FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Bc. Karla Strachoňová

Semiparametric Analysis of Nested Case-Control Design

Department of Probability and Mathematical Statistics

Supervisor of the master thesis: doc. Mgr. Michal Kulich, Ph.D.
Study programme: Mathematics
Study branch: Probability, mathematical statistics and econometry

Prague 2022

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In .... date ....

Author's signature

I would like to express my gratitude to doc. Mgr. Michal Kulich, Ph.D. for his advice, kindness and professional guidance that helped me to finish this work. I would like to dedicate this thesis to my family and especially my parents who supported me financially as well as mentally throughout my studies. I would like to thank my boyfriend and my friends who made my studies manageable with their constant help and emotional support.

Title: Semiparametric Analysis of Nested Case-Control Design

Author: Bc. Karla Strachoňová

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. Mgr. Michal Kulich, Ph.D., Department of Probability and Mathematical Statistics

Abstract: Studying rare diseases often deals with small percentage of cases requiring a large amount of subjects in the medical study. The common analysis by the Cox proportional hazards model may be very time-consuming and financially inefficient. Nested case-control design presents a sampling method offering fewer data needed for the analysis while keeping the estimator of the Cox model consistent and asymptotic normal. In this thesis, we introduce nested case-control design, we describe in detail the method for sampling controls for cases, we present the partial likelihood and the maximum partial likelihood estimator of the regression parameter and we prove the consistency and the asymptotic normality of the estimator. Then, we introduce the counter-matching design as an extension of the nested case-control design and the pseudolikelihood approach under nested case-control design. In the last chapter, we perform a simulation study comparing the four designs. The contribution of this thesis is the detailed introduction to nested case-control design and its alternatives, more detailed proofs of the asymptotic properties of the maximum partial likelihood of the regression parameter of nested case-control design and the comparison of the four approaches through the simulation study.

Keywords: nested case-control design, survival analysis, Cox model, counter-matching, pseudolikelihood

Název práce: Semiparametrická analýza vnořené studie případů a kontrol

Autor: Bc. Karla Strachoňová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: doc. Mgr. Michal Kulich, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Při studiu vzácných chorob pozorujeme malé procento případů, což vyžaduje velké množství subjektů ve studii. Obvyklý způsob analýzy těchto dat pomocí Coxova modelu proporcionálních rizik může být v tomto případě časově a finančně neefektivní. Metoda vnořené studie případů a kontrol umožňuje snížit množství dat potřebnou pro analýzu a přitom zachovat konzistenci a asymptotickou normalitu odhadnutého parametru Coxova modelu. V této práci představujeme vnořenou studii případů a kontrol, detailně popisujeme metodu výběrů kontrol pro případy, uvádíme parciální věrohodnost a maximálně parciálně věrohodný odhad regresního parametru a dokazujeme jeho konzistenci a asymptotickou normalitu. Poté představujeme metodu ”counter-matching” jako rozšíření již známé vnořené studie a uvádíme též ”pseudo-věrohodnostní” přístup k tomuto problému. V poslední kapitole této práce provádíme simulační studii, jež porovnává tyto čtyři představené přístupy. Příspěvkem této práce je detailní představení vnořené studie případů a kontrol a jejích dvou alternativních metod, dále podrobněji provedené důkazy asymptotických vlastností maximálně parciálně věrohodného odhadu regresního parametru vnořené studie a nakonec porovnání všech čtyř přístupů pomocí simulační studie.

Klíčová slova: vnořená studie případů a kontrol, analýza přežití, Coxův model, counter-matching, pseudověrohodnost

Contents

Introduction

Time-to-event data is a type of data measuring the time until a specific event, such as an occurrence of a disease, death, failure of a machine or device, occurrence of an insurance claim, etc. The analysis of time-to-event data in medical applications is called *survival analysis*. In the medical field, the main focus of it is studying an occurrence of a specific disease. Medical studies usually last for a few years, therefore the disease of interest might not develop for some individuals during that time, so there is censoring involved inevitably. This type of data requires specialized statistical methods and is commonly analyzed by the Cox proportional hazards model introduced by Cox [1972].

The aim of this thesis is to introduce the nested case-control design and some of its alternatives. The nested case-control design, first introduced by Cunliffe et al. [1977], is one of the most popular designs used in practice when studying rare diseases. Studying such diseases may be time and money inefficient, since only a small percentage of the subjects get the disease of interest during the time of the ongoing study, which causes the necessity of having a large amount of subjects. The nested case-control design presents a method for sampling only a few controls (individuals who have not had the disease yet) for each observed case (individual who got the disease of interest) at the time the case's disease occurred. This reduction of the number of subjects in the study decreases the time and money needed to measure all of the necessary covariates of the model.

In the first chapter of the thesis, we present some theoretical foundations of the Cox proportional hazards model for time-to-event data as described by Andersen and Gill [1982] and we make a brief introduction to the martingale theory written by Fleming and Harrington [1991]. The second chapter is the core of the thesis where we describe the nested case-control design, which can be analyzed by partial likelihood, similarly to the Cox model introduced in the first chapter. We then present the maximum partial likelihood estimator of the regression parameter and prove its asymptotic properties (consistency and asymptotic normality) in a similar way to the one used by Goldstein and Langholz [1992]. In the third chapter, we present two alternatives to the full cohort and nested case-control design: counter-matching (introduced by Borgan et al. [1995]) and the pseudo-likelihood approach (presented by Samuelsen [1997]). In the fourth chapter, we conduct a simulation study in which we model the risk of developing lung cancer of smoking cigarettes. We then use all four approaches described in the thesis to estimate the regression parameter and to compare the accuracy of the designs for different expected prevalences of the disease.

1. Theoretical foundations

1.1 Cox proportional hazards model

When working with time-to-event data from medical studies, there is a censoring involved inevitably. We need not observe the event for some individual, because the event might never occur whilst observing the subject. The reasons of this situation might be the individual's death (caused by a different event than the one defined as the failure), his decision to leave the study or the end of the study itself. The time of any of these situations is called the censoring time.

We consider two continuous random variables: $T \geq 0$ is the failure time (the time of the studied event) and $C \geq 0$ is the censoring time. If $C < T$, the event was not observed. Let $X = \min(T, C)$ be the censored failure time and $\delta = \mathbb{1}\{T \leq C\}$ the failure indicator. During a study, we also collect other information about the individuals in a form of a d-dimensional vector of covariates $\boldsymbol{Z}(t)$, which may be time dependent. The data are then in a form of n independent observations $(X_1, \delta_1, \boldsymbol{Z}_1(t)), \ldots, (X_n, \delta_n, \boldsymbol{Z}_n(t))$, $t \in [0, \tau]$, where τ is a fixed sufficiently large finite time. Let us also define right-continuous process $N_i(t) = \mathbb{1}\{T_i \leq t, \delta_i = 1\}$, $i \in \{1, \ldots, n\}$, which starts at zero and jumps to one if the failure time is observed. Denote $Y_i(t) = \mathbb{1}\{X_i \geq t\}$, $i \in \{1, \ldots, n\}$, a left-continuous at-risk process, which starts at one and jumps to zero if the censored failure time occures.

There are multiple ways to describe the distribution of the failure time. A hazard function is one of those possibilities, which is also a subject of interest in the Cox proportional hazard (Cox PH) model described by Cox [1972] and later by Andersen and Gill [1982]. A hazard function expresses the risk of having an event at a certain time given that the event had not occurred before. A cumulative hazard function then expresses the risk of having the event before a certain time.

Definition 1 (Hazard and cumulative hazard function). *The hazard function of a failure time T is*

$$\lambda(t) = \lim_{h \to 0+} \frac{1}{h} \mathsf{P}[t \leq T < t + h | T \geq t], \qquad t \in [0, \tau]$$

and its cumulative hazard function is

$$\Lambda(t) = \int_0^t \lambda(s) ds.$$

It is important to realize that the hazard function fully specifies the distribution of the failure time T. More specifically, its density may be written as

$$f(t) = \lambda(t) e^{-\Lambda(t)},$$

since T is a non-negative random variable.

Now by conditioning the hazard function on the covariates $\boldsymbol{Z}$, we may define the following.

Definition 2 (Conditional (cumulative) hazard function)**.** *The conditional hazard function of a failure time T is*

$$\lambda(t|\boldsymbol{Z}) = \underset{h\to 0+}{lim} \frac{1}{h} \mathsf{P}[t \leq T < t+h | T \geq t, \boldsymbol{Z}(t)]$$

and its conditional cumulative hazard function is

$$\Lambda(t|\boldsymbol{Z}) = \int_0^t \lambda(s|\boldsymbol{Z}) ds.$$

We now define the independent censoring condition, which is an important condition for further inferences.

Definition 3 (Independent censoring condition)**.** *Let T be the failure time and C the censoring time. We say that the independent censoring condition holds if*

$$\underset{h\to 0+}{lim} \frac{1}{h} \mathsf{P}[t \leq T < t+h | T \geq t, \boldsymbol{Z}(t)] = \underset{h\to 0+}{lim} \frac{1}{h} \mathsf{P}[t \leq T < t+h | T \geq t, C \geq t, \boldsymbol{Z}(t)].$$

This condition simply means that the conditional hazard function does not depend on the fact whether or not the individual has been censored. If the variables T and C are conditionally independent given $\boldsymbol{Z}$, then the independent censoring condition holds. This is however only a sufficient condition. If the independent censoring condition holds, it does not mean that the variables are conditionally independent.

For censored data, the most popular model to use is a model for the conditional hazard function proposed by Cox [1972]. We define this model in the next definition.

Definition 4 (Cox proportional hazards model)**.** *We say that the observations $(X_i, \delta_i, \boldsymbol{Z}_i(t))$, $i = 1, \dots, n$, satisfy the Cox proportional hazards model if*

a) the observations are mutually independent,

b) the conditional hazard function has the form

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t) e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}(t)}.$$

The function $\lambda_0(t)$ is an unknown and unspecified hazard function of a subject with all covariates equal to zero and it is called the baseline hazard function. The vector $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is an unknown parameter vector of regressions coefficients.

The Cox model is a semiparametric model because there is a condition on the form of the association between the covariate and the hazard function, however there are no assumptions about the baseline hazard. Therefore, we cannot estimate the parameter by the maximum likelihood methods and we use a modification proposed by Cox [1972] called the partial likelihood, which does not depend on the baseline hazard.

Definition 5 (Partial likelihood in the Cox model)**.** *The function*

$$L_{Cox}(\boldsymbol{\beta}) = \prod_{i=1}^{n}\prod_{s>0}\left\{\frac{Y_i(s)e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_i(s)}}{\sum_{j=1}^{n} Y_j(s)e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(s)}}\right\}^{\Delta N_i(s)} \tag{1.1}$$

is the partial likelihood function and the value

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{argmax} L_{Cox}(\boldsymbol{\beta})$$

is called the maximum partial likelihood estimator (MPLE) of the regression parameter in the Cox proportional hazards model.

In practice, the MPLE is obtained by maximizing the log partial likelihood. For an easier notation, define for $k = 0, 1, 2$,

$$S_n^{(k)}(\boldsymbol{\beta}, t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\boldsymbol{Z}_i^{\otimes k}(t)e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_i(t)},$$

where for a vector $\boldsymbol{a}$ we have $\boldsymbol{a}^{\otimes 0} = 1$, $\boldsymbol{a}^{\otimes 1} = \boldsymbol{a}$ and $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^\top$. Surely,

$$\frac{\partial}{\partial \boldsymbol{\beta}} S_n^{(0)}(\boldsymbol{\beta}, t) = \boldsymbol{S}_n^{(1)}(\boldsymbol{\beta}, t),$$

$$\frac{\partial}{\partial \boldsymbol{\beta}^\top} \boldsymbol{S}_n^{(1)}(\boldsymbol{\beta}, t) = \mathbb{S}_n^{(2)}(\boldsymbol{\beta}, t).$$

We can now write the log partial likelihood as

$$l_{Cox}(\boldsymbol{\beta}) = \log L_{Cox}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\int_0^\tau \left[\boldsymbol{\beta}^\top \boldsymbol{Z}_i(s) - \log n S_n^{(0)}(\boldsymbol{\beta}, s)\right] dN_i(s).$$

We need to differentiate this expression with respect to $\boldsymbol{\beta}$ to obtain the score statistic

$$\boldsymbol{U}_n(\boldsymbol{\beta}) = \frac{\partial l_{Cox}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}\int_0^\tau \left[\boldsymbol{Z}_i(s) - \frac{\boldsymbol{S}_n^{(1)}(\boldsymbol{\beta}, s)}{S_n^{(0)}(\boldsymbol{\beta}, s)}\right] dN_i(s).$$

Then the MPLE $\hat{\boldsymbol{\beta}}$ is the solution of the system of equations

$$\boldsymbol{U}_n(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}.$$

Numerically, the solution of the system of equations is obtained by the Newton-Raphson algorithm which looks as follows. Choose an initial value $\hat{\boldsymbol{\beta}}^{(0)} = \boldsymbol{0}$ and iterate

$$\hat{\boldsymbol{\beta}}^{(r+1)} = \hat{\boldsymbol{\beta}}^{(r)} + \left[n\mathcal{I}_n(\hat{\boldsymbol{\beta}}^{(r)})\right]^{-1} \boldsymbol{U}_n(\hat{\boldsymbol{\beta}}^{(r)}) \tag{1.2}$$

until convergence. Here $\mathcal{I}_n$ is the observed information matrix which we can calculate as

$$\begin{aligned}\mathcal{I}_n(\boldsymbol{\beta}) &= -\frac{1}{n}\frac{\partial \boldsymbol{U}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left[\frac{\mathbb{S}_n^{(2)}(\boldsymbol{\beta}, s)}{S_n^{(0)}(\boldsymbol{\beta}, s)} - \left[\frac{\boldsymbol{S}_n^{(1)}(\boldsymbol{\beta}, s)}{S_n^{(0)}(\boldsymbol{\beta}, s)}\right]^{\otimes 2}\right] dN_i(s).\end{aligned}$$

The solution $\hat{\boldsymbol{\beta}}$ exists and it is unique because $\boldsymbol{Z}_i, i = 1, \dots, n$ are all independent, hence, the observed information matrix is non-singular and therefore positive definite at all $\boldsymbol{\beta}$. This means that $\frac{\partial \boldsymbol{U}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top}$ is negative definite, hence log partial likelihood $l_{Cox}(\boldsymbol{\beta})$ is strictly concave and therefore has only one global maximum.

The MPLE $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}_0$ with an asymptotic normal distribution under specific assumptions, which were proposed by Andersen and Gill [1982]:

A.1 The data are observed on a time interval $[0, \tau]$, where $\tau \in (0, \infty)$ is fixed, and for all $i \in \{1, \dots, n\}$ and for some $\delta > 0$: $\mathsf{P}[Y_i(\tau) = 1] > \delta$.

A.2 The baseline hazard λ_0 is a deterministic function on $[0, \tau]$ and $\Lambda_0(\tau) = \int_0^\tau \lambda_0(t)dt < \infty$.

A.3 Let $\{\mathcal{F}_t\}_{t \in [0,\tau]}$ be a right-continuous filtration on $(\Omega, \mathcal{F}, \mathsf{P})$ defined as

$$\mathcal{F}_t = \sigma\{N_i(s), Y_i(s+), \boldsymbol{Z}_i(s), s \in [0, t], i = 1, \dots, n\}.$$

We assume that $\boldsymbol{Z}_i$ are bounded $\mathcal{F}_t$-predictable processes and that the independent censoring condition holds.

A.4 The variables T_i, C_i, $i \in \{1, \dots, n\}$, are independent given the covariate $\boldsymbol{Z}_i$ and the processes $\boldsymbol{Z}_i, Y_i$, $i \in \{1, \dots, n\}$, are independent.

A.5 There exists a neighborhood $\mathcal{B}$ of the true parameter $\boldsymbol{\beta}_0$ and functions $s^{(k)}, k = 0, 1, 2$, defined on $\mathcal{B} \times [0, \tau]$ such that

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}, t \in [0,\tau]} \|S_n^{(k)}(\boldsymbol{\beta}, t) - s^{(k)}(\boldsymbol{\beta}, t)\| \xrightarrow[n \to \infty]{\mathcal{P}} 0$$

for each $k = 0, 1, 2$. Here for $a \in \mathbb{R}$ we have $\|a\| = |a|$, for a vector $\boldsymbol{a} \in \mathbb{R}^d$ we denote $\|\boldsymbol{a}\| = \max\limits_{j=1,\dots,d}(|a_j|)$ and for a matrix $\mathbb{a} \in \mathbb{R}^{d \times d}$, $\|\mathbb{a}\| = \max\limits_{j,l=1,\dots,d}(|a_{jl}|)$.

A.6 The functions $s^{(k)}$ are bounded on $\mathcal{B} \times [0, \tau]$ and also $s^{(0)}$ is bounded away from 0 on $\mathcal{B} \times [0, \tau]$. The family $\{s^{(k)}(\boldsymbol{\beta}, t) : t \in [0, \tau]\}$ is equicontinuous at $\boldsymbol{\beta}_0$.

A.7 For all $\boldsymbol{\beta} \in \mathcal{B}$ and $t \in [0, \tau]$, it holds that

$$\frac{\partial s^{(0)}(\boldsymbol{\beta}, t)}{\partial \boldsymbol{\beta}} = \boldsymbol{s}^{(1)}(\boldsymbol{\beta}, t) \text{ and } \frac{\partial \boldsymbol{s}^{(1)}(\boldsymbol{\beta}, t)}{\partial \boldsymbol{\beta}^\top} = \mathbb{s}^{(2)}(\boldsymbol{\beta}, t).$$

A.8 The Fisher information matrix

$$\mathcal{I}(\boldsymbol{\beta}_0, t) = \int_0^t \left[\frac{\mathbb{s}^{(2)}(\boldsymbol{\beta}_0, s)}{s^{(0)}(\boldsymbol{\beta}_0, s)} - \left[\frac{\boldsymbol{s}^{(1)}(\boldsymbol{\beta}_0, s)}{s^{(0)}(\boldsymbol{\beta}_0, s)} \right]^{\otimes 2} \right] s^{(0)}(\boldsymbol{\beta}_0, s) d\Lambda_0(s)$$

is positive definite at $t = \tau$.

The last assumption **A.8** is the one that ensures the regularity of the Fisher information matrix. The assumption **A.3** about the bounded covariates is in place only to simplify the proofs and does not need to hold in order to attain the asymptotic properties of the MPLE that we await.

Theorem 1 (Asymptotic properties of the MPLE $\hat{\boldsymbol{\beta}}$)**.** *Under the assumptions **A.1** - **A.5**,*

$$\hat{\boldsymbol{\beta}} \xrightarrow[n\to\infty]{\mathcal{P}} \boldsymbol{\beta}_0$$

and

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\beta}_0, \tau)).$$

1.2 Martingale theory

In the subsequent chapters, we will find out that it is useful to view censored data as random processes because we will be able to use the martingale theory to prove some of the asymptotic properties of the semiparametric estimators. Therefore, in this subchapter, we will introduce some of the most important definitions and theorems of the theory of counting processes and martingales by Fleming and Harrington [1991].

Let us assume that we are working on a probability space $(\Omega, \mathcal{F}, \mathsf{P})$. We first introduce a random (stochastic) process on a time interval $[0, \tau]$, $\tau \in (0, \infty)$ fixed and non-random.

Definition 6 (Random process)**.** *A real random process X on an interval $[0, \tau]$ is a family of random variables $X = \{X(t), t \in [0, \tau]\}$, where*

$$X(t) : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}), \quad \forall t \in [0, \tau].$$

Because a random process is evolving over time, we need to define a family of σ-algebras which capture the events of the random process and its dynamic.

Definition 7 (Filtration)**.** *Let $\{\mathcal{F}_t, t \in [0, \tau]\}$ be a family of σ-algebras on a probability space $(\Omega, \mathcal{F}, P)$. We say that $\{\mathcal{F}_t\}$ is a filtration if $\forall 0 \leq s \leq t \leq \tau$:*

$$\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F}.$$

Definition 8 (Process adapted on a filtration)**.** *Let $\{\mathcal{F}_t\}$ be a filtration and let $X = \{X(t), t \in [0, \tau]\}$ be a random process. We say that X is adapted on a filtration $\{\mathcal{F}_t\}$, if $X(t)$ is $\mathcal{F}_t$-measurable, $\forall t \in [0, \tau]$.*

In the survival analysis, we very often count the times when there had been some event in a particular group of subjects and we use so called *counting process* for this. It is a random process which counts for every time the number of events that have happened before that time or right at that time.

Definition 9 (Counting process)**.** *Let $\{\mathcal{F}_t\}$ be a filtration. A random process $N = \{N(t), t \in [0, \tau]\}$ is an $\mathcal{F}_t$-counting process if*

a) $N(0) \overset{\text{a.s.}}{=} 0$,

b) $N(t) \overset{\text{a.s.}}{<} \infty$, $\forall t \in [0, \tau]$,

c) the trajectories of N are right-continuous, piecewise constant with jumps of size 1, *d) N is $\mathcal{F}_t$-adapted.*

Because there are jumps only of size 1, the part c) means that there are never two events that occur at the same time.

We also need to define random processes called *martingales* because they have some good properties which are useful later in the text. A martingale is a process for which the conditional expectation of the future value is equal to the last value that we have information about, regardless of all the previous values.

Definition 10 ($\mathcal{F}_t$-(sub)martingale)**.** *Let $\{\mathcal{F}_t\}$ be a filtration and let X be a right-continuous $\mathcal{F}_t$-adapted random process with left-hand limits, $\mathsf{E}\,|X(t)| < \infty$, $\forall t \in [0, \tau]$. We say that X is*

a) an $\mathcal{F}_t$-martingale if $\mathsf{E}[X(t)|\mathcal{F}_s] \overset{a.s.}{=} X(s)$, $\forall 0 \leq s \leq t \leq \tau$,

b) an $\mathcal{F}_t$-submartingale if $\mathsf{E}[X(t)|\mathcal{F}_s] \overset{a.s.}{\geq} X(s)$, $\forall 0 \leq s \leq t \leq \tau$.

Remark. An integrable $\mathcal{F}_t$-counting process is an $\mathcal{F}_t$-submartingale.

Theorem 2 (Sum of martingales)**.** *Let $\mathcal{F}_t$ be a filtration and let $M_i, i = 1, \ldots, n$ be $\mathcal{F}_t$-martingales. Then $\sum_{i=1}^{n} M_i$ is also an $\mathcal{F}_t$-martingale.*

Martingales have one important property formulated in the Doob-Meyer decomposition theorem later in this chapter. In order to be able to write this theorem, we need to define so called *predictable σ-algebra* and *predictable random process.* A predictable process is such that is measurable with respect to the predictable σ-algebra, which is generated by all left-continuous adapted processes. The exact definition is given in the following.

Definition 11 (Predictable σ-algebra)**.** *Let $\{\mathcal{F}_t\}$ be a filtration and let X be a random process. A predictable σ-algebra $\mathcal{P}(\mathcal{F}_t)$ is the smallest σ-algebra containing sets of type $\{0\} \times A$, $A \in \mathcal{F}_0$, and $(s, t] \times A$, $A \in \mathcal{F}_s$, $0 \leq s < t \leq \tau$.*

Definition 12 (Predictable process)**.** *Let $\{\mathcal{F}_t\}$ be a filtration, let X be a random process and let $\mathcal{P}(\mathcal{F}_t)$ be a predictable σ-algebra. We say that X is $\mathcal{F}_t$-predictable if it is $\mathcal{P}(\mathcal{F}_t)$-measurable.*

Claim 3. *Let $\{\mathcal{F}_t\}$ be a filtration. Any left-continuous $\mathcal{F}_t$-adapted process is $\mathcal{F}_t$-predictable.*

This next theorem formulates a very useful property of martingales and predictable processes and it will be used many times in the next chapters of this thesis.

