INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Additive Hazards Regression with Incomplete Covariate Data

by

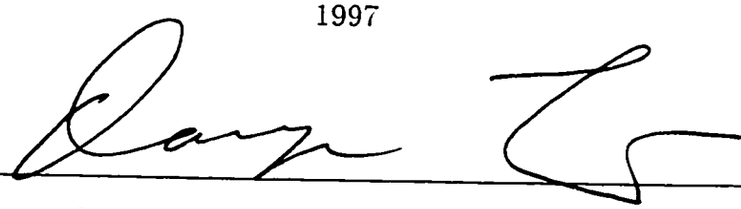
Michal Kulich

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington

1997

Approved by 

(Chairperson of Supervisory Committee)

Program Authorized
to Offer Degree Biostatistics

Date December 4, 1997

In presenting this dissertation in partial fulfillment of the requirements for the Doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to University Microfilms, 300 North Zeeb Road, P.O. Box 1346, Ann Arbor, Michigan 48106-1346, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature Michael Fulich

Date Dec. 4., 1997

University of Washington

Abstract

Additive Hazards Regression with Incomplete Covariate Data

by Michal Kulich

Chairperson of Supervisory Committee: *Professor Danyu Lin*

Department of Biostatistics

This dissertation addresses two incomplete covariate data problems in the additive hazards (AH) regression model for failure time data. Both are examples of two-phase designs where some covariate is measured only on a subset of the total sample. The first is the case-cohort design, where the covariates are ascertained on all the failures and on a randomly selected subcohort. We propose an estimator for the AH regression parameter under the case-cohort design, prove its consistency and asymptotic normality and discuss its practical use. The other case we consider is the errors-in-variables design, where the true covariate is observed only on a validation set selected by independent Bernoulli sampling, but a surrogate covariate is available for all subjects. Under these circumstances, we define the corrected score (CS) estimator for the AH regression parameter and show that it is consistent and asymptotically normal. Our approach works under mild regularity conditions, does not impose any parametric distributional assumptions on the covariates and allows for surrogates that are biased for the true covariate of interest. The CS estimator is easily generalized to multiple covariates. We describe its application in several situations involving both discrete and continuous covariates, investigate its behavior by simulation studies and illustrate its use on a real-life example.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I am deeply grateful to Prof. Danyu Lin for introducing me to this research area, directing me and helping me to proceed towards the end. He was always available and willing to give his advice whenever I needed. It was a pleasure for me to work with him. I would also like to thank the other members of my supervisory committee: Prof. Jon Wellner and Prof. Bill Barlow, who also served in the reading committee and contributed many helpful ideas; Prof. Norm Breslow, whose unceasing interest in my work has been a great encouragement for me; Prof. David Yanez, who was always willing to help; and Prof. Michelle Williams, my Graduate School Representative.

I would also like to express my gratitude to the National Wilms Tumor Study Group and Norm Breslow, in particular, for giving me access to their data, which served as an elucidating example in several chapters of the dissertation.

During the years I spent in Seattle I enjoyed the company of fellow students who made the occassional hard times easier to survive: My thanks to Martha, Juanjuan, Jinko, Brad, Rich, Sharon, and Jim. But most important for me was having Monika, my wife, at my side. As I was finishing this dissertation, Monika gave birth to our first-born son Damián, a wonderful gift that will change our lives for ever.

I do not want to forget those who made it possible for me to get this far (in both abstract and literal senses of the word): I am grateful to my parents; to those who taught me statistics in Prague, especially Prof. Marie Hušková and Prof. Jiří Anděl; and to those who convinced me that biostatistics is the most exciting science in the world, especially Prof. Herman Callaert, Prof. Alan Agresti, and Prof. Thomas Fleming.

DEDICATION

Monice, Damiánovi a Myšklibkám.

Chapter 1

INTRODUCTION

The publication of D. R. Cox's paper on proportional hazards regression (Cox, 1972) represented a real breakthrough in the statistical analysis of censored failure time data. Cox's paper introduced a convenient method for estimating the association of a vector of covariates with time to failure. The Cox model works under rather general conditions, is easy to apply, and estimates parameters that are easily interpretable as log relative risks. It has become very popular in biomedical research: today it is considered the standard method for the statistical analysis of both randomized and observational studies with failure time endpoints in medical research and epidemiology. Since 1972, several alternative regression models for failure time data have appeared, among them the additive hazards model (Breslow and Day, 1987, p. 182; Cox and Oakes, 1984, p. 74; Lin and Ying, 1994). Although the Cox model remains the most widely used failure time regression model in biostatistics, the alternative models possess some appealing characteristics that make them potentially useful in many situations.

The availability of regression models for censored failure time data has had a tremendous impact on biomedical research. It enabled investigators to extract much more information from a study with a failure time endpoint than it was possible with the statistical methods used prior to 1972. It became feasible to conduct studies addressing more delicate scientific questions. However, many important questions in medical research and epidemiology remained hard to answer because they required

conducting too large and expensive studies. Usually, the number of subjects to be included in a biomedical study is chosen so that a sufficient power is achieved against a prespecified clinically meaningful alternative. When the failure rate is low or when the alternative is hard to distinguish from the hypothesis, the sample size skyrockets along with the study budget. The cost of conducting the study particularly soars if the main financial burden does not lie in recruiting and following the subjects, but in the necessity to take one or several fairly expensive measurements on each enrolled subject. Examples of such a costly covariate are a complicated laboratory test or a thorough and reliable dietary assessment. The expensive covariate may be a confounder, which must be accounted for lest the main covariate's effect is distorted, or, in observational studies, it may be the main covariate of interest itself.

If such a problem arises in practice, the investigators sometimes measure the expensive covariate only on a small subset of the study subjects. Experimental designs based on this principle are called *two-phase designs*. At the first phase, study subjects are selected from the population of interest; at the second phase, a random sample from the first-phase subjects is chosen in a certain way and the expensive covariate is measured on that sample. Only the second-phase subjects thus have complete covariate information; the extent of missingness in the first-phase subjects may vary from no covariates observed at all to all observed but one. The two-phase design can also be viewed as a missing data problem where covariate data are missing by design. The fact that the missingness mechanism is completely known greatly facilitates statistical analysis.

The cost-effectiveness of a two-phase design depends on the costs of enrollment and follow-up relative to the cost of covariate assessment. A two-phase design does not decrease the total number of study subjects. On the contrary, to match the power of a simple one-phase study, a two-phase study must enroll extra subjects. However, if covariate assessment is much more expensive than enrollment and follow-up, a two-phase study can achieve substantial savings by reducing the number of costly

covariate measurements. In many cases, the savings will be only slightly offset by the increase in sample size.

The sampling schemes used for the selection of the second-phase subjects and the statistical methods used for the analysis of a two-phase study may differ case by case. However, they should always minimize the loss of efficiency due to the incompleteness of the data. An example of a clearly inefficient approach is the complete-case analysis based on the second-phase subjects selected by simple random sampling. Better results could be achieved by employing a more efficient second-phase sampling method or by exploiting the information provided by the study subjects not selected into the second-phase sample. For example, there often exists an imperfect surrogate for the costly covariate that can be easily and cheaply measured on all subjects. The information contained in the surrogate covariate may be used to improve the precision of the whole analysis.

In this dissertation, two special cases of two-phase designs in survival framework are considered: the case-cohort design and the errors-in-variables design. In *the case-cohort design*, the second-phase sample consists of all the subjects that are observed to fail during the study plus *the subcohort*, a random sample from the whole first-phase population. The covariates of the subjects not selected into the second phase sample are considered unknown. The subcohort may be selected by simple random sampling or a surrogate may be used to increase efficiency by altering the subcohort sampling probabilities in a suitable way. The analysis of the case-cohort design uses only the second-phase data.

The second design to be discussed here is *the errors-in-variables* design. It is applicable whenever there exists a cheap surrogate covariate, which is related to the expensive true covariate of interest. The surrogate covariate can be regarded as the true covariate measured with error: hence errors-in-variables. It is assumed that the surrogate is available for all the first-phase subjects and the true covariate is measured on a second-phase sample obtained by simple random sampling. The

second-phase sample is called *the validation set.* The validation set is used to estimate the association between the surrogate and the true covariate. That facilitates the use of the nonvalidation data, where the true covariate itself is unobserved, for the estimation of regression parameters.

The errors-in-variables design belongs to the wide class of covariate measurement error problems. There is a vast literature on covariate measurement errors but the problem is rarely introduced in the context of two-phase designs because most methods for dealing with measurement error do not require the existence of a validation set. Instead, it is often assumed that there exist multiple independent observations of the surrogate on each subject, or that the association between the surrogate and the true covariate (error variance, for example) is known from an extraneous source. However, the existence of a validation set makes relying on arbitrary extraneous information unnecessary and allows for the use of surrogates that are not unbiased for the true covariate. That is why the definition of the errors-in-variables design explicitely puts the covariate measurement error problem into the two-phase design context.

This dissertation considers the case-cohort and errors-in-variables designs in the setting of the additive hazards (AH) regression model. Like the Cox model, the AH model is a semiparametric regression model for censored failure time data. The regression parameters in the AH model can be interpreted as risk differences, or expected excess events per unit of time. Since medical researchers, especially epidemiologists, are sometimes primarily interested in estimating and interpreting risk differences rather than relative risks, the additive hazards model represents an important tool for the analysis of many such studies. It can be also useful when the Cox model does not fit the data well, and it may provide a revealing supplementary insight into the data even if the Cox model remains the preferred method of analysis. It is therefore important that statistical methods for analyzing data with the AH model should be developed.

This dissertation helps to expand the area where the additive hazards model can

be applied by demonstrating that the case-cohort estimation under the AH model is as easy and convenient as under the Cox model and that, unlike for the Cox model, there exists a simple and relatively general estimating method for the errors-in-variables design under the AH model. The main goal of the dissertation is to develop consistent estimators of the additive hazards regression parameters under the two specific two-phase designs, to derive their asymptotic distributions, and to investigate their behavior in finite samples.

In Chapter 2, the known results about the Cox model and the additive hazards model with complete data are briefly summarized and basic notation is introduced. Chapter 3 reviews the current methods for estimating the Cox model parameters under the case-cohort design and provides a short summary of previous work on Cox model estimation when covariates are subject to measurement error. The case-cohort design for the AH model is the topic of Chapter 4. There it is explained how the case-cohort estimator is derived, what its asymptotic distribution is and how the asymptotic variance can be estimated. It is discussed how to select the sub-cohort to increase efficiency and the AH case-cohort estimator is compared to the Cox case-cohort estimator in terms of asymptotic relative efficiency with respect to the full-data estimators. The small-sample behavior of the AH case-cohort estimator is investigated through simulation studies. An analysis of a real-life data set concludes Chapter 4. Chapter 5 is devoted to the errors-in-variables design under the AH model. It begins with a set of conditions imposed on the surrogate covariate and defines an estimator assuming that the model includes only the true covariate and that the association between the true and surrogate covariates is known. The asymptotic distribution of the proposed estimator is derived and a consistent estimator of the asymptotic variance is introduced. The initial assumptions are relaxed in the subsequent sections of the chapter. Examples of applications in certain special cases are followed by the results of simulation studies. An analysis of a data set is also included. Chapter 6 summarizes the results obtained previously, discusses the

usefulness of the proposed methods, and indicates possible directions for further research on the topic. The Appendix explains the mathematical notation that we use and includes an overview of important symbols that appear in the dissertation.

