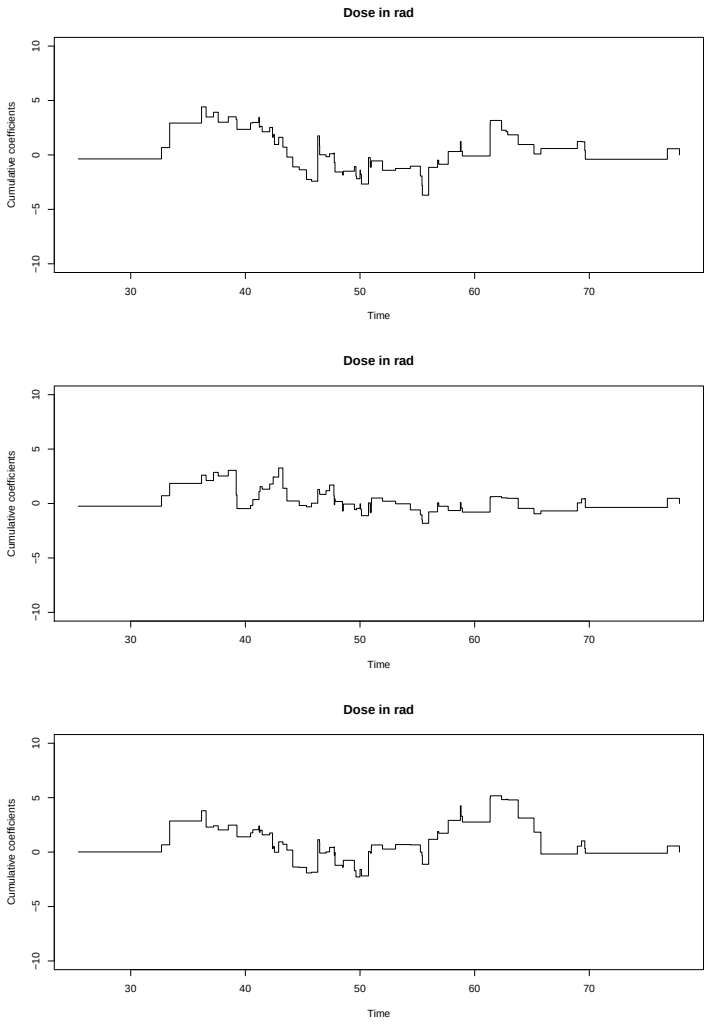

Figure 4.5: Score process against time for Dose in radiation in the Cox model for cohort, nested case-control with simple random sampling and counter-matched sampling using $m - 1 = 2$ controls

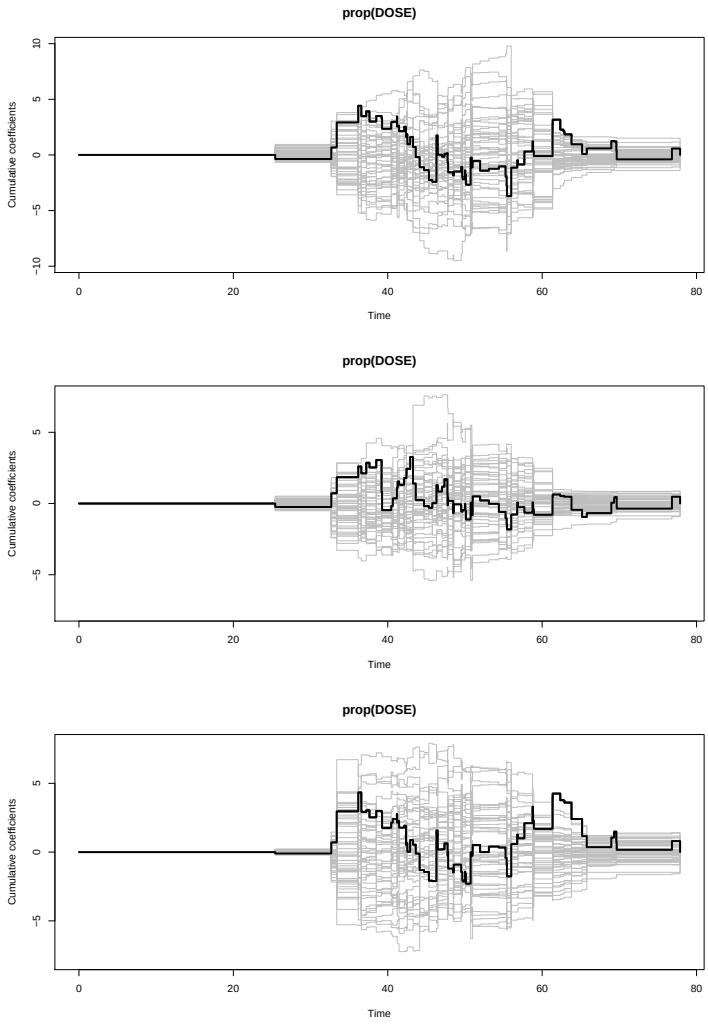

Figure 4.6: Simulation of score process against time for Dose in radiation in the Cox model for cohort, nested case-control with simple random sampling and counter-matched sampling using $m - 1 = 2$ controls

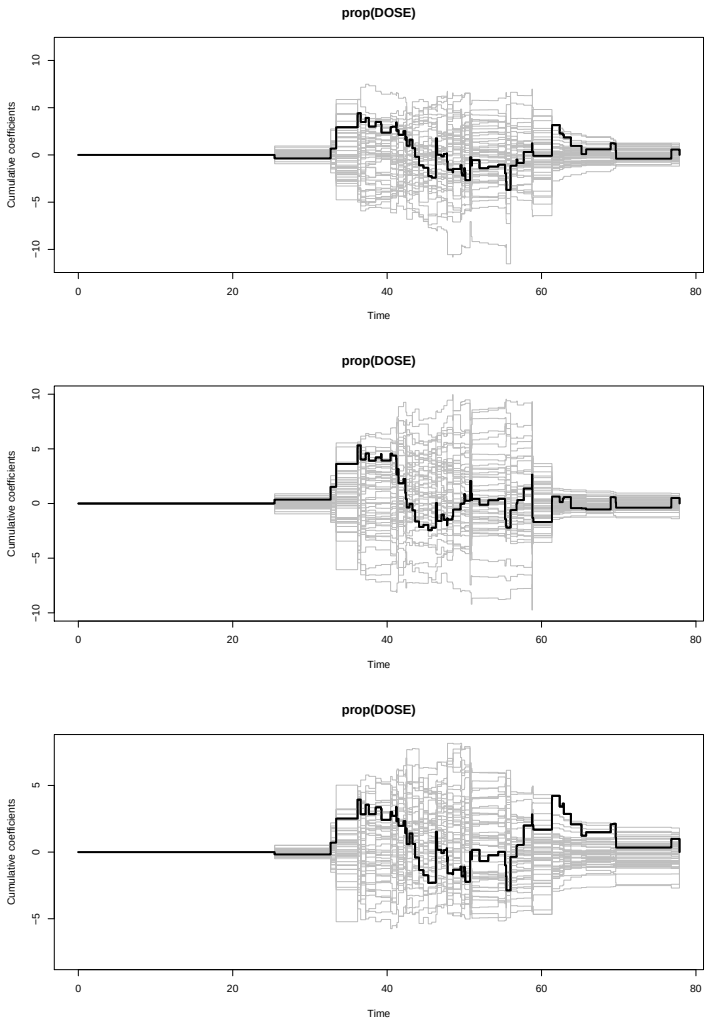

Figure 4.7: Simulation of score process against time for Dose in radiation in the Cox model for cohort, nested case-control with simple random sampling and counter-matched sampling using $m - 1 = 8$ controls

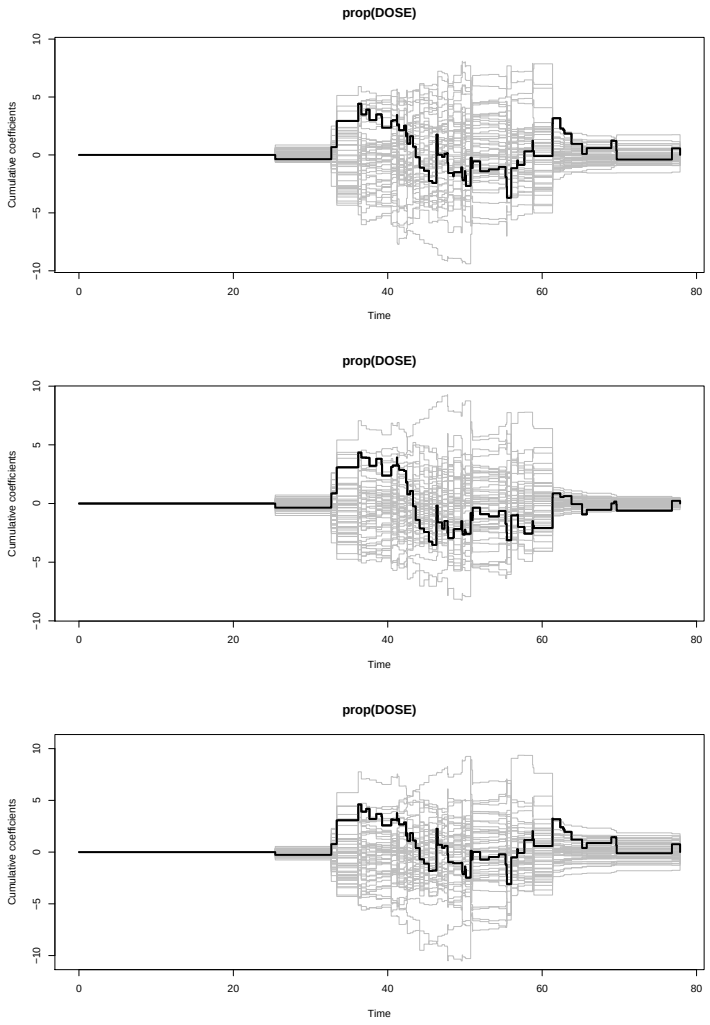

Figure 4.8: Simulation of score process against time for Dose in radiation in the Cox model for cohort, nested case-control with simple random sampling and counter-matched sampling using $m - 1 = 14$ controls

Chapter 5

Nested case-control studies in simulations

In Chapter 4 we have discussed nested case-control designs and model checking techniques based on real data from the Radiation and breast cancer study. In this chapter we are going to extend on that by simulating different situations, where we can define the distributions of covariates, hazard rates, survival times and levels of censoring, etc. In this way we will see the overall performance of model checking for nested case-control data and how it works in practice. In Section 5.1 we look at check of the log-linearity for a good model where data simulation method is shown, followed by model checking results illustrated in details. In Section 5.2 we first simulate a non-linear model, and then focus on checking the log-linearity for a wrong model. Further, in Section 5.3 we move on to the check of the proportionality for a good model by following a similar structure with Section 5.1. Finally, the check of proportionality for a wrong model is discussed in the last section.

5.1 Check the log-linearity of a good model

5.1.1 Data simulation for a good model

In this simulation scenario, we select $n = 2000$ as the number of cohort individuals, and two covariates x_1, x_2 are extracted for the fit of the Cox model. Covariate x_1 is a numeric type which is uniformly distributed on $(-1, 1)$, while x_2 is binary and takes the values -1 or 1 with $P = 0.5$ such that two numbers are in the same proportion. The hazard rate $\alpha(t|\mathbf{x}_i)$ is calculated by (2.5), where $\mathbf{x}_i = (x_{i1}, x_{i2})^T$, and the risk function is an exponential type. We also choose the baseline hazard α_0 as 1 and regression coefficients as $\boldsymbol{\beta} = (0.5, 0.5)^T$. Thus we obtain the simulation Cox model with covariates

x_{i1} and x_{i2} for individual i that takes the form

$$\alpha(t|\mathbf{x}_i) = \exp\{0.5x_{i1} + 0.5x_{i2}\}, \tag{5.1}$$

where

$$x_{i1} \sim U[-1, 1]$$

and

$$x_{i2} = \begin{cases} 1 & \text{with probability } 0.5 \\ -1 & \text{with probability } 0.5 \end{cases}$$

Then we generate the survival time T_i^0 and censoring time C_i by random sampling from the exponential distribution with rate parameter $\alpha(t|\mathbf{x}_i)$ and 10 respectively, where $\alpha(t|\mathbf{x}_i)$ is the hazard rate calculated from (5.1). Finally we find the censored lifetime $\widetilde{T}_i$ by selecting the minimum one between survival time and censoring time, that is, $\widetilde{T}_i = min(T_i^0, C_i)$. The reason why we choose these rate parameter settings is because we want to keep the event rate at around 10%. In addition we obtain the censoring indicator denoted by $D_i = I(\widetilde{T}_i = T_i^0)$, which takes value 1 if the event of interest has been observed and otherwise 0 due to censoring.

5.1.2 Simple random sampling

We will use the simulation model (5.1) to plot the cumulative martingale residuals against x_1 for both the cohort and the nested case-control design with simple random sampling of controls. We start with one simulation and $m-1 = 2$ controls for simple random sampling. The result is shown in Figure 5.1. We can see that two plots are quite similar, and that the observed cumulative martingale residuals are fluctuating around 0 and wrapped by most of the simulated lines. We also obtain the P-values corresponding to the two plots as 0.554 and 0.597 respectively. But at this stage we only have one simulation of data. In order to evaluate the accuracy of model checking for nested case-control data, we need to have more simulations where we can find the distribution of the P-values and make comparison with full cohort.