Theorem 4 (Integral w.r.t. a martingale is a martingale)**.** *Let $\{\mathcal{F}_t\}$ be a filtration, let M be an $\mathcal{F}_t$-martingale, $\Delta M(0) \overset{a.s.}{=} 0$, and let X be a bounded $\mathcal{F}_t$-predictable process. Then the integral $\int_0^t X(s) dM(s)$ is an $\mathcal{F}_t$-martingale, $t \in [0, \tau]$.*

Now we can finally formulate the Doob-Meyer decomposition, one of the main properties of martingales. It claims that every right-continuous non-negative submartingale may be decomposed to a sum of a right-continuous martingale and a right-continuous non-decreasing predictable process.

Theorem 5 (Doob-Meyer decomposition)**.** *Let X be a right-continuous non-negative $\mathcal{F}_t$-submartingale. Then there exists a unique right-continuous non-decreasing $\mathcal{F}_t$-predictable process A (compensator) and a right-continuous $\mathcal{F}_t$-martingale M such that:*

a) $\mathsf{E}\, A(t) < \infty,\ \forall t \in [0, \tau]$,

b) $A(0) \overset{a.s.}{=} 0$,

c) $X(t) \overset{a.s.}{=} M(t) + A(t),\ \forall t \in [0, \tau]$.

We will now use the Doob-Meyer decomposition theorem in the situation which we have described in Chapter 1.1. Let us define a filtration

$$\mathcal{F}_t = \sigma\{N_i(s), Y_i(s+), 0 \leq s \leq t, i = 1, \dots, n\}, t \in [0, \tau],$$

where $N_i(t)$ and $Y_i(t)$ are random processes defined in Chapter 1.1. Then $N_i(t)$ is a counting process with respect to this filtration.

Theorem 6. *Define the right-continuous $\mathcal{F}_t$-predictable process*

$$\mathrm{A}_i(t) = \int_0^t Y_i(s) d\Lambda_i(s), i \in \{1, \dots, n\},$$

where Λ_i is a cumulative hazard function. The process

$$M_i(t) = N_i(t) - \mathrm{A}_i(t), i \in \{1, \dots, n\}$$

is $\mathcal{F}_t$-martingale if and only if the independent censoring condition holds.

In other words, if the independent censoring condition holds, we know the form of the compensator of the counting process $N(t)$. Any counting process can be expressed by its intensity with which the number of new events increases. By Theorem 6 it is now clear that the intensity process is cumulated into the compensator. Let us now present the exact definition of an intensity process.

Definition 13 (Intensity process)**.** *Let $\{\mathcal{F}_t\}$ be a filtration and N be a counting process adapted to this filtration. We say that a random process $\{\alpha(t), t \in [0, \tau]\}$ is an intensity process of the counting process N if*

$$\alpha(t) = \mathop{lim}\limits_{h \to 0+} \frac{1}{h} \mathsf{E}[N(t + h) - N(t) | \mathcal{F}_t].$$

The cumulative intensity of a counting process N is a process $\{\mathrm{A}(t), t \in [0, \tau]\}$ such that

$$\mathrm{A}(t) = \int_0^t \alpha(u) du.$$

This is a definition of a time-dependent intensity which we will be working with during this thesis, however, intensity of a counting process may be also constant in time. For example, for Poisson counting process N with time-independent intensity α we have for $k \in \mathbb{N}$,

$$\mathsf{P}[N(t) - N(s) = k] = e^{-\alpha \cdot (t-s)} \frac{[\alpha \cdot (t - s)]^k}{k!}.$$

For Poisson counting process N with time-dependent intensity $\alpha(\cdot)$ and cumulative intensity $\mathrm{A}(\cdot)$ we have that

$$\mathsf{P}[N(t) - N(s) = k] = e^{-(\mathrm{A}(t)-\mathrm{A}(s))}\frac{[\mathrm{A}(t) - \mathrm{A}(s)]^k}{k!}.$$

If M is a right-continuous $\mathcal{F}_t$-martingale, then M^2 is a right-continuous $\mathcal{F}_t$-submartingale because of the Jensen inequality. Then there exists a unique decomposition as we know from Theorem 5. We now define the compensator for this decomposition.

Definition 14 (Predictable variation process)**.** *Let M be a right-continuous $\mathcal{F}_t$-martingale, $\mathsf{E}\, M^2(t) < \infty$, $\forall t \in [0, \tau]$. Then $\langle M \rangle$ is called the predictable variation process of M, which is the compensator for the $\mathcal{F}_t$-submartingale M^2, i.e. $M^2 - \langle M \rangle$ is a right-continuous $\mathcal{F}_t$-martingale.*

In a similar way, we define a predictable covariaton process of two martingales as a compensator of their product.

Definition 15 (Predictable covariation process)**.** *Let M, N be right-continuous $\mathcal{F}_t$-martingales, $\mathsf{E}\, M^2(t) < \infty$, $\mathsf{E}\, N^2(t) < \infty$, $\forall t \in [0, \tau]$. Then*

$$\langle M, N \rangle = \frac{1}{4}\Big[\langle M + N \rangle - \langle M - N \rangle\Big]$$

is called the predictable covariation process of M and N, i.e. $M \cdot N - \langle M, N \rangle$ is a right-continuous $\mathcal{F}_t$-martingale.

The notation $\langle M \rangle$ is only a short version of $\langle M, M \rangle$ as both of the notations mean the same thing. There are many properties of predictable variation and covariation processes. We will introduce only those which will be useful in the next chapters.

Theorem 7. *Let $\{\mathcal{F}_t\}$ be a filtration and $A(t)$ be a continuous compensator of a counting process $N(t)$, i.e. $M(t) = N(t) - A(t)$ is an $\mathcal{F}_t$-martingale. Then*

$$\langle M \rangle(t) = A(t).$$

Theorem 2 gives us the condition on when the sum of martingales is a martingale. Now we introduce a condition on the product of martingales as it gives us the knowledge about the predictable covariation process of two martingales. At first, let us define an orthogonality of martingales.

Definition 16 (Orthogonal martingales)**.** *Let $\{\mathcal{F}_t\}$ be a filtration and M_1, M_2 two $\mathcal{F}_t$-martingales. We say that M_1 and M_2 are orthogonal if and only if*

$$\langle M_1, M_2 \rangle(t) = 0, \quad \forall t \in [0, \tau].$$

Theorem 8 (Condition on orthogonality of martingales)**.** *Let $\{\mathcal{F}_t\}$ be a filtration, $N_i(t)$ $\mathcal{F}_t$-counting processes and $A_i(t)$ its compensators, i.e. $M_i(t) = N_i(t) - A_i(t)$ are $\mathcal{F}_t$-martingales. If $N_i(t)$ have all distinct times of jumps, then $M_i(t), M_j(t)$ are orthogonal, $i \neq j$, i.e. $\langle M_i, M_j \rangle(t) = 0$, $i \neq j$.*

From Definition 15, we can now see the condition on the product of martingales being a martingale.

Theorem 9 (Product of martingales)**.** *Let $\{\mathcal{F}_t\}$ be a filtration and $M_1(t), M_2(t)$ two $\mathcal{F}_t$-martingales. If $M_1(t), M_2(t)$ are orthogonal, then $M_1(t) \cdot M_2(t)$ is an $\mathcal{F}_t$-martingale.*

By Theorem 4, we know that an integral with respect to a martingale from a bounded predictable process is also a martingale. The next theorem calculates the predictable covariation of two such martingales.

Theorem 10 (Predictable covariation process for martingale integrals)**.** *Let $\{\mathcal{F}_t\}$ be a filtration, H_i bounded $\mathcal{F}_t$-predictable processes and M_i $\mathcal{F}_t$-martingales such that $\mathsf{E}(M_i)^2 < \infty$, $i = 1, 2$. Then for $t \in [0, \tau]$,*

$$\left\langle \int_0^t H_1(s)dM_1(s), \int_0^t H_2(s)dM_2(s) \right\rangle = \int_0^t H_1(s)H_2(s)d\langle M_1, M_2\rangle(s).$$

Denote $G_i(t) = \int_0^t H_i(s)dM_i(s)$. It is useful to realize that since G_i are martingales with H_i predictable bounded processes and M_i martingales, then

$$\operatorname{var}[G_i(t)] = \mathsf{E}\,\langle G_i\rangle(t) = \mathsf{E}\int_0^t H_i^2(s)d\langle M_i, M_i\rangle(s)$$

and

$$\operatorname{cov}[G_i(t), G_j(t)] = \mathsf{E}\,\langle G_i, G_j\rangle(t) = \mathsf{E}\int_0^t H_i(s)H_j(s)d\langle M_i, M_j\rangle(s).$$

At some point of this thesis, we will prove asymptotic normality of a given estimator. Since we will again use the knowledge of the martingale theory to do so, we need to present the central limit theorem for sums of martingale integrals. Before we do that, we define the Gaussian process, which will be the subject of the central limit theorem.

Definition 17 (Centered Gaussian process)**.** *A random process $X = \{X(t), t \in [0, \tau]\}$ is called a centered Gaussian process if*

a) $X(0) \overset{a.s.}{=} 0$,

b) $\mathsf{E}\,X(t) = 0, t \in [0, \tau]$,

c) joint distribution of the increments is multivariate normal.

Theorem 11 (Central limit theorem for the sums of martingale integrals)**.** *Let us assume that:*

- $\{N_i^{(n)}, i = 1, \dots, n\}$ *is a multivariate counting process with respect to the probability space* $(\Omega, \mathcal{F}, \mathsf{P})$,
- $A_i^{(n)}$ *is a continuous compensator of* $N_i^{(n)}$, *i.e.* $M_i^{(n)} = N_i^{(n)} - A_i^{(n)}$,
- $H_{ji}^{(n)}, i = 1, \dots, n, j = 1, \dots, d$ *is a bounded $\mathcal{F}_t$-predictable process on* $[0, \tau]$.

Denote

$$U_{ji}^{(n)}(t) = \int_0^t H_{ji}^{(n)}(s)dM_i^{(n)}(s) \quad \text{and} \quad U_j^{(n)}(t) = \sum_{i=1}^n U_{ji}^{(n)}(t).$$

For any $\varepsilon > 0$ denote

$$U_{ji,\varepsilon}^{(n)}(t) = \int_0^t H_{ji}^{(n)}(s)\mathbb{1}\{|H_{ji}^{(n)}(s)| > \varepsilon\}dM_i^{(n)}(s) \quad \text{and} \quad U_{j,\varepsilon}^{(n)}(t) = \sum_{i=1}^{n} U_{ji,\varepsilon}^{(n)}(t).$$

We know from the martingale theory that for $j, k = 1, \ldots, d$,

$$\langle U_j^{(n)}, U_k^{(n)}\rangle(t) = \sum_{i=1}^{n} \int_0^t H_{ji}^{(n)}(s)H_{ki}^{(n)}(s)dA_i^{(n)}(s)$$

and

$$\langle U_{j,\varepsilon}^{(n)}, U_{k,\varepsilon}^{(n)}\rangle(t) = \sum_{i=1}^{n} \int_0^t H_{ji}^{(n)}(s)H_{ki}^{(n)}(s)\mathbb{1}\{|H_{ji}^{(n)}(s)| > \varepsilon\}\mathbb{1}\{|H_{ki}^{(n)}(s)| > \varepsilon\}dA_i^{(n)}(s).$$

Let for $t \in [0, \tau]$ and $j, k \in \{1, \ldots, d\}$ the following two conditions hold:

1. $\langle U_j^{(n)}, U_k^{(n)}\rangle(t) \xrightarrow[n\to\infty]{\mathcal{P}} c_{jk}(t) < \infty$, *where c_{jk} are continuous functions on $[0, \tau]$,*

2. $\forall \varepsilon > 0$ $j = 1, \ldots, d$*:* $\langle U_{j,\varepsilon}^{(n)}, U_{j,\varepsilon}^{(n)}\rangle(t) \xrightarrow[n\to\infty]{\mathcal{P}} 0.$

Then

$$(U_1^{(n)}, \ldots, U_d^{(n)}) \Longrightarrow (X_1, \ldots, X_d),$$

where $X_1, \ldots, X_d$ are dependent centered Gaussian processes with independent increments with covariance functions $cov[X_j(s), X_k(t)] = c_{jk}(s)$ $j, k \in \{1, \ldots, d\}$, $0 \leq s \leq t \leq \tau$. The symbol "$\Longrightarrow$" denotes weak convergence.

2. Nested case-control design

The nested case-control design is one of the most popular designs to use in practice when studying rare diseases. In case of rare diseases, we need to include a large number of people in the study to obtain at least a small amout of individuals who get the disease of interest. Then there are too many individuals who have not got the disease or who have been censored during the study and it would cost a lot of money, effort and time to measure all the covariates on all the participants in the study. We say that the individual who got the disease of interest at a certain time "failed" at that time and we call him a *case*. The individuals who have not failed yet are called *controls*. Hence, an individual is called a control from the beginning of the study until his (potential) observed failure when he becomes a case. By conducting the nested case-control design, we select only certain amount of controls for each case. This reduces the amount individuals needed for the analysis and therefore we perform only a few necessary covariate measures, which allows us to minimize the cost and duration of the study.

The way we select controls for the analysis is as follows. For each case at its failure time we select without replacement a random sample of controls of size $m - 1$ from the set of subjects who have not been censored yet and have not failed before that time including this failure time. Here m is an arbitrary number, which is usually somewhere between 2 and 6. The set of subjects who have not been censored yet and have not failed before a specific time t is called the *risk set* at time t because the individuals in that set are "at risk" of getting the disease. At each failure time we then obtain so called *sampled risk set* which consists of the case and $m - 1$ sampled controls. This method may be also described as performing a matched case-control design by matching the individuals on their at-risk status at a particular time.

The nested case-control data can be analyzed by a Cox regression model and its regression parameters can be estimated by maximizing the partial likelihood for nested case-control data. The methods of this analysis are modified from those introduced in Chapter 1.1. In this chapter, we will, based on Goldstein and Langholz [1992], introduce the model for nested case-control data, the partial likelihood and the maximum partial likelihood estimator and we prove the consistency and asymptotic normality of this estimator.

2.1 Introduction to the design

Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space, n be a number of individuals in the population of interest and, for $i \in \{1, \dots, n\}$, let T_i be failure time, C_i censoring time, X_i censored failure time, and δ_i failure indicator for the i-th individual. For a fixed finite time τ, let N_i, Y_i, $\boldsymbol{Z}_i$, $i \in \{1, \dots, n\}$, be random processes on the probability space $(\Omega, \mathcal{F}, \mathsf{P})$ and on a time interval $t \in [0, \tau]$ defined as in Chapter 1.1. Specifically, time 0 is understood as the beginning of the study and time τ as the end. Define

$$\mathcal{F}_t = \sigma\{N_i(s), Y_i(s+), \boldsymbol{Z}_i(s), 0 \le s \le t, i = 1, \dots, n\}, \quad t \in [0, \tau],$$

a filtration on the probability space $(\Omega, \mathcal{F}, \mathsf{P})$.

Let α_i denote the intensity process of the counting process N_i as defined in Definition 13. This definition may be further rewritten as follows:

$$\begin{aligned}\alpha_i(t|\boldsymbol{Z}_i) &= \lim_{h\to 0+}\frac{1}{h}\,\mathsf{E}\left[N_i(t+h)-N_i(t)|\mathcal{F}_t\right]\\ &= \lim_{h\to 0+}\frac{1}{h}\,\mathsf{P}[N_i(t+h)-N_i(t)=1|\mathcal{F}_t]\\ &= \lim_{h\to 0+}\frac{1}{h}\,\mathsf{P}[t\leq T_i<t+h|\mathcal{F}_t].\end{aligned}$$

Under the independent censoring condition (Definition 3), the conditional hazard function $\lambda_i(t|\boldsymbol{Z}_i)$ equals

$$\lim_{h\to 0+}\frac{1}{h}\,\mathsf{P}[t\leq T_i<t+h|X_i\geq t,\boldsymbol{Z}_i(t)].$$

Hence, by multiplying the at-risk process $Y_i(t)=\mathbb{1}\{X_i\geq t\}$ with the conditional hazard function of the Cox proportional hazards model, we get the intensity of the counting process N_i as

$$\alpha_i(t)\equiv\alpha_i(t|\boldsymbol{Z}_i)=Y_i(t)\lambda_i(t|\boldsymbol{Z}_i)=Y_i(t)\lambda_0(t)e^{\boldsymbol{\beta}_0^\top\boldsymbol{Z}_i(t)},\tag{2.1}$$

where $\boldsymbol{Z}_i$ is the covariate vector for the i-th individual, $\boldsymbol{\beta}_0$ is a fixed vector of regression coefficients in $\mathbb{R}^d$, $d\in\mathbb{N}$ and $\lambda_0(t)$ is the baseline hazard function, which is unknown and unspecified. The cumulative intensity of N_i may be written as

$$\mathrm{A}_i(t)\equiv\mathrm{A}_i(t|\boldsymbol{Z}_i)=\int_0^t\alpha_i(s|\boldsymbol{Z}_i)ds.$$

Due to the Doob-Meyer decomposition (Theorem 6), it is useful to realize that $M_i=N_i-\mathrm{A}_i$ is an $\mathcal{F}_t$-martingale, i.e. the cumulative intensity is a compensator of its counting process.

We will now present the procedure of creating the nested case-control data from the observations $(X_i,\delta_i,\boldsymbol{Z}_i(t))$, i.e. from the processes Y_i, N_i, $\boldsymbol{Z}_i$. For a time t define

$$\mathcal{R}(t)=\{i:Y_i(t+)=1\}$$

a *risk set* containing all the individuals who are at risk right after time t and $n(t)=|\mathcal{R}(t)|$ the number of individuals in this risk set. Here $Y_i(t+)$ is understood as an indicator whether or not the i-th individual is at risk right after time t. Because Y_i is left-continuous, the $t+$ in the definition of $\mathcal{R}(t)$ is necessary to ensure that the individual who became a case or was censored at time t does not belong to $\mathcal{R}(t)$.

Define $X_0'=0$ and $X_1',X_2',\dots$ an ordered collection of observed censoring and failure times and $\bar{Y}(t)=\sum_{i=1}^n Y_i(t)$. The process $\bar{Y}(t)$ is left-continuous piecewise constant with an initial value equal to n and jumps of size -1 which occur at times X_k'. In other words, the value of the process $\bar{Y}(t)$ at time X_k' jumps from $n-k+1$ to $n-k$. For every $k\geq 1$, we have a risk set $\mathcal{R}(X_k')$ which consists of individuals who have not failed and have not been censored before or at time X_k'. This means that the individual denoted as i_k who failed at time X_k' is not included in the risk set $\mathcal{R}(X_k')$. Denote $\mathcal{P}_{m,i}(\mathcal{R}(X_k'))$ a set of all subsets of $\mathcal{R}(X_k')$ of size m which include the i-th individual. Clearly, if $i\in\mathcal{R}(X_k')$, then

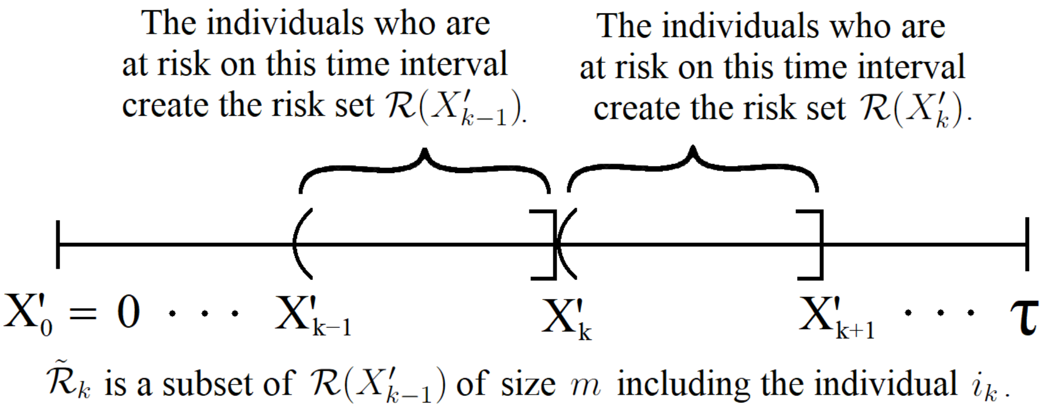
The individuals who are at risk on this time interval create the risk set $\mathcal{R}(X'_{k-1})$.
The individuals who are at risk on this time interval create the risk set $\mathcal{R}(X'_k)$.
$X'_0 = 0$
$\cdots$
X'_{k-1}
X'_k
X'_{k+1}
$\cdots$
τ
$\tilde{\mathcal{R}}_k$ is a subset of $\mathcal{R}(X'_{k-1})$ of size m including the individual i_k.

Figure 2.1: Denotion of risk sets and the sampled risk set when sampling the nested case-control data.

$\mathcal{P}_{m,i}(\mathcal{R}(X'_k))$ includes $\binom{n(X'_k)-1}{m-1}$ sets, where $n(X'_k) = |\mathcal{R}(X'_k)|$. If $i \notin \mathcal{R}(X'_k)$, then $\mathcal{P}_{m,i}(\mathcal{R}(X'_k)) = \emptyset$.

The selection of controls for each case is done as follows. Let us assume that for a particular j the time X'_j is a failure time, i.e. the i_j-th individual becomes a case at time X'_j. Then we independently and uniformly choose one set from $\mathcal{P}_{m,i_j}(\mathcal{R}(X'_{j-1}))$ and we denote this set as $\tilde{\mathcal{R}}_{j,i_j} \equiv \tilde{\mathcal{R}}_j$. This is called the *sampled risk set* and it consists of the individual i_j and $m-1$ controls who were at risk at the failure time X'_j. We repeat this process for every $j \geq 1$ for which X'_j is a failure time. The collection of all cases and their selected controls creates the nested case-control data. For better understanding, the sampling process is demonstrated in Figure 2.1.

This design allows any individual who had been selected as a control to be selected again for another case or to become a case itself in the future. This is because the matching of the controls to a case at any failure time is based on the at-risk status Y of the individuals at that time and not on the failure indicator δ.

Since we select a fixed number of controls for each case, we need to think about a certain situation that may occur. Suppose that there are less than $m-1$ available controls for a case observed at the end of the study. Then, we select the controls who are available even though there are fewer of them or none at all. The estimation of the parameters would still make sense since this situation is the one of the Cox PH model where we do not sample controls for a case. For an easier notation, we further consider the situation where there are $m-1$ controls available for each case.

In order to estimate the regression coefficients of this model, we need to write the form of the nested case-control partial likelihood function. The situation of the nested case-control data is a situation of censored data where we model the conditional hazard function and the model is semiparametric, therefore we cannot use the usual likelihood function and we need to use the partial likelihood.