Before we proceed to the next chapter, let us illustrate the practical use of two-phase designs on a real-life example.

Example 1.1 (Wilms Tumor Study). Since 1969, the National Wilms Tumor Study Group (NWTSG) has been conducting a series of randomized clinical trials to investigate the natural history of Wilms tumor and to identify the best treatment regimens for different subgroups of patients. Wilms tumor is a rare kidney cancer occurring only in children, with a relatively good prognosis when treated with combination chemotherapy. The largest studies conducted by NWTSG so far have been NWTSG-3 (closed for enrollment in 1986) and NWTSG-4 (closed for enrollment in 1993). The results of NWTSG-3 were previously published by D'Angio *et al.* (1989), Breslow *et al.* (1991), and by Green *et al.* (1994b); a subanalysis of a part of the NWTSG-4 data appeared in Green *et al.* (1994a).

The most important prognostic factors for death or recurrence in Wilms tumor patients include histologic type of the tumor, which can be roughly classified into favorable (FH) and unfavorable (UH), and stage, which can be regarded as a measure of tumor spread. Stage I corresponds to localized disease, stages II and III to regional disease, and stage IV to metastatic disease. See Breslow *et al.* (1991) for a formal definition of stages I–III. In NWTSG studies, histologic type was first evaluated by a local pathologist in the institution that treated a particular patient. Tumor tissue samples were then sent to the NWTS Pathology Center where histologic type was reevaluated by an experienced pathologist. By the protocol, the reevaluation was done for all patients for which sufficient information had been submitted. The institutional and central evaluations of histology agreed in most, but not all, cases. For obvious reasons, central histology can be regarded as much more precise and reliable than

institutional histology.

Since histologic type and tumor stage are very strong and independently acting risk factors, any statistical analysis of time to death or relapse should adjust for both. However, this requires reexamining all the samples for histologic type in the central laboratory, as was done in NWTSG studies. This is expensive and sometimes unfeasible, either because of the cost of the laboratory analyses or because of limited facilities or a lack of experienced personnel. Such financial or other restrictions may seriously limit the size of the study. If such a problem arises, a two-phase design may be the solution, provided that statistical methods for its analysis are available.

Suppose that a two-phase design was to be applied to a NWTSG study. For all the patients, the tissue samples would be collected and histologic type would be assessed by the local pathologist. The patients would be randomized to treatment arms as before. In addition, they would be also randomly divided into two groups: the members of the second-phase sample and the others. Only the tissue samples of the second-phase patients would be submitted to the central laboratory for a reevaluation.

For example, in the case-cohort design, the second-phase sample would consist of all the patients that experienced a failure (death or relapse) plus a random sample of a certain proportion of the first-phase subjects. Because it is not known which patient experienced a failure until the end of the study, the tissue samples of all the subjects would have to be frozen for a later analysis. So the case-cohort design can be used only if such a procedure is practically feasible.

If the errors-in-variables design is to be applied, it is natural to treat central histology as the true covariate and institutional histology as the surrogate. All the first-phase subject have the surrogate covariate observed. At the second phase, a random validation set is selected, and the tissue samples of the validation set subjects are submitted to the central laboratory. Since the validation subjects have both central and institutional measurements available, it is easy to estimate the probability of unfavorable central histology conditionally on a given value of institutional

histology. The knowledge of this association can then be used to recover the relevant information from the nonvalidation subjects.

Both case-cohort and errors-in-variables designs would substantially reduce the number of laboratory analyses performed by the Central Pathology Lab. The cost of conducting the NWTSG studies would thus be lowered. However, simple and efficient statistical methods to analyze the data collected in a two-phase study need to be developed first.

Chapter 2

FAILURE TIME REGRESSION

2.1 *Failure time data, censoring, basic notation*

Suppose that a subject is followed in time until a failure occurs. The time to failure T can be regarded as a non-negative random variable with an unknown distribution function F. We are interested in certain features of the failure time distribution F. If independent replicates of T are available, estimation of F does not pose much difficulty. Yet because the subjects cannot be always followed until all experience a failure, failure time data are often censored by the time the observation is terminated and T may not be observed directly. So let the censoring time C be the time when the follow-up ends if failure does not occur earlier. The observed data then consist of the censored failure time $X = \min(T, C)$ and the failure indicator $\Delta = \mathbb{1}(T \leq C)$. Here, $\mathbb{1}(\cdot)$ is the indicator function, which is equal to 1 if the event in the parenthesis occurs and to 0 otherwise. If $\Delta = 1$, we have $X = T$ and so T is observed exactly. If $\Delta = 0$, we only know that $T > X$.

The distribution of T can be described by the distribution function F or by the density $f = F'$, if it exists. However, in the presence of censoring it is more convenient to describe the distribution of T in a different way. Let

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

The function $\lambda(t)$ is called the hazard function. It is the instantaneous probability of failure at time t given that the failure has not occured prior to time t. If F is

continuous and the density f exists, the hazard function satisfies

$$\lambda(t) = \frac{f(t)}{1 - F(t)} = -\frac{d \ln[1 - F(t)]}{dt}.$$

In that case, $F(t) = 1 - \exp\{-\Lambda(t)\}$, where $\Lambda(t) \equiv \int_0^t \lambda(s)\, ds$ is the cumulative hazard function. Thus, the hazard function uniquely specifies the distribution of T and the same is true even if the distribution F is not continuous.

Sometimes it is not the distribution of T itself that we are most interested in, but the effect of certain factors on that distribution. So let $\boldsymbol{Z}$ be a p-vector of covariates and suppose that the main question is: How does $\boldsymbol{Z}$ affect the failure time distribution? As indicated earlier, one might as well ask: How does $\boldsymbol{Z}$ affect the hazard function? An answer to this question could be provided by a model selected from the general class of regression models defined by

$$\lambda(t|\boldsymbol{Z}) = \Phi(\lambda_0(t), \boldsymbol{\beta}_0^{\mathrm{T}} \boldsymbol{Z}), \tag{2.1}$$

where

$$\lambda(t|\boldsymbol{Z}) = \lim_{h \to 0+} h^{-1} \mathrm{P}\left[t \leq T < t + h \mid T \geq t, \boldsymbol{Z}\right].$$

The left hand side of (2.1) is the hazard function conditional on $\boldsymbol{Z}$; it is the hazard for a subject with covariate vector $\boldsymbol{Z}$. The right hand side specifies how the covariates affect the hazard. The p-vector $\boldsymbol{\beta}_0$ contains the parameters that descibe the association between $\boldsymbol{Z}$ and $\lambda(t|\boldsymbol{Z})$ and $\lambda_0(t)$ is the baseline hazard—the hazard for a subject with $\boldsymbol{Z} = \boldsymbol{0}$. The function Φ links the two parts together. To make the interpretation of λ_0 as the baseline hazard correct, Φ must satisfy $\Phi(x, 0) = x$. Evaluating the effect of $\boldsymbol{Z}$ on T is now equivalent to estimating the parameter vector $\boldsymbol{\beta}_0$.

When the baseline hazard λ_0 is known up to a finite number of parameters, we obtain a parametric regression model and the parameter vector $\boldsymbol{\beta}_0$ can be estimated by maximum likelihood. When λ_0 is left completely unspecified, we arrive at a semiparametric model. Here are two examples:

Example 2.1 (Cox model). Setting $\Phi(x, y) = x\,\mathrm{e}^{y}$ in (2.1) and leaving λ_0 unspecified, we get

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t)\exp\{\boldsymbol{\beta}_0^{\mathsf{T}}\boldsymbol{Z}\}.$$

This is the Cox proportional hazards (PH) model with exponential link. The basic assumption behind the PH model is that the hazard ratio $\lambda(t|\boldsymbol{z}_1)/\lambda(t|\boldsymbol{z}_2)$ is constant over time. Each parameter under the Cox model can be interpreted as the log relative risk per a unit change in the covariate.

Example 2.2 (Additive hazards model). Another semiparametric model is defined by $\Phi(x, y) = x + y$, that is

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t) + \boldsymbol{\beta}_0^{\mathsf{T}}\boldsymbol{Z}.$$

This equation defines the additive hazards (AH) model. The AH model assumes that the difference in the hazards for two given values of $\boldsymbol{Z}$ is constant over time. The AH model parameter is interpretable as the risk difference per a unit change in the corresponding covariate. That is, the AH parameter estimates the difference (between two subjects with a unit difference in the covariate) in expected numbers of failures occuring within a unit of time.

Before we proceed to discuss the Cox model and the additive hazards model in more detail, we introduce some convenient notation and a couple of assumptions. Let $N(t) = \mathbb{1}(X \le t, \Delta = 1)$ be the 0-1 process counting the number of failures observed prior to or at time t. Let $Y(t) = \mathbb{1}(X \ge t)$ be the 0-1 process indicating whether or not the subject is at risk for failure at time t. Notice that the pair (X, Δ) is equivalent to $(N(t), Y(t); t \ge 0)$. Denote by τ, $0 < \tau < \infty$, the time when observation ends. Write $\Lambda_0(t) = \int_0^t \lambda_0(s)\,ds$ for the cumulative baseline hazard. In addition, we allow the covariate vector $\boldsymbol{Z}$ to be time-dependent. More precisely, we assume that $\boldsymbol{Z}(\cdot)$ is a left-continuous process with right-hand limits. Finally, we assume throughout

this thesis that censoring time is conditionally independent of failure time, given the covariate process.

2.2 The Cox model

Suppose that $(X_i, \Delta_i, \boldsymbol{Z}_i(t), 0 \leq t \leq X_i)$, $i = 1, \ldots, n$, are independent replicates of $(X, \Delta, \boldsymbol{Z}(t))$. The subjects are indexed by i and n is the total sample size. Let the hazard for the failure time T follow the model

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t) \exp\{\boldsymbol{\beta}_0^{\mathsf{T}} \boldsymbol{Z}(t)\},$$

where $\lambda_0(\cdot)$ is an unknown and unspecified baseline hazard function. Cox (1972, 1975) suggested that the p-vector of parameters $\boldsymbol{\beta}_0$ be estimated by maximizing the partial likelihood

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{0 \leq s \leq \tau} \left[\frac{Y_i(s) \exp\{\boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{Z}_i(s)\}}{\sum_{j=1}^{n} Y_j(s) \exp\{\boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{Z}_j(s)\}} \right]^{\Delta N_i(s)},$$

where $\Delta N_i(s) = N_i(s) - N_i(s-)$. Let us define

$$S^{(0)}(\boldsymbol{\beta}, t) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t) \exp\{\boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{Z}_i(t)\},$$

$$\boldsymbol{S}^{(1)}(\boldsymbol{\beta}, t) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Z}_i(t) Y_i(t) \exp\{\boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{Z}_i(t)\},$$

and

$$\mathbb{S}^{(2)}(\boldsymbol{\beta}, t) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Z}_i^{\otimes 2}(t) Y_i(t) \exp\{\boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{Z}_i(t)\},$$

where $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^{\mathsf{T}}$, and set

$$\overline{\boldsymbol{Z}}(\boldsymbol{\beta}, t) = \frac{\boldsymbol{S}^{(1)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)}.$$

Then the partial likelihood score vector $U_{PH}(\boldsymbol{\beta}) \equiv \partial \ln L(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$ can be written as

$$U_{PH}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\tau} \left[\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}(\boldsymbol{\beta}, t) \right] dN_i(t).$$