We run 500 times of data simulation, and hence obtain the histograms of P-values, as shown in Figure 5.2. We see that for both the cohort and the nested case-control design with simple random sampling of controls, the P-values are equally high and in the same distribution, indicating that log-linearity is satisfied. Further, we look at the effective significance level, as shown in Table 5.1, which indicates the proportion of P-values less than or equal to 5%. It reveals that cohort and the simple random design are giving a quite close result. But it needs to be pointed out that the approximations of the P-values are not so perfect, since the number of P-values lower than 5% is too small, while in contrast those over 95% is too large. As a result we may not get enough P-values below 5%. This indicates that the sampling

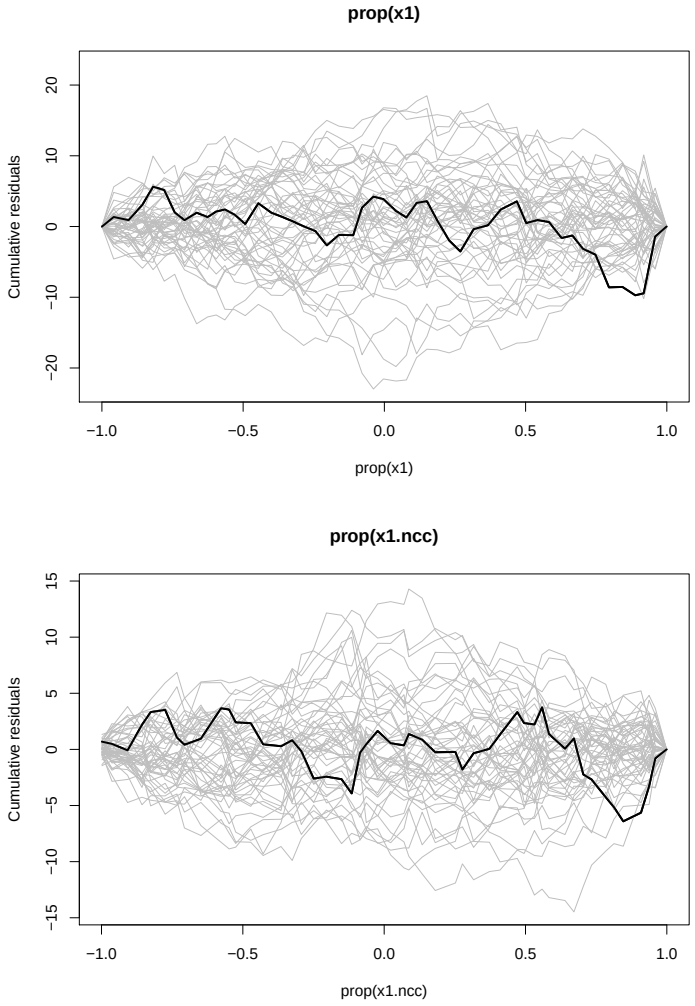

Figure 5.1: One simulation of cumulative martingale residuals against x_1 in the good Cox model where $\boldsymbol{\beta} = (0.5, 0.5)^T$ for cohort and nested case-control data with simple random sampling using $m - 1 = 2$ controls

procedure does not manage to get the exact null distribution for the test statistic, for in that case we would get uniform P-values when the model is correctly specified.

5.1.3 Counter-matched sampling

We will have a look at the performance of counter-matched sampling design. We start by applying the same simulation model (5.1) where x_1 is the variable of interest. Based on this we have the individuals stratified according to $x_1 \leq 0$ or $x_1 > 0$. Since x_1 is uniformly distributed between -1 and 1, the number of individuals that belongs to the two strata should be ap-

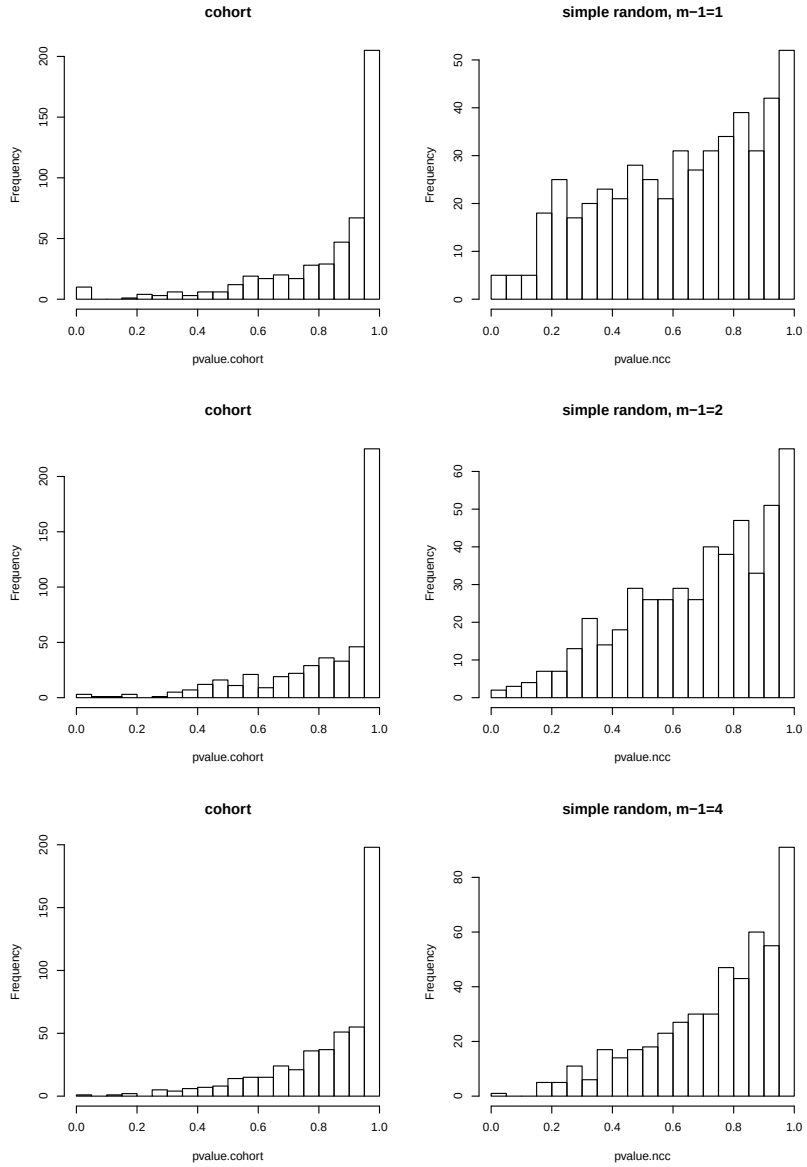

Figure 5.2: Histograms of P-values over 500 simulations of cumulative martingale residuals against x_1 in the good Cox model where $\boldsymbol{\beta} = (0.5, 0.5)^T$ for cohort and nested case-control data with simple random sampling using $m - 1 = 1$, $m - 1 = 2$, $m - 1 = 4$ controls

Table 5.1: Proportion of P-values $\leq 5\%$ over 500 simulations of cumulative martingale residuals against x_1 in the good Cox model where $\boldsymbol{\beta} = (0.5, 0.5)^T$ for cohort and nested case-control data with simple random sampling using $m-1=1$, $m-1=2$, $m-1=4$ controls

	controls	P-values$\leq 5\%$
cohort		0.02
simple random	$m-1=1$	0.01
cohort		0.006
simple random	$m-1=2$	0.004
cohort		0.002
simple random	$m-1=4$	0.002

proximately the same. So we obtain the plots of the cumulative martingale residuals against x_1 for both the cohort and the counter-matching design in Figure A.1 in the Appendix, where we do one simulation with $m-1=1$ control for counter-matched sampling. Note that the upper plot representing cumulative martingale residuals for cohort is the same as that one in Figure 5.1. As what we have expected, the two curves resemble each other. The observed cumulative martingale residuals are within the area covered by simulated lines. The P-values corresponding to the two plots are 0.520 and 0.472 respectively.

For evaluating the performance of model checking for using nested case-control data with counter-matched sampling, we will run 500 simulations to find the distribution of the P-values and then compare it with full cohort. According to Figure A.2 in the Appendix, the P-values corresponding to cohort and the nested case-control design with counter-matched sampling of controls are almost identically distributed. Note that the P-values should have been uniformly distributed when the model is correctly specified. When it comes to the effective significance level, as in Table 5.2, we see that the proportions of P-values less than or equal to 5% are quite small and even zero. Although the log-linearity is satisfied, the approximations of the P-values are still not perfect and we do not get enough P-values below 5%.

5.2 Check the log-linearity of a wrong model

5.2.1 Data simulation for a wrong model

In the previous section we have shown that the Cox model checking for nested case-control data works fine when the simulated risk function satisfies log-linearity. But we want to make sure that the model checking for nested case-control is able to detect deviations from log-linearity. Then we need to

Table 5.2: Proportion of P-values$\leq$ 5% over 500 simulations of cumulative martingale residuals against x_1 in the good Cox model where $\boldsymbol{\beta} = (0.5, 0.5)^T$ for cohort and nested case-control data with counter-matched sampling using $m-1=1$, $m-1=3$, $m-1=5$ controls

	controls	P-values$\leq$ 5%
cohort		0.002
counter-matching	$m-1=1$	0.002
cohort		0.000
counter-matching	$m-1=3$	0.004
cohort		0.000
counter-matching	$m-1=4$	0.000

simulate a non-linear risk function that ends up in a wrong model, and to see if the failure still can be detected under the nested case-control design. To obtain such a wrong model, it is suggested that we change the expression (5.1) into a quadratic form with for example $\boldsymbol{\beta} = (0.1, 1.0, 0.5)^T$. Thus we have

$$\alpha(t|\mathbf{x}_i) = \exp\left\{0.1x_{i1} + 1.0x_{i1}^2 + 0.5x_{i2}\right\}, \tag{5.2}$$

where

$$x_{i1} \sim U[-1, 1]$$

$$x_{i2} = \begin{cases} 1 & \text{with probability } 0.5 \\ -1 & \text{with probability } 0.5 \end{cases}$$

We also try other values of $\boldsymbol{\beta}$ where the log-linear model fails in varying degrees. So we select the regression coefficients $(0.5, 0.8, 0.5)^T$ and $(0.8, 0.6, 0.8)^T$ for the hazard. Then we make a plot of the hazards (5.1) and (5.2) together with these two hazards, as shown in Figure 5.3. We see that the hazard plot of the correctly specified model is a straight line on the log-scale, but the hazards for the wrong models are obviously deviating from the straight line. They are in parabolic shapes and hence non-linear.

5.2.2 Simple random sampling

We will now check the non-linearity of the Cox model for simulated nested case-control data with simple random sampling. To begin with, we look at the plot of cumulative martingale residuals with one only simulation. According to Figure 5.4, for all Cox models with non-linear hazard settings, the log-linearity seems to be violated since the supremums of the dark lines are higher than most of the sampled grey lines.

Further, we will focus on the proportion of P-values below 5% to see if the model checking for the simple random sampling design can disclose the result

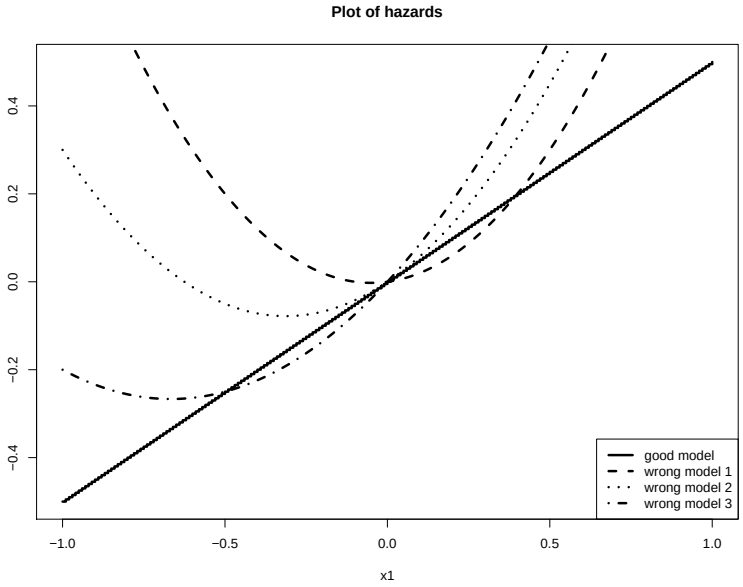

Figure 5.3: Plot of hazards (5.1) for the good model and three wrong models on log-scale with regression coefficients $(0.1, 1.0, 0.5)^T$, $(0.5, 0.8, 0.5)^T$ and $(0.8, 0.6, 0.8)^T$

of non log-linearity. Figures 5.5, 5.6 and 5.7 give the histograms of P-values with respect to three non-linear Cox models using different number of controls. The simple random design seems to be effective since the distributions of the paired histograms are quite similar. Finally, from Table 5.3 we see that in the cohort study the proportions of P-values below 5% are all greater than 0.05. To be more specific, the first group accounts for the highest proportion that is around 0.52, followed by the other two groups being around 0.16 and 0.07 respectively. When it comes to the simple random design, we find that the wrong models can be detected as easily as for full cohort in the first two groups of settings even with 1 control per case. For the last group of simulations, however, due to the fact that the proportions of P-values below 5% are just over 0.05 for both full cohort and simple random sampling, the non-linearity can be equally hard to detect. To sum up, the proportion of P-values below 5% is about the same for cohort and nested case-control data, indicating that model checking for the simple random design works fairly effective in various situations.

5.2.3 Counter-matched sampling

We are going to verify if the non log-linear model can be detected when using nested case-control data with counter-matched sampling. Recall the wrong model (5.2) together with two regression coefficients $(0.5, 0.8, 0.5)^T$ and $(0.8, 0.6, 0.8)^T$ for the hazard. By re-using these wrong models, we obtain the

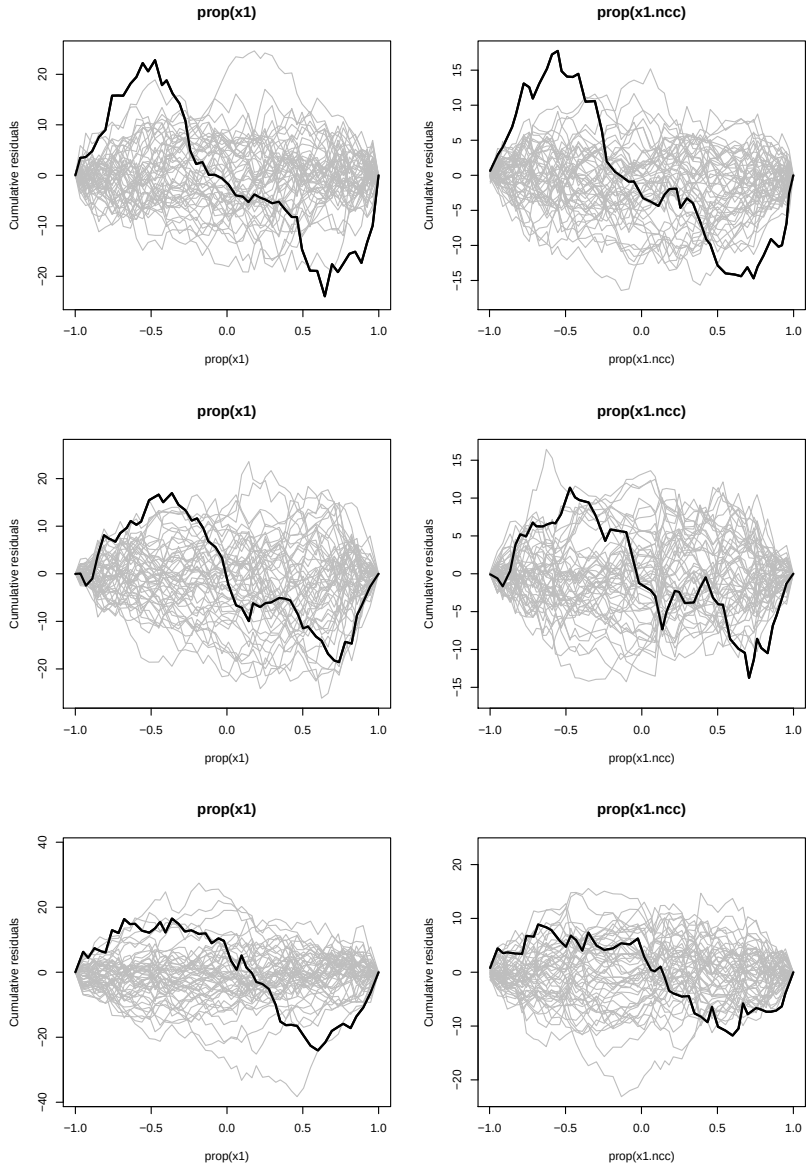

Figure 5.4: One simulation of cumulative martingale residuals against x_1 in three non-linear Cox models where $\boldsymbol{\beta} = (0.1, 1.0, 0.5)^T$ (upper panel), $\boldsymbol{\beta} = (0.5, 0.8, 0.5)^T$ (middle panel) and $\boldsymbol{\beta} = (0.8, 0.6, 0.8)^T$ (lower panel) for cohort and nested case-control data with simple random sampling using $m - 1 = 2$ controls