The partial likelihood may be created the same way as the Cox partial likelihood in Definition 5. The numerator of the terms of the product is the hazard of the case and the denominator is a sum of hazards of the controls and the case from the numerator. In this case, the controls are those from the sampled risk set at the failure time of the case in the numerator, i.e. the denominator is a sum over all subjects in the sampled risk set $\tilde{\mathcal{R}}_k$ for the observed case i_k. We take the product over the individuals who became cases during the study, which is also done in the Cox partial likelihood because $\Delta N_i(s) = 1$ if and only if the i-th individual failed at time s. Also, $Y_i(s)$ as an at-risk indicator process, is left-continuous, and since the product is taken over the failed individuals, the at-risk indicator in the numerator is always equal to one and in the denominator it is replaced by the sum over the sampled risk set. The nested case-control partial likelihood may then be written as

$$L(\boldsymbol{\beta}) = \prod_{k:\delta_k=1} \left\{ \frac{e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_{i_k}(X'_k)}}{\sum_{j\in\tilde{\mathcal{R}}_k} e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(X'_k)}} \right\}. \tag{2.2}$$

It is important to realize that we can rewrite the nested case-control partial likelihood function into the form which is very similar to the original partial likelihood equation (1.1) in the Cox proportional hazards model. Define the sampling indicators η_{ij} as

$$\eta_{ij}(t) = \sum_{k\geq 1} \mathbb{1}\{j \in \tilde{\mathcal{R}}_{k,i}\}\mathbb{1}\{X'_{k-1} < t \leq X'_k\}, \;\; t \in [0,\tau], \tag{2.3}$$

where $\eta_{ij}(0) = 0$. This indicator expresses whether or not the j-th individual is a selected control for a case i who failed at time X'_k. More precisely, if the i-th individual fails at time X'_k for any $k \geq 1$ and the j-th individual is selected as its control, then the sampling function η_{ij} is equal to 1 on a time interval $(X'_{k-1}, X'_k]$ and is equal to 0 otherwise. If the i-th individual is not a case at all during the study then this definition still makes sense as we have not defined the set $\tilde{\mathcal{R}}_{k,i}$ only for i as a case specifically. However, we will only use the sampling functions in situations where the i-th individual is a case in the study.

Using the sampling functions, we can write the nested case-control partial likelihood as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{s\in[0,\tau]} \left\{ \frac{e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_i(s)}}{\sum_{j=1}^{n} \eta_{ij}(s) e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(s)}} \right\}^{\Delta N_i(s)}. \tag{2.4}$$

From the form (2.4) of the partial likelihood function, we can proceed in finding the MPLE of $\boldsymbol{\beta}$ in a very similar way as the one that we have shown it in Chapter 1.1. At first, define for $k = 0, 1, 2$, $i \in \{1, \ldots, n\}$,

$$S^{(k)}_{n,i}(\boldsymbol{\beta}, t) = \frac{1}{n} \sum_{j=1}^{n} \eta_{ij}(t) \boldsymbol{Z}_j^{\otimes k}(t) e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(t)},$$

where clearly

$$\frac{\partial}{\partial \boldsymbol{\beta}} S^{(0)}_{n,i}(\boldsymbol{\beta}, t) = \boldsymbol{S}^{(1)}_{n,i}(\boldsymbol{\beta}, t),$$

$$\frac{\partial}{\partial \boldsymbol{\beta}^\top} \boldsymbol{S}^{(1)}_{n,i}(\boldsymbol{\beta}, t) = \mathbb{S}^{(2)}_{n,i}(\boldsymbol{\beta}, t).$$

Now we can write the log partial likelihood as

$$l(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^\tau \Big[\boldsymbol{\beta}^\top \boldsymbol{Z}_i(s) - \log n S_{n,i}^{(0)}(\boldsymbol{\beta}, s)\Big] dN_i(s).$$

By differentiating this expression with respect to $\boldsymbol{\beta}$, we obtain the score statistic

$$\boldsymbol{U}_n(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \int_0^\tau \left[\boldsymbol{Z}_i(s) - \frac{\boldsymbol{S}_{n,i}^{(1)}(\boldsymbol{\beta}, s)}{S_{n,i}^{(0)}(\boldsymbol{\beta}, s)}\right] dN_i(s). \tag{2.5}$$

Then the MPLE $\hat{\boldsymbol{\beta}}$ is a solution of the system of equations

$$\boldsymbol{U}_n(\hat{\boldsymbol{\beta}}) = \mathbf{0}.$$

Numerically, the solution is obtained by the Newton-Raphson algorithm described by (1.2), where the observed information matrix $\mathcal{I}_n$ may be calculated as

$$\begin{aligned}\mathcal{I}_n(\boldsymbol{\beta}) =& -\frac{1}{n}\frac{\partial \boldsymbol{U}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} \\ =& \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau \left[\frac{\mathbb{S}_{n,i}^{(2)}(\boldsymbol{\beta}, s)}{S_{n,i}^{(0)}(\boldsymbol{\beta}, s)} - \left[\frac{\boldsymbol{S}_{n,i}^{(1)}(\boldsymbol{\beta}, s)}{S_{n,i}^{(0)}(\boldsymbol{\beta}, s)}\right]^{\otimes 2}\right] dN_i(s).\end{aligned} \tag{2.6}$$

In further inferences, we will need to work with the score statistic and the information matrix as if they were processes. Let us take the integral in the defined expressions (2.5) and (2.6) from 0 to t for $t \in [0, \tau]$ to create such processes:

$$\begin{aligned}\boldsymbol{U}_n(\boldsymbol{\beta}, t) =& \sum_{i=1}^{n} \int_0^t \left[\boldsymbol{Z}_i(s) - \frac{\boldsymbol{S}_{n,i}^{(1)}(\boldsymbol{\beta}, s)}{S_{n,i}^{(0)}(\boldsymbol{\beta}, s)}\right] dN_i(s), \\ \mathcal{I}_n(\boldsymbol{\beta}, t) =& \frac{1}{n}\sum_{i=1}^{n} \int_0^t \left[\frac{\mathbb{S}_{n,i}^{(2)}(\boldsymbol{\beta}, s)}{S_{n,i}^{(0)}(\boldsymbol{\beta}, s)} - \left[\frac{\boldsymbol{S}_{n,i}^{(1)}(\boldsymbol{\beta}, s)}{S_{n,i}^{(0)}(\boldsymbol{\beta}, s)}\right]^{\otimes 2}\right] dN_i(s).\end{aligned}$$

The only difference in the procedure between the original Cox proportional hazards model and the Cox model for the nested case-control data is in the functions $S_n(\boldsymbol{\beta}, t)$, resp. $S_{n,i}(\boldsymbol{\beta}, t)$, i.e. in the at-risk indicator functions $Y_i(t)$ and the sampling functions $\eta_{ij}(t)$. In the case of nested case-control data, we need to distinguish between all people at risk and those who have been sampled as controls for the i-th failed individual. Therefore the functions $S_{n,i}$ are different for all cases i.

We can see a similarity with the stratified Cox model, which is an alternative to the Cox model in case that the proportional hazards assumption does not hold for some of the covariates. In this situation, we fit different hazard functions for different levels of that covariate, which we modify to a categorical variable V with values $1, \dots, q$, if it was previously a continuous variable. For different values of V (strata), we introduce different baseline hazards λ_{0j} and so the hazard functions depending on the strata take the form

$$\lambda(t|\boldsymbol{Z}, V = j) = \lambda_{0j}(t) e^{\boldsymbol{\beta}^\top \boldsymbol{Z}(t)}.$$

The observed data are in a form of $(N_{ji}(t), Y_{ji}(t), \boldsymbol{Z}_{ji}(t))$, $t \in [0, \tau]$, where $j = 1, \dots, q$ denotes the strata and $i = 1, \dots, n_j$ denotes the subject within the strata, since there are n_j individuals in the j-th strata. The partial likelihood is taken as a product of partial likelihood functions for each stratum and has the form

$$L_S(\boldsymbol{\beta}) = \prod_{j=1}^{q} \prod_{i=1}^{n_j} \prod_{s \in [0,\tau]} \left[\frac{Y_{ji}(s) e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_{ji}(s)}}{\sum_{k=1}^{n_j} Y_{jk}(s) e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_{jk}(s)}} \right]^{\Delta N_{ji}(s)}. \tag{2.7}$$

We can look at the nested case-control data as "stratified by a case" to which the controls belong. More precisely, for each observed failed individual we select $m-1$ at-risk individuals as controls and create the sampled risk set $\tilde{\mathcal{R}}_k$ for the k-th observed failure time. Let us assume that there are q observed failure times during the study. Then the nested case-control data may be viewed as being composed of q strata of m individuals in each and the partial likelihood is also being taken as a product of partial likelihood functions of each stratum. This is because the sum in the denominator is, using the sampling functions, only taken over the sampled risk set. The same happens for the stratified Cox model and its partial likelihood (2.7). The numerator is taken for the case because $\Delta N_{ji}(s) = 1$ if and only if the i-th individual in the j-th strata is a case, and the denominator is taken over the individuals who were at risk before the failure time of this case within the same stratum j. The difference is that in the nested case-control design, the "strata" have been created in a way that all of the controls were at-risk when they were selected whilst the stratified version still needs to take the possible censored or failed individuals into account with the at-risk indicators in the denominator. Also, the nested case-control design assumes that there is only one case in each stratum while the stratified version has no such assumption.

By continuing with the adjustments of partial likelihood (2.7), we get a similar form of the score statistic as in the nested case-control design as

$$\boldsymbol{U}_n(\boldsymbol{\beta}) = \sum_{j=1}^{q} \sum_{i=1}^{n_j} \int_0^\tau \left[\boldsymbol{Z}_{ji}(t) - \frac{\boldsymbol{S}_{n,ji}^{(1)}(\boldsymbol{\beta}, t)}{S_{n,ji}^{(0)}(\boldsymbol{\beta}, t)} \right] dN_{ji}(t),$$

where for $k = 0, 1$, we define

$$S_{n,ji}^{(k)}(\boldsymbol{\beta}, t) = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ji}(t) \boldsymbol{Z}_{ji}^{\otimes k}(t) e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_{ji}(t)}.$$

2.2 Consistency of the MPLE of the regression parameter

We will be working under assumptions which will allow us to prove asymptotic properties of the MPLE of the regression parameter in the nested case-control design. To the assumptions **A.1** - **A.4** for the Cox proportional hazards model from Chapter 1.1, we add the following one:

A.5* Let $\boldsymbol{Z}_Y$ be a random vector with the same distribution as a random vector $\boldsymbol{Z}$ conditionally on $Y = 1$, i.e.

$$\mathsf{P}[\boldsymbol{Z}_Y(t) \in B] = \mathsf{P}[\boldsymbol{Z}(t) \in B | Y(t) = 1]$$

for any Borel set $B \in \mathcal{B}(\mathbb{R}^d)$. The matrix $\mathbb{V}(t) = \mathrm{var}[\boldsymbol{Z}_Y(t)]$,

$$\mathbb{V} \equiv \int_0^\tau \mathbb{V}(t)\lambda_0(t)dt$$

is a positive definite matrix.

While the terms $S_n^{(k)}$ converge to deterministic functions $s^{(k)}$ in the analysis of the Cox model (assumption **A.5**), it is not so in the case of nested case-control design because the number of controls selected for each case is always $m-1$, hence it does not change even with a growing number of participants in the study. Therefore, the terms $S_{n,i}^{(k)}$ converge to zero when $n \to \infty$, since they are sums of m finite terms divided by n. If we were about to divide the sum by m and not by n, we would get a mean which would not converge since m does not change. The information matrix of the Cox model has terms which are the limits of $S_n^{(k)}$ and since $S_{n,i}^{(k)}$ all converge to zero, we need the assumption **A.5*** to introduce the matrix $\mathbb{V}$ and assume its positive definiteness. Later in the thesis, we will show that this assumption is sufficient to prove the positive definiteness of another matrix Γ, which will be the inverse of the asymptotic variation of the MPLE $\hat{\boldsymbol{\beta}}$.

The problem of $S_{n,i}^{(k)}$ not converging is also the reason why we need to present the following lemmas which establish the conditions for convergence of the terms $\frac{S_{n,i}^{(k)}}{S_{n,i}^{(0)}}$, $k = 1, 2$, when considered in a combination with other processes and taken as a sum over all subjects. We will be using the results of those lemmas many times in the upcoming theorems. We will only prove the first of the lemmas since the other could be proven in a similar way.

Lemma 12. *Let us fix an arbitrary $s \in [0, \tau]$ and write $Y = Y(s), \boldsymbol{Z} = \boldsymbol{Z}(s)$, $p = p(s) = P[Y(s) = 1] > 0$, and let $\mathcal{R}(s) = \mathcal{R} = \{i : Y_i = 1\}$ be the risk set at time s, $\mathcal{P}_m(\mathcal{R})$ a set of all subsets of $\mathcal{R}$ of size m and $\mathcal{P}_{m,i}(\mathcal{R})$ a set of all subsets of $\mathcal{R}$ of size m which include the i-th individual. Let $\tilde{\mathcal{R}}_i$ be independently and uniformly chosen from $\mathcal{P}_{m,i}(\mathcal{R})$ and let the sampling be conditionally independent from the covariate information given the at-risk indicators.*

For any set $T \in \mathcal{P}_m(\mathcal{R})$, define a vector

$$\boldsymbol{w}(T) = \frac{\sum_{j\in T} \boldsymbol{Z}_j e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j}}{\sum_{j\in T} e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j}},$$

where $\boldsymbol{w}(\emptyset) = \boldsymbol{0}$. Define $Y_T = \prod_{j\in T} Y_j$ and $B_i = e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i}$. Define a sequence

$$\boldsymbol{S}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{w}(\tilde{\mathcal{R}}_i) Y_i B_i.$$

Then

$$\boldsymbol{S}_n \xrightarrow[n\to\infty]{\mathcal{P}} \boldsymbol{q} = p \cdot \mathsf{E}\left[\boldsymbol{w}(U)\frac{1}{m}\sum_{j\in U} B_j \Big| Y_U = 1\right], \tag{2.8}$$

where $U = \{1, \ldots, m\}$.

Remark. The set of indexes $U = \{1, \dots, m\}$ denotes the first m individuals of the study dataset. Since the observations are independent and identically distributed, the subjects taken as a set U could be any arbitrary m subjects from the study, therefore there may be any number of cases in U.

Proof. Define a σ-algebra $\mathcal{G} = \sigma\{\boldsymbol{Z}_i, Y_i, i = 1, \dots, n\}$. Let us remind that $\boldsymbol{Z}_i$ is a d-dimensional vector, i.e. $\boldsymbol{Z}_i = (Z_{i1}, \dots, Z_{id})^\top$. In order to prove the convergence, it is sufficient to show that there exists a sequence $\boldsymbol{Q}_n$ such that $\forall k \in \{1, \dots, d\}$:

a) $\operatorname{var}[S_{n,k}|\mathcal{G}] \xrightarrow[n\to\infty]{\mathcal{P}} 0$,

b) $\frac{\mathsf{E}\,[S_{n,k}|\mathcal{G}]}{Q_{n,k}} \xrightarrow[n\to\infty]{\mathcal{P}} 1$,

c) $\boldsymbol{Q}_n \xrightarrow[n\to\infty]{\mathcal{P}} \boldsymbol{q}$,

where $S_{n,k}$, resp. $Q_{n,k}$, is the k-th component of $\boldsymbol{S}_n$, resp. $\boldsymbol{Q}_n$. The random part of $\boldsymbol{S}_n$ conditionaly on $\mathcal{G}$ is the sampling of the sets $\tilde{\mathcal{R}}_i$. This is why we want to use a sequence of random vectors $\boldsymbol{Q}_n$, which will only be random in the variables Y_i and $\boldsymbol{Z}_i$.

We know that $\boldsymbol{S}_n$ is an average of n independent random vectors which are identically distributed, because the data are identically distributed and the sampling of $\tilde{\mathcal{R}}_i$ is independent from $\boldsymbol{Z}_i$ and Y_i. Since $\boldsymbol{Z}_i$ are bounded, $\mathsf{E}\,[\boldsymbol{w}(\tilde{\mathcal{R}}_i)Y_iB_i|\mathcal{G}]$ is finite for all $i \in \{1, \dots, n\}$. Therefore, part a) assures that all the components of $\boldsymbol{S}_n$ converge to their conditional mean according to the law of large numbers. In order to show that parts a)$-$c) are enough to prove this lemma, assume that $\mathsf{E}\,[S_{n,k}|\mathcal{G}] \xrightarrow[n\to\infty]{\mathcal{P}} s_k \neq q_k$, where q_k is the k-th component of the vector $\boldsymbol{q}$. Part c) assures that $Q_{n,k} \xrightarrow[n\to\infty]{\mathcal{P}} q_k$, so then

$$\frac{\mathsf{E}\,[S_{n,k}|\mathcal{G}]}{Q_{n,k}} \xrightarrow[n\to\infty]{\mathcal{P}} \frac{s_k}{q_k} \neq 1,$$

which violates part b). This means that by proving a)$-$c), we prove the lemma.
a) Define $Z(T) = \max_{j\in T} |\boldsymbol{Z}_j|$, where $|\boldsymbol{a}| = \sqrt{\boldsymbol{a}^\top \boldsymbol{a}}$ is the Euclidean norm of any vector $\boldsymbol{a}$. Then $0 \leq |\boldsymbol{w}(T)| \leq Z(T)$, because clearly

$$\begin{aligned} \left| \frac{\sum_{j\in T} \boldsymbol{Z}_j e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j}}{\sum_{j\in T} e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j}} \right| &\leq \max_{j\in T} |\boldsymbol{Z}_j| \\ \left| \sum_{j\in T} \boldsymbol{Z}_j e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j} \right| &\leq \max_{j\in T} |\boldsymbol{Z}_j| \cdot \sum_{j\in T} e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j} \\ \sqrt{\sum_{j\in T} \boldsymbol{Z}_j^\top e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j} \cdot \sum_{j\in T} \boldsymbol{Z}_j e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j}} &\leq \max_{j\in T} \sqrt{\boldsymbol{Z}_j^\top \boldsymbol{Z}_j} \sum_{j\in T} e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j}. \end{aligned}$$

Every component of $\boldsymbol{S}_n$ is an average of n independent random variables. As we have just proven, we can restrict $\boldsymbol{w}^\top(\tilde{\mathcal{R}}_i)\boldsymbol{w}(\tilde{\mathcal{R}}_i)$ from above by $Z^2(\tilde{\mathcal{R}}_i)$ and $\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i$ by $|\boldsymbol{\beta}_0|Z(\tilde{\mathcal{R}}_i)$. Therefore, we can restrict the conditional variance of $S_{n,k}$ from above as

$$\operatorname{var}[S_{n,k}|\mathcal{G}] \leq \frac{1}{n^2} \sum_{i=1}^{n} \mathsf{E}\,[Y_i\, e^{2\cdot|\boldsymbol{\beta}_0|\cdot Z(\tilde{\mathcal{R}}_i)} Z^2(\tilde{\mathcal{R}}_i)|\mathcal{G}],$$

which goes to zero as n goes to infinity, because of the assumption that $\boldsymbol{Z}$ is bounded. This proves part a).

b) We can rewrite $\boldsymbol{w}(\tilde{\mathcal{R}}_i)$ as $\sum_{T \in \mathcal{P}_{m,i}(\mathcal{R})} \boldsymbol{w}(T)\mathbb{1}\{T = \tilde{\mathcal{R}}_i\}$, therefore

$$\mathsf{E}\left[\boldsymbol{w}(\tilde{\mathcal{R}}_i)Y_i e^{\beta_0^\top \boldsymbol{Z}_i}|\mathcal{G}\right] = \sum_{T \in \mathcal{P}_{m,i}(\mathcal{R})} \boldsymbol{w}(T)Y_i e^{\beta_0^\top \boldsymbol{Z}_i}\,\mathsf{E}\left[\mathbb{1}\{T = \tilde{\mathcal{R}}_i\}|\mathcal{G}\right].$$

Denote $\mathcal{Y} = \sigma\{Y_i, i = 1, \ldots, n\}$. Because the sampling of the risk sets is conditionally independent from the covariate information given the at-risk indicators, we can write

$$\begin{aligned} \mathsf{E}\left[\mathbb{1}\{T = \tilde{\mathcal{R}}_i\}|\mathcal{G}\right] &= \mathsf{E}\left[\mathbb{1}\{T = \tilde{\mathcal{R}}_i\}|\mathcal{Y}\right] \\ &= \mathsf{P}[T = \tilde{\mathcal{R}}_i|\mathcal{Y}] \\ &= \frac{1}{\binom{|\mathcal{R}|-1}{m-1}}\mathbb{1}\{i \in T\}Y_T. \end{aligned} \tag{2.9}$$

By summing the terms (2.9) over $T \in \mathcal{P}_{m,i}(\mathcal{R})$, the indicators $\mathbb{1}\{i \in T\}$ and Y_T are clearly equal to 1. This leads to

$$\begin{aligned} \mathsf{E}\left[\boldsymbol{S}_n|\mathcal{G}\right] &= \frac{1}{n}\frac{1}{\binom{|\mathcal{R}|-1}{m-1}}\sum_{i=1}^{n}\sum_{T \in \mathcal{P}_{m,i}(\mathcal{R})} \boldsymbol{w}(T)e^{\beta_0^\top \boldsymbol{Z}_i} \\ &= \frac{1}{n}\frac{1}{\binom{|\mathcal{R}|-1}{m-1}}\sum_{|T|=m}\{\boldsymbol{w}(T)Y_T\sum_{i \in T} e^{\beta_0^\top \boldsymbol{Z}_i}\}, \end{aligned} \tag{2.10}$$

where the second equality holds because the two sums over i and T may be rewritten as a sum over all T of size m if adding the indicator Y_T to the sum to make sure that all the subjects from T are at risk and then summing over those subjects in T.

Next, define the sequence $\boldsymbol{Q}_n$ as

$$\boldsymbol{Q}_n = \frac{1}{mp^{m-1}\binom{n}{m}}\sum_{|T|=m}\{\boldsymbol{w}(T)Y_T\sum_{i \in T} e^{\beta_0^\top \boldsymbol{Z}_i}\}.$$

The random parts of this random vector are the variables Y_i and $\boldsymbol{Z}_i$. Clearly $\frac{|\mathcal{R}|}{n} \xrightarrow[n\to\infty]{\mathcal{P}} p$, hence for any $a < |\mathcal{R}|$,

$$\frac{n-a}{|\mathcal{R}|-a} \xrightarrow[n\to\infty]{\mathcal{P}} \frac{1}{p}.$$

Then for all $k = 1, \ldots, d$,

$$\begin{aligned} \frac{\mathsf{E}\left[S_{n,k}|\mathcal{G}\right]}{Q_{n,k}} &= \frac{1}{n}\frac{1}{\binom{|\mathcal{R}|-1}{m-1}}mp^{m-1}\binom{n}{m} \\ &= \frac{(n-1)!}{(m-1)!(n-m)!}\frac{(m-1)!(|\mathcal{R}|-m)!}{(|\mathcal{R}|-1)!}p^{m-1} \\ &= \frac{(n-1)!(|\mathcal{R}|-m)!}{(|\mathcal{R}|-1)!(n-m)!}p^{m-1} \\ &= \frac{(n-1)\ldots(n-m+1)}{(|\mathcal{R}|-1)\ldots(|\mathcal{R}|-m+1)}p^{m-1} \xrightarrow[n\to\infty]{\mathcal{P}} \frac{1}{p^{m-1}}p^{m-1} = 1. \end{aligned}$$

This proves part b).
c) First, we need to look at the variance of $\boldsymbol{Q}_n$.

$$\mathrm{var}(\boldsymbol{Q}_n) = const \cdot \frac{1}{\binom{n}{m}^2} \sum_{\substack{|T|=m\\|S|=m}} \mathrm{cov}\left\{\boldsymbol{w}(T)Y_T \sum_{i\in T} e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i}, \boldsymbol{w}(S)Y_S \sum_{i\in S} e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i}\right\}.$$

The covariance in the sum is equal to zero if the sets T, S are disjoint because of the assumed independence between the observations. There are $\binom{n}{2m}\binom{2m}{m}$ pairs of such disjoint sets, because there are $\binom{n}{2m}$ possibilities of choosing $2m$ subjects for these two sets together and there are $\binom{2m}{m}$ possibilities how to divide those subjects into sets T and S. This means that the sum has at most $\binom{n}{m}^2 - \binom{n}{2m}\binom{2m}{m}$ nonzero terms, which are, nevertheless, finite, since we assume that $\boldsymbol{Z}_i$ are bounded. And since

$$\begin{aligned}\frac{1}{\binom{n}{m}^2}\left[\binom{n}{m}^2 - \binom{n}{2m}\binom{2m}{m}\right] &= 1 - \frac{\binom{n}{2m}\binom{2m}{m}}{\binom{n}{m}^2}\\ &= 1 - \frac{(n-m)!^2}{(n-2m)!n!}\\ &= 1 - \frac{(n-m)\dots(n-2m+1)(n-2m)!(n-m)!}{(n-2m)!n\dots(n-m+1)(n-m)!}\\ &\approx 1 - \frac{n^m}{n^m} \xrightarrow[n\to\infty]{} 0,\end{aligned}$$

we get that $\mathrm{var}(\boldsymbol{Q}_n) \xrightarrow[n\to\infty]{\mathcal{P}} 0$. Therefore, $\boldsymbol{Q}_n$ converges to its mean by the law of large numbers. Since

$$mp^{m-1}\boldsymbol{Q}_n = \frac{1}{\binom{n}{m}} \sum_{|T|=m} \{\boldsymbol{w}(T)Y_T \sum_{i\in T} e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i}\} \tag{2.11}$$

is a form of an average and the mean of all the summands is the same, the mean of $\boldsymbol{Q}_n$ can be expressed by replacing (2.11) with a mean using the general set U introduced in a remark earlier. And since $\mathsf{E}\ Y_U = \mathsf{P}[Y_U = 1] = p^m$, we write

$$\begin{aligned}\mathsf{E}\ \boldsymbol{Q}_n &= \frac{1}{mp^{m-1}} \mathsf{E}\left\{\boldsymbol{w}(U)Y_U \sum_{i\in U} e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i}\right\}\\ &= \frac{p\,\mathsf{E}\left\{\boldsymbol{w}(U)Y_U \frac{1}{m}\sum_{i\in U} e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i}\right\}}{\mathsf{E}\ Y_U}\\ &= \frac{p\,\mathsf{E}\left\{\boldsymbol{w}(U)Y_U \frac{1}{m}\sum_{i\in U} e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i}\right\}}{\mathsf{P}[Y_U = 1]}\\ &= p\,\mathsf{E}\left\{\boldsymbol{w}(U)Y_U \frac{1}{m}\sum_{i\in U} e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i} | Y_U = 1\right\} = \boldsymbol{q}.\end{aligned}$$

This proves part c) of this lemma.