The maximum partial likelihood estimator $\widehat{\boldsymbol{\beta}}_{PH}$ is the solution of $\boldsymbol{U}_{PH}(\boldsymbol{\beta}) = 0$. It can be found by a Newton-Raphson iterative procedure. Notice that only the subjects whose failures are observed contribute a term to the PL score, while censored subjects' data affect $\overline{\boldsymbol{Z}}(\boldsymbol{\beta}, t)$ only. Breslow (1972) proposed a cumulative baseline hazard estimator for the Cox model, which takes the form

$$\widehat{\Lambda}_0(t) = \int_0^t \frac{\sum_{i=1}^n dN_i(s)}{\sum_{i=1}^n Y_i(s) \exp\{\widehat{\boldsymbol{\beta}}_{PH}{}^{\mathsf{T}} \boldsymbol{Z}_i(s)\}}.$$

The theory of counting process martingales is extremely convenient for investigating the properties of estimators and score functions in failure time regression. Andersen and Gill (1982) used martingale theory to develop a rigorous asymptotic theory for the Cox partial likelihood score and the corresponding estimator. For a thorough description of the use of counting process martingales in censored data, see Fleming and Harrington (1991) or Andersen *et al.* (1993). For the Cox model, the martingale theory asserts that

$$M_i(t) \equiv N_i(t) - \int_0^t Y_i(s) \exp\{\boldsymbol{\beta}_0^{\mathsf{T}} \boldsymbol{Z}_i(s)\} \, d\Lambda_0(s)$$

is a martingale with respect to the filtration $\mathcal{F}_t = \sigma\{N_i(s), Y_i(s), \boldsymbol{Z}_i(s), 0 \le s \le t, i = 1, \ldots, n\}$. It can be shown (see Fleming and Harrington, 1991, Chapter 4) that

$$U_{PH}(\boldsymbol{\beta}_0) = \sum_{i=1}^{n} \int_0^{\tau} \left[\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}(\boldsymbol{\beta}_0, t) \right] dM_i(t)$$

(which implies that $\mathrm{E}\, U_{PH}(\boldsymbol{\beta}_0) = 0$) and

$$\begin{aligned} \Sigma_{PH}(\boldsymbol{\beta}_0) \equiv \mathrm{E}\left[\frac{1}{n} U_{PH}^{\otimes 2}(\boldsymbol{\beta}_0) \right] &= -\frac{1}{n} \mathrm{E}\, \frac{\partial}{\partial \boldsymbol{\beta}^{\mathsf{T}}} U_{PH}(\boldsymbol{\beta}_0) \\ &= \int_0^{\tau} \mathrm{E}\left[\frac{\mathbb{S}^{(2)}(\boldsymbol{\beta}_0, t)}{S^{(0)}(\boldsymbol{\beta}_0, t)} - \overline{\boldsymbol{Z}}^{\otimes 2}(\boldsymbol{\beta}_0, t) \right] S^{(0)}(\boldsymbol{\beta}_0, t) \, d\Lambda_0(t). \end{aligned}$$

Andersen and Gill (1982) introduced a set of regularity conditions for the consistency and asymptotic normality of the Cox estimator. Under their conditions, they proved that

$$\widehat{\boldsymbol{\beta}}_{PH} \to_p \boldsymbol{\beta}_0,$$
$$n^{-1/2} U_{PH}(\boldsymbol{\beta}_0) \to_d \mathrm{N}_p(\mathbf{0}, \Sigma_{PH}(\boldsymbol{\beta}_0)), \quad \text{and}$$
$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{PH} - \boldsymbol{\beta}_0) \to_d \mathrm{N}_p(\mathbf{0}, \Sigma_{PH}^{-1}(\boldsymbol{\beta}_0)).$$

Andersen and Gill also showed that, when $(X_i, \Delta_i, \boldsymbol{Z}_i)$ are independent and identically distributed replicates of $(X, \Delta, \boldsymbol{Z})$, the regularity conditions reduce to

$$\Lambda_0(\tau) < \infty, \tag{2.2}$$
$$\mathrm{P}[Y(\tau) = 1] > 0, \tag{2.3}$$
$$\Sigma_{PH}(\boldsymbol{\beta}_0) > 0, \tag{2.4}$$
$$\mathrm{E} \sup_{\substack{\boldsymbol{\beta} \in \mathcal{B} \\ 0 \le t \le \tau}} \left[Y(t) |\boldsymbol{Z}(t)| \exp\{\boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{Z}(t)\} \right] < \infty, \tag{2.5}$$

where $\mathcal{B}$ is some neighborhood of $\boldsymbol{\beta}_0$.

We conclude the review of the Cox model by noting that the Cox partial likelihood estimator achieves semiparametric efficiency. Proofs of this fact can be found in Begun *et al.*(1983) or in Klaassen (1989).

2.3 *The additive hazards model*

Under the additive hazards (AH) model, the conditional hazard function for the failure time T given the covariate vector $\boldsymbol{Z}$ takes the form

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t) + \boldsymbol{\beta}_0^{\mathsf{T}} \boldsymbol{Z}(t), \tag{2.6}$$

where $\lambda_0(\cdot)$ is an unknown and unspecified baseline hazard function. The additive hazards model was proposed by many authors, among them Aalen (1980), Cox and Oakes (1984, p. 74) and Breslow and Day (1987, p. 182). However, semiparametric

methods for estimation in the AH model were not available until Lin and Ying (1994). Motivated by the form of the Cox partial likelihood score, these authors developed a pseudoscore function for the AH model which defines a consistent and asymptotically normal estimator.

To explain Lin and Ying's additive hazards estimator, we start by introducing the at-risk covariate average

$$\overline{\boldsymbol{Z}}(t) \equiv \frac{\sum_{i=1}^{n} \boldsymbol{Z}_i(t) Y_i(t)}{\sum_{i=1}^{n} Y_i(t)}.$$

Notice that $\overline{\boldsymbol{Z}}(t)$ is just the average of the covariates for the subjects at risk at time t. Lin and Ying defined the AH pseudoscore as

$$U_A(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\tau} [\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}_i(t)] \, [dN_i(t) - \boldsymbol{Z}_i(t)^{\mathsf{T}} \boldsymbol{\beta} Y_i(t) \, dt]. \tag{2.7}$$

Since

$$\sum_{i=1}^{n} [\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}(t)] Y_i(t) = \mathbf{0}, \tag{2.8}$$

it is easy to see that the pseudoscore satisfies

$$U_A(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\tau} [\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}(t)] \, dN_i(t) - \sum_{i=1}^{n} \int_0^{\tau} [\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}(t)]^{\otimes 2} \boldsymbol{\beta} Y_i(t) \, dt, \tag{2.9}$$

and

$$U_A(\boldsymbol{\beta}_0) = \sum_{i=1}^{n} \int_0^{\tau} [\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}(t)] \, [dN_i(t) - Y_i(t) \, d\Lambda_0(t) - \boldsymbol{Z}_i(t)^{\mathsf{T}} \boldsymbol{\beta}_0 Y_i(t) \, dt]. \tag{2.10}$$

Equation (2.9) implies that the AH estimator $\widehat{\boldsymbol{\beta}}_A$ defined by $U_A(\widehat{\boldsymbol{\beta}}_A) = \mathbf{0}$ takes on the explicit form

$$\widehat{\boldsymbol{\beta}}_A = \left\{ \sum_{i=1}^{n} \int_0^{\tau} [\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}(t)]^{\otimes 2} Y_i(t) \, dt \right\}^{-1} \left\{ \sum_{i=1}^{n} \int_0^{\tau} [\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}(t)] \, dN_i(t) \right\}.$$

The counting process martingale for $N_i(t)$ in the AH model is given by

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) \, d\Lambda_0(s) - \int_0^t \boldsymbol{Z}_i(s)^{\mathsf{T}} \boldsymbol{\beta}_0 Y_i(s) \, ds.$$

It follows from (2.10) that $U_A(\boldsymbol{\beta}_0) = \sum_{i=1}^n \int_0^\tau [\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}(t)]\, dM_i(t)$ and thus $U_A(\boldsymbol{\beta}_0)$ is a martingale integral.

Lin and Ying (1994) argued that the counting process arguments of Andersen and Gill (1982) can be used to show that $n^{-1/2}U_A(\boldsymbol{\beta}_0)$ is asymptotically normal with mean zero and covariance matrix which can be consistently estimated by

$$\widehat{\Sigma}_A \equiv \frac{1}{n}\sum_{i=1}^n \int_0^\tau [\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}(t)]^{\otimes 2} dN_i(t), \tag{2.11}$$

and that $\sqrt{n}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0)$ converges weakly to p-variate normal distribution with zero mean and covariance matrix that can be consistently estimated by $\widehat{\mathbb{D}}_A^{-1}\widehat{\Sigma}_A\widehat{\mathbb{D}}_A^{-1}$, where

$$\widehat{\mathbb{D}}_A \equiv \frac{1}{n}\sum_{i=1}^n \int_0^\tau [\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}(t)]^{\otimes 2} Y_i(t)\, dt$$

is the negative partial derivative of the pseudoscore. At the end of this chapter, we prove these assertions in the special case of independent and identically distributed data.

The cumulative baseline hazard in the AH model can be consistently estimated by

$$\widehat{\Lambda}_0(t) \equiv \int_0^t \frac{\sum_{i=1}^n dN_i(s)}{\sum_{j=1}^n Y_j(s)} - \int_0^t \overline{\boldsymbol{Z}}(s)^\mathsf{T}\widehat{\boldsymbol{\beta}}_A\, ds, \tag{2.12}$$

an estimator proposed by Lin and Ying and motivated by the Breslow estimator for the Cox cumulative baseline hazard.

It is useful to note that there are several important distinctions between the additive hazards model and the Cox model. First, the form of the hazard equation (2.6) that defines the AH model imposes an implicit constraint on the covariates: it requires that $\boldsymbol{Z}(t)^\mathsf{T}\boldsymbol{\beta}_0 \geq -\lambda_0(t)$ for all t. Thus, the covariates are necessarily bounded from one side. Second, the cumulative baseline hazard estimator defined by (2.12) may be non-monotone. Lin and Ying suggested that, should such a case arise, the estimator be augmented by taking the supremum: $\widehat{\Lambda}^*(t) = \sup_{0\leq s\leq t}\widehat{\Lambda}_0(s)$. They also

considered the model $\lambda_0(t|\boldsymbol{Z}) = \lambda_0(t) + \exp\{\boldsymbol{\beta}_0^{\mathsf{T}}\boldsymbol{Z}(t)\}$. which does not suffer from this drawback; however. the parameters lose their convenient interpretation. Third. unlike the Cox partial likelihood score. the AH pseudoscore includes one term for each subject no matter if the subject fails or is censored. Finally, the estimator $\widehat{\boldsymbol{\beta}}_A$ is not semiparametric efficient. Lin and Ying demonstrated that it acheves full efficiency if and only if $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\lambda_0(t)$ is constant. However, the loss in efficiency should be small provided that $\boldsymbol{\beta}_0$ is small and the baseline hazard is approximately constant.