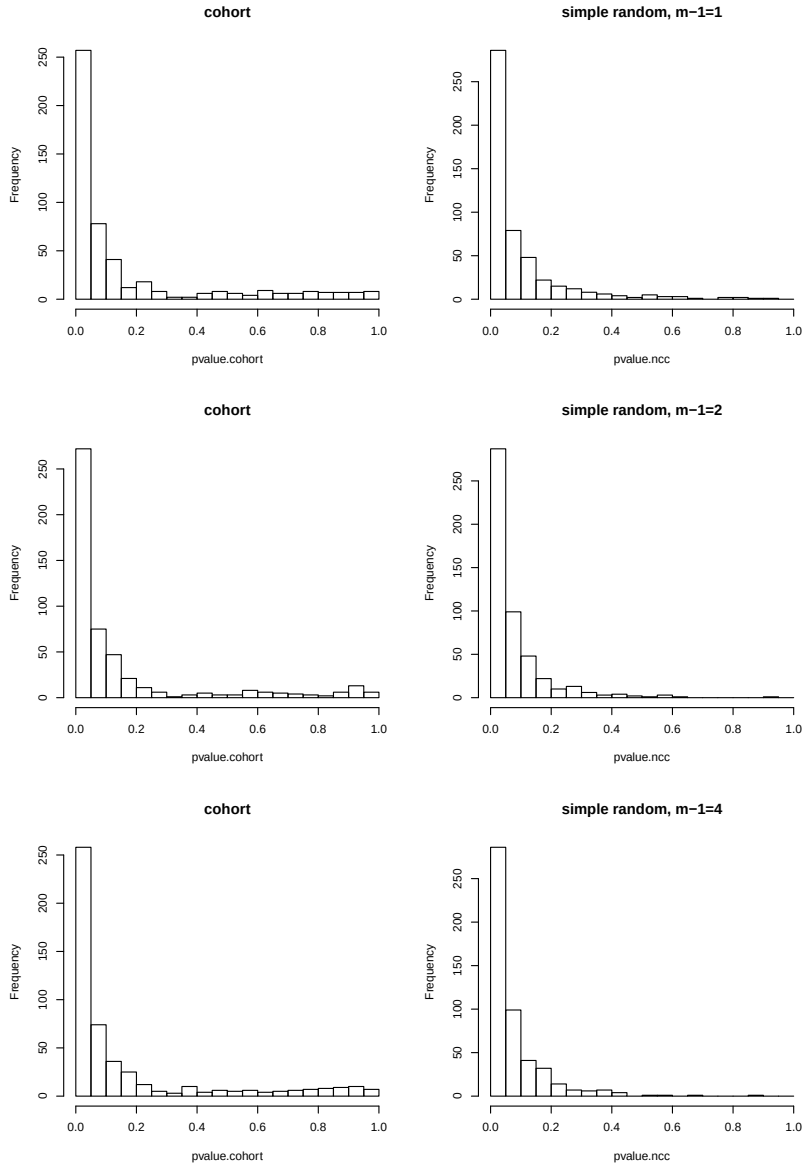

Figure 5.5: Histograms of P-values over 500 simulations of cumulative martingale residuals against x_1 in the non-linear Cox model where $\boldsymbol{\beta} = (0.1, 1.0, 0.5)^T$ for cohort and nested case-control data with simple random sampling using $m-1=1$, $m-1=2$, $m-1=4$ controls

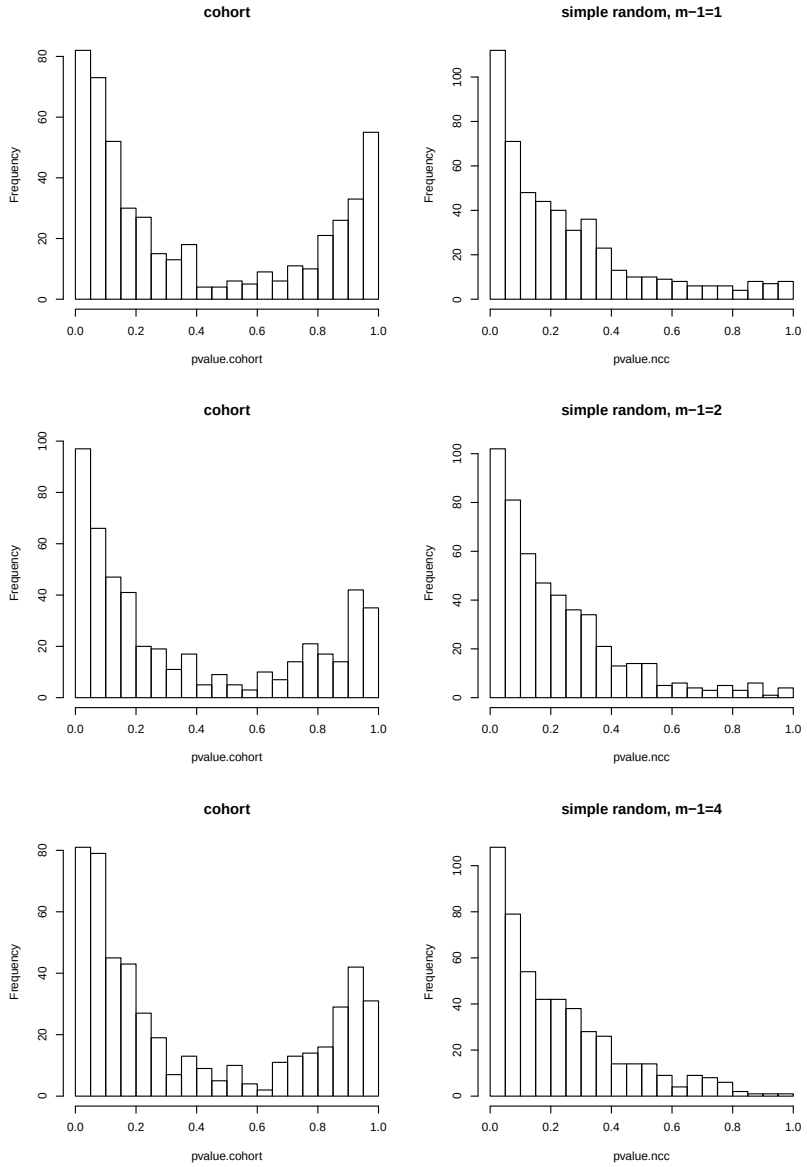

Figure 5.6: Histograms of P-values over 500 simulations of cumulative martingale residuals against x_1 in the non-linear Cox model where $\boldsymbol{\beta} = (0.5, 0.8, 0.5)^T$ for cohort and nested case-control data with simple random sampling using $m-1=1$, $m-1=2$, $m-1=4$ controls

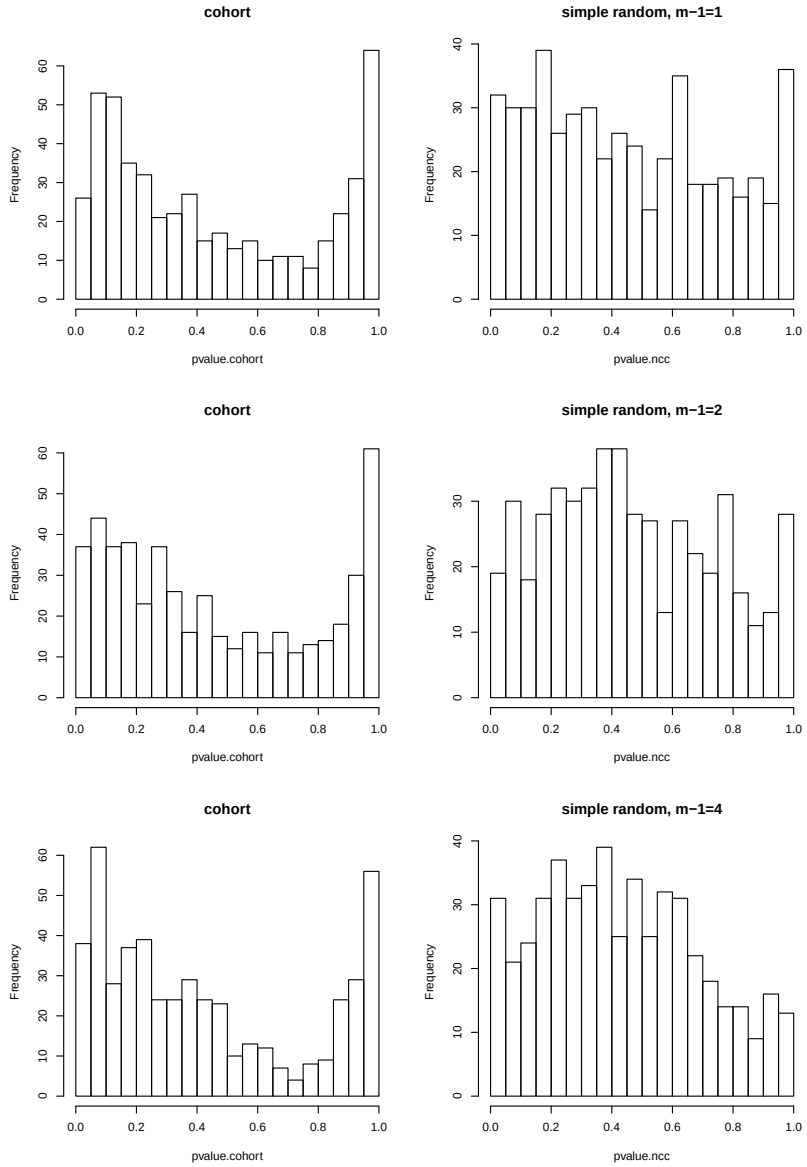

Figure 5.7: Histograms of P-values over 500 simulations of cumulative martingale residuals against x_1 in the non-linear Cox model where $\boldsymbol{\beta} = (0.8, 0.6, 0.8)^T$ for cohort and nested case-control data with simple random sampling using $m-1=1$, $m-1=2$, $m-1=4$ controls

Table 5.3: Proportion of P-values$\leq$ 5% over 500 simulations of cumulative martingale residuals against x_1 in three non-linear Cox models where $\boldsymbol{\beta} = (0.1, 1.0, 0.5)^T$, $(0.5, 0.8, 0.5)^T$ and $(0.8, 0.6, 0.8)^T$ for cohort and nested case-control data with simple random sampling using $m-1=1$, $m-1=2$, $m-1=4$ controls

	controls	P-values$\leq$ 5%
cohort		0.514
simple random	$m-1=1$	0.572
cohort		0.544
simple random	$m-1=2$	0.574
cohort		0.516
simple random	$m-1=4$	0.572
cohort		0.164
simple random	$m-1=1$	0.224
cohort		0.194
simple random	$m-1=2$	0.204
cohort		0.162
simple random	$m-1=4$	0.216
cohort		0.052
simple random	$m-1=1$	0.064
cohort		0.074
simple random	$m-1=2$	0.038
cohort		0.076
simple random	$m-1=4$	0.062

one simulation plots of cumulative martingale residuals. As shown in Figure A.3 in the appendix, all of the observation curves are in sharp fluctuations with high supremums. Due to the non-linear hazards, the log-linearity is likely to have been violated.

Similar to what we did earlier for simple random sampling, we will look at the proportion of P-values below 5% in order to evaluate the performance of nested case-control data with counter-matched sampling. The histograms of P-values for three non-linear Cox models using different number of controls are illustrated in Figures A.4, A.5 and A.6 in the appendix. Then, by looking at the result collected in Table 5.4, we find that the proportion of P-values below 5% is about the same for cohort and nested case-control data with counter-matched sampling using $m-1=3$ or $m-1=5$ controls. This implies that when using at least 3 controls per case, the non log-linearity should be detected as easily as for the full cohort. However, if we only use $m-1=1$ control for the nested case-control data with counter-matched sampling, though it still manages to work when the log-linearity model fails badly, in some cases it can be difficult to conclude exactly the same model checking result as cohort especially when the model is at the edge of the violating log-linearity.

5.3 Check the proportionality of a good model

5.3.1 Simple random sampling

We are at this point starting to perform the check of proportionality for simulated nested case-control data. We use the same simulation model (5.1) and plot the score process against time for x_1 and x_2 in the Cox model for both the cohort and the nested case-control design with simple random sampling of controls. The plots with one simulation using $m-1=2$ controls for simple random sampling are given in Figure 5.8. We can see that both of the two paired plots are not similar, this is due to the randomness of sampling, it is required to have more simulations to find the distribution of the P-values regarding the score processes.