□

Corollary. In the special case for $\boldsymbol{\beta} = \boldsymbol{\beta}_0$,

$$\mathsf{E}\left[\boldsymbol{S}_n|\mathcal{G}\right] = \frac{1}{n}\sum_{i=1}^{n} Y_i \boldsymbol{Z}_i e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i}.$$

Proof. From the equation (2.10), we write

$$\begin{aligned}
\mathsf{E}\left[\boldsymbol{S}_n|\mathcal{G}\right] =& \frac{1}{n}\frac{1}{\binom{|\mathcal{R}|-1}{m-1}} \sum_{|T|=m} \left\{\boldsymbol{w}(T) Y_T \sum_{i\in T} B_i\right\} \\
=& \frac{1}{n}\frac{1}{\binom{|\mathcal{R}|-1}{m-1}} \sum_{T\in\mathcal{P}_m(\mathcal{R})} \left\{\frac{\sum_{i\in T}\boldsymbol{Z}_i B_i}{\sum_{i\in T} B_i}\sum_{i\in T} B_i\right\} \\
=& \frac{1}{n}\frac{1}{\binom{|\mathcal{R}|-1}{m-1}} \sum_{T\in\mathcal{P}_m(\mathcal{R})} \sum_{i\in T} \boldsymbol{Z}_i B_i \\
=& \frac{1}{n}\frac{1}{\binom{|\mathcal{R}|-1}{m-1}} \binom{|\mathcal{R}|-1}{m-1} \sum_{i\in\mathcal{R}} \boldsymbol{Z}_i B_i \\
=& \frac{1}{n}\sum_{i\in\mathcal{R}} \boldsymbol{Z}_i B_i \\
=& \frac{1}{n}\sum_{i=1}^{n} Y_i \boldsymbol{Z}_i e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i}.
\end{aligned}$$

Here, the second equation holds due to the fact that the sum is taken over all $|T| = m$ and the terms of the sum contain Y_T which ensures that all of the subjects in the set T are at risk. Therefore, we may remove Y_T and take the sum over all $T \in \mathcal{P}_m(\mathcal{R})$. The fourth equation stems from the fact that the double sum $\sum_{T\in\mathcal{P}_m(\mathcal{R})}\sum_{i\in T}$ may be rewritten as $\binom{|\mathcal{R}|-1}{m-1}\sum_{i\in\mathcal{R}}$, since by taking all $i \in T$ and all $T \in \mathcal{P}_m(\mathcal{R})$, every $i \in \mathcal{R}$ is used in the double sum exactly $\binom{|\mathcal{R}|-1}{m-1}$ times because there are this many subsets of $\mathcal{R}$ of a size m including i.

□

We may formulate three very similar lemmas to Lemma 12 which will be just as important and can be proven the same way.

Lemma 13. *Let all assumptions of Lemma 12 hold. Let* $\boldsymbol{B}_i = \boldsymbol{Z}_i^\top e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i}$. *Then (2.8) holds.*

Lemma 14. *Let all assumptions of Lemma 12 hold. For any set* $T \in \mathcal{P}_m(\mathcal{R})$, *define*

$$\mathbb{w}(T) = \left[\frac{\sum_{j\in T}\boldsymbol{Z}_j e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j}}{\sum_{j\in T} e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j}}\right]^{\otimes 2},$$

where $\mathbb{w}(\emptyset) = \mathbb{0}$. *Let* $B_i = e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i}$. *Then (2.8) holds.*

Lemma 15. *Let all assumptions of Lemma 12 hold. For any set* $T \in \mathcal{P}_m(\mathcal{R})$, *define*

$$\mathbb{w}(T) = \frac{\sum_{j\in T}\boldsymbol{Z}_j^{\otimes 2} e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j}}{\sum_{j\in T} e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j}},$$

where $\mathbb{w}(\emptyset) = \mathbb{0}$. *Let* $B_i = e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i}$. *Then (2.8) holds.*

Remark. It is important to realize that for $(\gamma,\rho)\in\{(1,1),(1,2),(2,1)\}$,

$$w(\tilde{\mathcal{R}}_i)=\left[\frac{S^{(\gamma)}_{n,i}(\boldsymbol{\beta},s)}{S^{(0)}_{n,i}(\boldsymbol{\beta},s)}\right]^{\otimes\rho},$$

where $(\gamma,\rho)=(1,1)$ in Lemmas 12 and 13, $(\gamma,\rho)=(1,2)$ in Lemma 14 and $(\gamma,\rho)=(2,1)$ in Lemma 15. This is because for any $k=0,1,2$, $S^{(k)}_{n,i}$ is a sum over the sampled controls for the i-th subject who is a case including the case, therefore it is a sum over the set $\tilde{\mathcal{R}}_i$ if this set was chosen to be the sampled risk set for the i-th individual. We can also see that $\boldsymbol{w}(\tilde{\mathcal{R}}_i)$ for $(\gamma,\rho)=(1,1)$ is a part of the score statistic (2.5) and $\mathbb{w}(\tilde{\mathcal{R}}_i)$ for $(\gamma,\rho)=(1,2)$ and $(\gamma,\rho)=(2,1)$ are parts of the observed information matrix (2.6), which is why those lemmas will be very useful in proving the asymptotic properties of the MPLE.

Theorem 16 (The consistency of the MPLE)**.** *The maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}$ is a consistent estimator of the regression parameter.*

Proof. Define the following processes:

$$\begin{aligned}
l_n(\boldsymbol{\beta},t)=&logL(\boldsymbol{\beta},t)\\
=&\sum_{i=1}^{n}\int_0^t\left[\boldsymbol{\beta}^\top\boldsymbol{Z}_i(s)-\log nS^{(0)}_{n,i}(\boldsymbol{\beta},s)\right]dN_i(s),\\
H_i(t)=&(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top\boldsymbol{Z}_i(t)-log\frac{S^{(0)}_{n,i}(\boldsymbol{\beta},t)}{S^{(0)}_{n,i}(\boldsymbol{\beta}_0,t)},\\
X_n(\boldsymbol{\beta},t)=&\frac{1}{n}\Big[l_n(\boldsymbol{\beta},t)-l_n(\boldsymbol{\beta}_0,t)\Big]\\
=&\frac{1}{n}\sum_{i=1}^{n}\int_0^t H_i(s)dN_i(s),\\
A_n(\boldsymbol{\beta},t)=&\frac{1}{n}\sum_{i=1}^{n}\int_0^t H_i(s)d\mathrm{A}_i(s).
\end{aligned}$$

Here, $\mathrm{A}_i(t)=\int_0^t\alpha_i(s)$ is the cumulative intensity of the i-th subject and we denote it A_i as a large alpha letter. Clearly, H_i are bounded and $\mathcal{F}_t$-predictable because we assume that $\boldsymbol{Z}_i$ are bounded and $\mathcal{F}_t$-predictable in the assumption **A.3**. According to Theorem 6, $N_i-\mathrm{A}_i=M_i$ is an $\mathcal{F}_t$-martingale and an integral of a predictable bounded process with respect to a martingale is also a martingale due to Theorem 4, therefore $\int_0^t H_i(s)dM_i(s)$ is a martingale. Also, a sum of martingales is a martingale due to Theorem 2, therefore $\forall\boldsymbol{\beta}$,

$$X_n(\boldsymbol{\beta},t)-A_n(\boldsymbol{\beta},t)=\frac{1}{n}\sum_{i=1}^{n}\int_0^t H_i(s)dM_i(s)$$

is an $\mathcal{F}_t$-martingale. This means that

$$\mathsf{E}\left[X_n(\boldsymbol{\beta},t)-A_n(\boldsymbol{\beta},t)\right]=0$$

and if we prove that $\mathrm{var}[X_n(\boldsymbol{\beta},t)-A_n(\boldsymbol{\beta},t)]\xrightarrow[n\to\infty]{\mathcal{P}}0$, then for a special case $t=\tau$, we get that

$$X_n(\boldsymbol{\beta},\tau)-A_n(\boldsymbol{\beta},\tau)\xrightarrow[n\to\infty]{\mathcal{P}}0. \tag{2.12}$$

Let us look at the variance. For easier notation, denote $G_i(t) = \int_0^t H_i(s)dM_i(s)$. Then

$$\begin{aligned} &\mathrm{var}[X_n(\boldsymbol{\beta},t) - A_n(\boldsymbol{\beta},t)] \\ &\quad = \mathrm{var}\left[\frac{1}{n}\sum_{i=1}^{n} G_i(t)\right] \\ &\quad = \frac{1}{n^2}\left[\sum_{i=1}^{n} \mathrm{var}[G_i(t)] + \sum_{i\neq j} \mathrm{cov}[G_i(t), G_j(t)]\right]. \end{aligned}$$

We can write

$$\begin{aligned} \mathrm{var}[G_i(t)] &= \mathsf{E}\left\langle \int_0^t H_i(s)dM_i(s)\right\rangle \\ &= \mathsf{E}\int_0^t H_i^2(s)d\langle M_i, M_i\rangle(s) \\ &= \mathsf{E}\int_0^t H_i^2(s)d\mathrm{A}_i(s) \\ &= \mathsf{E}\int_0^t H_i^2(s)Y_i(s)\lambda_0(s)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(s)}ds \leq K < \infty, \end{aligned}$$

where K is a constant. The second equality holds due to Theorem 10 and the third equality holds due to Theorem 7. The estimation from above by a constant K is possible, because H_i, Y_i and $\boldsymbol{Z}_i$ are bounded and we compute the mean of an integral composed of those variables. In a similar way we may rewrite

$$\begin{aligned} \mathrm{cov}[G_i(t), G_j(t)] &= \mathsf{E}\left\langle \int_0^t H_i(s)dM_i(s), \int_0^t H_j(s)dM_j(s)\right\rangle \\ &= \mathsf{E}\int_0^t H_i(s)H_j(s)d\langle M_i, M_j\rangle(s) \\ &= 0, \end{aligned}$$

where the second equality again holds due to Theorem 10 and the third equality holds due to Theorem 8, by which $M_i, M_j, i \neq i$ are orthogonal since $N_i, N_j, i \neq j$ have distinct event times. Therefore, $\langle M_i, M_j\rangle(s) = 0$ by Definition 16. Together, we get that

$$\begin{aligned} \mathrm{var}[X_n(\boldsymbol{\beta},t) - A_n(\boldsymbol{\beta},t)] &= \frac{1}{n^2}\left[\sum_{i=1}^{n} \mathrm{var}[G_i(t)] + \sum_{i\neq j} \mathrm{cov}[G_i(t), G_j(t)]\right] \\ &\leq \frac{1}{n^2}\Big[n\cdot K + n(n-1)\cdot 0\Big] \\ &= \frac{1}{n}\cdot K \xrightarrow[n\to\infty]{\mathcal{P}} 0. \end{aligned}$$

Therefore, (2.12) holds, hence $A_n(\boldsymbol{\beta},\tau)$ and $X_n(\boldsymbol{\beta},\tau)$ have the same limit.

We now need to differentiate the function $A_n(\boldsymbol{\beta},\tau)$ with respect to $\boldsymbol{\beta}$ to find the point which maximizes this function. Using the equation (2.1), we get

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\beta}} A_n(\boldsymbol{\beta},\tau) =&\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left[\boldsymbol{Z}_i(s) - \frac{\sum_{j=1}^n \eta_{ij}(s)\boldsymbol{Z}_j(s)e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(s)}}{\sum_{j=1}^n \eta_{ij}(s)e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(s)}}\right] Y_i(s)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(s)}\lambda_0(s)ds\\ =&\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left[\boldsymbol{Z}_i(s) - \frac{\boldsymbol{S}_{n,i}^{(1)}(\boldsymbol{\beta},s)}{S_{n,i}^{(0)}(\boldsymbol{\beta},s)}\right] Y_i(s)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(s)}\lambda_0(s)ds\\ =&\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left[\boldsymbol{Z}_i(s) - \boldsymbol{w}(\tilde{\mathcal{R}}_i(s))\right] Y_i(s)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(s)}\lambda_0(s)ds.\end{aligned}$$

Using Lemma 12, we know that the sequence $\boldsymbol{S}_n$ at a point s is

$$\boldsymbol{S}_n(s) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{w}(\tilde{\mathcal{R}}_i(s))Y_i(s)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(s)} \xrightarrow[n\to\infty]{\mathcal{P}} \boldsymbol{q}(\boldsymbol{\beta},s),$$

where

$$\boldsymbol{q}(\boldsymbol{\beta},s) = p(s)\cdot \mathsf{E}\left[\frac{\sum_{j\in U}\boldsymbol{Z}_j(s)e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(s)}}{\sum_{j\in U}e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(s)}}\frac{1}{m}\sum_{j\in U}e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_j(s)}\middle| Y_U = 1\right].$$

Next, using the Corollary of Lemma 12, we get that

$$E[\boldsymbol{S}_n(s)|\mathcal{G}] = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i(s)Y_i(s)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(s)} \xrightarrow[n\to\infty]{\mathcal{P}} \boldsymbol{q}(\boldsymbol{\beta}_0,s),$$

where

$$\begin{aligned}\boldsymbol{q}(\boldsymbol{\beta}_0,s) =&p(s)\cdot \mathsf{E}\left[\frac{\sum_{j\in U}\boldsymbol{Z}_j(s)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_j(s)}}{\sum_{j\in U}e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_j(s)}}\frac{1}{m}\sum_{j\in U}e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_j(s)}\middle| Y_U = 1\right]\\ =&p(s)\cdot \mathsf{E}\left[\frac{1}{m}\sum_{j\in U}\boldsymbol{Z}_j(s)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_j(s)}\middle| Y_U = 1\right].\end{aligned}$$

Therefore

$$\frac{\partial}{\partial \boldsymbol{\beta}} A_n(\boldsymbol{\beta},\tau) \xrightarrow[n\to\infty]{\mathcal{P}} \int_0^\tau [\boldsymbol{q}(\boldsymbol{\beta}_0,s) - \boldsymbol{q}(\boldsymbol{\beta},s)]\alpha_0(s)ds.$$

This means that $\frac{\partial}{\partial \boldsymbol{\beta}} A_n(\boldsymbol{\beta},\tau)$ converges in probability to zero for $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, hence $\boldsymbol{\beta}_0$ maximizes $A_n(\boldsymbol{\beta},\tau)$. We know that $\hat{\boldsymbol{\beta}}$ maximizes $X_n(\boldsymbol{\beta},\tau)$, because $\hat{\boldsymbol{\beta}}$ maximizes the partial likelihood and $X_n(\boldsymbol{\beta},\tau) = \frac{1}{n}\Big[l_n(\boldsymbol{\beta},\tau) - l_n(\boldsymbol{\beta}_0,\tau)\Big]$. We have already proven that $X_n(\boldsymbol{\beta},\tau)$ and $A_n(\boldsymbol{\beta},\tau)$ have the same limit, which means that $\hat{\boldsymbol{\beta}} \xrightarrow[n\to\infty]{\mathcal{P}} \boldsymbol{\beta}_0$.

□

2.3 Asymptotic normality of the MPLE of the regression parameter

Now, having proved that the MPLE is a consistent estimator of the regression parameter, we will concentrate on its asymptotic distribution. First, we need to present matrix Γ which will be shown to be the inverse of the asymptotic variance of the MPLE. In order to do so, we use the random vector $\boldsymbol{Z}_Y$ defined in assumption **A.5***, which will be used in situations which need to be conditioned on the element $[Y = 1]$ while using the results from previous theorems to prove new ones.

Further, denote $\mathbb{Z}_{Y,U} = (\boldsymbol{Z}_{Y,1}, \dots, \boldsymbol{Z}_{Y,m})$, where $\boldsymbol{Z}_{Y,i}$, $i \in U$, are independent copies of $\boldsymbol{Z}_Y$. Let

$$\mathsf{P}[\boldsymbol{Z} = \boldsymbol{Z}_{Y,j} | \mathbb{Z}_{Y,U}] \equiv p_j = \frac{e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_{Y,j}}}{\sum_{i \in U} e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_{Y,i}}} \tag{2.13}$$

be the probability of $\boldsymbol{Z}_{Y,j}$ among all the vectors in $\mathbb{Z}_{Y,U}$ defined as a ratio of the hazard of the j-th individual and the sum of hazards of all the individuals in the set U. Then

$$\begin{aligned}\operatorname{var}[\boldsymbol{Z} | \mathbb{Z}_{Y,U}] &= \sum_{j \in U} \boldsymbol{Z}_{Y,j}^{\otimes 2} p_j - \Big[\sum_{j \in U} \boldsymbol{Z}_{Y,j} p_j\Big]^{\otimes 2} \\ &= \sum_{j \in U} \Big[\boldsymbol{Z}_{Y,j} - \bar{\boldsymbol{Z}}\Big]^{\otimes 2} p_j,\end{aligned}$$

where $\bar{\boldsymbol{Z}} = \sum_{j \in U} p_j \boldsymbol{Z}_{Y,j}$. Let us remind that $p(t) = \mathsf{P}[Y(t) = 1]$ and define a matrix

$$\Gamma(\boldsymbol{\beta}, t) = \mathsf{E}\left\{\int_0^t p(s) \frac{1}{m} \sum_{j \in U} e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_{Y,j}(s)} \operatorname{var}[\boldsymbol{Z}(s) | \mathbb{Z}_{Y,U}] \lambda_0(s) ds\right\}, \tag{2.14}$$

and finally define the matrix $\Gamma \equiv \Gamma(\boldsymbol{\beta}_0, \tau)$.

From assumption **A.5***, the matrix $\mathbb{V} = \int_0^\tau \operatorname{var}[\boldsymbol{Z}_Y(s)] \lambda_0(s) ds$ is positive definite. We need to prove that this implicates that Γ is also a positive definite matrix, because then it is invertible, which is needed when proving the asymptotic distribution of the MPLE.

Lemma 17. *If the matrix $\mathbb{V}$ is positive definite, then the matrix Γ is also positive definite.*

Proof. Suppose that Γ is not positive definite, then $\exists \boldsymbol{a} \in \mathbb{R}^d, \boldsymbol{a} \neq \boldsymbol{0}$, such that $\boldsymbol{a}^\top \Gamma \boldsymbol{a} = 0$. Because

$$\boldsymbol{a}^\top \Gamma \boldsymbol{a} = \\ \mathsf{E}\left\{\int_0^\tau p(s) \frac{1}{m} \sum_{j \in U} e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_{Y,j}(s)} \Big[\sum_{i \in U} \boldsymbol{a}^\top (\boldsymbol{Z}_{Y,i}(s) - \bar{\boldsymbol{Z}}(s))^{\otimes 2} \boldsymbol{a} p_i(s)\Big]_{\boldsymbol{\beta} = \boldsymbol{\beta}_0} \lambda_0(s) ds\right\},$$

and for almost every $s \in [0, \tau]$ we have

$$p(s) \frac{1}{m} \sum_{j \in U} e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_{Y,j}(s)} > 0,$$

then necessarily at all $s \in [0, \tau]$

$$\Big[\sum_{i \in U} \boldsymbol{a}^\top \Big(\boldsymbol{Z}_{Y,i}(s) - \bar{\boldsymbol{Z}}(s)\Big)^{\otimes 2} \boldsymbol{a} p_i(s)\Big]_{\boldsymbol{\beta} = \boldsymbol{\beta}_0} \overset{\text{a.s.}}{=} 0. \tag{2.15}$$

Instead of $\boldsymbol{a}^\top\Big(\boldsymbol{Z}_{Y,i}(s)-\bar{\boldsymbol{Z}}(s)\Big)^{\otimes 2}\boldsymbol{a}$, we can write $\Big(\boldsymbol{a}^\top(\boldsymbol{Z}_{Y,i}(s)-\bar{\boldsymbol{Z}}(s)\Big)^2$. It is clear that $\Big[\boldsymbol{a}^\top\Big(\boldsymbol{Z}_{Y,i}(s)-\bar{\boldsymbol{Z}}(s)\Big)\Big]^2 p_i(s)$ are all non-negative terms and from (2.15) their sum is almost surely equal to zero. Therefore, we get that

$$\forall i \in U: \;\; \boldsymbol{a}^\top\Big(\boldsymbol{Z}_{Y,i}(s)-\bar{\boldsymbol{Z}}(s)\Big)=0.$$

Hence,

$$\boldsymbol{a}^\top\boldsymbol{Z}_{Y,j}(s)\overset{\text{a.s.}}{=}\boldsymbol{a}^\top\boldsymbol{Z}_{Y,i}(s),\ i,j\in U,$$

and therefore $\boldsymbol{a}^\top\mathbb{V}\boldsymbol{a}=0$, because $\boldsymbol{a}^\top\operatorname{var}[\boldsymbol{Z}_Y(s)]\boldsymbol{a}=0$ for almost all $s\in[0,\tau]$. This contradicts the assumption that $\mathbb{V}$ is positive definite.

□

Since we have showed that Γ is an invertible matrix, let us prove in the following that its inverse is the variance of the asymptotic distribution of MPLE $\hat{\boldsymbol{\beta}}$. For further inference, define the norm of a matrix $\mathbb{D}\in\mathbb{R}^{d\times d}$ as

$$\|\mathbb{D}\|=\sup_{|\boldsymbol{a}|\le 1}|\mathbb{D}\boldsymbol{a}|.$$

Let us remind that the observed information matrix redefined as a process is

$$\mathcal{I}_n(\boldsymbol{\beta},t)=\frac{1}{n}\sum_{i=1}^{n}\int_0^t\left[\frac{\mathbb{S}^{(2)}_{n,i}(\boldsymbol{\beta},s)}{S^{(0)}_{n,i}(\boldsymbol{\beta},s)}-\left[\frac{\boldsymbol{S}^{(1)}_{n,i}(\boldsymbol{\beta},s)}{S^{(0)}_{n,i}(\boldsymbol{\beta},s)}\right]^{\otimes 2}\right]dN_i(s).$$

This next Theorem gives us a consistent estimator of Γ.