Asymptotic distribution of the AH estimator

Here we present short proofs of Lin and Ying's asymptotic results under the iid assumption. So, suppose that $(X_i, \Delta_i, \boldsymbol{Z}_i(t), 0 \le t \le X_i)$, $i = 1, \ldots, n$, are independent replicates of $(X, \Delta, \boldsymbol{Z}(t))$. Recall that the data are collected on the time interval $[0, \tau]$, $\tau < \infty$, so that $\mathrm{P}[X > \tau] = 0$. Define

$$\pi_k(t) \equiv \mathrm{E}\, \boldsymbol{Z}^{\otimes k}(t) Y(t), \qquad k = 0, 1, 2,$$

and denote $\boldsymbol{e}(t) = \boldsymbol{\pi}_1(t)/\pi_0(t)$. Since the data are iid, $\pi_k(t)$ is also the limit in probability of $n^{-1}\sum_{i=1}^n \boldsymbol{Z}_i^{\otimes k}(t) Y_i(t)$. Consider the following regularity conditions:

Condition 2.1. $\Lambda_0(\tau) < \infty$.

Condition 2.2. $\mathrm{P}[Y_i(\tau) = 1] > 0$.

Condition 2.3. $\mathrm{E} \sup_{0 \le t \le \tau} |Y(t)\boldsymbol{Z}^{\otimes 2}(t)\boldsymbol{\beta}_0^{\mathsf{T}}\boldsymbol{Z}(t)| < \infty$.

Condition 2.4. The matrix $\Sigma_A(\boldsymbol{\beta}_0) \equiv \mathrm{E}\int_0^\tau [\boldsymbol{Z}(t) - \boldsymbol{e}(t)]^{\otimes 2} dN(t)$ is positive definite.

Conditions 2.1–2.4 are similar to Andersen and Gill's (1982) regularity conditions for the Cox model with iid data (see Equations (2.2)–(2.5)). Conditions 2.1 and 2.2 constrain the data to a finite time interval, with probability of observing a failure being positive throughout the interval. They are probably not necessary; Andersen and Gill succeeded in relaxing them provided that the covariates were bounded. Condition 2.3 is a boundedness condition on the covariates. It assures that the matrix

Σ_A defined in Condition 2.4 exists. Condition 2.4 itself is necessary for consistency and asymptotic normality of the AH estimator.

By Condition 2.3, $\pi_k(t)$, $k = 0, 1, 2$, are uniformly bounded on $[0, \tau]$. By Condition 2.2, $\pi_0(t)$ is bounded away from zero on $[0, \tau]$. Taking these facts into account and using a result from Andersen and Gill, we obtain the following lemma:

Lemma 2.1. *Let the data be iid and let Conditions 2.2 and 2.3 hold. Then*

(i) $\sup_{0 \le t \le \tau} \left| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Z}_i^{\otimes k}(t) Y_i(t) - \pi_k(t) \right| \to_p 0$, $k = 0, 1, 2$;

(ii) $\sup_{0 \le t \le \tau} |\overline{\boldsymbol{Z}}(t) - \boldsymbol{e}(t)| \to_p 0$.

Proof. (i) Andersen and Gill used the strong law of large numbers for separable Banach spaces to show a similar result for the Cox model (Andersen and Gill, 1982, Corollary III.2). The sufficient condition for the Cox model is (2.5), which is analogous to Condition 2.3. The proof given by Anderesen and Gill, though a little more general than needed here, applies without a change.

(ii) We have

$$\overline{\boldsymbol{Z}}(t) - \boldsymbol{e}(t) = \frac{n^{-1} \sum \boldsymbol{Z}_i(t) Y_i(t)}{n^{-1} \sum Y_i(t)} - \frac{\boldsymbol{\pi}_1(t)}{\pi_0(t)}$$
$$= \frac{1}{\pi_0(t)} \left[\frac{1}{n} \sum \boldsymbol{Z}_i(t) Y_i(t) - \boldsymbol{\pi}_1(t) \right] + \frac{1}{n} \sum \boldsymbol{Z}_i(t) Y_i(t) \left[\left(\tfrac{1}{n} \sum Y_i(t) \right)^{-1} - \pi_0^{-1}(t) \right].$$

This is bounded in absolute value by

$$\frac{1}{\pi_0(t)} \left| \frac{1}{n} \sum \boldsymbol{Z}_i(t) Y_i(t) - \boldsymbol{\pi}_1(t) \right| + \left| \frac{1}{n} \sum \boldsymbol{Z}_i(t) Y_i(t) \right| \left| \left(\tfrac{1}{n} \sum Y_i(t) \right)^{-1} - \pi_0^{-1}(t) \right|.$$

The supremum of the the whole expression above converges in probability to zero by Part (i) of this lemma and by Condition 2.2. In fact, Condition 2.2 implies not only that $\pi_0(t) > 0$ for $0 \le t \le \tau$ but also $\mathrm{P}[\sum Y_i(t) = 0] \to_p 0$ as $n \to \infty$ for any $0 \le t \le \tau$. This finishes the proof. □

We proceed to prove asymptotic normality of the normalized AH pseudoscore evaluated at the true parameter $\boldsymbol{\beta}_0$.

Lemma 2.2. *Let Conditions 2.1-2.4 hold. Then*

$$\frac{1}{\sqrt{n}} U_A(\boldsymbol{\beta}_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{\psi}_i^{(A)}(\boldsymbol{\beta}_0) + o_P(1), \tag{2.13}$$

where

$$\widetilde{\psi}_i^{(A)}(\boldsymbol{\beta}_0) \equiv \int_0^\tau [\boldsymbol{Z}_i(t) - \boldsymbol{e}(t)]\, dM_i(t)$$

are iid random vectors. Consequently,

$$\frac{1}{\sqrt{n}} U_A(\boldsymbol{\beta}_0) \to_d \mathrm{N}_p(\mathbf{0}, \Sigma_A(\boldsymbol{\beta}_0)),$$

where

$$\Sigma_A(\boldsymbol{\beta}_0) \equiv \operatorname{var} \widetilde{\psi}_i^{(A)}(\boldsymbol{\beta}_0) = \mathrm{E} \int_0^\tau [\boldsymbol{Z}_i(t) - \boldsymbol{e}(t)]^{\otimes 2} dN_i(t).$$

Proof. To prove (2.13), it suffices to show

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^\tau [\overline{\boldsymbol{Z}}(t) - \boldsymbol{e}(t)]\, dM_i(t) \to_p 0.$$

The left hand side is a martingale integral with predictable variance function

$$\int_0^\tau [\overline{\boldsymbol{Z}}(t) - \boldsymbol{e}(t)]^{\otimes 2} \frac{1}{n} \sum_{i=1}^{n} dN_i(t).$$

This is bounded in absolute value by

$$\sup_{0 \le t \le \tau} |\overline{\boldsymbol{Z}}(t) - \boldsymbol{e}(t)|^{\otimes 2} \frac{1}{n} \sum_{i=1}^{n} \Delta_i,$$

where the supremum converges to zero in probability by Lemma 2.1(ii). Since $n^{-1} \sum \Delta_i$ converges in probability to a constant, (2.13) is proven.

It follows that $n^{-1/2} U_A(\boldsymbol{\beta}_0)$ has the same limiting distribution as $n^{-1/2} \sum \widetilde{\psi}_i^{(A)}(\boldsymbol{\beta}_0)$. But $\widetilde{\psi}_i^{(A)}(\boldsymbol{\beta}_0)$, $i = 1, \ldots, n$, are iid random vectors with zero mean (they are martingale integrals!) and finite positive definite covariance matrix $\Sigma_A(\boldsymbol{\beta}_0)$. Hence their normalized sum must converge in law to the desired normal distribution. □

It remains to establish consistency and asymptotic normality of the AH estimator. This is done in the last lemma.

Lemma 2.3. *Under Conditions 2.1–2.4, $\widehat{\boldsymbol{\beta}}_A$ is consistent and*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0) \to_d \mathrm{N}_p(\mathbf{0}, \mathbb{D}_A^{-1}\Sigma_A(\boldsymbol{\beta}_0)\mathbb{D}_A^{-1}),$$

where

$$\mathbb{D}_A = \mathrm{E}\int_0^\tau [\boldsymbol{Z}(t) - \boldsymbol{e}(t)]^{\otimes 2} Y(t)\, dt.$$

Proof. Differentiating the observed pseudoscore, we get

$$\begin{aligned} -n^{-1}\frac{\partial U_A(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\mathsf{T}} =& \frac{1}{n}\sum_{i=1}^n \int_0^\tau [\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}(t)]\boldsymbol{Z}_i^\mathsf{T}(t) Y_i(t)\, dt \\ =& \frac{1}{n}\sum_{i=1}^n \int_0^\tau [\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}(t)]^{\otimes 2} Y_i(t)\, dt \\ =& \frac{1}{n}\sum_{i=1}^n \int_0^\tau [\boldsymbol{Z}_i(t) - \boldsymbol{e}(t)]^{\otimes 2} Y_i(t)\, dt + L_n, \end{aligned} \tag{2.14}$$

where

$$\begin{aligned} L_n =& \frac{1}{n}\sum_{i=1}^n \int_0^\tau [\boldsymbol{e}(t) - \overline{\boldsymbol{Z}}(t)]^{\otimes 2} Y_i(t)\, dt \\ &+ \frac{1}{n}\sum_{i=1}^n \int_0^\tau [\boldsymbol{Z}_i(t) - \boldsymbol{e}(t)][\boldsymbol{e}(t) - \overline{\boldsymbol{Z}}(t)]^\mathsf{T} Y_i(t)\, dt \\ &+ \frac{1}{n}\sum_{i=1}^n \int_0^\tau [\boldsymbol{e}(t) - \overline{\boldsymbol{Z}}(t)][\boldsymbol{Z}_i(t) - \boldsymbol{e}(t)]^\mathsf{T} Y_i(t)\, dt. \end{aligned}$$

We used the identity (2.8) to obtain the second equality in (2.14). Notice that it is admissible to interchange differentiation and integration without any conditions. It follows from the fact that $U_A(\boldsymbol{\beta})$ may be written in such a way that $\boldsymbol{\beta}$ is completely outside the integrals. The remainder term L_n is $o_P(1)$ by Lemma 2.1(ii). Since the leading part is an average of iid terms,

$$-n^{-1}\frac{\partial U_A(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\mathsf{T}} \to_p \mathrm{E}\int_0^\tau [\boldsymbol{Z}_i(t) - \boldsymbol{e}(t)]^{\otimes 2} Y_i(t)\, dt = \mathbb{D}_A.$$

Using the Taylor expansion to decompose $U_A(\widehat{\boldsymbol{\beta}}_A) - U_A(\boldsymbol{\beta}_0)$, we get

$$U_A(\widehat{\boldsymbol{\beta}}_A) - U_A(\boldsymbol{\beta}_0) = \frac{\partial U_A(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^\mathsf{T}}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0).$$

where $\boldsymbol{\beta}^*$ is on the line segment between $\widehat{\boldsymbol{\beta}}_A$ and $\boldsymbol{\beta}_0$. Since $n^{-1}\partial U_A(\boldsymbol{\beta}^*)/\partial \boldsymbol{\beta}^\mathsf{T}$ is constant in $\boldsymbol{\beta}^*$ and converges in probability to $-\mathbb{D}_A$, and since $U_A(\widehat{\boldsymbol{\beta}}_A) = 0$, we have

$$\frac{1}{\sqrt{n}}U_A(\boldsymbol{\beta}_0) = \mathbb{D}_A\sqrt{n}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0) + o_P(1).$$

By Lemma 2.2, the left hand side converges in law to $\mathrm{N}_p(\mathbf{0}, \Sigma_A(\boldsymbol{\beta}_0))$. Hence, $\sqrt{n}(\widehat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_0)$ converges in law to $\mathrm{N}_p(\mathbf{0}, \mathbb{D}_A^{-1}\Sigma_A(\boldsymbol{\beta}_0)\mathbb{D}_A^{-1})$. Consistency of $\boldsymbol{\beta}_0$ follows from asymptotic normality. □

Chapter 3

INCOMPLETE COVARIATE DATA IN FAILURE TIME REGRESSION

3.1 *The two-phase design as a missing data problem*

Suppose that we have a first-phase sample: a random sample of subjects (indexed by $i = 1, \ldots, n$) selected from an infinite general population of interest. We would like to use the first-phase sample plus some additional information for making inference about certain population characteristics. During the first phase, we observe the data $\{\mathcal{D}_{1i}, i = 1, \ldots, n\}$. In failure time regression, $\mathcal{D}_{1i}$ may include the failure and censoring information and some of the covariates. Suppose, however, that the first-phase data are not sufficient for correct inference, e.g., because some important covariates are not observed. So, we draw a second-phase sample $\mathcal{V} \subset \{1, \ldots, n\}$ and collect additional data $\mathcal{D}_{2i}$ for each $i \in \mathcal{V}$. We assume that the complete data $\{\mathcal{D}_{1i} \cup \mathcal{D}_{2i}, i = 1, \ldots, n\}$ include all the information necessary for making inference about the desired general population characteristics. However, we observe only $\{\mathcal{D}_{1i}, i = 1, \ldots, n\} \cup \{\mathcal{D}_{2i}, i \in \mathcal{V}\}$ and $\{\mathcal{D}_{2i}, i \notin \mathcal{V}\}$ is missing.