We run 100 simulations of the score process, and hence obtain the histograms of P-values, as shown in Figure 5.9. It can be seen that for both the cohort and the nested case-control design with simple random sampling of controls, the P-values have similar distributions. Note that here the distributions of the P-values look quite uniform, indicating that the simulations are as good as they should be. Further, we look at the effective significance level, as shown in Table 5.5, which shows the proportion of P-values less than or equal to 5%. It proves that proportionality is satisfied and can be correctly checked by using nested case-control data with simple random sampling. Moreover,

Table 5.4: Proportion of P-values$\leq 5\%$ over 500 simulations of cumulative martingale residuals against x_1 in three non-linear Cox models where $\boldsymbol{\beta} = (0.1, 1.0, 0.5)^T$, $(0.5, 0.8, 0.5)^T$ and $(0.8, 0.6, 0.8)^T$ for cohort and nested case-control data with counter-matched sampling using $m-1=1$, $m-1=3$, $m-1=5$ controls

	controls	P-values$\leq 5\%$
cohort		0.520
counter-matching	$m-1=1$	0.298
cohort		0.524
counter-matching	$m-1=3$	0.500
cohort		0.540
counter-matching	$m-1=5$	0.534
cohort		0.212
counter-matching	$m-1=1$	0.132
cohort		0.202
counter-matching	$m-1=3$	0.156
cohort		0.186
counter-matching	$m-1=5$	0.190
cohort		0.090
counter-matching	$m-1=1$	0.020
cohort		0.088
counter-matching	$m-1=3$	0.048
cohort		0.066
counter-matching	$m-1=5$	0.056

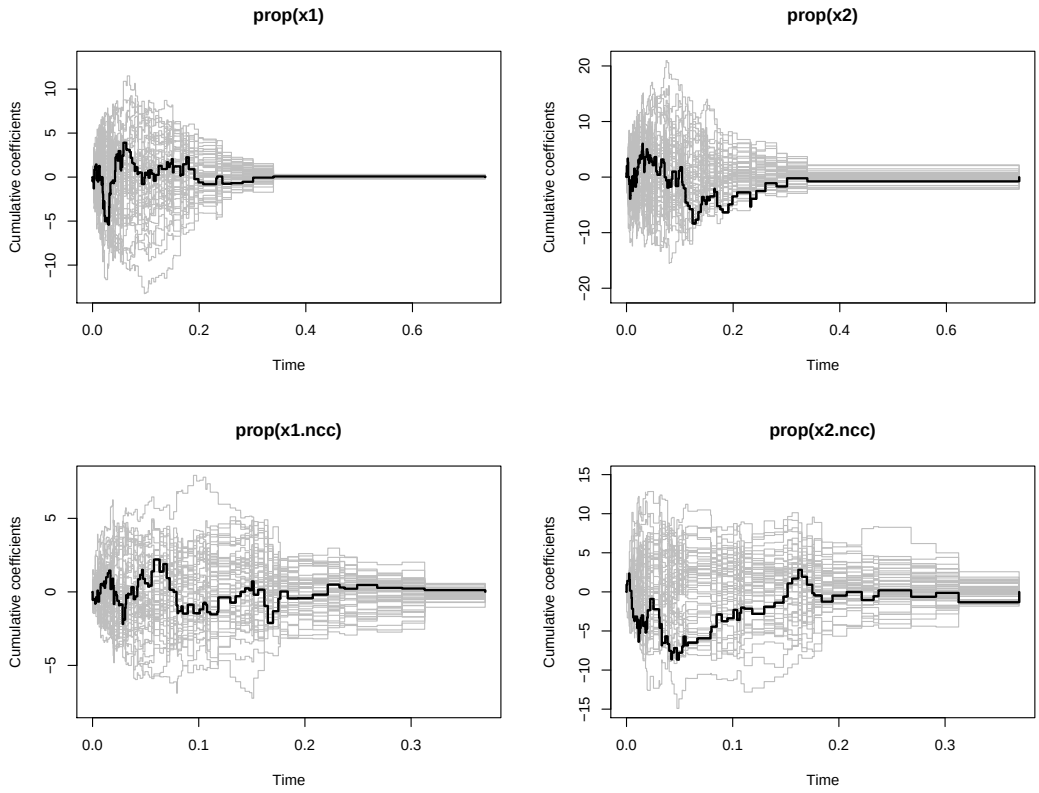

Figure 5.8: One simulation of score process against time for x_1 and x_2 in the good Cox model where $\boldsymbol{\beta} = (0.5, 0.5)^T$ for cohort and nested case-control data with simple random sampling using $m - 1 = 2$ controls

it ought to be mentioned here that ideally we should have chosen to run 500 simulations like we did for cumulative martingale residuals. However, these simulations turn out to be quite slow, which is the reason for the low number of simulations that we choose.

5.3.2 Counter-matched sampling

We will have a look at the check of proportionality for nested case-control data with counter-matched sampling. Using the same simulation model (5.1) with individuals stratified according to $x_1 \leq 0$ or $x_1 > 0$. Since x_1 is uniformly distributed, the number of individuals that belongs to two strata should be approximately the same. So we obtain one simulation plots of score process against time for x_1 and x_2 in the Cox model for cohort and nested case-control data with counter-matched sampling using $m - 1 = 1$ control in Figure A.7 in the appendix. As a result, the two paired curves resemble each other.

Then we run 100 simulations of the score process, and the distribution of the

Table 5.5: Proportion of P-values $\leq 5\%$ over 100 simulations of score process against time for x_1 and x_2 in the good Cox model where $\boldsymbol{\beta} = (0.5, 0.5)^T$ for cohort and nested case-control data with simple random sampling using $m-1=1$, $m-1=2$, $m-1=4$ controls

	controls	x_1: P-values$\leq 5\%$	x_2: P-values$\leq 5\%$
cohort		0.06	0.06
simple random	$m-1=1$	0.09	0.04
cohort		0.04	0.01
simple random	$m-1=2$	0.06	0.04
cohort		0.03	0.04
simple random	$m-1=4$	0.07	0.05

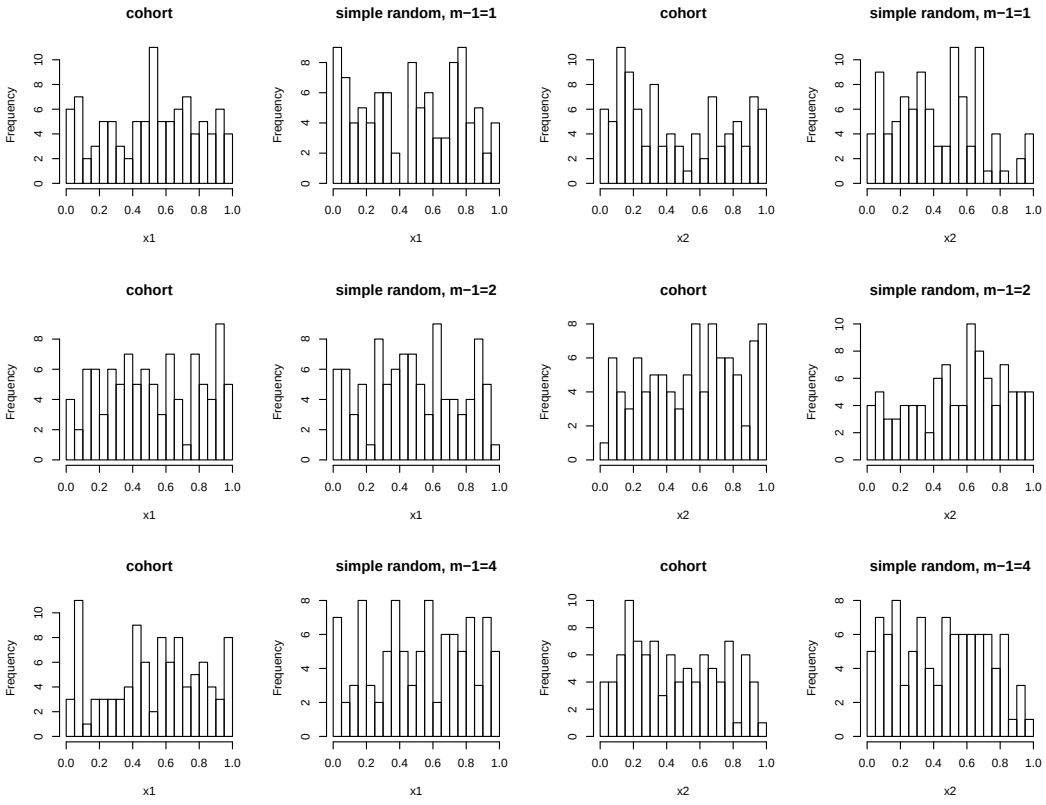

Figure 5.9: Histograms of P-values over 100 simulations of score process against time for x_1 and x_2 in the good Cox model where $\boldsymbol{\beta} = (0.5, 0.5)^T$ for cohort and nested case-control data with simple random sampling using $m-1=1$, $m-1=2$, $m-1=4$ controls

Table 5.6: Proportion of P-values $\leq 5\%$ over 100 simulations of score process against time for x_1 and x_2 in the good Cox model where $\boldsymbol{\beta} = (0.5, 0.5)^T$ for cohort and nested case-control data with counter-matched sampling using $m-1=1$, $m-1=3$, $m-1=5$ controls

	controls	x_1: P-values$\leq 5\%$	x_2: P-values$\leq 5\%$
cohort		0.08	0.00
counter-matching	$m-1=1$	0.06	0.03
cohort		0.01	0.05
counter-matching	$m-1=3$	0.02	0.03
cohort		0.07	0.04
counter-matching	$m-1=5$	0.08	0.03

P-values is shown in Figure A.8 in the appendix, which look quite uniform. Overall, the P-values corresponding to cohort and the nested case-control design with counter-matched sampling of controls are quite similar. In terms of the effective significance level, according to Table 5.6, we find that the proportions of P-values less than or equal to 5% are equally small. This shows that the proportionality assumption for Cox model is satisfied and correctly checked by nested case-control data with counter-matched sampling. In addition, as we have explained earlier, here we only choose to run 100 simulations because they are fairly time-consuming, and it would be quite a challenge to be able to run as many as 500 simulations.

5.4 Check the proportionality of a wrong model

5.4.1 Data simulation for a wrong model

We will look at the simulation for a model where proportionality fails. To do this, we first select x_1 and x_2 the same way as in Section 5.1.1. Further, if $x_{i2} = -1$, we draw uncensored survival time T_i^0 from exponential distribution with hazard $\exp\{0.5x_{i1} + 0.5x_{i2}\}$; while if $x_{i2} = 1$, we draw T_i^0 from Weibull distribution with hazard $kt^p \exp\{0.5x_{i1} + 0.5x_{i2}\}$, where k and p are parameters that satisfy $kt^p = 1$ when $t = 0.05$. In this case, the hazard ratio is no longer a fixed constant, but a value that is dependent on time. The reason why we make this choice is because we want to make sure that the event rate on the average is kept at about the same size as the one simulated for the good model (5.1). We make a plot of the hazards by choosing $x_{i1} = 0$ with three groups of parameters $(k, p) = (20, 1)$, $(0.05^{-0.5}, 0.5)$ and $(0.05^{-0.2}, 0.2)$. As shown in Figure 5.10, we can see that all lines are intersecting at the point where $t = 0.05$. The horizontal line corresponds to the hazard of the good model while the rest of lines represent the hazards of the wrong models.

More specifically, the line with the highest slope where $(k, p) = (20, 1)$ is going all the way up, which implies a strong non-proportional effect. The other two lines also reveal different levels of non-proportionality, with the one with respect to $(k, p) = (0.05^{-0.5}, 0.5)$ being slightly stronger.

To this end we select the regression coefficients as $\boldsymbol{\beta} = (0.5, 0.5)^T$, also define $k = 20$ and $p = 1$, thus we obtain a wrong model that takes the form

$$\alpha(t|\mathbf{x}_i) = \begin{cases} \exp\{0.5x_{i1} - 0.5\} & x_{i2} = -1 \\ 20t \exp\{0.5x_{i1} + 0.5\} & x_{i2} = 1 \end{cases}$$

where

$$x_{i1} \sim U[-1, 1]$$

$$x_{i2} = \begin{cases} 1 & \text{with probability } 0.5 \\ -1 & \text{with probability } 0.5 \end{cases}$$

Note that the censoring time C_i, censored survival time $\widetilde{T}_i$ and censoring indicator D_i are generated following exactly the same steps as in Section 5.1.1. Further, we would also like to consider some situations where the deviation from proportionality is not so clear. By defining $k = 0.05^{-0.5}$ and $p = 0.5$, the wrong model becomes

$$\alpha(t|\mathbf{x}_i) = \begin{cases} \exp\{0.5x_{i1} - 0.5\} & x_{i2} = -1 \\ 0.05^{-0.5} t^{0.5} \exp\{0.5x_{i1} + 0.5\} & x_{i2} = 1 \end{cases}$$

Finally we also choose $k = 0.05^{-0.2}$ and $p = 0.2$, thus we obtain a wrong model

$$\alpha(t|\mathbf{x}_i) = \begin{cases} \exp\{0.5x_{i1} - 0.5\} & x_{i2} = -1 \\ 0.05^{-0.2} t^{0.2} \exp\{0.5x_{i1} + 0.5\} & x_{i2} = 1 \end{cases}$$

5.4.2 Simple random sampling

We are now going to have a check of the proportionality of the wrong model simulated in Section 5.4.1 for cohort and nested case-control data with simple random sampling using different number of controls. We start off by obtaining the plot of one simulation of score process against time for x_1 and x_2 in three non-proportional Cox models. According to Figure 5.11, we observe that for all three groups x_1 seems to be proportional. However, x_2 seems to be non-proportional, especially in the first group, where the deviation from proportionality for x_2 is clearly stronger than the other two groups. This is also revealed in Figures 5.12, 5.13 and 5.14, which are histograms of P-values with respect to three wrong models using different number of controls.

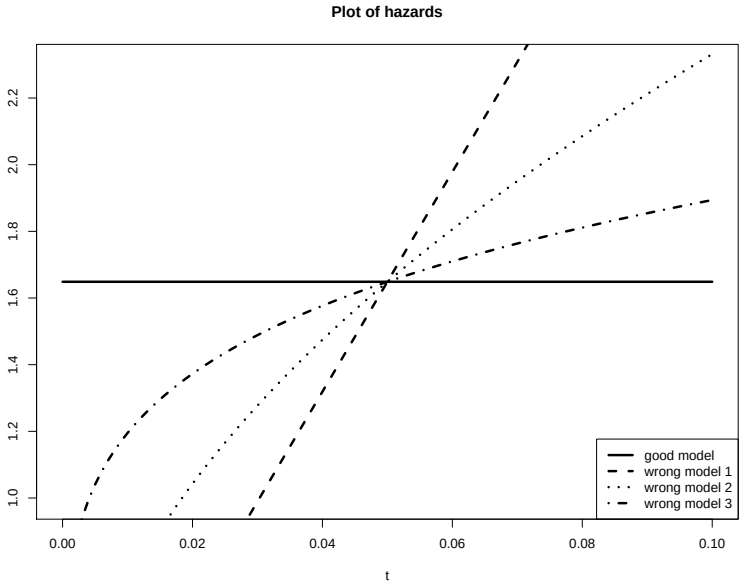

Figure 5.10: Plot of hazards (5.1) for the good model and three wrong models on log-scale where $\boldsymbol{\beta} = (0.5, 0.5)^T$ with $(k, p) = (20, 1)$, $(0.05^{-0.5}, 0.5)$ and $(0.05^{-0.2}, 0.2)$

Further we look at Table 5.7, it can be seen that for the first group of parameters, the proportions of P-values $\leq 5\%$ of x_1 and x_2 for both cohort and nested case-control data are hovering around 0.05 and 0.97 respectively, thus the model checking for nested case-control data has correctly disclosed that there is a strong non-proportional effect of x_2 but not of x_1. When it comes to the other two groups of parameters, apparently the medium and low non-proportional effect of x_2 can be detected by model checking for nested case-control data with simple random sampling as easily as cohort when at least $m - 1 = 2$ controls are used.