Theorem 18. *For any consistent estimator $\boldsymbol{\beta}^*$ of $\boldsymbol{\beta}_0$,*

$$\mathcal{I}_n(\boldsymbol{\beta}^*,\tau)\xrightarrow[n\to\infty]{\mathcal{P}}\Gamma.$$

Proof. Denote

$$\mathbb{D}_i(\boldsymbol{\beta},t)=\frac{\mathbb{S}^{(2)}_{n,i}(\boldsymbol{\beta},t)}{S^{(0)}_{n,i}(\boldsymbol{\beta},t)}-\left[\frac{\boldsymbol{S}^{(1)}_{n,i}(\boldsymbol{\beta},t)}{S^{(0)}_{n,i}(\boldsymbol{\beta},t)}\right]^{\otimes 2}$$

the integrands in the observed information matrix. From the assumption of bounded vectors $\boldsymbol{Z}_i$, we know that

$$\|\mathbb{D}_i(\boldsymbol{\beta},s)\|<\infty,\ s\in[0,\tau],i\in\{1,\dots,n\},$$

and since

$$\frac{\partial}{\partial t}\mathcal{I}_n(\boldsymbol{\beta},t)=\frac{1}{n}\sum_{i=1}^{n}\int_0^t\mathbb{D}_i(\boldsymbol{\beta},s)dN_i(s),$$

then also

$$\left\|\frac{\partial}{\partial t}\mathcal{I}_n(\boldsymbol{\beta},t)\right\|<\infty,$$

therefore $\mathcal{I}_n$ is Lipschitz continuous, which means that $\exists K>0$ such that

$$\left\|\mathcal{I}_n(\boldsymbol{\beta}^*,\tau)-\mathcal{I}_n(\boldsymbol{\beta}_0,\tau)\right\|\le K|\boldsymbol{\beta}^*-\boldsymbol{\beta}_0|.$$

Because we assume that $\boldsymbol{\beta}^* \xrightarrow[n\to\infty]{\mathcal{P}} \boldsymbol{\beta}_0$, it is sufficient to show that

$$\mathcal{I}_n(\boldsymbol{\beta}_0, \tau) \xrightarrow[n\to\infty]{\mathcal{P}} \Gamma.$$

We know that

$$\int_0^t \mathbb{D}_i(\boldsymbol{\beta}_0, s) dN_i(s) - \int_0^t \mathbb{D}_i(\boldsymbol{\beta}_0, s) d\mathrm{A}_i(s), \;\; i = 1, \dots, n,$$

are $\mathcal{F}_t$-martingales from the Doob-Meyer decomposition theorem and because $\mathbb{D}_i(\boldsymbol{\beta}_0, s)$ are clearly bounded and $\mathcal{F}_t$-predictable. By taking a form of an average of these martingales

$$\frac{1}{n}\sum_{i=1}^n \int_0^t \mathbb{D}_i(\boldsymbol{\beta}_0, s) dN_i(s) - \frac{1}{n}\sum_{i=1}^n \int_0^t \mathbb{D}_i(\boldsymbol{\beta}_0, s) d\mathrm{A}_i(s),$$

we get that this difference converges in probability to its mean, which is zero since it is a martingale. The first part of this difference is equal to $\mathcal{I}(\boldsymbol{\beta}_0, t)$, so it is sufficient to show that

$$\frac{1}{n}\sum_{i=1}^n \int_0^t \mathbb{D}_i(\boldsymbol{\beta}_0, s) d\mathrm{A}_i(s) \xrightarrow[n\to\infty]{\mathcal{P}} \Gamma.$$

This holds because of Lemmas 14 and 15. More specifically, we can write

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n & \int_0^t \mathbb{D}_i(\boldsymbol{\beta}_0, s) d\mathrm{A}_i(s) \\ &= \frac{1}{n}\sum_{i=1}^n \int_0^t \left\{ \frac{\mathbb{S}^{(2)}_{n,i}(\boldsymbol{\beta}_0, s)}{S^{(0)}_{n,i}(\boldsymbol{\beta}_0, s)} - \left[\frac{\boldsymbol{S}^{(1)}_{n,i}(\boldsymbol{\beta}_0, s)}{S^{(0)}_{n,i}(\boldsymbol{\beta}_0, s)}\right]^{\otimes 2} \right\} Y_i(s) B_i(s) \lambda_0(s) ds,\end{aligned}$$

where $B_i(s) = e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(s)}$. Due to Lemma 15,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n & \int_0^t \frac{\mathbb{S}^{(2)}_{n,i}(\boldsymbol{\beta}_0, s)}{S^{(0)}_{n,i}(\boldsymbol{\beta}_0, s)} Y_i(s) B_i(s) \lambda_0(s) ds \\ & \xrightarrow[n\to\infty]{\mathcal{P}} \mathsf{E}\left\{ \int_0^t p(s) \frac{\sum_{j\in U} \boldsymbol{Z}_j^{\otimes 2}(s) e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(s)}}{\sum_{j\in U} e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(s)}} \frac{1}{m} \sum_{j\in U} B_j(s)\lambda_0(s) ds | Y_U(s) = 1 \right\},\end{aligned} \tag{2.16}$$

and due to Lemma 14,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n & \int_0^t \left[\frac{\boldsymbol{S}^{(1)}_{n,i}(\boldsymbol{\beta}_0, s)}{S^{(0)}_{n,i}(\boldsymbol{\beta}_0, s)}\right]^{\otimes 2} Y_i(s) B_i(s) \lambda_0(s) ds \\ & \xrightarrow[n\to\infty]{\mathcal{P}} \mathsf{E}\left\{ \int_0^t p(s) \left[\frac{\sum_{j\in U} \boldsymbol{Z}_j(s) e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(s)}}{\sum_{j\in U} e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(s)}}\right]^{\otimes 2} \frac{1}{m} \sum_{j\in U} B_j(s)\lambda_0(s) ds | Y_U(s) = 1 \right\}.\end{aligned} \tag{2.17}$$

The difference of 2.16 and 2.17 at point $t = \tau$ gives us the matrix Γ (2.14) by also taking the random variable $\boldsymbol{Z}_Y$ instead of $\boldsymbol{Z}$ and by realizing that

$$\begin{aligned}\mathrm{var}[\boldsymbol{Z}|\mathbb{Z}_{Y,U}] &= \sum_{j\in U} \boldsymbol{Z}_{Y,j}^{\otimes 2} p_j - \Big[\sum_{j\in U} \boldsymbol{Z}_{Y,j} p_j\Big]^{\otimes 2} \\ &= \frac{\sum_{j\in U} \boldsymbol{Z}_{Y,j}^{\otimes 2} e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_{Y,j}}}{\sum_{j\in U} e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_{Y,j}}} - \left[\frac{\sum_{j\in U} \boldsymbol{Z}_{Y,j} e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_{Y,j}}}{\sum_{j\in U} e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_{Y,j}}}\right]^{\otimes 2},\end{aligned}$$

where the probabilities p_j are defined in (2.13). This completes the proof.

□

Lemma 17 and Theorem 18 are the basics we need to prove the following theorem about the asymptotic distribution of the MPLE.

Theorem 19 (The asymptotic normality of the MPLE)**.** *It holds for $\hat{\boldsymbol{\beta}}$, the MPLE of $\boldsymbol{\beta}_0$, that*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(\mathbf{0}, \Gamma^{-1}).$$

Proof. First, let us write the Taylor expansion of $\boldsymbol{U}_n(\hat{\boldsymbol{\beta}}, \tau)$ around $\boldsymbol{\beta}_0$ as

$$\begin{aligned}
\boldsymbol{U}_n(\hat{\boldsymbol{\beta}}, \tau) =& \ \boldsymbol{U}_n(\boldsymbol{\beta}_0, \tau) + \frac{\partial \boldsymbol{U}_n(\boldsymbol{\beta}^*, \tau)}{\partial \boldsymbol{\beta}^\top}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\\
\boldsymbol{U}_n(\hat{\boldsymbol{\beta}}, \tau) - \boldsymbol{U}_n(\boldsymbol{\beta}_0, \tau) =& - n\mathcal{I}_n(\boldsymbol{\beta}^*, \tau)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\\
\frac{1}{\sqrt{n}}\boldsymbol{U}_n(\boldsymbol{\beta}_0, \tau) =& \ \mathcal{I}_n(\boldsymbol{\beta}^*, \tau)\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),
\end{aligned}$$

where $\boldsymbol{\beta}^*$ is consistent for $\boldsymbol{\beta}_0$. We used the fact that $\boldsymbol{U}_n(\hat{\boldsymbol{\beta}}, \tau) = \mathbf{0}$. It is sufficient to show that

$$\frac{1}{\sqrt{n}}\boldsymbol{U}_n(\boldsymbol{\beta}_0, \tau) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(\mathbf{0}, \Gamma)$$

because we already know from Theorem 18 that $\mathcal{I}_n(\boldsymbol{\beta}^*, \tau) \xrightarrow[n\to\infty]{\mathcal{P}} \Gamma$. Using the positive definiteness (and therefore the invertibility) of matrix Γ, proven in Lemma 17, we get the required statement. Define

$$\boldsymbol{E}_i(t) = \frac{S_{n,i}^{(1)}(\boldsymbol{\beta}_0, t)}{S_{n,i}^{(0)}(\boldsymbol{\beta}_0, t)} = \frac{\sum_{j=1}^n \eta_{ij}(t)\boldsymbol{Z}_j(t)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_j(t)}}{\sum_{j=1}^n \eta_{ij}(t)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_j(t)}}$$

and

$$\boldsymbol{E}(t) = \frac{S_n^{(1)}(\boldsymbol{\beta}_0, t)}{S_n^{(0)}(\boldsymbol{\beta}_0, t)} = \frac{\sum_{i=1}^n Y_i(t)\boldsymbol{Z}_i(t)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(t)}}{\sum_{i=1}^n Y_i(t)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(t)}}.$$

Let us write $\boldsymbol{U}_n(\boldsymbol{\beta}_0, t)$ in terms of $\boldsymbol{E}_i$ and $\boldsymbol{E}$.

$$\begin{aligned}
\boldsymbol{U}_n(\boldsymbol{\beta}_0, t) =& \sum_{i=1}^n \int_0^t [\boldsymbol{Z}_i(s) - \boldsymbol{E}_i(s)]dN_i(s)\\
=& \sum_{i=1}^n \left\{ \int_0^t [\boldsymbol{Z}_i(s) - \boldsymbol{E}(s)]dN_i(s) + \int_0^t [\boldsymbol{E}(s) - \boldsymbol{E}_i(s)]dN_i(s) \right\}\\
=& \sum_{i=1}^n \left\{ \int_0^t [\boldsymbol{Z}_i(s) - \boldsymbol{E}(s)]dM_i(s) + \int_0^t [\boldsymbol{Z}_i(s) - \boldsymbol{E}(s)]d\mathrm{A}_i(s) \right.\\
&\left. + \int_0^t [\boldsymbol{E}(s) - \boldsymbol{E}_i(s)]dM_i(s) + \int_0^t [\boldsymbol{E}(s) - \boldsymbol{E}_i(s)]d\mathrm{A}_i(s) \right\}.
\end{aligned}$$

We used that $M_i(t) = N_i(t) - \mathrm{A}_i(t)$ due to the Doob-Meyer decomposition (Theorem 5). We know that $\sum_{i=1}^n \int_0^t [\boldsymbol{Z}_i(s) - \boldsymbol{E}(s)]d\mathrm{A}_i(s) = \mathbf{0}$, because $\boldsymbol{E}$ can be interpreted as the expectation of the covariate vector $\boldsymbol{Z}_i(t)$ of the i-th individual.

By merging together the integrals which are with respect to the martingale and by multiplying the equality by $\frac{1}{\sqrt{n}}$, we get a sum of two processes

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\boldsymbol{U}_n(\boldsymbol{\beta}_0, t)\\
&\quad=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^t[\boldsymbol{Z}_i(s)-\boldsymbol{E}_i(s)]dM_i(s)+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^t[\boldsymbol{E}(s)-\boldsymbol{E}_i(s)]d\mathrm{A}_i(s)\\
&\quad=\boldsymbol{C}^{(n)}(t)+\boldsymbol{D}^{(n)}(t). \qquad (2.18)
\end{aligned}$$

(A) First let us focus on the process $\boldsymbol{C}^{(n)}(t)$ and its convergence in probability. It is easy to realize that this process is an $\mathcal{F}_t$-martingale, because $\boldsymbol{Z}_i$ are predictable processes by the assumption **A.3** and $\boldsymbol{E}_i$ are also predictable by their definition. By Theorem 4, the integral is a martingale and due to Theorem 2, the process $\boldsymbol{C}^{(n)}(t)$ is a martingale. This means that the mean of $\boldsymbol{C}^{(n)}(t)$ is zero. To determine its predictable variation process, it is important to realize that $\langle M_i, M_j\rangle(t) = 0$, because $M_i(t), M_j(t), i \neq j$, are orthogonal $\mathcal{F}_t$-martingales. Therefore, we can write

$$\begin{aligned}
\langle\boldsymbol{C}^{(n)}\rangle(t) &=\frac{1}{n}\sum_{i=1}^{n}\int_0^t[\boldsymbol{Z}_i(s)-\boldsymbol{E}_i(s)]^{\otimes 2}d\langle M_i, M_i\rangle(s)\\
&=\int_0^t\frac{1}{n}\sum_{i=1}^{n}[\boldsymbol{Z}_i(s)-\boldsymbol{E}_i(s)]^{\otimes 2}d\mathrm{A}_i(s)\\
&=\int_0^t\frac{1}{n}\sum_{i=1}^{n}[\boldsymbol{Z}_i(s)-\boldsymbol{E}_i(s)]^{\otimes 2}Y_i(s)e^{\boldsymbol{\beta}_0^\top\boldsymbol{Z}_i(s)}\lambda_0(s)ds,
\end{aligned}$$

where the second equality stems from Theorem 7. We can write further that

$$[\boldsymbol{Z}_i(s)-\boldsymbol{E}_i(s)]^{\otimes 2}=\boldsymbol{Z}_i^{\otimes 2}(s)-\boldsymbol{Z}_i(s)\boldsymbol{E}_i^\top(s)-\boldsymbol{E}_i(s)\boldsymbol{Z}_i^\top(s)+\boldsymbol{E}_i^{\otimes 2}(s).$$

Now, let us determine the convergence of all the terms separately and look at them without the baseline hazard function $\lambda_0(s)$ at the moment.

- The first term is $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i^{\otimes 2}(s)Y_i(s)e^{\boldsymbol{\beta}_0^\top\boldsymbol{Z}_i(s)}$. Let us remind that $p(s) = \mathsf{P}[Y(s)=1]$ and $p_i(s)=\frac{e^{\boldsymbol{\beta}^\top\boldsymbol{Z}_{Y,i}(s)}}{\sum_{j\in U}e^{\boldsymbol{\beta}^\top\boldsymbol{Z}_{Y,j}(s)}}$ as it was defined in the previous text. We can use the law of large numbers and write

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i^{\otimes 2}(s)Y_i(s)e^{\boldsymbol{\beta}_0^\top\boldsymbol{Z}_i(s)} &\xrightarrow[n\to\infty]{\mathcal{P}} p(s)\,\mathsf{E}\left\{\boldsymbol{Z}^{\otimes 2}(s)e^{\boldsymbol{\beta}_0^\top\boldsymbol{Z}(s)}|Y(s)=1\right\}\\
&=p(s)\,\mathsf{E}\left\{\boldsymbol{Z}_Y^{\otimes 2}(s)e^{\boldsymbol{\beta}_0^\top\boldsymbol{Z}_Y(s)}\right\}\\
&=p(s)\,\mathsf{E}\left\{\Big[\sum_{i\in U}\boldsymbol{Z}_{Y,i}^{\otimes 2}(s)p_i(s)\Big]\cdot\Big[\frac{1}{m}\sum_{j\in U}e^{\boldsymbol{\beta}_0^\top\boldsymbol{Z}_{Y,j}(s)}\Big]\right\}.
\end{aligned}$$

  Here we used the fact that the mean of terms containing $\boldsymbol{Z}_Y$ may be rewritten as a mean of the average of the terms containing $\boldsymbol{Z}_{Y,i}, i\in U$.

- The second and the third term may be both written in a very similar way as they only differ in the order. The second term

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i(s)\boldsymbol{E}_i^\top(s)Y_i(s)e^{\boldsymbol{\beta}_0^\top\boldsymbol{Z}_i(s)}$$

converges due to Lemma 13 with $\boldsymbol{B}_i(s) = \boldsymbol{Z}_i(s)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(s)}$ to

$$p(s)\,\mathsf{E}\left\{\frac{1}{m}\sum_{j\in U}\boldsymbol{B}_j(s)\boldsymbol{w}^\top(U)|Y_U(s)=1\right\}$$

and the third term

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{E}_i(s)\boldsymbol{Z}_i^\top(s)Y_i(s)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(s)}$$

converges from the same reason with $\boldsymbol{B}_i(s) = \boldsymbol{Z}_i^\top(s)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(s)}$ to

$$p(s)\,\mathsf{E}\left\{\boldsymbol{w}(U)\frac{1}{m}\sum_{j\in U}\boldsymbol{B}_j(s)|Y_U(s)=1\right\}.$$

As the next adjustments would be the same for both cases, let us show them only for the third term:

$$\begin{aligned}
&p(s)\,\mathsf{E}\left\{\boldsymbol{w}(U)\frac{1}{m}\sum_{j\in U}\boldsymbol{B}_j(s)|Y_U(s)=1\right\}\\
&\qquad=p(s)\,\mathsf{E}\left\{\frac{\sum_{j\in U}\boldsymbol{Z}_{Y,j}(s)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_{Y,j}(s)}}{\sum_{j\in U}e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_{Y,j}(s)}}\frac{1}{m}\sum_{j\in U}\boldsymbol{Z}_{Y,j}^\top(s)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_{Y,j}(s)}\right\}\\
&\qquad=p(s)\,\mathsf{E}\left\{\left[\frac{\sum_{j\in U}\boldsymbol{Z}_{Y,j}(s)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_{Y,j}(s)}}{\sum_{i\in U}e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_{Y,i}(s)}}\right]^{\otimes 2}\frac{1}{m}\sum_{j\in U}e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_{Y,j}(s)}\right\}\\
&\qquad=p(s)\,\mathsf{E}\left\{\left[\sum_{j\in U}\boldsymbol{Z}_{Y,j}(s)p_j(s)\right]^{\otimes 2}\frac{1}{m}\sum_{j\in U}e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_{Y,j}(s)}\right\}.
\end{aligned}$$

- In the fourth term we use Lemma 14 with $B_i(s) = e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(s)}$:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{E}_i^{\otimes 2}(s)Y_i(s)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(s)}\\
&\qquad\xrightarrow[n\to\infty]{\mathcal{P}}p(s)\,\mathsf{E}\left\{\left[\frac{\sum_{j\in U}\boldsymbol{Z}_j(s)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_j(s)}}{\sum_{j\in U}e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_j(s)}}\right]^{\otimes 2}\frac{1}{m}\sum_{j\in U}e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_j(s)}|Y_U(s)=1\right\}\\
&\qquad=p(s)\,\mathsf{E}\left\{\left[\frac{\sum_{j\in U}\boldsymbol{Z}_{Y,j}(s)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_{Y,j}(s)}}{\sum_{j\in U}e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_{Y,j}(s)}}\right]^{\otimes 2}\frac{1}{m}\sum_{j\in U}e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_{Y,j}(s)}\right\}\\
&\qquad=p(s)\,\mathsf{E}\left\{\left[\sum_{j\in U}\boldsymbol{Z}_{Y,j}(s)p_j(s)\right]^{\otimes 2}\frac{1}{m}\sum_{j\in U}e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_{Y,j}(s)}\right\},
\end{aligned}$$

  which is the same limit as it was for the second and third term.

Now by adding those terms together, we end up with

$$\begin{aligned}
&\langle \boldsymbol{C}^{(n)}\rangle(t)\xrightarrow[n\to\infty]{\mathcal{P}}\\
&\quad\int_0^t p(s)\,\mathsf{E}\left\{\left[\sum_{i\in U}\boldsymbol{Z}_{Y,i}^{\otimes 2}(s)p_i(s)-\left[\sum_{j\in U}\boldsymbol{Z}_{Y,j}(s)p_j(s)\right]^{\otimes 2}\right]\frac{1}{m}\sum_{j\in U}e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_{Y,j}(s)}\right\}\lambda_0(s)ds\\
&=\int_0^t p(s)\,\mathsf{E}\left\{\mathrm{var}[\boldsymbol{Z}(s)|\boldsymbol{Z}_{Y,U}]\frac{1}{m}\sum_{j\in U}e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_{Y,j}(s)}\right\}\lambda_0(s)ds\\
&=\Gamma(\boldsymbol{\beta}_0,t).
\end{aligned}\tag{2.19}$$

Now we would like to use Theorem 11 to determine the asymptotic distribution of $\boldsymbol{C}^{(n)}$. For any vector $\boldsymbol{a}_i$, denote a_{ji} its j-th component. Define

$$H_{ji}^{(n)}(s) = \frac{1}{\sqrt{n}}[Z_{ji}(s) - E_{ji}(s)],$$

which is clearly a bounded $\mathcal{F}_t$-predictable process. Then

$$C_j^{(n)}(t) = \int_0^t \sum_{i=1}^n H_{ji}^{(n)}(s) dM_i^{(n)}(s),$$
$$\langle C_j^{(n)}, C_k^{(n)} \rangle(t) = \int_0^t \sum_{i=1}^n H_{ji}^{(n)}(s) H_{ki}^{(n)}(s) d\mathrm{A}_i^{(n)}(s)$$

and for $\varepsilon > 0$,

$$\langle C_{j,\varepsilon}^{(n)}, C_{k,\varepsilon}^{(n)} \rangle(t) = \int_0^t \sum_{i=1}^n H_{ji}^{(n)}(s) H_{ki}^{(n)}(s) \mathbb{1}\{|H_{ji}^{(n)}(s)| > \varepsilon\} \mathbb{1}\{|H_{ki}^{(n)}(s)| > \varepsilon\} d\mathrm{A}_i^{(n)}(s).$$

From (2.19), we may say that

$$\langle C_j^{(n)}, C_k^{(n)} \rangle(t) \xrightarrow[n\to\infty]{\mathcal{P}} \int_0^t c_{jk}(s) ds, \;\; j,k \in \{1,\ldots,d\},$$

where $c_{jk}(s)$ are some continuous functions. Therefore the first assumption of Theorem 11 has been verified. The second assumption

$$\langle C_{j,\varepsilon}^{(n)}, C_{j,\varepsilon}^{(n)} \rangle(t) \xrightarrow[n\to\infty]{\mathcal{P}} 0$$

is easily verified since $\boldsymbol{Z}_i$ are bounded $\mathcal{F}_t$-predictable processes (**A.3**) and they are independent (**A.4**), therefore for any $j \in \{1,\ldots,d\}$, $H_{ji}^{(n)}$ are bounded, $\mathcal{F}_t$-predictable and independent for different $i = 1,\ldots,n$.