Thus, a two-phase design can be regarded as a special case of a general missing data problem. While the general missing data theory has to make various assumptions about the missingness process that are hard to verify (missing at random, missing completely at random—see Rubin, 1976), in a two-phase design we usually know exactly how the second-phase sample $\mathcal{V}$ was selected. We use the observed data and the knowledge of the second-phase sampling mechanism to estimate what the values of certain statistics of interest would have been if the complete data were observed

and we base the inference on the estimated statistics. This is our general approach to two-phase designs.

In the rest of this chapter, we discuss the case-cohort design in general, review the existing methods for estimating the Cox model parameters under the case-cohort design, and summarize different approaches to covariate measurement errors in general and their applications to the Cox model in particular.

3.2 The case-cohort design

3.2.1 General discussion

The case-cohort design was first proposed by Prentice (1986). It is a special case of a two-phase design, where the second-phase sample $\mathcal{V}$ consists of all the failures (cases) and the subcohort. The subcohort is a random sample from the first-phase subjects regardless of their failure status. The information on the first-phase subjects not selected into $\mathcal{V}$ is very limited: it is assumed that only their failure status is known and, sometimes, that their at-risk status is also available. Even if more information is observed on these subjects (some covariates, say), it cannot be directly used by the case-cohort estimator. On the other hand, the second-phase subjects have complete failure, censoring and covariate information.

The case-cohort design has an appealing motivation in the Cox model case, for which it has been originally proposed. In the Cox model, it is only the failures that contribute directly to the partial likelihood score. The covariates of the censored observations (controls) serve only as a baseline to which the covariates of the failures are compared. So, when the covariates of all the failures are known, no term is lost from the partial likelihood. When the failure rate is low, the first phase sample consists of a relatively small number of failures and a large number of controls. Like in the case-control study, little efficiency is gained by including a large number of controls per failure and hence little efficiency is lost when some of the controls are

excluded because their covariates are not observed.

It follows that the case-cohort design is especially useful for large cohort studies of rare diseases. In such cases, it loses little efficiency compared to the complete-data estimator and it achieves the largest savings in terms of covariate measurements. That has already been demostrated by Prentice (1986).

The case-control design suffers from several drawbacks. Here is a list of the most important ones:

- The timing of covariate assessment may pose a significant problem in the case-cohort design. It is impossible to assess the covariates of the cases until it is known which subjects fail and hence the cases' covariates are ascertained retrospectively. On the other hand, the subcohort is often identified at the start of the study and followed more closely than the non-subcohort subjects. That may lead to substantial bias due to differential covariate assessment, especially if time-dependent covariates are used. A case-cohort study must be designed very carefully to avoid such bias. Occassionally, this problem can be entirely avoided: as an example, imagine that blood samples are drawn from all the subjects at the beginning of the study, stored in a freezer, and analyzed once it becomes known that a particular subject is a member of the second-phase sample. In most cases, however, such a procedure is unfeasible.

- The case-cohort study may lose efficiency due to heavy censoring at the end of the follow-up. The censoring thins out the subcohort and thus the subjects that fail later are compared to a relatively small number of controls still remaining at risk. Several authors (Samuelsen, 1989; Lin and Ying, 1993) have proposed a remedy for this problem: a new subcohort is to be selected later in time, from the subjects still remaining at risk.

- The case-cohort design assumes no covariate data are observed during the first

phase. However, in most cases at least some covariates are observed on all first-phase subjects; yet the analysis of the case-cohort study cannot directly utilize any of these data. This is the most substantial drawback of the case-cohort design. The missing data approach of Lin and Ying (1993) makes use of the partial information, but the efficiency of their method is unclear.

3.2.2 The case-cohort design under the Cox model

In this section, we review the existing methods for the estimation of the Cox model parameters under the case-cohort design. We make use of the notation introduced in Chapter 2. We start with the first results published by Prentice (1986), summarize briefly the asymptotic theory of Self and Prentice (1988) and note some later proposals and improvements considered by Lin and Ying (1993) and Samuelsen (1989).

In his 1986 paper, Prentice defined the case-cohort design as follows: let the subcohort be identified by binary selection indicators ξ_i so that $\mathrm{P}[\xi_i = 1] = n_C/n$, $\sum_{i=1}^{n} \xi_i = n_C$ and the subcohort is $\{i : \xi_i = 1\}$. This means that the subcohort is a simple random sample of the first-phase subjects with a fixed size n_C. Prentice assumed that the covariate history of the ith subject at time t is known if and only if $\xi_i = 1$ or $N_i(t) = 1$, that is, if and only if the subject is a subcohort member or has experienced a failure at time t. To formalize this, let us introduce the time-dependent indicator $\varrho_i(t) = 1 - (1 - \xi_i)(1 - \Delta N_i(t))$. Then the covariate history $\{\boldsymbol{Z}_i(s), 0 \leq s \leq t\}$ is known if and only if $\varrho(t) = 1$.

Prentice proposed using the following pseudoscore to estimate the vector of Cox model parameters $\boldsymbol{\beta}_0$:

$$U_P(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\infty} [\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}_P(\boldsymbol{\beta}, t)] dN_i(t),$$

where

$$\overline{\boldsymbol{Z}}_P(\boldsymbol{\beta}, t) = \frac{\sum_{i=1}^{n} \varrho_i(t) \boldsymbol{Z}_i(t) \exp\{\boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{Z}_i(t)\} Y_i(t)}{\sum_{i=1}^{n} \varrho_i(t) \exp\{\boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{Z}_i(t)\} Y_i(t)}.$$

Prentice's case-cohort pseudoscore has the same form as the complete-data partial likelihood score, except that the complete-data weighted covariate average $\overline{\boldsymbol{Z}}(\boldsymbol{\beta}, t)$ is replaced by an estimator $\overline{\boldsymbol{Z}}_P(\boldsymbol{\beta}, t)$ based only on the observed second-phase data: $\overline{\boldsymbol{Z}}_P(\boldsymbol{\beta}, t)$ is a weighted average of covariates of the subject that failed at t and of the subcohort members at risk at time t. Notice that the covariates of the subjects that failed prior to t are not included in $\overline{\boldsymbol{Z}}_P(\boldsymbol{\beta}, t)$ even though they are known.

Prentice showed that $\mathrm{E}\, U_P(\boldsymbol{\beta}_0) = 0$ and calculated var $U_P(\boldsymbol{\beta}_0)$. He gave a heuristic argument for consistency and asymptotic normality of the estimator $\widehat{\boldsymbol{\beta}}_P$ defined by $U_P(\widehat{\boldsymbol{\beta}}_P) = 0$. He also proposed an estimator for the cumulative baseline hazard:

$$\widehat{\Lambda}_0(t) = \frac{n_C}{n} \int_0^t \frac{\sum_{i=1}^n dN_i(s)}{\sum_{i=1}^n \xi_i \exp\{\boldsymbol{\beta}^\mathsf{T} \boldsymbol{Z}_i(s)\} Y_i(s)}.$$

Self and Prentice (1988) provided a rigorous asymptotic theory for a slightly modified case-cohort estimator when the data are observed on a finite time interval $[0, \tau]$. Like Prentice (1986), they also assumed that the subcohort is selected by simple random sampling of a fixed size n_C such that the subcohort proportion n_C/n converges to α as $n \to \infty$. They considered the pseudoscore

$$U_{SP}(\boldsymbol{\beta}) = \sum_{i=1}^n \int_0^\infty [\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}_{SP}(\boldsymbol{\beta}, t)] dN_i(t),$$

where

$$\overline{\boldsymbol{Z}}_{SP}(\boldsymbol{\beta}, t) = \frac{\sum_{i=1}^n \xi_i \boldsymbol{Z}_i(t) \exp\{\boldsymbol{\beta}^\mathsf{T} \boldsymbol{Z}_i(t)\} Y_i(t)}{\sum_{i=1}^n \xi_i \exp\{\boldsymbol{\beta}^\mathsf{T} \boldsymbol{Z}_i(t)\} Y_i(t)},$$

which differs from $\overline{\boldsymbol{Z}}_P(\boldsymbol{\beta}, t)$ by not including any cases in the weighted covariate average unless they belong to the random subcohort. Self and Prentice formulated regularity conditions under which the case-cohort pseudoscore can be expressed as a sum of two asymptotically independent terms, the first of which is the full-cohort partial likelihood score and the second reflects the error due to subcohort sampling. They showed that

$$\frac{1}{\sqrt{n}} U_{SP}(\boldsymbol{\beta}_0) \to_d \mathrm{N}(\mathbf{0}, \Sigma_{PH}(\boldsymbol{\beta}_0) + \Sigma^*(\boldsymbol{\beta}_0))$$

and

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{SP} - \boldsymbol{\beta}_0) \rightarrow_d \mathrm{N}(\mathbf{0}, \Sigma_{PH}^{-1}(\boldsymbol{\beta}_0) + \Sigma_{PH}^{-1}(\boldsymbol{\beta}_0)\Sigma^*(\boldsymbol{\beta}_0)\Sigma_{PH}^{-1}(\boldsymbol{\beta}_0)).$$

where $\widehat{\boldsymbol{\beta}}_{SP}$ is the solution of $U_{SP}(\widehat{\boldsymbol{\beta}}_{SP}) = 0$ and $\Sigma^*(\boldsymbol{\beta}_0)$ is the extra score variance due to the incompleteness of the data.

Self and Prentice also proposed a consistent estimator for the extra score variance matrix $\Sigma^*(\boldsymbol{\beta}_0)$ and for the variance of the case-cohort estimator. They demonstrated that their estimator was asymptotically equivalent to Prentice's and calculated its asymptotic relative efficiency with respect to the complete-data partial likelihood estimator in a simple case.