5.4.3 Counter-matched sampling

We look at the proportionality test of a wrong Cox model for nested case-control data with counter-matched sampling. First of all, we obtain Figure A.9 in the appendix which illustrates one simulation of score process against time for x_1 and x_2 in three non-proportional Cox models simulated in Section 5.4.1. Similarly, x_1 seems to satisfy the proportionality in all three models, but x_2 has different degrees of non-proportional effect. Histograms of P-values for three non-proportional Cox models for cohort and nested case-control data with counter-matched sampling are shown in Figures A.10, A.11 and A.12 in the appendix.

The first group of values in Table 5.8 reveals that the model check of the

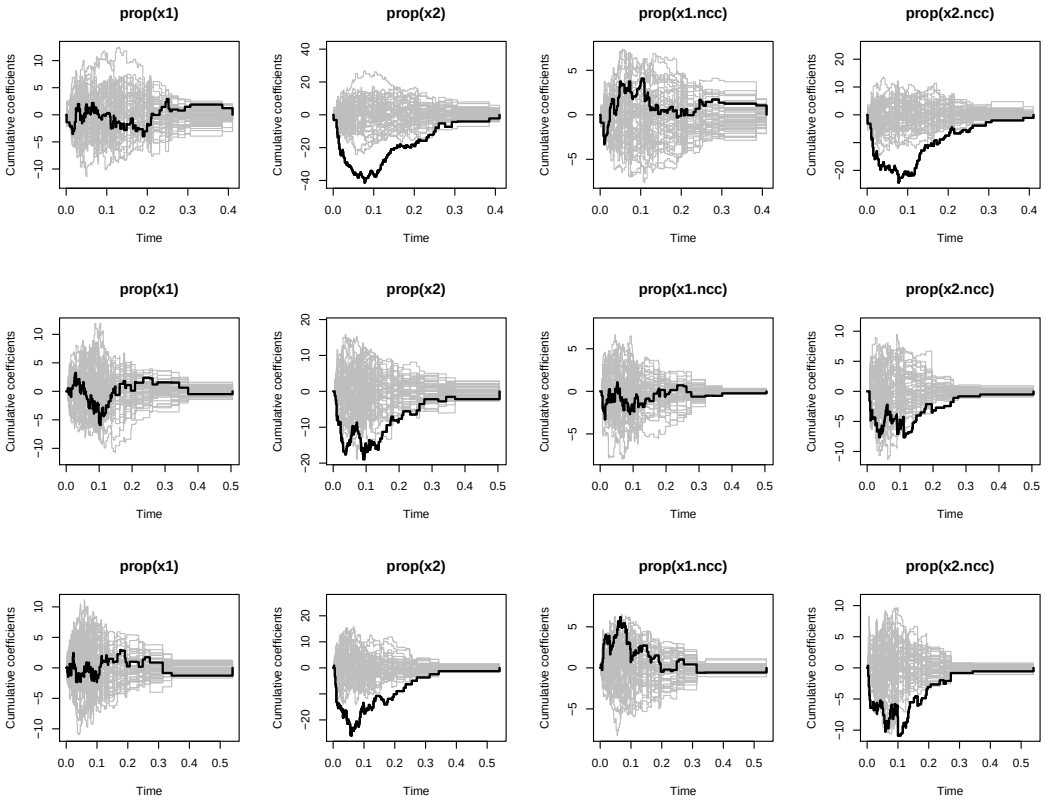

Figure 5.11: One simulation of score process against time for x_1 and x_2 in three non-proportional Cox models where $\boldsymbol{\beta} = (0.5, 0.5)^T$ with $(k, p) = (20, 1)$ (upper panel), $(0.05^{-0.5}, 0.5)$ (middle panel) and $(0.05^{-0.2}, 0.2)$ (lower panel) for cohort and nested case-control data with simple random sampling using $m - 1 = 2$ controls

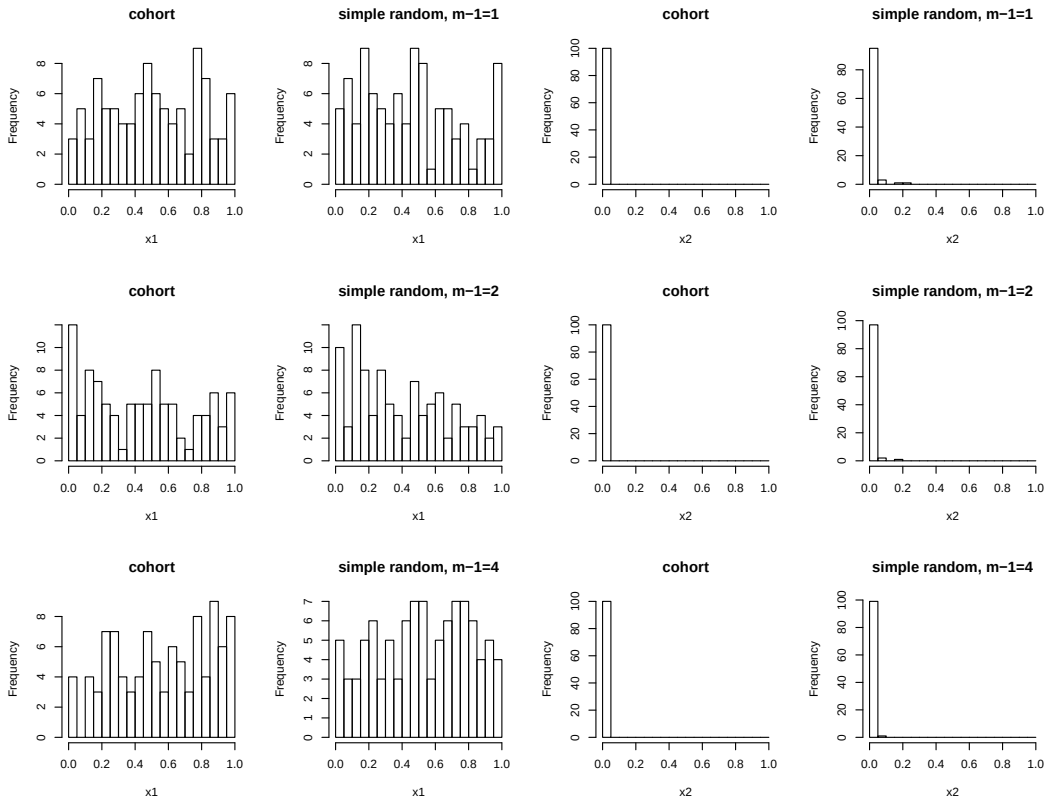

Figure 5.12: Histograms of P-values over 100 simulations of score process against time for x_1 and x_2 in the non-proportional Cox model where $\boldsymbol{\beta} = (0.5, 0.5)^T$ with $(k, p) = (20, 1)$ for cohort and nested case-control data with simple random sampling using $m - 1 = 1$, $m - 1 = 2$, $m - 1 = 4$ controls

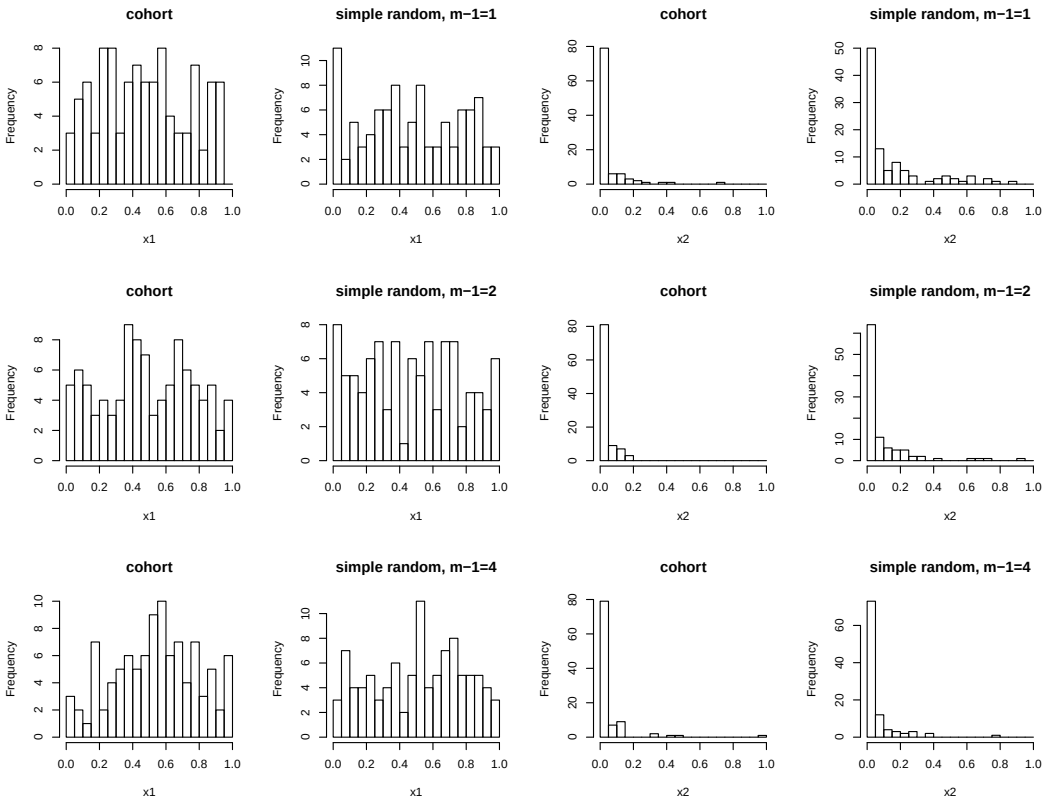

Figure 5.13: Histograms of P-values over 100 simulations of score process against time for x_1 and x_2 in the non-proportional Cox model where $\boldsymbol{\beta} = (0.5, 0.5)^T$ with $(k, p) = (0.05^{-0.5}, 0.5)$ for cohort and nested case-control data with simple random sampling using $m-1 = 1$, $m-1 = 2$, $m-1 = 4$ controls

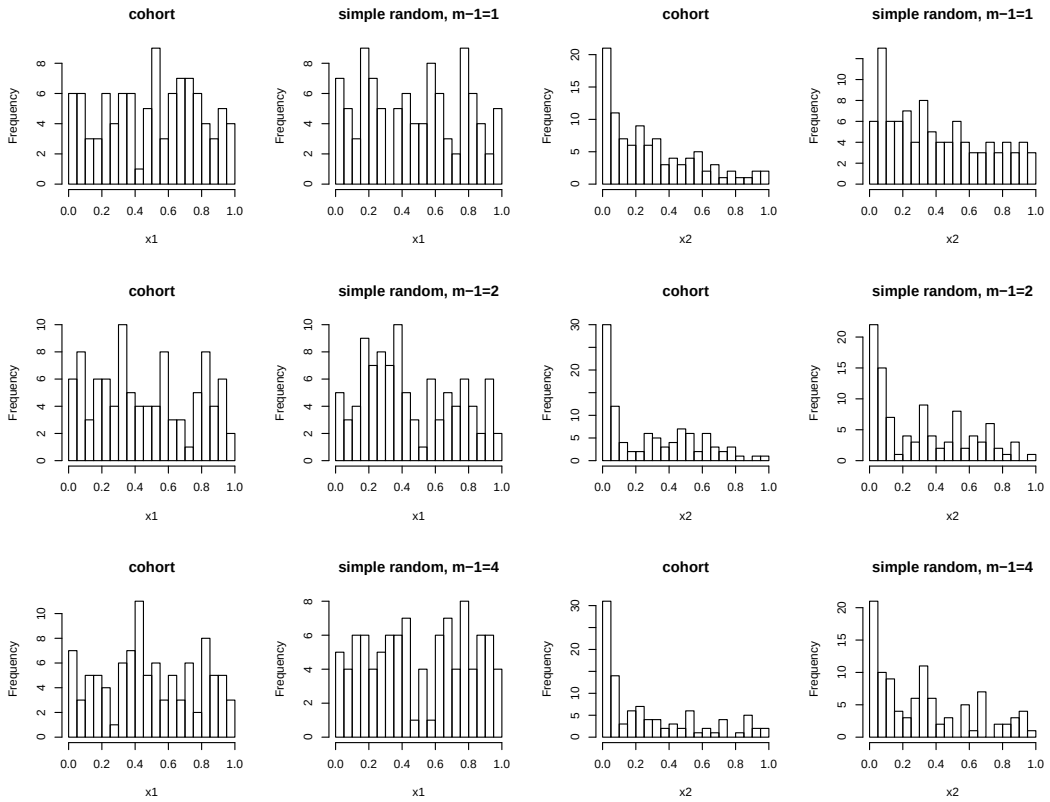

Figure 5.14: Histograms of P-values over 100 simulations of score process against time for x_1 and x_2 in the non-proportional Cox model where $\boldsymbol{\beta} = (0.5, 0.5)^T$ with $(k, p) = (0.05^{-0.2}, 0.2)$ for cohort and nested case-control data with simple random sampling using $m-1 = 1$, $m-1 = 2$, $m-1 = 4$ controls

Table 5.7: Proportion of P-values $\leq 5\%$ over 100 simulations of score process against time for x_1 and x_2 in three non-proportional Cox models where $\boldsymbol{\beta} = (0.5, 0.5)^T$ with $(k, p) = (20, 1)$, $(0.05^{-0.5}, 0.5)$ and $(0.05^{-0.2}, 0.2)$ for cohort and nested case-control data with simple random sampling using $m - 1 = 1$, $m - 1 = 2$, $m - 1 = 4$ controls

	controls	x_1: P-values$\leq 5\%$	x_2: P-values$\leq 5\%$
cohort		0.03	1
simple random	$m - 1 = 1$	0.05	0.95
cohort		0.12	1
simple random	$m - 1 = 2$	0.10	0.97
cohort		0.04	1
simple random	$m - 1 = 4$	0.05	0.99
cohort		0.03	0.79
simple random	$m - 1 = 1$	0.11	0.5
cohort		0.05	0.81
simple random	$m - 1 = 2$	0.08	0.64
cohort		0.03	0.79
simple random	$m - 1 = 4$	0.03	0.73
cohort		0.06	0.21
simple random	$m - 1 = 1$	0.07	0.06
cohort		0.06	0.30
simple random	$m - 1 = 2$	0.05	0.22
cohort		0.07	0.31
simple random	$m - 1 = 4$	0.05	0.21

Table 5.8: Proportion of P-values $\leq 5\%$ over 100 simulations of score process against time for x_1 and x_2 in three non-proportional Cox models where $\boldsymbol{\beta} = (0.5, 0.5)^T$ with $(k, p) = (20, 1)$, $(0.05^{-0.5}, 0.5)$ and $(0.05^{-0.2}, 0.2)$ for cohort and nested case-control data with counter-matched sampling using $m-1 = 1$, $m-1 = 3$, $m-1 = 5$ controls

	controls	x_1: P-values$\leq 5\%$	x_2: P-values$\leq 5\%$
cohort		0.05	1
counter-matching	$m-1=1$	0.13	0.90
cohort		0.07	1
counter-matching	$m-1=3$	0.18	1
cohort		0.06	1
counter-matching	$m-1=5$	0.18	1
cohort		0.08	0.82
counter-matching	$m-1=1$	0.12	0.49
cohort		0.05	0.85
counter-matching	$m-1=3$	0.08	0.65
cohort		0.03	0.83
counter-matching	$m-1=5$	0.05	0.74
cohort		0.05	0.21
counter-matching	$m-1=1$	0.07	0.10
cohort		0.09	0.21
counter-matching	$m-1=3$	0.11	0.18
cohort		0.02	0.23
counter-matching	$m-1=5$	0.02	0.25

strong non-proportional Cox model for nested case-control data with counter-matched sampling works very good. In terms of the second group where the model is in medium non-proportionality, we find that the counter-matching only gives a value 0.49 when $m-1=1$ control is used. In contrast with the value 0.82 that corresponds to the full cohort, it is quite obvious that the model checking result for the counter-matching design is not accurate enough. Despite of that, by comparing the other values in this group we are certain that the model checking for counter-matching design works equally good as cohort when at least $m-1=3$ controls are chosen. Eventually, the same problem applies to the last group as well. In the situation when the Cox model is slightly non-proportional, the model checking for counter-matching design cannot be able to work as well as cohort, unless there are at a minimum $m-1=3$ controls being selected.