Taking the value $t = \tau$, we get that

$$\boldsymbol{C}^{(n)}(\tau) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(\boldsymbol{0}, \Gamma).$$

(B) We have shown the convergence in distribution of the first part of (2.18). To complete the proof, we need to show that $\boldsymbol{D}^{(n)}(\tau) \xrightarrow[n\to\infty]{\mathcal{P}} 0$, where

$$\boldsymbol{D}^{(n)}(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^t [\boldsymbol{E}(s) - \boldsymbol{E}_i(s)] d\mathrm{A}_i(s)$$

is the second part of (2.18). For an easier notation, we write

$$\boldsymbol{D}^{(n)}(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^t \boldsymbol{d}_i(s) ds,$$

where $\boldsymbol{d}_i(s) = [\boldsymbol{E}(s) - \boldsymbol{E}_i(s)] Y_i(s) e^{\beta_0^\top \boldsymbol{Z}_i(s)} \lambda_0(s)$. Let us consider a σ-algebra

$$\mathcal{G} = \sigma\{\boldsymbol{Z}_i(s), Y_i(s), N_i(s), s \in [0,\tau], i = 1,\ldots,n\},$$

which contains all information from the study except that of sampling. We know that for $i \neq j$, $\boldsymbol{E}_i$ and $\boldsymbol{E}_j$ are conditionally independent given $\mathcal{G}$ because the only random parts of $\boldsymbol{E}_i$ and $\boldsymbol{E}_j$ are those of sampling if conditioned on $\mathcal{G}$. Then also $\boldsymbol{d}_i$ and $\boldsymbol{d}_j$ are conditionally independent given $\mathcal{G}$ and so

$$\mathsf{E}\left[\boldsymbol{d}_i^\top(s)\boldsymbol{d}_j(t)|\mathcal{G}\right] = \mathsf{E}\left[\boldsymbol{d}_i^\top(s)|\mathcal{G}\right]\mathsf{E}\left[\boldsymbol{d}_j(t)|\mathcal{G}\right],\ \ i \neq j.$$

Using Lemma 12 and its Corollary, we may determine $\sum_{i=1}^n \mathsf{E}\left[\boldsymbol{d}_i(s)|\mathcal{G}\right]$. Since we can generalize the lemma in a way where $\boldsymbol{w}(\tilde{\mathcal{R}}_i)$ may be replaced by $\boldsymbol{w}(\tilde{\mathcal{R}})$, it holds that according to the Corollary of Lemma 12,

$$\mathsf{E}\left[\frac{1}{n}\sum_{i=1}^n \boldsymbol{E}(s)Y_i(s)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(s)}\lambda_0(s)|\mathcal{G}\right] = \frac{1}{n}\sum_{i=1}^n Y_i(s)\boldsymbol{Z}_i(s)B_i(s)\lambda_0(s)$$

and

$$\mathsf{E}\left[\frac{1}{n}\sum_{i=1}^n \boldsymbol{E}_i(s)Y_i(s)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(s)}\lambda_0(s)|\mathcal{G}\right] = \frac{1}{n}\sum_{i=1}^n Y_i(s)\boldsymbol{Z}_i(s)B_i(s)\lambda_0(s),$$

and so their difference is equal to $\mathbf{0}$. Hence,

$$\sum_{i=1}^n \mathsf{E}\left[\boldsymbol{d}_i(s)|\mathcal{G}\right] = \mathbf{0}.$$

Then

$$\begin{aligned}
&\mathsf{E}\left[\|\boldsymbol{D}^{(n)}(\tau)\|^2|\mathcal{G}\right]\\
&\quad=\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n\int_0^\tau\int_0^\tau \mathsf{E}\left[\boldsymbol{d}_i^\top(s)\boldsymbol{d}_j(t)|\mathcal{G}\right]ds\,dt\\
&\quad=\frac{1}{n}\int_0^\tau\int_0^\tau\Big\{\sum_{i=1}^n \mathsf{E}\left[\boldsymbol{d}_i^\top(s)\boldsymbol{d}_i(t)|\mathcal{G}\right]+\sum_{i=1}^n\sum_{j=1}^n \mathsf{E}\left[\boldsymbol{d}_i^\top(s)|\mathcal{G}\right]\mathsf{E}\left[\boldsymbol{d}_j(t)|\mathcal{G}\right]\\
&\qquad\qquad -\sum_{i=1}^n \mathsf{E}\left[\boldsymbol{d}_i^\top(s)|\mathcal{G}\right]\mathsf{E}\left[\boldsymbol{d}_i(t)|\mathcal{G}\right]\Big\}ds\,dt\\
&\quad=\frac{1}{n}\int_0^\tau\int_0^\tau\Big\{\sum_{i=1}^n \mathsf{E}\left[\boldsymbol{d}_i^\top(s)\boldsymbol{d}_i(t)|\mathcal{G}\right]+\sum_{i=1}^n \mathsf{E}\left[\boldsymbol{d}_i^\top(s)|\mathcal{G}\right]\sum_{j=1}^n \mathsf{E}\left[\boldsymbol{d}_j(t)|\mathcal{G}\right]\\
&\qquad\qquad -\sum_{i=1}^n \mathsf{E}\left[\boldsymbol{d}_i^\top(s)|\mathcal{G}\right]\mathsf{E}\left[\boldsymbol{d}_i(t)|\mathcal{G}\right]\Big\}ds\,dt\\
&\quad=\frac{1}{n}\int_0^\tau\int_0^\tau\Big\{\sum_{i=1}^n \mathsf{E}\left[\boldsymbol{d}_i^\top(s)\boldsymbol{d}_i(t)|\mathcal{G}\right]+\mathbf{0}^\top\mathbf{0}-\sum_{i=1}^n \mathsf{E}\left[\boldsymbol{d}_i^\top(s)|\mathcal{G}\right]\mathsf{E}\left[\boldsymbol{d}_i(t)|\mathcal{G}\right]\Big\}ds\,dt\\
&\quad=\frac{1}{n}\sum_{i=1}^n\int_0^\tau\int_0^\tau\Big\{\mathsf{E}\left[\boldsymbol{d}_i^\top(s)\boldsymbol{d}_i(t)|\mathcal{G}\right]-\mathsf{E}\left[\boldsymbol{d}_i^\top(s)|\mathcal{G}\right]\mathsf{E}\left[\boldsymbol{d}_i(t)|\mathcal{G}\right]\Big\}ds\,dt. \qquad (2.20)
\end{aligned}$$

Now, the next step would be to estimate (2.20) from above by something converging to zero in probability. That would ensure that

$$\frac{1}{\sqrt{n}}\boldsymbol{U}_n(\boldsymbol{\beta}_0,\tau) = \boldsymbol{C}^{(n)}(\tau)+\boldsymbol{D}^{(n)}(\tau)\xrightarrow[n\to\infty]{\mathcal{D}}\mathcal{N}(\mathbf{0},\Gamma)$$

and the proof would be completed. This part of the proof will not be shown in this thesis due to its technical difficulty. It is conducted by Goldstein and Langholz [1992] using the large deviation argument formulated in Billingsley [1986].

□

In this chapter, we proved that the MPLE $\hat{\boldsymbol{\beta}}$ in the nested case-control design has the same asymptotic properties as the estimator in the Cox PH model. Therefore, the use of the nested case-control design is fully justified and is a good alternative to the Cox PH model given its financial advantages. The algorithms of estimating the regression parameters are the same for both approaches since their likelihoods differ only in the indicators Y_i and η_{ij}.

3. Other similar methods and their properties

In this chapter, we will present some alternatives and extensions to the nested case-control design described in the last chapter.

3.1 Counter-matching: a stratified nested case-control design

In Chapter 2, we described the simple nested case-control design, which is very popular to use when conducting studies about rare diseases. We collect detailed covariate information only for cases and some of the controls matched to them due to their at risk status at the moments of observed failures. Now suppose that there is an information known about all of the individuals in the study. For example, the exposure information could have been collected for everyone and the aim of the study is to assess the role of potential confounders or to study interactions of the exposure with other risk factors. These potential confounders or risk factors will be then collected in more detail for a small sample of subjects. Another example may be when the exposure information could only be gathered very crudely for everyone and the researchers would like to have more exact information about the exposure, which will be again collected for a small sample.

In this chapter, based on Borgan and Langholz [1995], we present the counter-matching method, which is a stratified version of the simple nested case-control design. We assume the same variables, processes and sets as defined in Chapter 2.1 and we assume Cox's proportional hazards model, so the conditional hazard function for the i-th subject equals

$$\lambda_i(t|\boldsymbol{Z}_i) = \lambda_0(t)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(t)},$$

where λ_0 is a baseline hazard function. Under the independent censoring condition (Definition 3), the intensity of the counting process N_i is

$$\alpha_i(t) \equiv \alpha_i(t|\boldsymbol{Z}_i) = Y_i(t)\lambda_i(t|\boldsymbol{Z}_i). \tag{3.1}$$

Let us present the procedure of creating the counter-matching design. For every $k \geq 1$, we have a risk set $\mathcal{R}(X'_k)$. For each j such that X'_j is a failure time, we classify each subject from the risk set $\mathcal{R}(X'_{j-1})$ into one of L strata, where the classification cannot be based on case-control status. To classify, we use the additional information which is known about all of the individuals. Since this information may be time-dependent, the strata of one individual may differ over time. Suppose now that there are $n_l(X'_{j-1})$ subjects in the l-th stratum. For $l = 1, \ldots, L$, we fix integers $m_l > 0$ and for every l we select without replacement m_l controls from the l-th stratum except for the stratum where the case at time X'_j was classified. We will only select $m_l - 1$ controls from this stratum and the case is selected automatically. Therefore, there are $\binom{n_l(X'_{j-1})}{m_l}$ possible combinations of selected controls for each stratum l, except for the case's stratum, where there

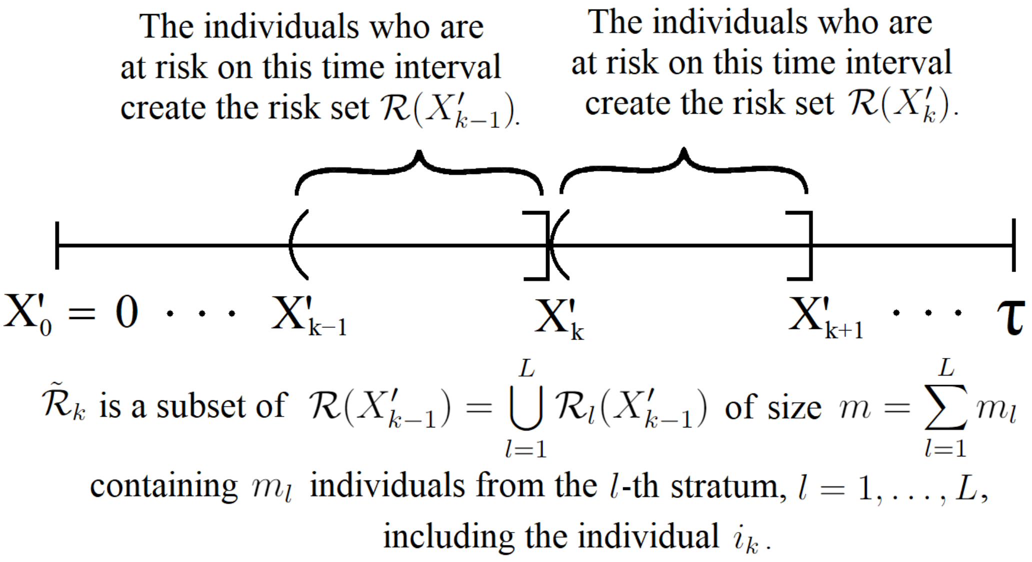
The individuals who are at risk on this time interval create the risk set $\mathcal{R}(X'_{k-1})$.
The individuals who are at risk on this time interval create the risk set $\mathcal{R}(X'_{k})$.
$X'_0 = 0$
X'_{k-1}
X'_k
X'_{k+1}
τ
$\tilde{\mathcal{R}}_k$ is a subset of $\mathcal{R}(X'_{k-1}) = \bigcup_{l=1}^{L} \mathcal{R}_l(X'_{k-1})$ of size $m = \sum_{l=1}^{L} m_l$
containing m_l individuals from the l-th stratum, $l = 1, \dots, L$,
including the individual i_k.

Figure 3.1: Denotion of risk sets and the sampled risk set when sampling the counter-matching data.

are $\binom{n_l(X'_{j-1})-1}{m_l-1}$ such combinations. The entire set of a case i_j at time X'_j and $\sum_{l=1}^{L} m_l - 1$ selected controls is called the sampled risk set at time X'_j, denoted as $\tilde{\mathcal{R}}_{j,i_j} \equiv \tilde{\mathcal{R}}_j$. The collection of all the cases and their selected controls creates the counter-matching data. For an easier understanding, the sampling process is demonstrated in Figure 3.1.

The name "counter-matching" indicates the opposite of matching. If we were about to perform the matched case-control study, we would draw controls from the same stratum as the case and therefore compare cases and controls within that stratum. Since we would like to control the effect of potential confounders, we would choose them to be the stratifying variables. Clearly, the effects of those stratifying variables could no longer be estimated. On the other hand, when performing the counter-matching design, the strata are determined by variables of interest or by proxies for such variables. By sampling the controls from all the strata, we maximize the variation of exposure in the analysis. It is also possible to estimate the effect of the stratifying variables.

Let us present an example of counter-matching used in practice. This example is similar to one introduced by Langholz and Clayton [1994]. Assume that we want to assess the risk of breast cancer from using different types of hormonal contraceptives. The women might be at first asked by a mailed questionnaire, what type of contraceptive they were using (pill, patch, injection, intrauterine device, vaginal ring,...). This information is likely to be accurate, however, the women might not know the exact name of the brand and therefore the composition of the contraceptive they were using. So at first, we stratify all the women according to the type of the contraceptive and we sample some of them from each stratum and personally ask them to give us more information about the contraceptive, from which we are able to write down the exact composition and determine what substances, hormones or what combinations of dosages increase the risk of breast cancer.

We will now derive the partial likelihood of counter-matching. Define $G_i(t) \in \{1, \dots, L\}$ a process which denotes the sampling stratum for the i-th subject at time t, i.e. $G_i(X'_j)$ denotes the stratum of the i-th subject when he was classified as a part of the risk set $\mathcal{R}(X'_{j-1})$. For $l \in 1, \dots, L$, define a risk set

$$\mathcal{R}_l(t) = \{i : Y_i(t+) = 1, G_i(t) = l\} \subset \mathcal{R}(t)$$

and $n_l(t) = |\mathcal{R}_l(t)|$. These are the risk sets for different strata, hence, they are disjunct and clearly

$$\bigcup_{l=1}^{L} \mathcal{R}_l(t) = \mathcal{R}(t).$$

For $i \in \{1, \dots, n\}$ define

$$\mathcal{P}_i(t) = \{\boldsymbol{r} \subset \mathcal{R}(t) : i \in \boldsymbol{r}, |\boldsymbol{r} \cap \mathcal{R}_l(t)| = m_l,\ l = 1, \dots, L\}$$

a set of all possible sampled risk sets which include the i-th individual and fulfill the process of sampling the controls, i.e. from each stratum there are m_l subjects in the final sampled risk set. Let us determine how many of such sets there are in $\mathcal{P}_i(t)$. Since the i-th individual has to be included in all the sampled risk sets in $\mathcal{P}_i(t)$, we can write

$$\begin{aligned}|\mathcal{P}_i(t)| =& \binom{n_1(t)}{m_1} \cdots \binom{n_{G_i(t)}(t) - 1}{m_{G_i(t)} - 1} \cdots \binom{n_L(t)}{m_L} \\ =& \frac{m_{G_i(t)}}{n_{G_i(t)}(t)} \prod_{l=1}^{L} \binom{n_l(t)}{m_l} \\ =& w_i(t)^{-1} \cdot C(t),\end{aligned}$$

where $w_i(t) = \frac{n_{G_i(t)}(t)}{m_{G_i(t)}}$ and $C(t) = \prod_{l=1}^{L} \binom{n_l(t)}{m_l}$.

Let $\mathcal{H}_t$ be a filtration on a probability space $(\Omega, \mathcal{F}, \mathsf{P})$ which includes the same information as $\mathcal{F}_t$ and is augmented by the sampling information. For each $\boldsymbol{r} \subset \{1, \dots, n\}$, $|\boldsymbol{r}| = \sum_{l=1}^{L} m_l$, define the counting process $N_{(i,\boldsymbol{r})}(t)$ counting the number of times in $[0, t]$ when the i-th individual fails and $\boldsymbol{r}$ is chosen to be the sampled risk set. Then we may write

$$\begin{aligned}\mathsf{P}[\Delta N_{(i,\boldsymbol{r})}(t) = 1|\mathcal{H}_t] =& \mathsf{P}[\Delta N_i(t) = 1, \tilde{\mathcal{R}}(t) = \boldsymbol{r}|\mathcal{H}_t] \\ =& \mathsf{P}[\Delta N_i(t) = 1|\mathcal{H}_t] \times \mathsf{P}[\tilde{\mathcal{R}}(t) = \boldsymbol{r}|\Delta N_i(t) = 1, \mathcal{H}_t]. \quad (3.2)\end{aligned}$$

We assume that the additional sampling information at any time t does not change the intensities of failures at this time. Therefore,

$$\mathsf{P}[\Delta N_i(t) = 1|\mathcal{H}_t] = \mathsf{P}[\Delta N_i(t) = 1|\mathcal{F}_t].$$

Also, it is simple to derive the second probability in (3.2) since it equals

$$\frac{1}{|\mathcal{P}_i(t)|} \mathbb{1}\{\boldsymbol{r} \in \mathcal{P}_i(t)\} = w_i(t) C(t)^{-1} \mathbb{1}\{\boldsymbol{r} \in \mathcal{P}_i(t)\}. \quad (3.3)$$

From the equations (3.1), (3.2) and (3.3), we get that the $\mathcal{H}_t$-intensity process of the counting process $N_{(i,\boldsymbol{r})}$ is

$$\begin{aligned}\alpha_{(i,\boldsymbol{r})}(t) \equiv& \alpha_{(i,\boldsymbol{r})}(t|\boldsymbol{Z})\\ =&\alpha_i(t)\,\mathsf{P}[\tilde{\mathcal{R}}(t)=\boldsymbol{r}|\Delta N_i(t)=1,\mathcal{H}_t]\\ =&Y_i(t)\lambda_0(t)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(t)}w_i(t)C(t)^{-1}\mathbb{1}\{\boldsymbol{r}\in\mathcal{P}_i(t)\}.\end{aligned}$$

Now, let us define $\alpha_{\boldsymbol{r}}(t)=\sum_{i\in\boldsymbol{r}}\alpha_{(i,\boldsymbol{r})}(t)$, the intensity process associated with the counting process counting the number of times the sampled risk set is $\boldsymbol{r}$. Then $\alpha_{(i,\boldsymbol{r})}(t)=\alpha_{\boldsymbol{r}}(t)\pi_t(i|\boldsymbol{r})$, where $\pi_t(i|\boldsymbol{r})$ is a´the conditional probability of the i-th individual failing at time t given $\mathcal{H}_t$ and given that there is a failed individual among those in $\boldsymbol{r}$ at time t. Then

$$\pi_t(i|\boldsymbol{r})=\frac{\alpha_{(i,\boldsymbol{r})}(t)}{\alpha_{\boldsymbol{r}}(t)}=\frac{Y_i(t)w_i(t)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(t)}\mathbb{1}\{\boldsymbol{r}\in\mathcal{P}_i(t)\}}{\sum_{j\in\boldsymbol{r}}Y_j(t)w_j(t)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_j(t)}\mathbb{1}\{\boldsymbol{r}\in\mathcal{P}_j(t)\}}.$$

Since the partial likelihood function is a product over all cases in the study of fractions, where the numerator is the hazard of the case and the denominator is a sum of hazards of the case and its sampled controls, we get that the partial likelihood for counter-matching $L_C(\boldsymbol{\beta})$ is a product of the probabilities $\pi_t(i|\boldsymbol{r})$, where the indicators Y_i and $\mathbb{1}\{\boldsymbol{r}\in\mathcal{P}_i(t)\}$ are no longer relevant, since the product is taken over the cases and the sum in the denominator can be replaced by a sum over $\tilde{\mathcal{R}}_k$. Therefore

$$\begin{aligned}L_C(\boldsymbol{\beta})=&\prod_{k:\delta_k=1}\left\{\frac{w_{i_k}(X_k')e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_{i_k}(X_k')}}{\sum_{j\in\tilde{\mathcal{R}}_k}w_j(X_k')e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(X_k')}}\right\}\\ =&\prod_{i=1}^n\prod_{s\in[0,\tau]}\left\{\frac{w_i(s)e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_i(s)}}{\sum_{j=1}^n\eta_{ij}(s)w_j(s)e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(s)}}\right\}^{\Delta N_i(s)},\end{aligned}\tag{3.4}$$

where η_{ij} are sampling indicators defined in (2.3) as

$$\eta_{ij}(t)=\sum_{k\geq1}\mathbb{1}\{j\in\tilde{\mathcal{R}}_{k,i}\}\mathbb{1}\{X_{k-1}'<t\leq X_k'\},\;\; t\in[0,\tau],\tag{3.5}$$

where $\eta_{ij}(0)=0$, X_k' are ordered observed censored failure times and $\tilde{\mathcal{R}}_{k,i}$ is a sampled risk set at time X_k' to which the i-th individual belongs.

The maximum partial likelihood estimator would be estimated in a similar way to the one introduced in Chapter 1.1 using the Newton Raphson algorithm described in (1.2). It can be proven by standard counting process and martingale methods that the partial likelihood (3.4) has basic likelihood properties. Details of this can be seen in Andersen and Gill [1982]. The maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}$ has asymptotically multivariate normal distribution with mean $\boldsymbol{\beta}_0$ and a covariate matrix, which may be estimated by the inverse of the observed information matrix. The proof of this would be similar to the one for the nested case-control design in Goldstein and Langholz [1992] in chapter 4 and 5.

3.2 Pseudolikelihood approach under a nested case-control design

Another option of dealing with nested case-control data is a pseudolikelihood approach introduced by Samuelsen [1997]. According to his article, it appears that we may improve the efficiency of the estimator of the regression paratemer by applying the pseudolikelihood method on the dataset. The idea is that we may obtain the conditional probability that an individual in the study will ever be chosen as a control in the nested case-control design. This probability is called the *inclusion probability* and it is conditioned on all the risk sets and all the cases at the failure times. Then, we weigh the individual likelihood contributions with the inclusion probabilities.

The main idea of this is the improvement of efficient data collection. Suppose that the covariates are constants or known functions of time. Then it would be efficient to use the data that we have already collected more than just once. For rare diseases, we select every observed case to be a part of the study, hence their inclusion probability is equal to one. Therefore, we measure all of the covariates for each case, and so it is efficient to use the case as a control for all the cases that were observed before its own failure time. The other controls would have different inclusion probabilities depending on their censoring times. The greater the censoring time, the higher probability of ever being selected as a control. Therefore, it is more efficient to select controls with higher inclusion probabilities and use them as controls more times. The main disadvantage of this approach is the fact that we may only make use of this when the complete covariate histories are known for all the cases and controls, which is not always the case when conducting medical studies.

Let us consider the same data, selection of controls and notation as described in Chapter 2. Assume Cox's proportional hazards model with the conditional hazards function presented in Definition 4.