Lin and Ying (1993) mainly dealt with estimation in the Cox model with irregularly missing covariates. They worked under the missing-completely-at-random assumption (Rubin, 1976). The case-cohort estimator of Self and Prentice can be obtained as a special case of Lin and Ying's estimator. Lin and Ying also proposed a different way to estimate the variance of the case-cohort estimator. However, it can be shown that their variance estimator is asymptotically equivalent to that introduced by Self and Prentice. Lin and Ying's approach easily accomodates augmenting the subcohort by selecting more subjects later in time and works under different regularity conditions that do not include finiteness of the observation interval.

Samuelsen in his unpublished PhD dissertation (Samuelsen, 1989) introduced several possible generalizations of the case-cohort design of Self and Prentice. He suggested that the assumption of simple random sampling of the subcohort could be relaxed in two ways: First, by making the subcohort selection indicators time-dependent. In this way it would be possible to use different subcohorts at different times of observation. His second proposal was to use general selection probabilities p_i for the inclusion of the first-phase subjects into the subcohort and to employ the Horvitz-Thompson weights (Horvitz and Thompson, 1951) in the case-cohort pseudoscore. Samuelsen also noticed that the entire covariate histories of all the failures

can be utilized, if they are known.

Samuelsen suggested a case-cohort estimator for a parametric failure time regression model, which made use of several of these improvements. However, he neither defined nor investigated an analogous estimator for the Cox model. It is nevertheless not dificult to do so; the estimator can be based on the following generalized case-cohort pseudoscore:

$$U_S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\infty} [\boldsymbol{Z}_i(t) - \overline{\boldsymbol{Z}}_S(\boldsymbol{\beta}, t)] dN_i(t), \tag{3.1}$$

where

$$\overline{\boldsymbol{Z}}_S(\boldsymbol{\beta}, t) = \frac{\sum_{i=1}^{n} \varrho_i \boldsymbol{Z}_i(t) \exp\{\boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{Z}_i(t)\} Y_i(t)}{\sum_{i=1}^{n} \varrho_i \exp\{\boldsymbol{\beta}^{\mathsf{T}} \boldsymbol{Z}_i(t)\} Y_i(t)}, \tag{3.2}$$

and $\varrho_i = \Delta_i + (1 - \Delta_i)\xi_i/p_i$. The estimator based on U_S allows for subject-specific sampling probabilities and includes all failures in the estimator for $\overline{\boldsymbol{Z}}(\boldsymbol{\beta}, t)$.

3.3 *Covariate measurement error in failure time regression*

In this section, we briefly explain how statistical models that account for covariate measurement error can be constructed and review the methods proposed for dealing with covariate mesurement error under the Cox model.

3.3.1 Measurement error in regression models

Carroll, Ruppert and Stefanski (1995; Chapter 1) provided a nice overview of various approaches to the statistical analysis of regression models with covariate measurement error. They were mainly concerned with nonlinear regression models for noncensored data but their classification applies to failure time regression models with covariare measurement error as well. Here we summarize their main arguments.

Let Y be the response variable, Z the true covariate and W the surrogate covariate, i.e., the true covariate measured with error. The underlying model we are interested in

describes the conditional distribution of the response given the true covariate $\mathcal{L}(Y|Z)$ in terms of an unknown parameter to be estimated. Even without any measurement error, the underlying model can be understood in two different ways: either we assume that Z_i, $i = 1, \ldots, n$, are realizations of a random variable with a certain parametric distribution (*structural modeling*); or we take Z_i's as random or fixed but, if they are random, we do not make any assumptions about their distribution (*functional modeling*). This distinction is important in covariate measurement error models as well as in classical linear models, where a similar distinction is made between random effects (structural modeling) and fixed effects (functional modeling).

Now suppose that the true covariate Z is not observed and that we observe W instead. To make inference about the undelying model $\mathcal{L}(Y|Z)$, we have to specify the measurement error model, that is, the association between W and Z. This can be done in two ways: we either specify $\mathcal{L}(W|Z)$, or $\mathcal{L}(Z|W)$. The models that work with $\mathcal{L}(W|Z)$ are called *error calibration models*. They treat the surrogate as a random variable when the true covariate is held fixed. Examples of error calibration models are the classical error model $W = Z + \varepsilon$ or the general error calibration model $W = \gamma_0 + \gamma_1 Z + \varepsilon$, where γ_0 and γ_1 are fixed parameters and ε is a random error. The models that specify $\mathcal{L}(Z|W)$ are called *regression calibration models*. Here, the true covariate is random given the surrogate. The simplest regression calibration model is $Z = W + \varepsilon$; when regression parameters are introduced, we get the Berkson model $Z = \gamma_0 + \gamma_1 W + \varepsilon$.

Error calibration models may seem more appealing because they describe the measurement error more intuitively: the subject's true covariate is fixed but unknown and we can observe it only with an error. However, error calibration models are interesting from another point of view: in statistical analyses, it is customary to condition on observed data; and observed is W, not Z. In addition, the regression calibration model may better describe the underlying physical mechanism. As an example, suppose that W is the amount of a herbicide applied to a plant and Z is

the amount of the herbicide actually absorbed by the plant. In this example it makes perfect sense to model Z as a function of W and not vice versa.

In general, the choice between error calibration models and regression calibration models is a matter of convenience. Sometimes one or the other is more natural and therefore preferred (like regression calibration model in the herbicide example mentioned above), but mostly either of them may be used.

Once we have chosen the measurement error model, we must learn something about its parameters. The required information about the measurement error model can be obtained externally (another data set from independent sources) or internally: from an internal validation set, where Z is directly observable, or from repeated replications of W, if only the measurement error variance is needed. It is always preferable to use internal information about the measurement error process and to this end the validation set is the ideal instrument.

3.3.2 The Cox model with covariate measurement error

In this section we review the current methods for estimating the Cox model parameters when covariates are subject to measurement error. We start with the early results of Prentice (1982), who introduced the induced relative risk and studied its properties, summarize briefly several subsequent results based on the regression calibration approach (Pepe, Self and Prentice, 1989; Hughes, 1993; Zhou and Pepe, 1995; Wang *et al.*, 1997), and describe the corrected score approach of Nakamura (1992).

The first fundamental paper on covariate measurement errors in the Cox model was written by Prentice (1982). He worked with the Cox model hazard for failure time T defined by $\lambda(t \mid \boldsymbol{Z}) = \lambda_0(t)\exp\{\boldsymbol{\beta}^\mathsf{T}\boldsymbol{Z}(t)\}$ and assumed that $\boldsymbol{W}$ is a surrogate covariate for $\boldsymbol{Z}$ such that $\lambda(t \mid \boldsymbol{Z}, \boldsymbol{W}) = \lambda(t \mid \boldsymbol{Z})$. He showed that the induced hazard $\lambda(t \mid \boldsymbol{W})$ can be written as

$$\lambda(t \mid \boldsymbol{W}) = \lambda_0(t)\mathrm{E}\left[e^{\boldsymbol{\beta}^\mathsf{T}\boldsymbol{Z}(t)} \,\middle|\, T \geq t, \boldsymbol{W}\right],$$

where $\mathrm{E}\left[\exp\{\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{Z}(t)\} \mid T \geq t, \boldsymbol{W}\right]$ is called the induced relative risk. Since Prentice conditioned on the surrogate $\boldsymbol{W}$, he actually used a regression calibration model for the measurement error. However, the induced relative risk is difficult to use and there are two reasons why. First, the induced relative risk conditions on $T \geq t$ and hence it is a complicated function of β and λ_0. It is no longer possible to factor out the baseline hazard from the relative risk equation. Second, $\mathrm{E}\exp\{\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{Z}(t)\}$ is the moment generating function of $\mathcal{L}(\boldsymbol{Z})$, which means that the induced relative risk cannot be evaluated unless the distribution $\mathcal{L}(\boldsymbol{Z} \mid \boldsymbol{W})$ is completely known.

The induced relative risk can be used to define a pseudoscore analogous to the partial likelihood score. Denote

$$\widetilde{S}_i^{(0)}(\boldsymbol{\beta}, t) = Y_i(t)\mathrm{E}\left[e^{\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{Z}_i(t)} \,\middle|\, Y_i(t) = 1, W_i\right]$$

and

$$\widetilde{\boldsymbol{S}}_i^{(1)}(\boldsymbol{\beta}, t) = \frac{\partial}{\partial \boldsymbol{\beta}^{\mathsf{T}}}\widetilde{S}_i^{(0)}(\boldsymbol{\beta}, t).$$

Then

$$U_{IS}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\int_0^{\infty}\left[\frac{\widetilde{\boldsymbol{S}}_i^{(1)}(\boldsymbol{\beta}, t)}{\widetilde{S}_i^{(0)}(\boldsymbol{\beta}, t)} - \frac{\sum_{j=1}^{n}\widetilde{\boldsymbol{S}}_j^{(1)}(\boldsymbol{\beta}, t)}{\sum_{j=1}^{n}\widetilde{S}_j^{(0)}(\boldsymbol{\beta}, t)}\right] dN_i(t).$$

The induced score U_{IS} depends on the unknown baseline hazard through the induced relative risk, which is itself unknown, and so it is very difficult to evaluate. However, as Prentice showed, it may be used to derive a score test for $\boldsymbol{\beta} = \mathbf{0}$ because, under $\boldsymbol{\beta} = \mathbf{0}$, $U_{IS}(\mathbf{0})$ is independent of λ_0.

Prentice argued that $U_{IS}(\boldsymbol{\beta})$ can be evaluated approximately when the failure rate is low and $\boldsymbol{Z}$ is normally distributed given $\boldsymbol{W} = \boldsymbol{w}$, with mean $\boldsymbol{\mu}(\boldsymbol{w})$ and variance $\Sigma(\boldsymbol{w})$. He showed that whenever $\mathrm{P}[T > t \mid \boldsymbol{Z}] \approx 1$, the dependence of $\widetilde{S}_i^{(0)}(\boldsymbol{\beta}, t)$ and $\widetilde{\boldsymbol{S}}_i^{(1)}(\boldsymbol{\beta}, t)$ on λ_0 is negligible and therefore

$$\mathrm{E}\left[e^{\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{Z}(t)} \,\middle|\, T \geq t, \boldsymbol{W}\right] \approx \mathrm{E}\left[e^{\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{Z}(t)} \,\middle|\, \boldsymbol{W}\right] = e^{\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\mu}(\boldsymbol{W}) + \boldsymbol{\beta}^{\mathsf{T}}\Sigma(\boldsymbol{W})\boldsymbol{\beta}/2}.$$

This idea leads to an easy method to estimate $\boldsymbol{\beta}$. It was further investigated by Pepe, Self and Prentice (1989), who proposed an approximate partial likelihood for $\boldsymbol{\beta}$ when the failure rate is low and measurement error is normal.

Hughes (1993) studied the bias arising when the ordinary partial likelihood is used to estimate $\boldsymbol{\beta}$ in the presence of measurement error. He proposed an adjustment to the naive estimator based on the Cox partial likelihood that removes most of the bias when censorship is high.