To sum up, through studying the performance of the Cox model checking for simulated nested case-control data, we can reach the conclusion that checking both the log-linearity and proportionality of the correct Cox models for nested case-control design is working as good as full cohort, even by using

1 control per case. Nevertheless, when the Cox model is wrong, it is required to use no less than 2 or 3 controls for nested case-control design in order to correctly detect a non-linear and non-proportional effect.

Chapter 6

Discussion

In this final chapter of the thesis, we would like to give a brief summary of the conclusion together with some further discussion about the problems that we have encountered.

6.1 Conclusion

In Chapter 4 we have shown how the model checking methods of Lin et al. (1993) for cohort data based on cumulative residuals processes may be extended to nested case-control data. Further in Chapter 4 and 5 we have studied the performance of the model checking methods using cumulative sums of martingale-based residuals.

The model checking in Chapter 4 is based on a real data set, and the results of the log-linearity test and proportionality test given by the full cohort and the nested case-control studies are fairly close, indicating that model checking for nested case-control data is quite efficient. We also show that the larger number of controls selected for the nested case-control design, the more they get close to the cohort. But it is unnecessary to choose too many controls. According to the example given in Chapter 4, we see that 2 controls are already effective enough.

In Chapter 5 we generalize the model checking to simulated data sets, where both good and wrong Cox models are available, in order to see if model checking for nested case-control data is still able to give a similar result to the full cohort. More specifically, for the good models which satisfy the log-linearity and proportionality assumptions, the performance of the model checking for nested case-control data is quite good, even with 1 control per case. In terms of the wrong models, the model checking for nested case-control data can detect the large deviation from log-linearity and proportionality as easy as

for the cohort even if there is only 1 control being used. However, for those wrong models with medium or low violation, in order to keep the efficiency of the model checking for nested case-control data, 2 or more controls are required. Overall, the analysis shows that the model checking for nested case-control data and the cohort resemble each other in many different situations, which confirms that the extension of model checking techniques to nested case-control data is working fine.

6.2 Problems

Now we will have some discussion about the problems that we have come across. To begin with, in Chapter 4 we have given examples about model checking of the Cox model for nested case-control studies. A common problem that we encounter is that when selecting only 1 or 2 controls for nested case-control data, it is not always possible to obtain a similar P-value of the log-linearity test or proportionality test as for the full cohort. Sometimes they are quite close, but sometimes they look a bit different. This is partly due to the randomness of control selection. Moreover, in cases where too many events are observed in the cohort studies, it is likely that a nested case-control design with small number of controls will not be able to capture enough information. A good way to solve this problem is by adding more controls. As we see in Chapter 5, the nested case-control design turns out to be quite effective when using 4 or 5 controls per case.

A second problem is in Chapter 5 where we want to simulate the data such that the histograms of P-values in the Cox model for cohort and nested case-control data have uniform distributions when the model is correct. As a result we did not always manage to achieve that. To be more precise, at the start when we simulate data for the model checking of log-linearity, we end up in all non-uniform histograms where the majority of P-values are either two low or two high. But in the later part of the chapter where we work on the model checking of proportionality, the histograms look quite good and appear to be fairly uniform. Although the simulation of data is not always as good as expected, it still does not affect the conclusion since our goal in this chapter is not to obtain the perfect simulated data, but to compare the result of model checking for nested case-control data with that for full cohort data in various situations by using simulation.

Another problem worth mentioning is, when we are trying to calculate the P-values over 500 simulations of score process in Chapter 5, we find that it is taking an awful lot of time to run the R program. It is estimated that a high configuration laptop with i7 CPU and 8G RAM requires literally 20-25 hours to run these 500 simulations of score process. As a result, we have to adjust that by lowering the number of simulations to 100 in order to make it faster.

The cause of the slow running speed is that the *timereg* package that we used for model checking is slightly slow, plus there are many loops and samplings involved in the codes that we wrote. There should be some possible ways to optimize the structure of the coding. However, considering that it is not the goal of the thesis, at this stage we decide not to work on the enhancement of it. Besides, when it comes to real nested control-case studies, one does not actually need to run this R program many times. So we can imagine that in practice the time for running the model checking program is still reasonable. In other words, this problem might not be deserving special attention.

Appendix A

Plots

In the appendix we list some plots for Sections 5.1.3, 5.2.3, 5.3.2 and 5.4.3 of Chapter 5 which are relevant to nested case-control studies with counter-matched sampling. The reason why we put them here is because these plots seem to resemble those in the simple random sampling sections, we decide not to include them in the main text.

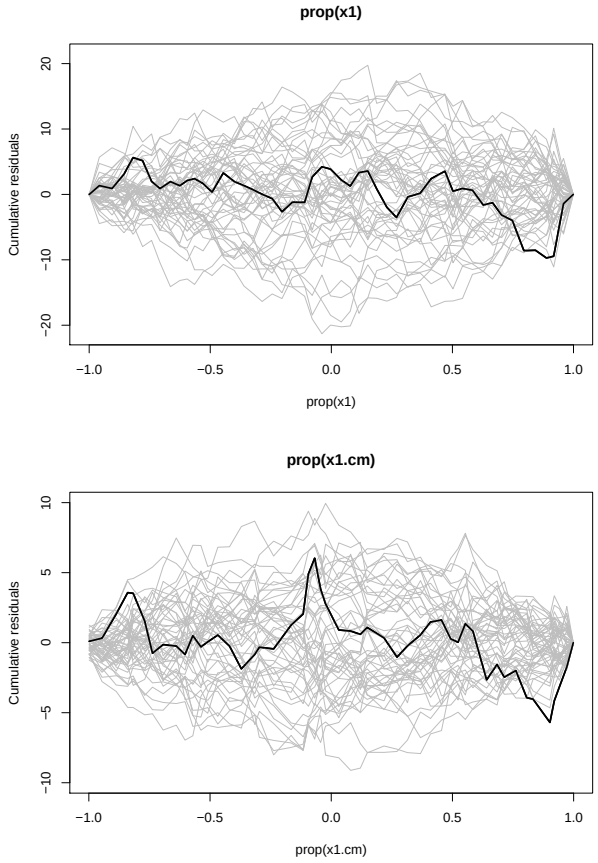

Figure A.1: One simulation of cumulative martingale residuals against x_1 in the good Cox model where $\boldsymbol{\beta} = (0.5, 0.5)^T$ for cohort and nested case-control data with counter-matched sampling using $m - 1 = 1$ control

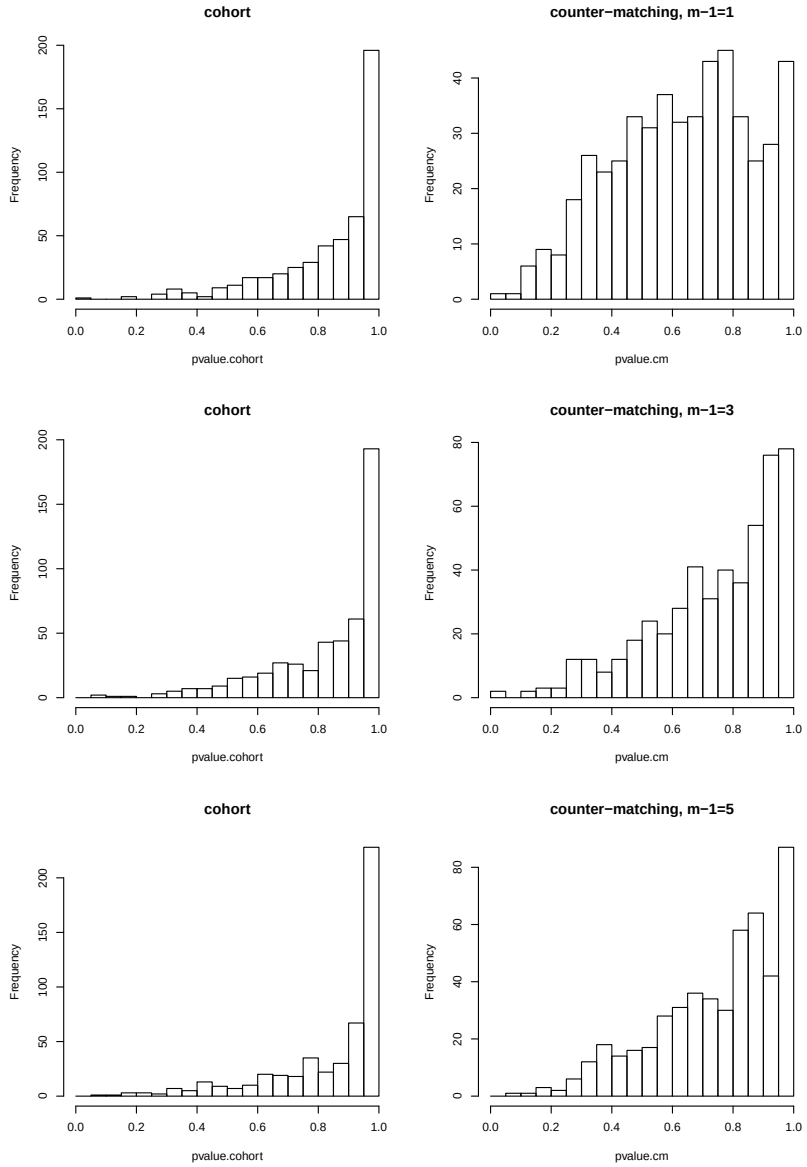

Figure A.2: Histograms of P-values over 500 simulations of cumulative martingale residuals against x_1 in the good Cox model where $\boldsymbol{\beta} = (0.5, 0.5)^T$ for cohort and nested case-control data with counter-matched sampling using $m-1=1$, $m-1=3$, $m-1=5$ controls

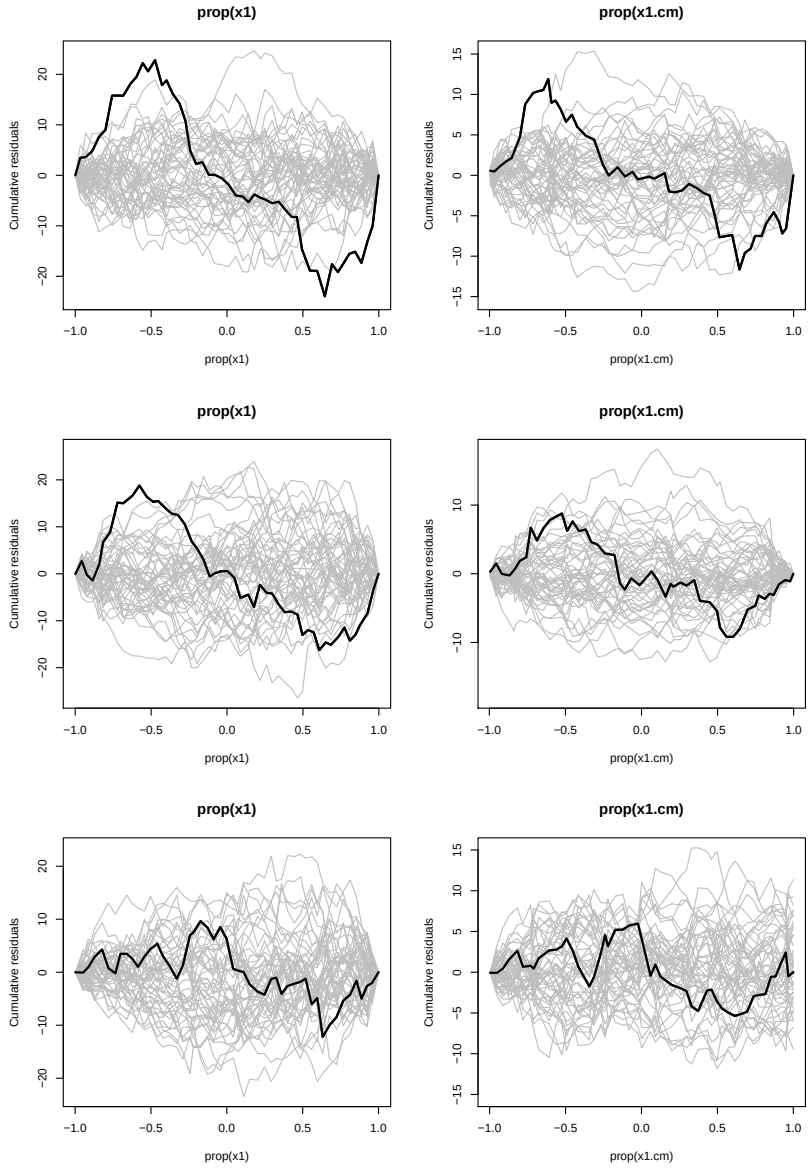

Figure A.3: One simulation of cumulative martingale residuals against x_1 in three non-linear Cox models where $\boldsymbol{\beta} = (0.1, 1.0, 0.5)^T$ (upper panel), $\boldsymbol{\beta} = (0.5, 0.8, 0.5)^T$ (middle panel) and $\boldsymbol{\beta} = (0.8, 0.6, 0.8)^T$ (lower panel) for cohort and nested case-control data with counter-matched sampling using $m - 1 = 1$ control