The probability of ever being included in the nested case-control study as a case or a control is

$$p_j = \begin{cases} 1, & \text{if } \delta_j = 1, \\ p_{0j}, & \text{if } \delta_j = 0, \end{cases} \tag{3.6}$$

where p_{0j} is the inclusion probability. Since p_{0j} is the probability of a subject ever being chosen as a control, it may be derived as one minus the probability of never being chosen. The probability of never being chosen is a product over all cases k for which the j-th subject was at risk at their failure time, i.e. for which $X'_k < X_j$, of probabilities that the subject was not chosen as a control for them. This probability that the j-th subject was not selected to be a control for the case i_k equals one minus the probability that it was selected, which equals $\frac{m-1}{n(\mathcal{R}(X'_k))}$, since we choose $m-1$ controls for each case from $n(\mathcal{R}(X'_k))$ potential subjects. Hence,

$$p_{0j} = 1 - \prod_k \Big\{1 - \frac{m-1}{n(\mathcal{R}(X'_k))}\Big\}, \tag{3.7}$$

where the product is taken over all k such that i_k is a case at time X'_k and $X'_k < X_j$. Here, X'_k are ordered observed censored failure times. Also, denote

V_{0j} an indicator that the j-th individual is ever selected as a control and define $V_j = \max(\delta_j, V_{0j})$ an indicator that the j-th individual is ever a case or a selected control in the study.

According to Samuelsen [1997] and his approach, we suggest fitting model which maximizes the pseudolikelihood. This pseudolikelihood is quite similar to the partial likelihood of the nested case-control design in 2.4, however, the individual contributions are weighed by the inverse of their inclusion probabilities and the denominator of the product is the sum over all selected controls and cases in the study. Define $\mathcal{R}_i$ a set of all individuals who were at risk at the event time of the i-th individual. Then,

$$\begin{aligned} L_P(\boldsymbol{\beta}) &= \prod_{i=1}^{n} \prod_{s \in [0,\tau]} \left\{ \frac{e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_i(s)}}{\sum_{j \in \mathcal{R}_i} \frac{V_j}{p_j} e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(s)}} \right\}^{\Delta N_i(s)} \\ &= \prod_{i=1}^{n} \prod_{s \in [0,\tau]} \left\{ \frac{e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_i(s)}}{\sum_{j=1}^{n} Y_j(s) \frac{V_j}{p_j} e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(s)}} \right\}^{\Delta N_i(s)}. \end{aligned} \tag{3.8}$$

The inverse of the probability of being included in the study does not need to be written in the numerator of the product, since there is always a case in the numerator and its probability of being included and indicator V_i are both equal to one. The sum over $\mathcal{R}_i$ in the denominator was replaced with the sum over all $j \in \{1, \ldots, n\}$ and with the indicator $Y_j(s)$, which assures that the sum is taken over the individuals who were at risk at the failure time of the i-th subject.

The maximum pseudolikelihood estimator would be estimated in a similar way to the one introduced in Chapter 1.1 using the Newton Raphson algorithm described in (1.2). Under certain assumptions, Samuelsen [1997] proved that the maximum pseudolikelihood estimator is consistent and asymptotically normal.

Let us compare the likelihoods L_{Cox} in (1.1) of the Cox PH model, L in (2.4) of the nested case-control design, L_C in (3.4) of the counter-matching design and the pseudolikelihood L_P in (3.8):

$$\begin{aligned} L_{Cox}(\boldsymbol{\beta}) &= \prod_{i=1}^{n} \prod_{s \in [0,\tau]} \left\{ \frac{e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_i(s)}}{\sum_{j=1}^{n} Y_j(s) e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(s)}} \right\}^{\Delta N_i(s)} \\ L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} \prod_{s \in [0,\tau]} \left\{ \frac{e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_i(s)}}{\sum_{j=1}^{n} \eta_{ij}(s) e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(s)}} \right\}^{\Delta N_i(s)} \\ L_C(\boldsymbol{\beta}) &= \prod_{i=1}^{n} \prod_{s \in [0,\tau]} \left\{ \frac{w_i(s) e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_i(s)}}{\sum_{j=1}^{n} \eta_{ij}(s) w_j(s) e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(s)}} \right\}^{\Delta N_i(s)} \\ L_P(\boldsymbol{\beta}) &= \prod_{i=1}^{n} \prod_{s \in [0,\tau]} \left\{ \frac{e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_i(s)}}{\sum_{j=1}^{n} Y_j(s) \frac{V_j}{p_j} e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j(s)}} \right\}^{\Delta N_i(s)}. \end{aligned}$$

The difference between L_{Cox} and L is in the indicators Y_j and η_{ij}, where in L_{Cox} the denominator is taken over all subjects at risk at the time of observed failure, while the denominator in L is only taken over $m - 1$ sampled controls and the observed case. The difference between L and L_C is in the weights $w_j(s)$, which compensate for the sampling. The likelihood L_P differs from L_C also by the weights, which are now the inverse of the inclusion probabilites, and the indicators

$\eta_{ij}(s)$ are replaced by $Y_j(s)V_j$. The question whether or not these alternatives to the nested case-control design assure more accuracy in estimating the regression parameters will be the subject of interest for the simulation study in the next chapter.

4. Simulation study

In this last chapter, we conduct a simulation study which allows us to compare the quality of estimation of the regression paratemers by the Cox proportional hazards model described in Chapter 1.1 and the three study designs described in Chapters 2, 3.1 and 3.2.

Let us remind that the model for all study designs is the same Cox proportional hazards model

$$\lambda_i(t|\boldsymbol{Z}_i) = \lambda_0(t)e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i(t)},$$

where λ_i is the conditional hazard of the i-th individual conditioned on the vector of regressors $\boldsymbol{Z}_i$, λ_0 is the baseline hazard function and $\boldsymbol{\beta}_0$ is a vector of regression parameters. We may consider time-varying regressors in the model, however, for the purpose of the simulation study, which is to compare different study designs, we will assume that all of the regressors are time-invariant, since the pseudolikelihood approach requires that.

4.1 Generating data for Cox PH model

First, we need to simulate time-to-event data suitable for the analysis by the Cox proportional hazards model. Later, from these datasets, we generate smaller datasets suitable for the other study designs.

4.1.1 Regressors

We will simulate a study investigating the effect of smoking cigarettes on the risk of getting lung cancer. Imagine that this study is conducted with men and women who are 60 to 70 years old at the beginning of the study which lasted 10 years and consider the following variables:

- *age* . . . the age at the beginning of the study,
- *abstinence* . . . the length in years of not smoking at the beginning of the study,
  - non-smokers . . . *abstinence* is equal to zero,
  - smokers . . . *abstinence* is equal to zero,
  - ex-smokers . . . *abstinence* equals the difference of subject's *age* and the age when the subject stopped smoking,
- *cumcig* . . . the cumulative number of smoked cigarettes in hundreds of thousands at the beginning of the study,
- *education* . . . the level of education with two categories:
  - 0 . . . middle school or less completed,
  - 1 . . . at least high school completed.

The variable *education* does not affect the risk of getting lung cancer, however, it is correlated with smoking. We assume that individuals with lower education tend to smoke more (for longer periods of time) than individuals with higher education. We will use this variable to determine the strata for the counter-matching design, as its value is very easy to obtain and we may know this information about every individual in the study.

The other three presented variables will be the regressors of the model:

$$\boldsymbol{Z} = (age, abstinence, cumcig)^\top.$$

We assume that by increasing variables *age* and *cumcig*, the risk of getting lung cancer also increases, while by increasing *abstinence* this risk decreases. Therefore, we choose the regression parameter to be $\boldsymbol{\beta}_0 = (0.1, -0.1, 1.0)^\top$, and then we get that

$$e^{\boldsymbol{\beta}_0} = (1.105, 0.905, 2.718)^\top.$$

This means that, for example, people who are 70 years old have 2.72 times higher risk of getting lung cancer than 60 year old people ($e^{10\cdot\beta_{01}} \approx 2.718$), a person who stopped smoking 15 years ago has 77.7% lower risk than a person who is still smoking ($e^{15\cdot\beta_{02}} \approx 0.223$) and a person who smoked 1 package of cigarettes (20 cigarettes) per day for 40 years, i.e. has already smoked 292 000 cigarettes, has 18.5 times higher risk of lung cancer than a non-smoker ($e^{2.92\cdot\beta_{03}} \approx 18.541$).

To simulate these regressors, we need to create the history of smoking for all the individuals. At first, we choose piecewise constant hazard functions describing the risk of starting smoking at a certain age for men and women separately. We assume that most people start smoking around their adolescence and early twenties, after which the risk of starting smoking decreases. Also, women tend to start smoking a few years later than men. In the same way, we create piecewise constant hazard functions for the risk of ending smoking. Here, we assume that women tend to stop when they become pregnant, while men tend to stop much later in life.

In addition, we need to include the effect of the variable *education* on the risk of starting smoking. We assume that *education* has the alternative distribution with parameter $p = \frac{1}{2}$ and that those with a lower level of education have a higher risk of starting smoking and a lower risk of ending smoking. Hence, we choose the piecewise constant hazard functions for the risk of beginning smoking for men and women with a lower level of education, which may be seen in Figure 4.1 as the two plots on the left, and we assume that the hazard functions for the risk of beginning smoking for subjects with a higher level of education are 30% of those with lower education. The same principle is applied to the risk of ending smoking, only this risk is much smaller for individuals with lower education. Therefore, the piecewise constant hazard functions of the risk of ending smoking for the subjects with a lower level of education, which may be seen as the two plots on the right in Figure 4.1, are assumed to be three times smaller than the hazard functions for subjects with a higher level of education.

We now generate the age of starting (*age begin*) and ending (*age end*) smoking from the piecewise constant hazard functions. To ensure that *age begin* is always smaller than *age end*, we generate *age end* by considering the hazard function of ending smoking to be zero until *age begin*. The hazard function is non-zero from this point and has the same chosen values as in the plots on the right in Figure 4.1

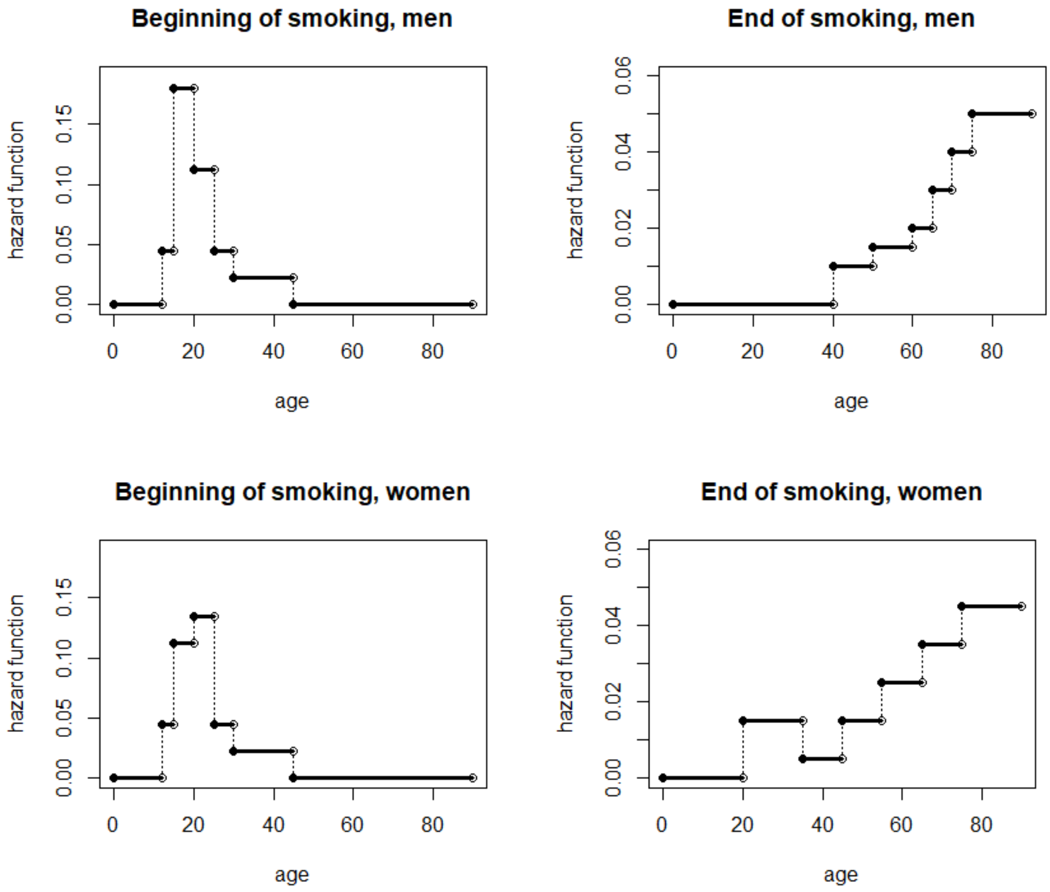

Figure 4.1: Piecewise constant hazard functions for beginning and ending smoking for men and women with lower level of education.

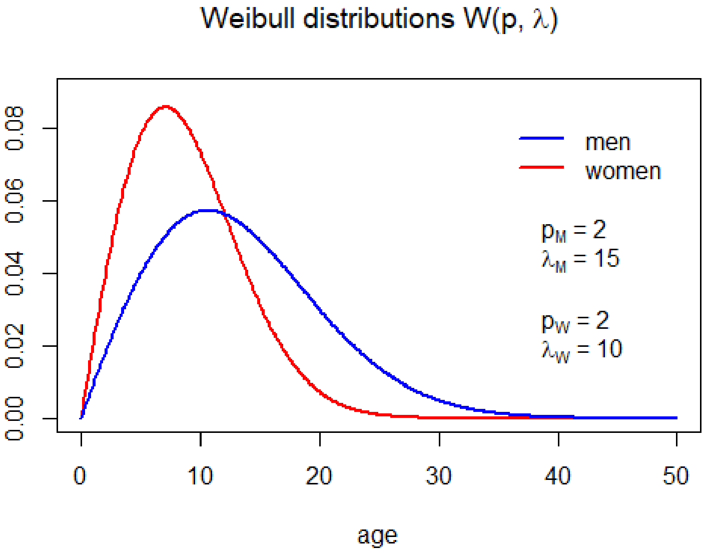

Figure 4.2: The distributions of the average of smoked cigarettes per day per men and women, where p_M, p_W are the shape parameters and λ_M, λ_W are the scale parameters.

for lower educated, resp. three times the values for higher educated individuals. Since the hazard functions of the beginning of smoking are set to zero after the age of 45, those who have not started smoking until then are pronounced lifetime non-smokers.

Now, we generate the variable *age* from the uniform distribution $Unif(60, 70)$ and create the variable *abstinence*. This variable equals zero for non-smokers and those who still smoke at the beginning of the study, otherwise, it equals the difference of *age* and *age end*. Next, we create a categorical variable *smoker*, which equals 0 if the individual never started smoking, it equals 1 if the individual is still a smoker at the age he entered the study, and it equals 2 if the individual stopped smoking before entering the study. With the chosen piecewise constant hazard functions, we get approximately 32% of smokers, 34% of ex-smokers and 34% of lifetime non-smokers.

The last variable to simulate is *cumcig*. At first, we generate a variable *cig*, which is an average of smoked cigarettes per day during the whole person's lifetime. We assume that it follows the Weibull distribution with parameters different according to sex since men tend to smoke on average slightly more cigarettes per day than women. These two distributions are drawn in Figures 4.2. By knowing these averages, we generate time-varying variable *cigtime*, where for every year of a smoking person we generate this variable from $Unif(a \cdot cig, b \cdot cig)$, where we choose $a = 0.5$ and $b = 1.5$. Therefore, we get the averages of smoked cigarettes per day which may differ every year while still being around the total average amount *cig*. By having this information, we obtain the time-varying variable *cumcigtime* as a cumulative sum of smoked cigarettes, i.e.

$$\begin{aligned}
cumcigtime[t] =& 0, \quad t = 0, 1, \ldots, age\,begin - 1, \\
cumcigtime[t] =& \sum_{i=age\,begin}^{t} 365 \cdot cigtime[i], \quad t = age\,begin, \ldots, age\,end - 1, \\
cumcigtime[t] =& cumcig[age\,end], \quad t = age\,end, \ldots age.
\end{aligned}$$

Because we want to compare the four designs described in all of the previous chapters, we need to have time-invariant regressors, because the pseudolikelihood analysis demands that. Therefore, the last regressor of the model will be *cumcig*, which is equal to $cumcigtime[age]$, where *age* is the age of the individual at the beginning of the study.

4.1.2 Failure and censoring times

Let us present a method of generating the failure times. We assume that the failure times T_i have piecewise exponential distribution with different rates on each one year time interval $[a_j, b_j) = [0, 1), [1, 2), [2, 3), \ldots, [9, 10)$. Therefore, for each $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, 10\}$, we generate

$$T_{ij} \sim Exp(\lambda_{ij}) + a_j,$$

where $\lambda_{ij} = \lambda_{0,j} \cdot e^{\boldsymbol{\beta}_0^\top \boldsymbol{Z}_i}$. Here, $\lambda_{0,j}$ is the baseline hazard on the j-th interval and we choose it to increase over time as $\lambda_{0,j} = 0.5j \cdot q$, where q is a constant. The selection of q will be discussed later in this subchapter. Then, $T_i = \min_j T_{ij}$ such

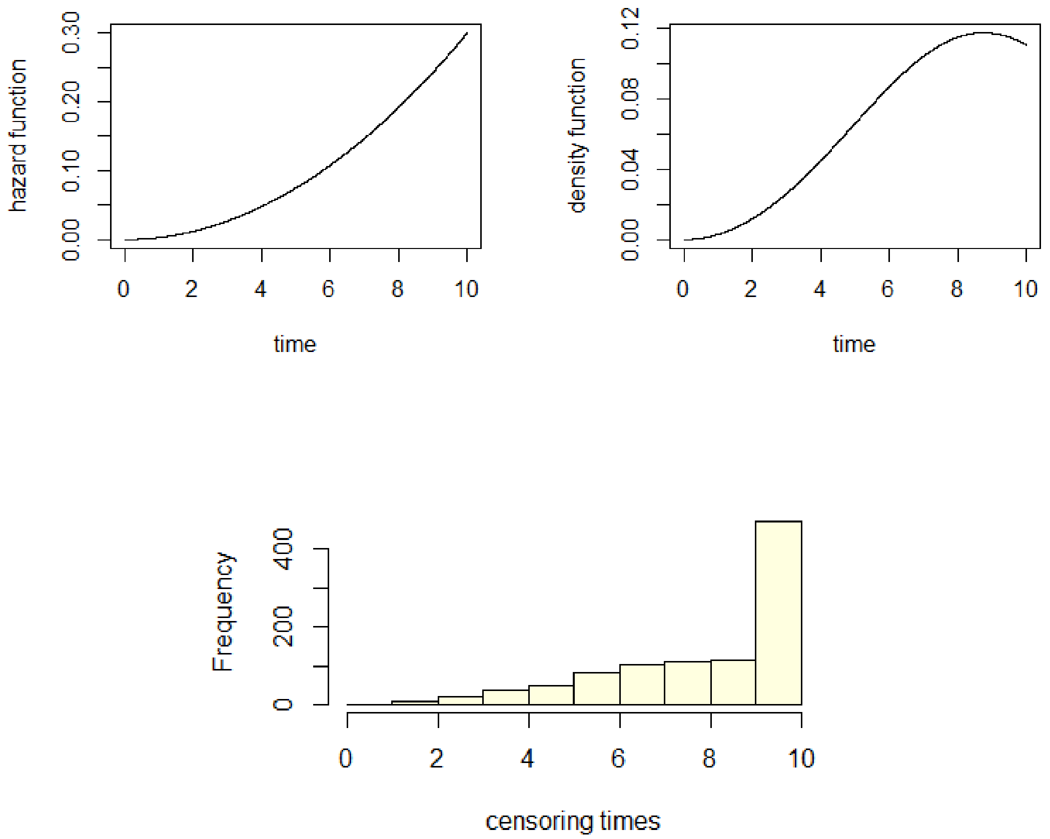

Figure 4.3: Hazard and density function of $W(3, 10)$ (Weibull distribution with shape parameter $p = 3$ and scale parameter $\lambda = 10$) on time interval $[0, 10]$ along with the histogram of generated censoring times $C_i = \min(C_{i,0}, 10)$, where $C_{i,0} \sim W(3, 10)$.

that $T_{ij} \in [a_j, b_j)$. If none of the generated T_{ij} falls into its correct interval, then we set $T_i = 11$.

Now, we generate the censoring times C_i, which we assume to have Weibull distribution with parameters chosen so that its hazard function is mildly increasing ($p > 1$) and the density function is such that the probability of being censored is the highest before the end of the study. Hence, we generate variable

$$C_{i,0} \sim W(3, 10),$$

and denote $C_i = \min(C_{i,0}, 10)$, since 10 is the ending time of the study. The hazard function and density function of $W(3, 10)$ on time interval $[0, 10]$ may be seen in Figure 4.3, as well as the histogram of generated censoring times C_i. This choice of parameters assures that there will never be a situation when we run out of controls selected for cases when conducting the nested-case control study. It may also reduce the chance of not having enough controls in each stratum for a late case when conducting the counter-matching design.

Lastly, let us discuss the choices of parameter q. This parameter affects the 10-year prevalence of the disease of interest. In Table 4.1, there are different choices of parameter q along with the prevalences, which follow from it. These prevalences were calculated from 1 000 simulated datasets. For example, for $q = 10^{-6}$, the expected prevalence of lung cancer in the 10-year study is around 5.8% with the standard deviation of 0.74%.

q	10-year prevalence (sd) [%]
10^{-7}	0.9 (0.28)
$2 \cdot 10^{-7}$	1.6 (0.38)
$5 \cdot 10^{-7}$	3.4 (0.54)
10^{-6}	5.8 (0.74)

Table 4.1: 10-year prevalence of disease by the choice of the parameter q as a part of the baseline hazard.

4.1.3 Visualization of simulated data

The visualization of one simulated dataset for the Cox PH model is demonstrated in Figure 4.4. The simulated data was performed for $n = 1000$ individuals with parameter $q = 10^{-6}$, i.e. the mean prevalence of 5.8%.

The top left plot is a boxplot of the association of *age* and lung cancer. It may be surprising that, due to this plot, there seems to be no such association between age and lung cancer, however, it is because the effect of *age* is way smaller than the effect of the other two regressors and in many cases the Cox PH model might result in a high p-value of this regressor.

On the other hand, there is a large association between *cumcig* and lung cancer as may be seen in the bottom left boxplot, as well as between *abstinence* and lung cancer in the top right boxplot. The boxplot for *abstinence* was only taken over the ex-smokers (those who stopped smoking before the study began) so that the plot is not misunderstood by having abstinence equal to zero for smokers and non-smokers.

The next plot in Figure 4.4 is the histogram of the censored failure times X. It shows that more censoring and failure times are observed over time and that there are many individuals censored by the ending time of the study, therefore, we might expect to have enough controls even for a late case.

The last plot is a boxplot of the association of *education* and *cumcig* at the bottom of Figure 4.4. We can see that more educated individuals tend to smoke less amount of cigarettes than those less educated. We may also look at Table 4.2, from which we may say that more non-smokers are among those with higher education and more smokers are among those with lower education. Also, more subjects stopped smoking than those who still smoke among the higher educated subjects, which is the opposite of the lower educated group.

	smoker		
education	non-smoker (0)	smoker (1)	ex-smoker (2)
lower (0)	64	279	157
higher (1)	259	61	180

Table 4.2: Contingency table of variables *smoker* and *education* for generated data with $n = 1\,000$ and the 10-year prevalence of 5.8%.