Zhou and Pepe (1995) noticed that the induced relative risk (and hence the induced score) can be estimated nonparametrically when all surrogate covariates are discrete and a validation set is available. Let the validation set be defined by $\{i : \xi_i = 1\}$, where ξ_i's are iid binary variables. The induced relative risk for validation subjects is simply $\exp\{\boldsymbol{\beta}^\top \boldsymbol{Z}_i\}$. The induced relative risk for a nonvalidation subject, $\mathrm{E}\left[e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_i} \,\middle|\, Y_i(t) = 1, \boldsymbol{W}_i\right]$, can be estimated by

$$\frac{\sum_{j=1}^{n} \xi_j Y_j(t) \mathbb{1}(\boldsymbol{W}_i = \boldsymbol{W}_j)\, e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_j}}{\sum_{j=1}^{n} \xi_j Y_j(t) \mathbb{1}(\boldsymbol{W}_i = \boldsymbol{W}_j)}.$$

The authors derived the limiting distribution of the resulting estimator and investigated its efficiency. The biggest drawback of this method is that it requires discrete surrogate covariates and may be unstable in small samples. Zhou and Wang (1995) worked out an extention to continuous surrogate covariates by using nonparametric kernel smoothing. Yet, because of that, a practical application of their method requires quite large sample sizes.

Wang, Hsu, Feng and Prentice (1997) developed a general regression calibration method for estimation in the Cox model with covariates subject to measurement error. They assumed that a validation set is available and specified a model for $\mathrm{E}[\boldsymbol{Z} \mid \boldsymbol{W}, \boldsymbol{U}]$, where $\boldsymbol{U}$ are some covariates that are always observed. They used this regression calibration model to estimate the missing $\boldsymbol{Z}$'s and to impute them into the partial likelihood score. The parameter estimator based on this procedure is, in general, biased, but the simulation studies conducted by the authors suggest that

the bias is small in most practical situations.

The methods reviewed so far are all based on regression calibration. Error calibration was applied by Nakamura (1992), who used the so-called corrected score method to derive an approximately unbiased estimator. Nakamura assumed that $\boldsymbol{W} = \boldsymbol{Z} + \boldsymbol{\varepsilon}$, where the measurement error $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \Sigma)$ and Σ is a known covariance matrix. The argument behind the corrected score method (see also Nakamura, 1990; Carroll, Ruppert and Stefanski, 1995, Chapter 6) goes as follows. Suppose that, when there is no measurement error, the score function $U(\boldsymbol{\beta}, Y, \boldsymbol{Z})$ yields a consistent estimator. It follows that $\mathrm{E}\, U(\boldsymbol{\beta}_0, Y, \boldsymbol{Z}) = \mathbf{0}$, where $\boldsymbol{\beta}_0$ is the true parameter. When $\boldsymbol{W}$ is observed instead of $\boldsymbol{Z}$, the score $U(\boldsymbol{\beta}, Y, \boldsymbol{W})$ in general yields a biased estimator since $U(\boldsymbol{\beta}, Y, \boldsymbol{W}) \neq \mathbf{0}$. However, if there exists a modified score $U^*(\boldsymbol{\beta}, Y, \boldsymbol{W})$ such that $\mathrm{E}\,[U^*(\boldsymbol{\beta}, Y, \boldsymbol{W}) \mid Y, \boldsymbol{Z}] = U(\boldsymbol{\beta}, Y, \boldsymbol{Z})$, then the estimator based on $U^*(\boldsymbol{\beta}, Y, \boldsymbol{W})$ should be consistent. That follows from $\mathrm{E}\, U^*(\boldsymbol{\beta}_0, Y, \boldsymbol{W}) = \mathrm{E}\, U(\boldsymbol{\beta}_0, Y, \boldsymbol{Z}) = \mathbf{0}$. The modified score U^* is called the corrected score.

The problem with the corrected score is that it may not exist. Unfortunately, the Cox partial likelihood score is an example of this phenomenon. However, even though the exact corrected score does not exist for the Cox model, Nakamura succeeded in deriving an approximate corrected score based on a second order Taylor expansion under the assumption of normally distributed measurement errors with known variance. The estimator based on Nakamura's approximate corrected score is inconsistent, but the magnitude of the bias is rather small. Nakamura did not prove asymptotic normality of his estimator, but derived its limiting variance through a heuristic argument.

Chapter 4

CASE-COHORT DESIGN FOR THE ADDITIVE HAZARDS MODEL

In this chapter, we define a case-cohort estimator for the additive hazards regression model and investigate its limiting and finite-sample properties. In the case-cohort design, covariate data are available for all failures and for the subjects selected into the subcohort. Unlike Prentice (1986) and Self and Prentice (1988), who assumed that the subcohort is obtained as a simple random sample of a fixed size from all the study subjects, we allow for completely general subcohort sampling and fix the subcohort size only asymptotically. In the next sections, we define the generalized case-cohort design and propose an estimator for the AH regression parameters. Then we derive limiting distribution of the estimator and develop a consistent estimator for the limiting variance matrix. We describe how the subcohort sampling probabilities should be selected to increase efficiency compared to simple random sampling. The efficiency gain in a special simple case is demonstrated by a numerical calculation. We also calculate the relative efficiency of the case-cohort estimator (relative to the full-data estimator) in the AH model and compare it to the relative efficiency calculated by Self and Prentice (1988) for the Cox model. The chapter is concluded by a simulation study and an application of the proposed methods to the NWTSG data set.

4.1 *Generalized case-cohort sampling*

In the case-cohort design, the second-phase sample includes all the failures and the subcohort: a random sample of all study subjects. As explained in Section 3.2.2, most authors assume that the subcohort is selected by simple random sampling and this assumption is not always reasonable. We will therefore consider a general subcohort sampling design defined by assigning an individual selection probability $p_i \in (0, 1]$ to each subject, similarly to the Samuelsen's proposal for the Cox model (see Section 3.2.2). The selection probabilities may be fixed pre-specified numbers, or they may depend on the covariates and other data observed on a subject during the first phase. The covariates that determine the sampling probabilities may themselves be a part of the AH model; or they may be completely extraneous. The decision on whether the selection probabilities are treated as fixed or random makes some difference to statistical inference. If the subjects' data are iid, and the sampling probabilities p_i depend on them, the p_i's may be also treated as iid random variables. However, if the p_i's are fixed numbers, they cannot be considered iid and some conditions are necessary to assure that the sequence $p_1, p_2, \ldots$ and the estimator based on it have good asymptotic properties.

Let (X, Δ) be the pair of censored failure time and failure indicator and let $\boldsymbol{Z}$ be the covariate p-vector. Although everything we do in this chapter applies to time-dependent covariates as well, we omit the argument t in $\boldsymbol{Z}(t)$ to make the notation simpler. We continue to assume that the observation period is the finite interval $[0, \tau]$. Denote by $\boldsymbol{V}$ the set of covariates observed at the first phase. Some components of $\boldsymbol{V}$ may be also a part of $\boldsymbol{Z}$, but $\boldsymbol{Z}$ should contain at least one covariate not included in $\boldsymbol{V}$. We assume that $(X_i, \Delta_i, \boldsymbol{Z}_i, \boldsymbol{V}_i)$, $i = 1, \ldots, n$, are n independent replicates of $(X, \Delta, \boldsymbol{Z}, \boldsymbol{V})$. When the selection probabilities p_i are determined by the first-phase data, we can write $p_i \equiv p(\boldsymbol{V}_i, X_i, \Delta_i)$, where $p(\cdot)$ is some function with values in $(0, 1]$.

When the sampling probabilities p_i are fixed, we define the subcohort selection indicator ξ_i as follows: let ξ_i be a binary random variable with $\mathrm{P}[\xi_i = 1] = 1 - \mathrm{P}[\xi_i = 0] = p_i$. We assume that each ξ_i is independent of X_i and Δ_i and also of ξ_j for $j \neq i$. The subcohort is defined as the set $\{i : \xi_i = 1\}$ and the total size of the subcohort is $n_C = \sum_{i=1}^{n} \xi_i$. This is a random variable; however, if we require that $n^{-1} \sum_{i=1}^{n} p_i \to \alpha$ as $n \to \infty$, then also $n_C/n \to \alpha$. So we do not fix the subcohort size but we fix the limiting proportion of subjects sampled into the subcohort. A simple random sample with random sample size is obtained by setting $p_i = \alpha$. When $p_i = p(\boldsymbol{V}_i, X_i, \Delta_i)$, we define ξ_i so that $p(\boldsymbol{V}_i, X_i, \Delta_i) = \mathrm{E}[\xi_i \mid \boldsymbol{V}_i, X_i, \Delta_i]$ and the independence of ξ_i and ξ_j should hold conditionally on the first-phase data observed on the two subjects $i \neq j$.

Apart from the components of $\boldsymbol{V}_i$ that do not enter the AH model equation, and apart from the sampling probabilities p_i, the observable data are $(X_i, \Delta_i, \boldsymbol{Z}_i, \xi_i)$ when $\xi_i = 1$ or $\Delta_i = 1$, and (X_i, Δ_i, ξ_i) when $\xi_i = 0$ and $\Delta_i = 0$. Notice that the whole covariate history is known for each of the cases. To estimate the AH regression parameters, we define a new pseudoscore function that uses only the observed data and mimics the AH pseudoscore U_A. This is the topic of the next section.

4.2 Definition of the case-cohort pseudoscore and estimator

Following the lines of Samuelsen (1989), let us define the weighted availability indicator ϱ_i as

$$\varrho_i = \Delta_i + (1 - \Delta_i)\frac{\xi_i}{p_i}.$$

The availability indicator is zero if the covariate data for the i-th subject are not available; it is one if the subject is a case (an observed failure); and it equals the inverse selection probability if the subject is not a case but has been selected into the subcohort. Weighting incomplete data by inverse selection probabilities is an old idea dating back to Horvitz and Thompson (1951). The indicator ϱ_i properly handles even the failing subjects selected into the subcohort because it assigns their

sampling probability implicitly to 1. Notice that $\mathrm{E}[\varrho_i \mid \Delta_i, \boldsymbol{Z}_i] = \mathrm{E}\,\varrho_i = 1$ because ξ_i is independent of Δ_i conditionally on $\boldsymbol{V}_i$.

We define the case-cohort pseudoscore by taking the full-data AH pseudoscore U_A and multiplying each contribution by the weighted availability indicator. In this way the contributions of the subjects with unobserved covariates are eliminated from the pseudoscore and the contributions of the remaining subjects are properly weighted by inverse selection probabilities. This is unlike the case-cohort design for the Cox model, where only the failures contribute to the partial likelihood score and hence there is no need to weight individual score contributions.

Let us first modify the at-risk covariate average as follows:

$$\overline{\boldsymbol{Z}}_H(t) = \frac{\sum_{i=1}^n \varrho_i \boldsymbol{Z}_i Y_i(t)}{\sum_{i=1}^n \varrho_i Y_i(t)}.$$

The case-cohort at-risk average $\overline{\boldsymbol{Z}}_H(t)$ is a weighted average of covariates of failures and subcohort members that are at risk at time t. Thus, subject with unknown covariates are not included in the average. Now the case-cohort pseudoscore is defined as

$$U_H(\boldsymbol{\beta}) = \sum_{i=1}^n \varrho_i \int_0^\infty [\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_H(t)] \left[dN_i(t) - \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\beta} Y_i(t)\, dt\right].$$

The AH full-data pseudoscore includes a term for each subject no matter if it is censored or not. Therefore, the case-cohort pseudoscore for the AH model inevitably loses some contributions originally included in the full-data AH pseudoscore. This was not the case with the case-cohort pseudoscore for the Cox model. It would be interesting to see whether this fact has any detrimental effect on the relative efficiency of the AH case-cohort design (this point is addressed in Section 4.5).