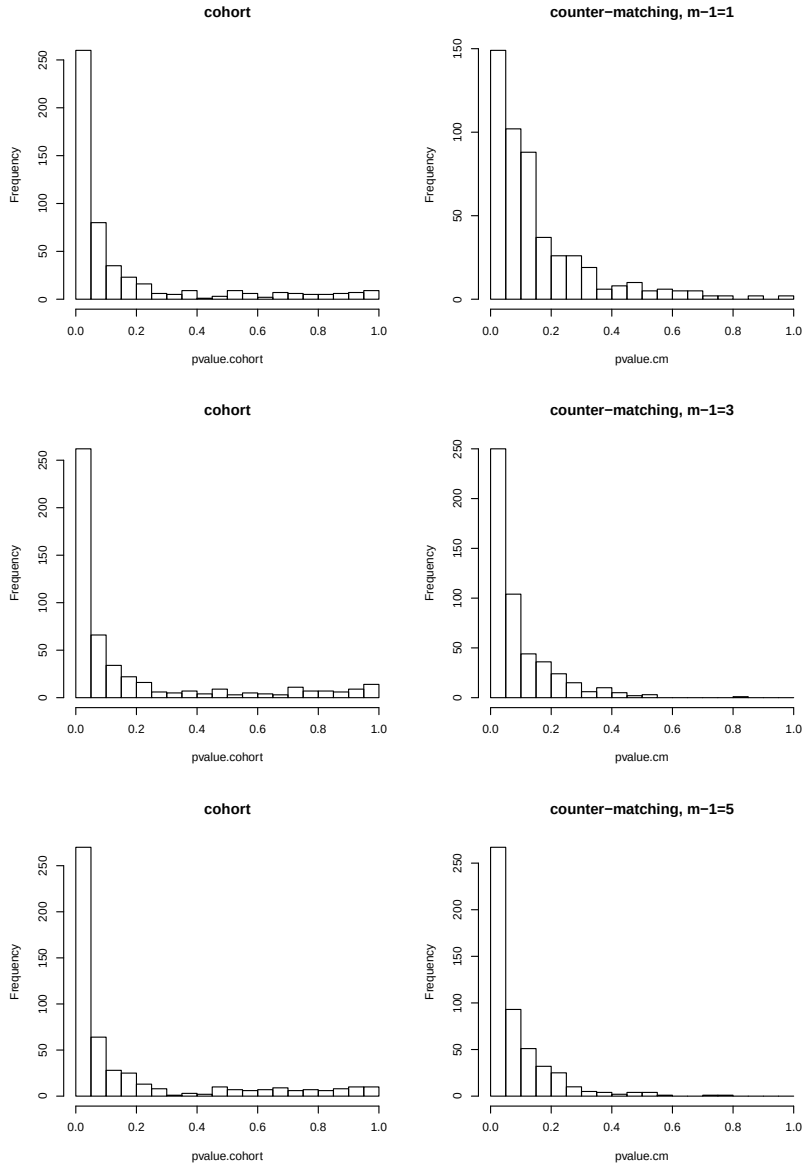

Figure A.4: Histograms of P-values over 500 simulations of cumulative martingale residuals against x_1 in the non-linear Cox model where $\boldsymbol{\beta} = (0.1, 1.0, 0.5)^T$ for cohort and nested case-control data with counter-matched sampling using $m-1=1$, $m-1=3$, $m-1=5$ controls

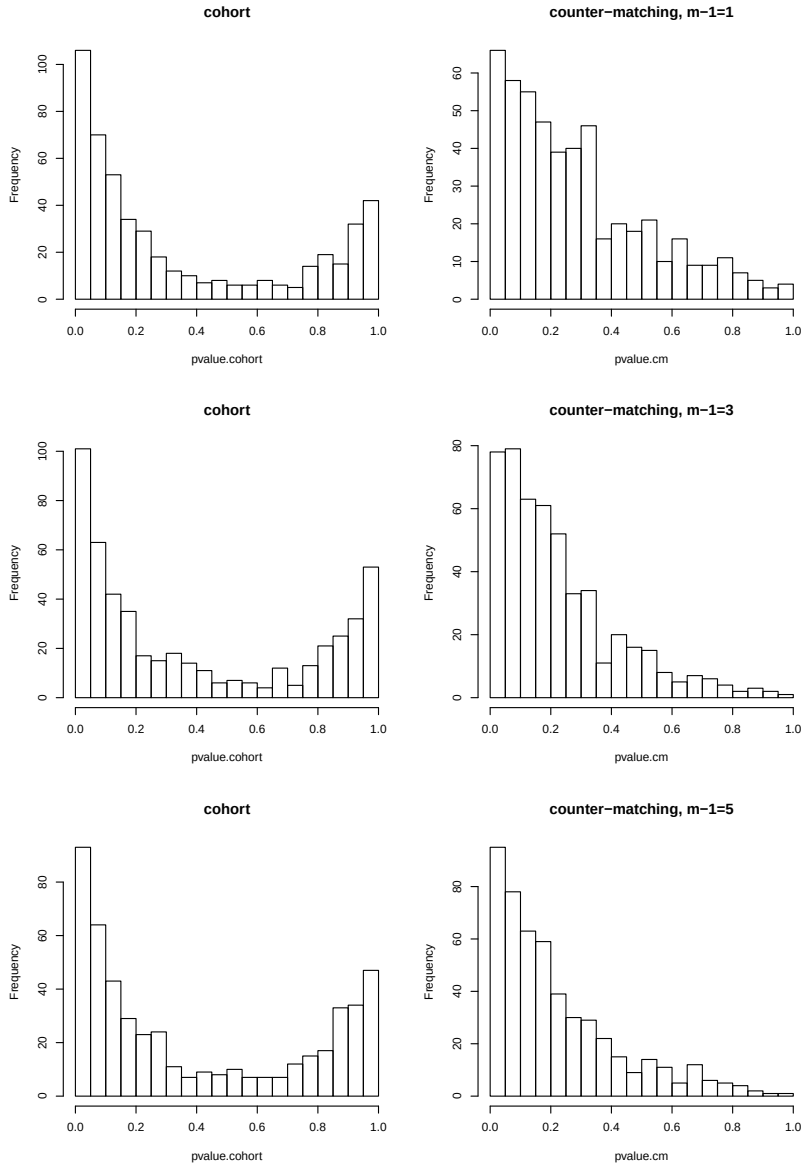

Figure A.5: Histograms of P-values over 500 simulations of cumulative martingale residuals against x_1 in the non-linear Cox model where $\boldsymbol{\beta} = (0.5, 0.8, 0.5)^T$ for cohort and nested case-control data with counter-matched sampling using $m-1=1$, $m-1=3$, $m-1=5$ controls

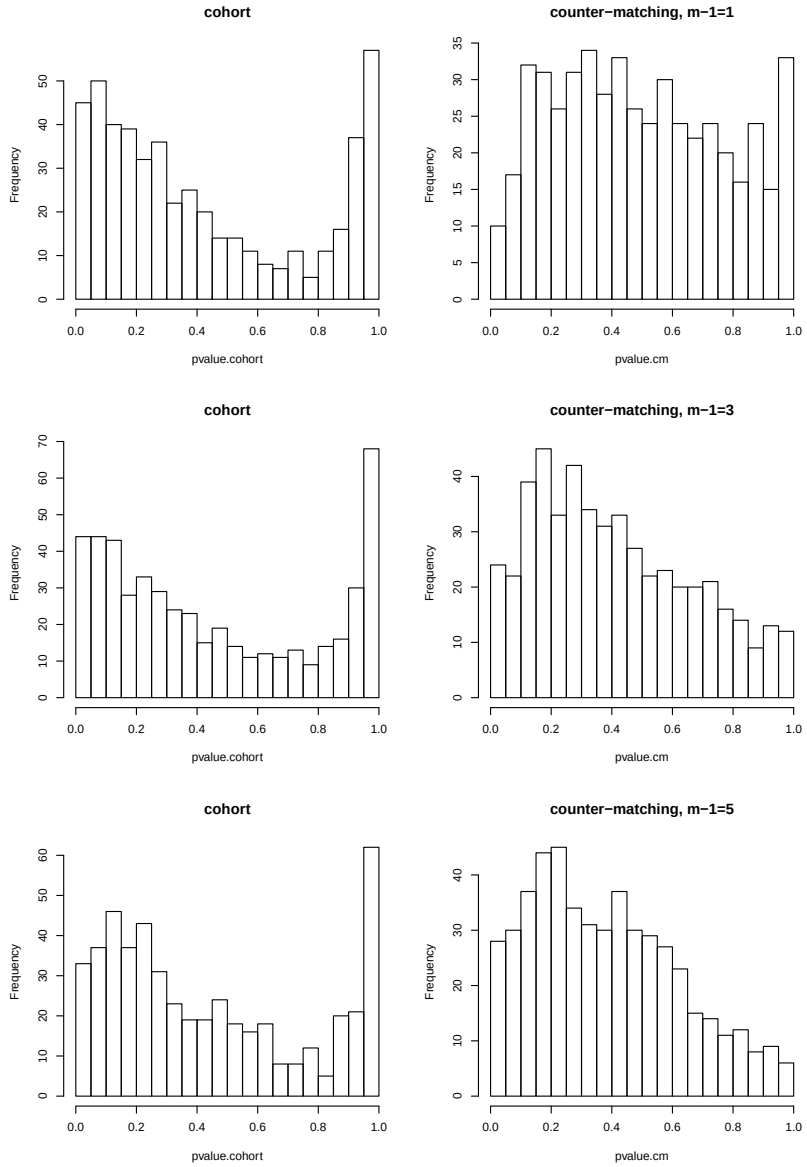

Figure A.6: Histograms of P-values over 500 simulations of cumulative martingale residuals against x_1 in the non-linear Cox model where $\boldsymbol{\beta} = (0.8, 0.6, 0.8)^T$ for cohort and nested case-control data with counter-matched sampling using $m-1=1$, $m-1=3$, $m-1=5$ controls

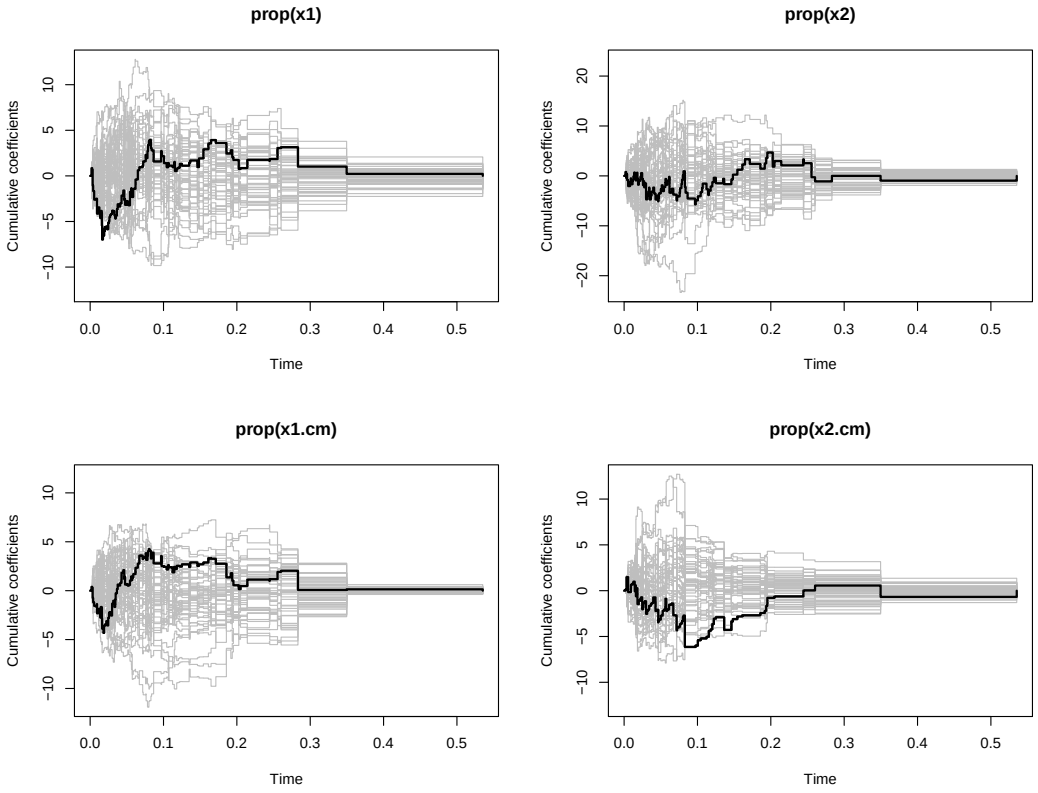

Figure A.7: One simulation of score process against time for x_1 and x_2 in the good Cox model where $\boldsymbol{\beta} = (0.5, 0.5)^T$ for cohort and nested case-control data with counter-matched sampling using $m - 1 = 1$ control

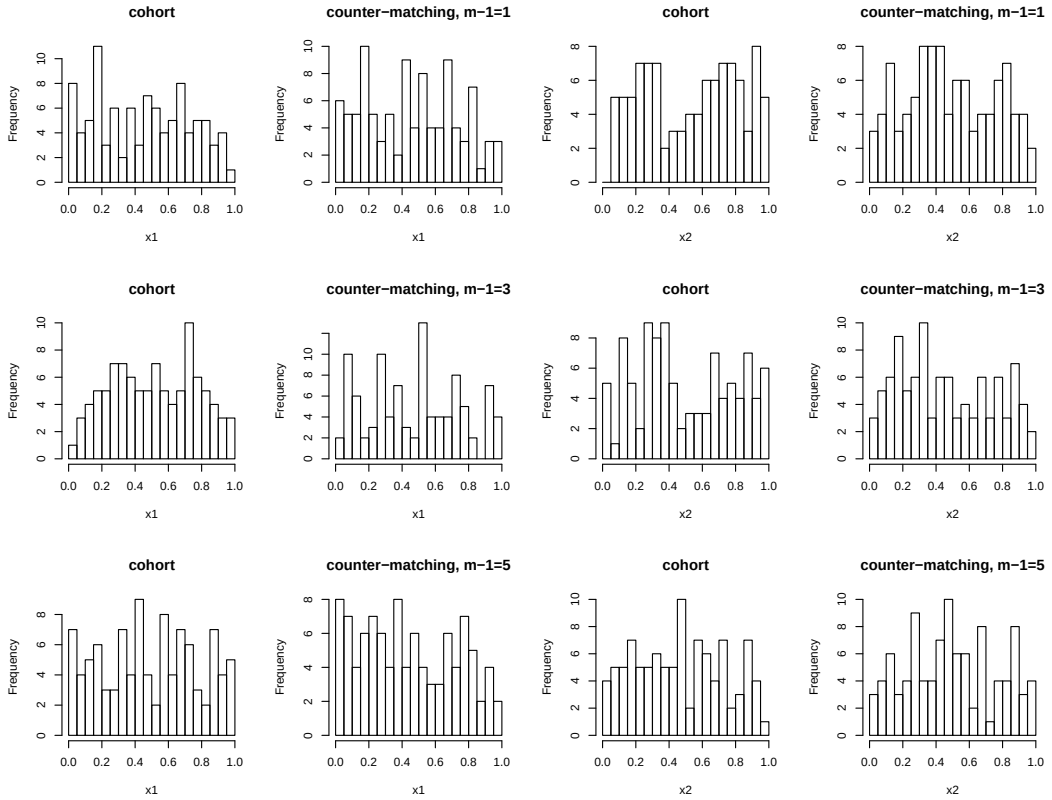

Figure A.8: Histograms of P-values over 100 simulations of score process against time for x_1 and x_2 in the good Cox model where $\boldsymbol{\beta} = (0.5, 0.5)^T$ for cohort and nested case-control data with counter-matched sampling using $m-1=1$, $m-1=3$, $m-1=5$ controls