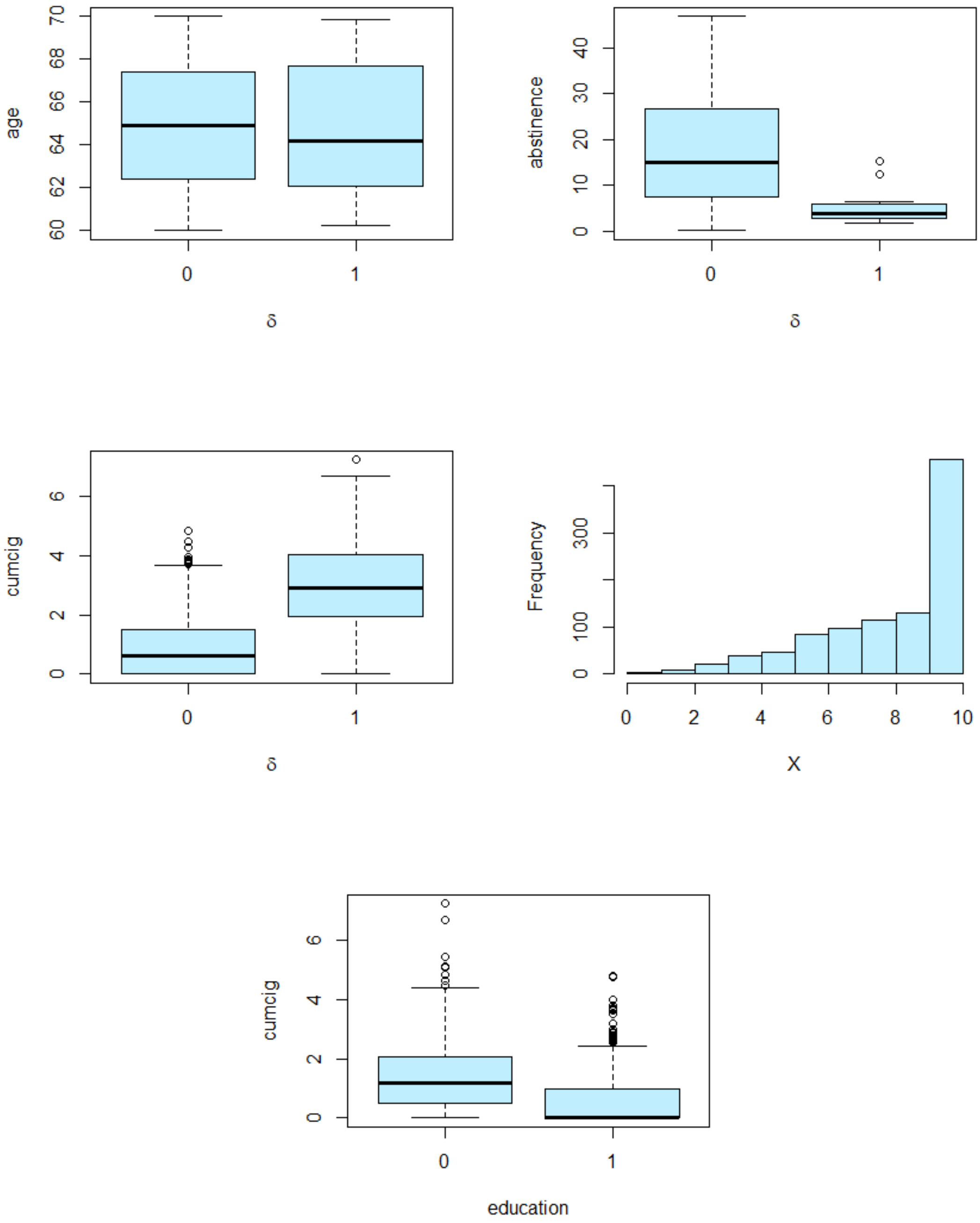

Figure 4.4: Visualization of one generated dataset of $n = 1\,000$ subjects with the 10-year prevalence of lung cancer of 5.8%. The boxplot of the association of *abstinence* and lung cancer is only taken over the subset of the data frame for those, who are ex-smokers.

4.2 Generating data for the nested case-control design and its alternatives

Creating a dataset for the nested case-control analysis from the dataset for the Cox analysis is very straightforward. For every occurring case, we select $m - 1$ individuals who are at risk at the failure time. We then save the collection of cases and their controls as a new dataset.

The dataset for the pseudolikelihood analysis is the same as the dataset for the nested case-control analysis. In addition, it contains the inclusion probabilities defined in (3.6) and (3.7) as

$$p_j = \begin{cases} 1, & \text{if } \delta_j = 1, \\ p_{0j}, & \text{if } \delta_j = 0, \end{cases}$$

where p_{0j} are calculated as

$$p_{0j} = 1 - \prod_k \Big\{ 1 - \frac{m-1}{n(\mathcal{R}(X'_k))} \Big\},$$

where the product is taken over all k such that i_k is a case at time X'_k and $X'_k < X_j$ and X'_k are ordered observed censored failure times.

When creating the dataset for counter-matching analysis, we use the known variable *education* according to which we determine two strata. This variable is not the regressor of the model, since it does not affect the risk of lung cancer, however, it was used to determine the beginning and end of smoking. Due to this and due to the bottom boxplot in Figure 4.4, it is correlated with variable *cumcig*. The sampling for the counter-matching design is then as follows. We choose $m_{CM} = (m_0, m_1)^\top$ and if the value of *education* of an observed case equals 0, we select $m_0 - 1$ individuals from those at risk whose *education* equals 0 and m_1 individuals from those at risk whose *education* equals 1, and vice versa. All these sampled controls along with the cases that they were sampled for create the dataset for counter-matching design.

4.3 Results of the simulations

We programmed the simulation in the statistical computing software R. We used function `coxph` from package `survival` (Therneau [2022]) and `distinct` from package `dplyr` (Wickham et al. [2022]).

We ran Monte Carlo simulations, each with $B = 10^3$ repetitions. For each simulation, we chose the parameters n (the number of subjects in the study) and q (to determine the prevalence of the disease). The parameter n was adapted to the parameter q so that the less percent of cases, the more subjects in the study are needed so that we have a reasonable amount of observations for each design. The results of the simulations may be seen in Tables 4.4 – 4.7, which are ordered from the lowest prevalence of 1.6% to the highest of 5.8%. The expected number of cases and the expected size of the dataset for different selections of the parameter m are listed in Table 4.3.

Each table contains the estimated $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)^\top$, which is the average of the estimators of each analysis. It also contains the average standard deviation (*average sd*), which is the average of the standard deviations of the estimators of $\boldsymbol{\beta}_0$ from all of the analyses, and the sample standard deviation (*sample sd*), which is defined as

$$\frac{1}{B-1} \sum_{b=1}^{B} (\hat{\beta}_{bj} - \hat{\beta}_j)^2, \quad j = 1, 2, 3,$$

	Table 4.4	Table 4.5	Table 4.6	Table 4.7
n	10 000	15 000	5 000	5 000
mean 10-year prevalence [%]	1.6	1.6	3.4	5.8
expected number of cases	160	240	170	290
dimension of dataset, $m = 2$	320	480	340	580
dimension of dataset, $m = 6$	960	1440	1020	1740
dimension of dataset, $m = 10$	1600	2400	1700	2900

Table 4.3: The list of the tables with simulation results, the expected number of cases and the expected size of the datasets for different selections of the parameter m.

where $\hat{\beta}_{bj}$ is the estimated β_{0j} in the b-th analysis and $\hat{\bar{\beta}}_j$ is the average of all of the estimated $\hat{\beta}_{bj}$. Lastly, the tables contain the estimated coverage probability, which is the number of times when the confidence interval covered the true value of the regression parameter divided by B, for example the estimated coverage probability of the first regression parameter is calculated as

$$\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{L_{b1} < \beta_{01} < U_{b1}\},$$

where L_{b1}, resp. U_{b1}, is the lower bound, resp. upper bound, of the confidence interval for the first regression parameter of the b-th analysis and β_{01} is the true value of the first parameter.

Let us summarize the results of the simulations, starting with the estimators. We can see that most of the estimators $\hat{\boldsymbol{\beta}}$ are within the range of 10% around the true value $\boldsymbol{\beta}_0$. The lowest deviations from the true value of $\boldsymbol{\beta}$ have the estimates in Table 4.7, where the highest deviation is 7% for CM design for $\mathbf{m} = (1, 1^\top)$ (-0.107). Most of the estimates from this table have the deviation from the true parameter lower than 4% which is a good result. Estimates $\hat{\beta}_2$ by NCC with $m = 2$ and CM with $\mathbf{m} = (1, 1)^\top$ in all the tables are the ones with the highest deviation from the true value, the highest of 15% in Table 4.4 and 4.6. From this, it seems that the simulation with the best results so far is the one with the prevalence of 5.8%. Let us notice that this Monte Carlo simulation presented us with the most cases. The second best simulation is the one with the prevalence of 1.6% and $n = 15\,000$ with the expected 240 cases.

Next, let us focus on the average and sample standard deviations. Ideally, these two should be of the same value. The biggest differences between the average and sample standard deviation are the ones for the pseudolikelihood approach, especially for $m = 2$. The biggest differences may be observed for the third parameter, the highest of 0.073 in Table 4.5 (0.113 and 0.040). The CM and NCC methods for one sampled control for each case give the second biggest differences in the standard deviations, again for the third parameter. The largest difference for the NCC design is the one in Table 4.4 of 0.019 (0.209 and 0.190) and for the CM it is the one in Table 4.6 of 0.017 (0.196 and 0.179). Those higher values of average standard deviations versus the sample standard deviations mean that the estimated variability is higher than the true variability. Summed up, the biggest differences between the average and sample standard deviation were given by all the designs with the lowest number of sampled controls for each case.

Let us evaluate the differences in the standard deviations between the different designs. It is not surprising that the lowest standard deviations were obtained by the Cox full data design, since it uses the largest amount of data. The second lowest standard deviations are for the pseudolikelihood method with 5 and 9 controls for each case. Next are the NCC and CM designs with 5 controls for each case. Clearly, the largest standard deviations were obtained for the designs using the least amount of data, i.e. NCC, CM and Pseudo with only one control for each case. The standard deviations of NCC and CM are very similar and tend to be slightly higher for CM design, especially for the first and second regression parameter. The standard deviations of the third parameter do not show much or any improvement at all for the CM approach versus the NCC method, which is the opposite of what we have expected. We may also notice that the lowest standard deviations from all the tables were obtained in Table 4.7 for 5.8% of cases and the highest ones for 1.6% of cases in Table 4.4. However, the standard deviations in Table 4.5 for 1.6% of cases were smaller than the ones in Table 4.6 for the prevalence of 3.4%. Again, it shows that the more cases we obtain in the study, the better the results. Hence, by adjusting the number n to the expected prevalence of the disease (by increasing n if the prevalence is low), we can get the same good results as if the prevalence of the disease was high.

Last but not least, let us look at the coverage probabilities of each estimator. Most of the probabilities are around 0.95 besides those of the pseudolikelihood method. The pseudolikelihood method gives the lowest coverage probabilities, the lowest of 0.441 in Table 4.5. The lowest probabilities are for the lowest number of sampled controls for each case and it increases when m increases as well. This is a poor result but it is something to be expected. If the disease of interest is rare and we have a large dataset with fewer cases, the inclusion probabilities of the other subjects are very low and their inversions too high, which may distort the estimation. The second lowest coverage probabilities are given by the CM method with $\mathbf{m} = (3,3)^\top$ and the NCC design with $m = 6$. The coverage probabilities of these two analyses are the highest for the prevalce of 1.6% and are the lowest for the prevalence of 5.8%.

Summed up, the results given in Tables 4.4 and 4.7 seem to be the most accurate and reliable. Also, the more sampled controls for each case, the lower the standard deviations of the estimators, and the lower the coverage probabilities. Also, given the obtained results, we did not show that the counter-matching design is any more efficient or accurate than the nested case-control design.

According to the simulation results, we may say that the nested case-control design is truly a great alternative to the basic Cox model when dealing with rare diseases as it was able to give us very accurate estimates. The results are more accurate by the full data design, however, given how much less observations was needed for the nested case-control design (as demonstrated in Table 4.3), there is no doubt that it is a very useful design in practice.

	Cox	**NCC**		**CM**		**Pseudo**		
m		2	6	$(1,1)^\top$	$(3,3)^\top$	2	6	10
$\hat{\beta}_1$	0.100	0.100	0.101	0.104	0.101	0.104	0.102	0.099
average sd	0.031	0.077	0.046	0.087	0.048	0.082	0.048	0.041
sample sd	0.029	0.076	0.044	0.081	0.045	0.031	0.030	0.029
coverage prob.	0.940	0.954	0.947	0.951	0.946	0.525	0.774	0.835
$\hat{\beta}_2$	-0.106	-0.115	-0.107	-0.114	-0.106	-0.110	-0.106	-0.107
average sd	0.031	0.058	0.036	0.053	0.036	0.056	0.037	0.035
sample sd	0.029	0.049	0.033	0.050	0.033	0.029	0.029	0.029
coverage prob.	0.953	0.945	0.950	0.964	0.948	0.726	0.906	0.919
$\hat{\beta}_3$	1.002	1.057	1.013	1.059	1.019	1.067	1.023	1.012
average sd	0.047	0.209	0.096	0.195	0.101	0.127	0.081	0.069
sample sd	0.048	0.190	0.095	0.189	0.096	0.050	0.049	0.048
coverage prob.	0.954	0.958	0.952	0.976	0.950	0.452	0.766	0.827

Table 4.4: The results of Monte Carlo simulation with $B = 1\,000$ repetitions with **8 divergences** for $\boldsymbol{n} = \mathbf{10\,000}$ subjects with expected **1.6% of cases** and with the regression parameter $\boldsymbol{\beta}_0 = (0.1, -0.1, 1.0)^\top$. The expression "Cox" means the full data analysis, "NCC" means the nested case-control sampling, "CM" stands for the counter-matching sampling and "Pseudo" is the pseudolikelihood approach. The number m stands for sampling $m-1$ controls for one case. The vector $\boldsymbol{m} = (m_0, m_1)^\top$ stands for sampling m_0 subjects from the first stratum and m_1 subjects from the second stratum.

	Cox	**NCC**		**CM**		**Pseudo**		
m		2	6	$(1,1)^\top$	$(3,3)^\top$	2	6	10
$\hat{\beta}_1$	0.101	0.103	0.101	0.106	0.102	0.102	0.102	0.101
average sd	0.024	0.064	0.038	0.067	0.038	0.069	0.038	0.034
sample sd	0.024	0.061	0.036	0.065	0.036	0.025	0.024	0.024
coverage prob.	0.953	0.940	0.945	0.954	0.938	0.506	0.795	0.839
$\hat{\beta}_2$	-0.104	-0.109	-0.105	-0.109	-0.105	-0.106	-0.105	-0.104
average sd	0.025	0.042	0.029	0.040	0.029	0.044	0.029	0.027
sample sd	0.023	0.038	0.027	0.039	0.027	0.023	0.023	0.023
coverage prob.	0.943	0.950	0.941	0.958	0.935	0.732	0.889	0.919
$\hat{\beta}_3$	1.001	1.034	1.012	1.043	1.013	1.049	1.015	1.007
average sd	0.039	0.155	0.081	0.159	0.082	0.113	0.070	0.058
sample sd	0.039	0.149	0.078	0.150	0.077	0.040	0.039	0.039
coverage prob.	0.946	0.955	0.944	0.960	0.936	0.441	0.717	0.789

Table 4.5: The results of Monte Carlo simulation with $B = 1\,000$ repetitions with **5 divergences** for $\boldsymbol{n} = \mathbf{15\,000}$ subjects with expected **1.6% of cases** and with the regression parameter $\boldsymbol{\beta}_0 = (0.1, -0.1, 1.0)^\top$. The expression "Cox" means the full data analysis, "NCC" means the nested case-control sampling, "CM" stands for the counter-matching sampling and "Pseudo" is the pseudolikelihood approach. The number m stands for sampling $m-1$ controls for one case. The vector $\boldsymbol{m} = (m_0, m_1)^\top$ stands for sampling m_0 subjects from the first stratum and m_1 subjects from the second stratum.

	Cox	**NCC**		**CM**		**Pseudo**		
m		2	6	$(1,1)^\top$	$(3,3)^\top$	2	6	10
$\hat{\beta}_1$	0.101	0.107	0.103	0.106	0.102	0.103	0.102	0.102
average sd	0.028	0.073	0.043	0.078	0.043	0.067	0.040	0.034
sample sd	0.028	0.070	0.040	0.074	0.041	0.029	0.028	0.028
coverage prob.	0.944	0.953	0.932	0.951	0.935	0.607	0.821	0.889
$\hat{\beta}_2$	-0.105	-0.113	-0.107	-0.115	-0.107	-0.108	-0.105	-0.105
average sd	0.027	0.045	0.032	0.049	0.033	0.045	0.031	0.029
sample sd	0.026	0.044	0.030	0.045	0.029	0.026	0.026	0.026
coverage prob.	0.949	0.961	0.943	0.954	0.936	0.787	0.924	0.939
$\hat{\beta}_3$	1.006	1.053	1.019	1.069	1.020	1.042	1.016	1.013
average sd	0.049	0.188	0.090	0.196	0.094	0.110	0.067	0.059
sample sd	0.049	0.178	0.090	0.179	0.090	0.051	0.050	0.049
coverage prob.	0.953	0.968	0.954	0.966	0.942	0.591	0.848	0.896

Table 4.6: The results of Monte Carlo simulation with $B = 1\ 000$ repetitions with **12 divergences** for $\boldsymbol{n} = \mathbf{5\ 000}$ subjects with expected **3.4% of cases** and with the regression parameter $\boldsymbol{\beta}_0 = (0.1, -0.1, 1.0)^\top$. The expression ”Cox” means the full data analysis, ”NCC” means the nested case-control sampling, ”CM” stands for the counter-matching sampling and ”Pseudo” is the pseudolikelihood approach. The number m stands for sampling $m-1$ controls for one case. The vector $\boldsymbol{m} = (m_0, m_1)^\top$ stands for sampling m_0 subjects from the first stratum and m_1 subjects from the second stratum.

	Cox	**NCC**		**CM**		**Pseudo**		
m		2	6	$(1,1)^\top$	$(3,3)^\top$	2	6	10
$\hat{\beta}_1$	0.100	0.105	0.101	0.104	0.100	0.102	0.101	0.100
average sd	0.021	0.050	0.032	0.053	0.031	0.042	0.027	0.023
sample sd	0.021	0.049	0.028	0.052	0.029	0.021	0.021	0.021
coverage prob.	0.957	0.949	0.922	0.957	0.934	0.682	0.872	0.919
$\hat{\beta}_2$	-0.102	-0.106	-0.103	-0.107	-0.103	-0.104	-0.102	-0.103
average sd	0.019	0.031	0.023	0.032	0.023	0.028	0.021	0.020
sample sd	0.018	0.030	0.020	0.030	0.020	0.018	0.018	0.018
coverage prob.	0.948	0.949	0.918	0.947	0.910	0.802	0.911	0.934
$\hat{\beta}_3$	1.002	1.026	1.007	1.026	1.005	1.018	1.006	1.004
average sd	0.039	0.139	0.070	0.131	0.070	0.075	0.049	0.044
sample sd	0.039	0.128	0.066	0.125	0.065	0.040	0.039	0.039
coverage prob.	0.949	0.948	0.937	0.954	0.937	0.695	0.880	0.920

Table 4.7: The results of Monte Carlo simulation with $B = 1\ 000$ repetitions with **3 divergences** for $\boldsymbol{n} = \mathbf{5\ 000}$ subjects with expected **5.8% of cases** and with the regression parameter $\boldsymbol{\beta}_0 = (0.1, -0.1, 1.0)^\top$. The expression ”Cox” means the full data analysis, ”NCC” means the nested case-control sampling, ”CM” stands for the counter-matching sampling and ”Pseudo” is the pseudolikelihood approach. The number m stands for sampling $m-1$ controls for one case. The vector $\boldsymbol{m} = (m_0, m_1)^\top$ stands for sampling m_0 subjects from the first stratum and m_1 subjects from the second stratum.

Conclusion

In this thesis, we have described in detail the nested case-control design as a useful sampling method when dealing with rare diseases and with possibly time-varying covariates. We have shown the procedure of creating the partial likelihood and the Newton-Raphson algorithm for obtaining the maximum partial likelihood estimator (MPLE). We have also proved the consistency of the MPLE and we have almost proved its asymptotic normality. We have not finished the proof because of its technical difficulties. Those asymptotic properties hold for $n \to \infty$. This is an impressive finding, because this means that the MPLE of nested case-control design has the same asymptotic properties as the MPLE of the basic Cox model given some assumptions.

In addition, we have extended the nested case-control design to a counter-matching design, which uses one variable known for all subjects in the study to stratify them into strata and adjusts the sampling process to this knowledge. This extention allows us to maximize the variation of exposure in the analysis. Unfortunately, we have not verified by the simulation study that this design provides more accurate results than the nested case-control design since the standard deviations of the counter-matching design were very similar to those of nested case-control design and the coverage probabilities were even lower in most situations. We have also tried to run a simulation with higher association between *education* and the ages of beginning and ending smoking to see whether or not the advantages of counter-matching would be more clear. We have assumed the hazard functions for beginning of smoking for the less educated group as ten times larger than for the higher educated, and the hazard functions for ending smoking were assumed ten times lower. We have seen some improvement in decreased standard deviations of the third estimated parameter, however, the difference still was not great and the association between *education* and *cumcig* was too large and definitely unreal in practice.

We have also presented another alternative to nested case-control design: the pseudolikelihood approach. In this design, we use an additional information in the form of the inclusion probability (the probability of being included in nested case-control sampling). The sampled controls are used for all the cases in the study while being weighted by the inverse of their inclusion probabilities. This design is only profitable when the regressors are time-invariant, which is its practical disadvantage. Unfortunately, we have not shown that this method is as good or even better than the nested case-control design. In fact, the coverage probabilities of the estimates were very low, especially with $m = 2$, and they were still under 0.95 in every simulated situation for $m = 6$ and $m = 10$.

There are many possible extensions available of this thesis. First would be to finish the proof of the asymptotic normality of the MPLE of nested case-control design, which was considered to be too technically demanding. Second would be to describe and test with the simulation study other alternative methods or extensions of nested case-control design. Also, there could be another simulation study done comparing all the designs, except for the pseudolikelihood method, with time-varying covariates.

Bibliography

P. K. Andersen and R. D. Gill. Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10:1100–1120, 1982.

P. Billingsley. *Probability and Measure.* Third edition. John Wiley Sons, Inc., United States of America, 1986. ISBN 0-471-00710-2.

Ø. Borgan and B. Langholz. Counter-matching: A stratified nested case-control sampling method. *Biometrics*, 82:69–79, 1995.

Ø. Borgan, L. Goldstein, and B. Langholz. Methods for the analysis of sampled cohort data in the cox proportional hazards model. *The Annals of Statistics*, 23:1749–1778, 1995.

D. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34:187–220, 1972.

S. V. Cunliffe, F. D. K. Lidell, J. C. McDonald, and D. C. Thomas. Addendum to methods of cohort analysis: Appraisal by application to asbestos mining. *Journal of the Royal Statistical Society, Series A*, 140:469–491, 1977.

T. R. Fleming and D. P. Harrington. *Counting Processes and Survival Analysis.* John Wiley & Sons, New Jersey, 1991.

L. Goldstein and B. Langholz. Asymptotic theory for nested case-control sampling in the cox regression model. *The Annals of Statistics*, 20:1903–1928, 1992.

B. Langholz and D. Clayton. Sampling strategies in nested case-control studies. *Environmental Health Perspectives*, 102:47 – 51, 1994.

S. O. Samuelsen. A pseudolikelihood approach to analysis of nested case-control studies. *Biometrics*, 84:379–394, 1997.

T. M. Therneau. *A Package for Survival Analysis in R*, 2022. URL `https://CRAN.R-project.org/package=survival`. R package version 3.5-0.

H. Wickham, R. François, L. Henry, and K. Müller. *A Grammar of Data Manipulation*, 2022. URL `https://CRAN.R-project.org/package=dplyr`. R package version 3.4-0.

List of Figures

List of Tables