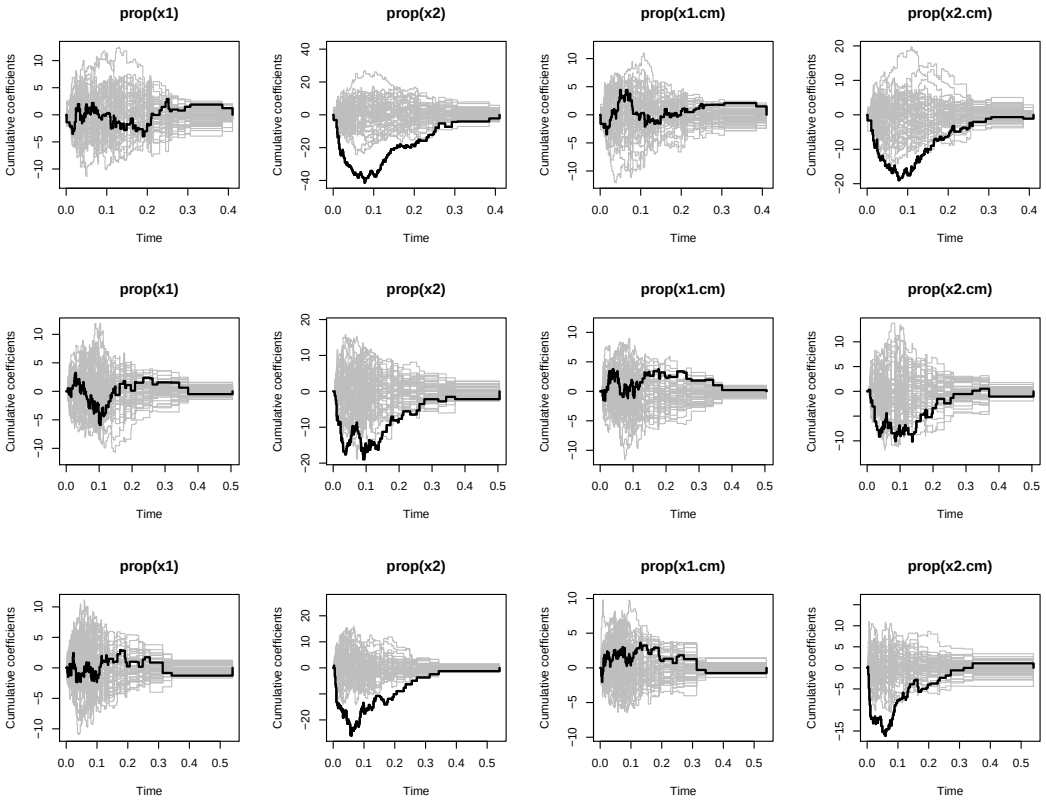

Figure A.9: One simulation of score process against time for x_1 and x_2 in three non-proportional Cox models where $\boldsymbol{\beta} = (0.5, 0.5)^T$ with $(k, p) = (20, 1)$ (upper panel), $(0.05^{-0.5}, 0.5)$ (middle panel) and $(0.05^{-0.2}, 0.2)$ (lower panel) for cohort and nested case-control data with counter-matched sampling using $m - 1 = 1$ control

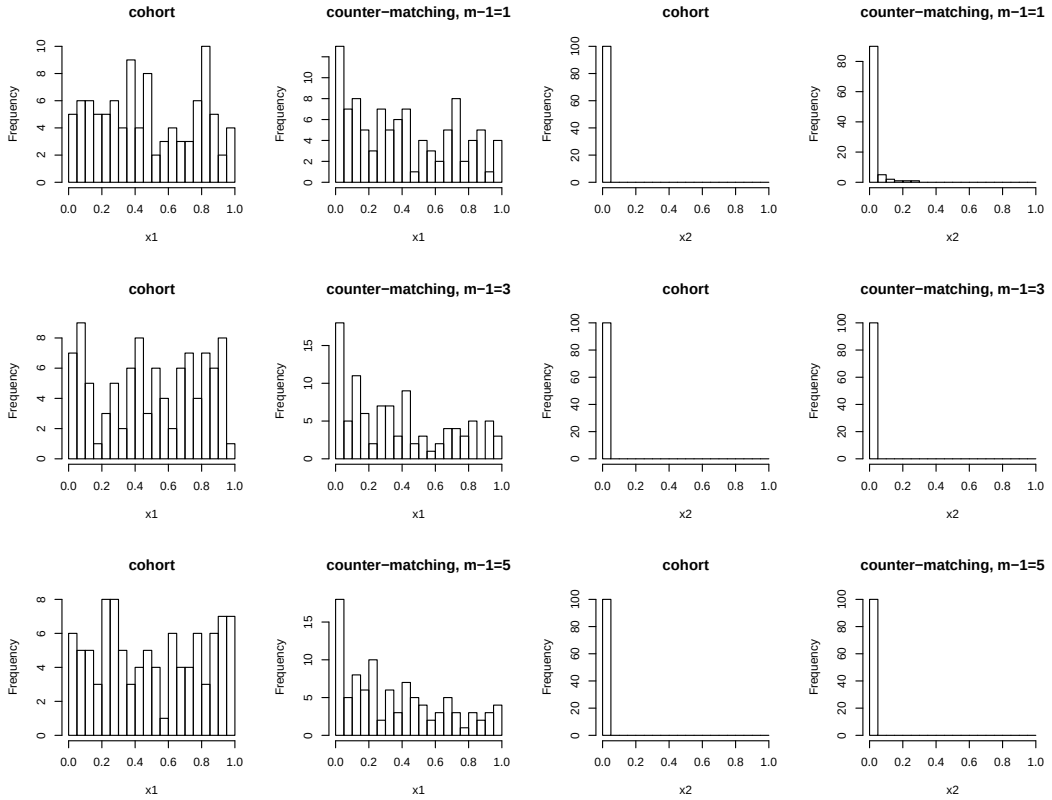

Figure A.10: Histograms of P-values over 100 simulations of score process against time for x_1 and x_2 in the non-proportional Cox model where $\boldsymbol{\beta} = (0.5, 0.5)^T$ with $(k, p) = (20, 1)$ for cohort and nested case-control data with counter-matched sampling using $m - 1 = 1$, $m - 1 = 3$, $m - 1 = 5$ controls

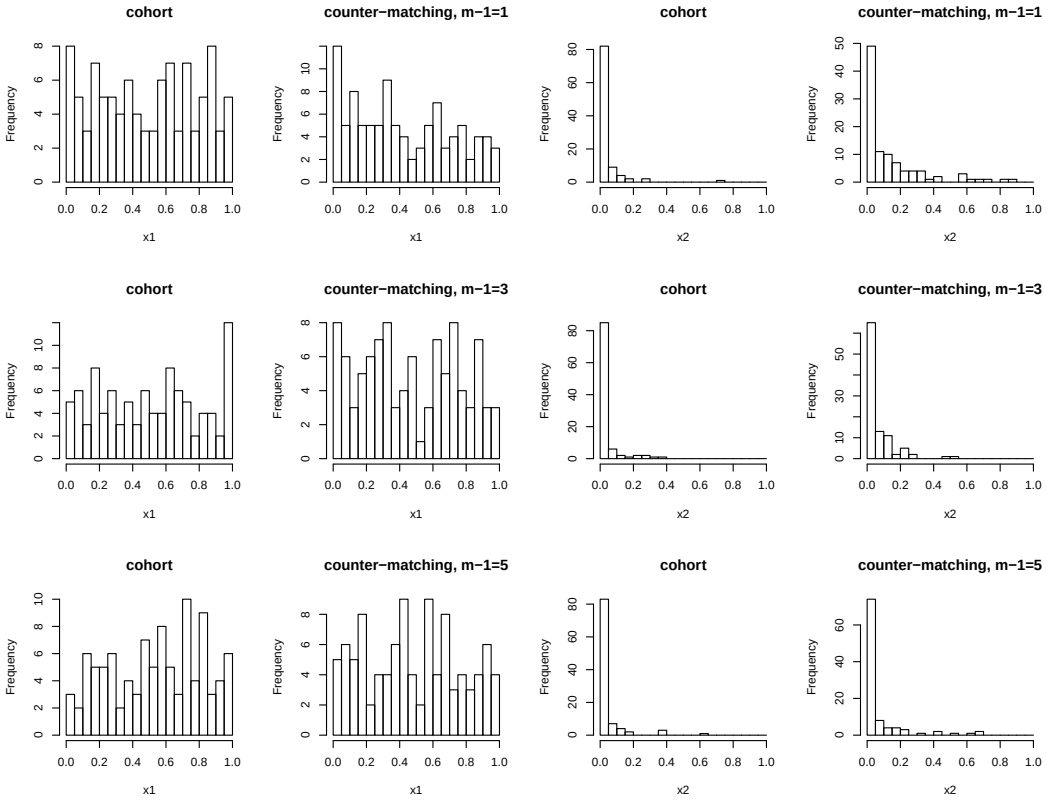

Figure A.11: Histograms of P-values over 100 simulations of score process against time for x_1 and x_2 in the non-proportional Cox model where $\boldsymbol{\beta} = (0.5, 0.5)^T$ with $(k, p) = (0.05^{-0.5}, 0.5)$ for cohort and nested case-control data with counter-matched sampling using $m-1=1$, $m-1=3$, $m-1=5$ controls

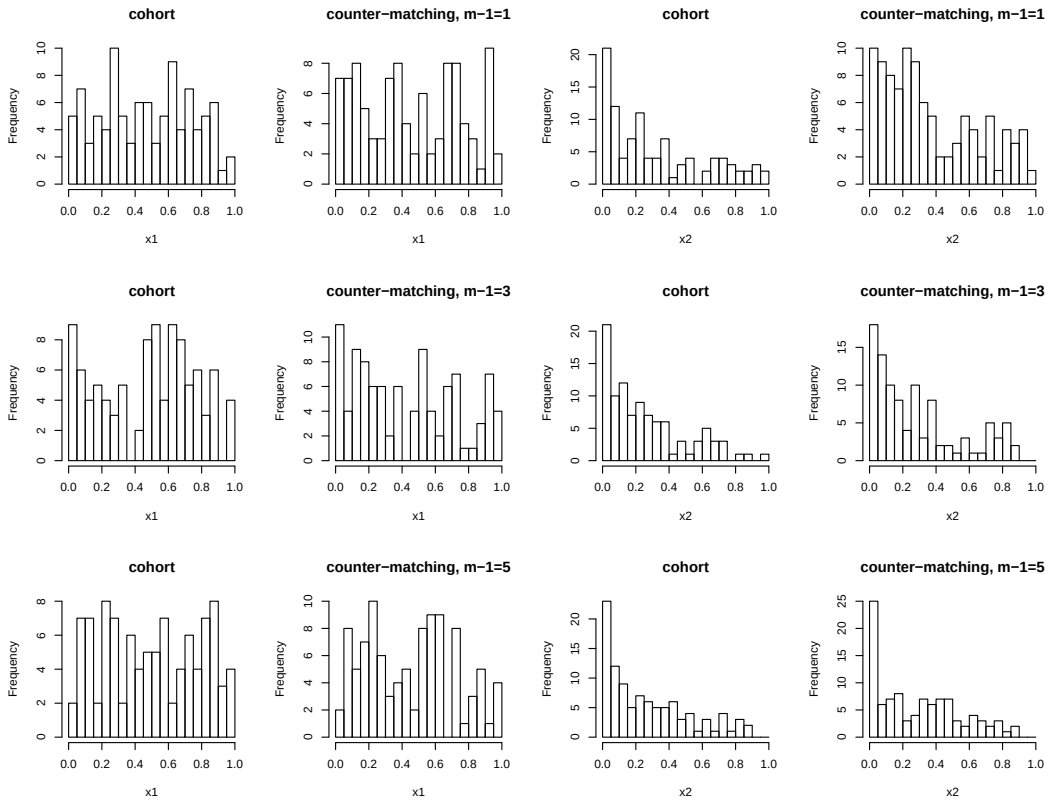

Figure A.12: Histograms of P-values over 100 simulations of score process against time for x_1 and x_2 in the non-proportional Cox model where $\boldsymbol{\beta} = (0.5, 0.5)^T$ with $(k, p) = (0.05^{-0.2}, 0.2)$ for cohort and nested case-control data with counter-matched sampling using $m - 1 = 1$, $m - 1 = 3$, $m - 1 = 5$ controls

Bibliography

Aalen, O. O., Borgan, Ø. and Gjessing, H. K. (2008). *Survival and Event History Analysis: A Process Point of View*. Springer-Verlag, New York.

Sauerbrei, W. and Royston, P. (1999). *Multivariate Model-building*. Wiley, New York.

Lin, D. Y., Wei, L. J. and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557-572.

Martinussen, T. and Scheike, T. H. (2006). *Dynamic Regression Models for Survival Data*. Springer-Verlag, New York.

Borgan, Ø. and Samuelsen, S. O. (2013). Nested case-control and case-cohort studies. In *Handbook of Survival Analysis* edited by Klein, J. P., Ibrahim, J. G., Scheike, T. and van Houwelingen, J.. Chapman and Hall, London.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis. Techniques for Censored and Truncated Data* (2nd ed.). Springer-Verlag, New York.

Borgan, Ø. and Langholz, B. (2007). Using martingale residuals to assess goodness-of-fit for sampled risk set data. In V. Nair (Ed.), *Advances in Statistical Modeling and Inference. Essays in Honor of Kjell A Doksum*, pp. 65-90. World Scientific Publishing, Singapore.

Hrubec, Z., Boice, Jr., J. D., Monson, R. R. and Rosenstein, M. (1989). Breast cancer after multiple chest fluoroscopies: second follow-up of Massachusetts women with tuberculosis. *Cancer Research* **49**, 229-234.