The case-cohort pseudoscore U_H satisfies the same basic identities as the full-data pseudoscore U_A. Indeed, since

$$\sum_{i=1}^n \varrho_i [\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_H(t)] Y_i(t) = \sum_{i=1}^n \varrho_i \boldsymbol{Z}_i Y_i(t) - \sum_{i=1}^n \varrho_i Y_i(t) \frac{\sum_{i=1}^n \varrho_i \boldsymbol{Z}_i Y_i(t)}{\sum_{i=1}^n \varrho_i Y_i(t)} = \boldsymbol{0}, \qquad (4.1)$$

we can write U_H as

$$U_H(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \int_0^\infty [\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_H(t)]\, dN_i(t) - \varrho_i \int_0^\infty [\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_H(t)]^{\otimes 2} \boldsymbol{\beta} Y_i(t)\, dt \right\}. \quad (4.2)$$

where we use the trivial identity $\varrho_i\, dN_i(t) = dN_i(t)$. It also follows from (4.1) that

$$U_H(\boldsymbol{\beta}_0) = \sum_{i=1}^{n} \varrho_i \int_0^\infty [\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_H(t)]\, dM_i(t),$$

where

$$M_i(t) \equiv N_i(t) - \int_0^t Y_i(s)\, d\Lambda_0(s) - \int_0^t \boldsymbol{Z}_i^{\mathsf{T}} \boldsymbol{\beta}_0 Y_i(s)\, ds$$

is the AH counting process martingale. It is, however, important to note that the integrand $\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_H(t)$ is not predictable and hence $U_H(\boldsymbol{\beta}_0)$ is not a martingale integral. Predictability of the integrand is violated because $\overline{\boldsymbol{Z}}_H(t)$ depends on the weighted availability indicators, which in turn are functions of Δ_i.

By (4.2), the estimator $\widehat{\boldsymbol{\beta}}_H$ defined by the estimating equation $U_H(\widehat{\boldsymbol{\beta}}_H) = 0$ takes the form

$$\widehat{\boldsymbol{\beta}}_H = \left\{ \sum_{i=1}^{n} \int_0^\infty \varrho_i [\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_H(t)]^{\otimes 2} Y_i(t)\, dt \right\}^{-1} \left\{ \sum_{i=1}^{n} \int_0^\infty [\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_H(t)]\, dN_i(t) \right\}.$$

Both U_H and $\widehat{\boldsymbol{\beta}}_H$ depend on the covariates only through the difference $\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_H(t)$. That implies that both are invariant with respect to linear transformations of the covariates and that the matrix

$$\widehat{\mathbb{D}}_H \equiv \sum_{i=1}^{n} \int_0^\infty \varrho_i [\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_H(t)]^{\otimes 2} Y_i(t)\, dt$$

is symmetric and positive definite unless $\overline{\boldsymbol{Z}}_H(t)$ is identical to $\boldsymbol{Z}_i$ at all times. These important properties hold only because of (4.1); alternative definitions of the at-risk average $\overline{\boldsymbol{Z}}_H(t)$, which do not satisfy (4.1), result in estimators with undesirable small- and large-sample behavior.

4.3 *Limiting distribution of the case-cohort estimator*

Before we proceed to deriving the asymptotic distributions of the case-cohort pseudoscore and the estimator, let us summarize the working assumptions and review the notation and regularity conditions introduced in Chapter 2.

We assume that the data are observed on the time interval $[0, \tau]$, $0 < \tau < \infty$ and that $(X_i, \Delta_i, \boldsymbol{Z}_i)$, $i = 1, \ldots, n$, are independent and identically distributed replicates of $(X, \Delta, \boldsymbol{Z})$. As in Section 2.3, denote

$$\pi_k(t) \equiv \mathrm{E}\, \boldsymbol{Z}^{\otimes k} Y(t), \qquad k = 0, 1, 2,$$

and define $e(t) = \boldsymbol{\pi}_1(t)/\pi_0(t)$. Recall that $\pi_k(t)$ is also the limit in probability of $n^{-1} \sum_{i=1}^n \boldsymbol{Z}_i^{\otimes k} Y_i(t)$ for each $k = 0, 1, 2$.

We will make use of the regularity conditions needed for the weak convergence of the full-data AH estimator. Introduced in Section 2.3, they were as follows:

Condition 2.1. $\Lambda_0(\tau) < \infty$.

Condition 2.2. $\mathrm{P}[Y_i(\tau) = 1] > 0$.

Condition 2.3. $\mathrm{E} \sup_{0 \le t \le \tau} |Y(t) \boldsymbol{Z}^{\otimes 2} \boldsymbol{\beta}_0^{\mathsf{T}} \boldsymbol{Z}| < \infty$.

Condition 2.4. The matrix $\Sigma_A(\boldsymbol{\beta}_0) \equiv \mathrm{E} \int_0^\tau [\boldsymbol{Z} - \boldsymbol{e}(t)]^{\otimes 2} dN(t)$ is positive definite.

Under Conditions 2.2 and 2.3, $\pi_k(t)$, $k = 0, 1, 2$, are uniformly bounded on $[0, \tau]$ and $\pi_0(t)$ is bounded away from zero on $[0, \tau]$. By Lemma 2.1,

$$\sup_{0 \le t \le \tau} \left| \frac{1}{n} \sum_{i=1}^n \boldsymbol{Z}_i^{\otimes k} Y_i(t) - \pi_k(t) \right| \to_p 0, \qquad k = 0, 1, 2,$$

and $\sup_{0 \le t \le \tau} |\overline{\boldsymbol{Z}}(t) - \boldsymbol{e}(t)| \to_p 0$.

We will need another regularity condition for the case-cohort design. It forces the subcohort selection probabilities to be bounded away from zero:

Condition 4.1. There exists a constant $p_0 > 0$ such that $p_i > p_0$ for $i = 1, 2, \ldots$.

We proceed by showing that the case-cohort at-risk average $\overline{\boldsymbol{Z}}_H(t)$ is uniformly consistent in the following sense:

Lemma 4.1. *Let Conditions 2.2 and 2.3 hold and let p_i be independent and identically distributed random variables satisfying Condition 4.1. Then*

$$\sup_{0\le t\le\tau} |\overline{\boldsymbol{Z}}_H(t) - \boldsymbol{e}(t)| \to_p 0.$$

Proof. Since $\sup_{0\le t\le\tau} |\overline{\boldsymbol{Z}}(t) - \boldsymbol{e}(t)| \to_p 0$, it suffices to show

$$\sup_{0\le t\le\tau} |\overline{\boldsymbol{Z}}_H(t) - \overline{\boldsymbol{Z}}(t)| \to_p 0.$$

By definition of $\overline{\boldsymbol{Z}}_H(t)$ and $\overline{\boldsymbol{Z}}(t)$,

$$\begin{aligned}
|\overline{\boldsymbol{Z}}(t)_H - \overline{\boldsymbol{Z}}(t)_H| &= \left| \frac{n^{-1}\sum \varrho_i \boldsymbol{Z}_i Y_i(t)}{n^{-1}\sum \rho_i Y_i(t)} - \frac{n^{-1}\sum \boldsymbol{Z}_i Y_i(t)}{n^{-1}\sum Y_i(t)} \right| \\
&\le \frac{1}{n^{-1}\sum Y_i(t)} \left| \frac{1}{n}\sum (1-\varrho_i)\boldsymbol{Z}_i Y_i(t) \right| \\
&\quad + \left| \frac{1}{n}\sum \varrho_i \boldsymbol{Z}_i Y_i(t) \right| \left| \left[\tfrac{1}{n}\sum \varrho_i Y_i(t) \right]^{-1} - \left[\tfrac{1}{n}\sum Y_i(t) \right]^{-1} \right|.
\end{aligned}$$

The supremum over t of the right-hand side converges in probability to zero provided that

$$\sup_{0\le t\le\tau} \frac{1}{n} \left| \sum_{i=1}^{n} (1-\varrho_i)\boldsymbol{Z}_i Y_i(t) \right| \to_p 0, \tag{4.3}$$

and

$$\sup_{0\le t\le\tau} \frac{1}{n} \left| \sum_{i=1}^{n} (1-\varrho_i) Y_i(t) \right| \to_p 0. \tag{4.4}$$

Notice that $1-\varrho_i = (1-\Delta_i)(1-\xi_i/p_i)$. Convergences in (4.3) and (4.4) obviously hold pointwise, since $\mathrm{E}(1-\varrho_i) = 0$. It remains to prove uniformity. But, since ξ_i are identically distributed, uniformity in both (4.3) and (4.4) can be proven in the same way as Lemma 2.1(i). □

When the sampling probabilities are fixed constants, we just assume that (4.3) and (4.4) hold so that Lemma 4.1 remains valid:

Condition 4.2. If $p_1, p_2, \ldots$ are fixed constants, they satisfy both (4.3) and (4.4).

Conditions 4.1 and 4.2 are assumed to be satisfied for the rest of this section. We show next that the normalized pseudoscore evaluated at $\boldsymbol{\beta}_0$ can be approximated by a sum of independent zero-mean terms.

Theorem 4.1. *Assume that Conditions 2.1-2.4 hold. Then*

$$\frac{1}{\sqrt{n}} U_H(\boldsymbol{\beta}_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widetilde{\psi}_i^{(A)}(\boldsymbol{\beta}_0) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(1 - \frac{\xi_i}{p_i}\right)(1 - \Delta_i) \boldsymbol{S}_i(\boldsymbol{\beta}_0) + o_P(1). \quad (4.5)$$

where

$$\widetilde{\psi}_i^{(A)}(\boldsymbol{\beta}_0) = \int_0^\tau [\boldsymbol{Z}_i - \boldsymbol{e}(t)]\, dM_i(t),$$

and

$$\boldsymbol{S}_i(\boldsymbol{\beta}_0) = \int_0^\tau [\boldsymbol{Z}_i - \boldsymbol{e}(t)] Y_i(t) [d\Lambda_0(t) + \boldsymbol{Z}_i^\top \boldsymbol{\beta}_0\, dt].$$

In addition, $\widetilde{\psi}_i^{(A)}(\boldsymbol{\beta}_0)$ *and* $(1 - \xi_i/p_i)(1 - \Delta_i)\boldsymbol{S}_i(\boldsymbol{\beta}_0)$ *have zero means and are uncorrelated.*

Theorem 4.1 approximates the case-cohort AH pseudoscore by a sum of the full-data pseudoscore U_A and an adjustment factor due to the missing covariates in the case-cohort design. The adjustment factor is independent of the full-data pseudoscore. This representation of the case-cohort pseudoscore is analogous to that derived by Self and Prentice (1988) for the Cox model. We present the proof of Theorem 4.1 under the iid assumption only. It relies on three technical lemmas, which we formulate at this point.

Lemma 4.2. *Let* $\{f_n\}$ *and* $\{g_n\}$ *be two sequences of bounded deterministic functions such that for some constant* τ,

- $\sup_{0\le t\le\tau}|f_n(t)-f(t)|\to 0$ *as* $n\to\infty$, *where* f *is a continuous function;*
- $\{g_n\}$ *are monotonic;*
- $g_n(t)\to g(t)$ *for* $t\in[0,\tau]$ *and some* g, *which is right-continuous at 0 and left-continuous at* τ.

Then

$$\sup_{0\le t\le\tau}\left|\int_0^t f_n(s)\,dg_n(s)-\int_0^t f(s)\,dg(s)\right|\to 0$$

as $n\to\infty$.

Proof. We can write $f_n=f_n^+-f_n^-$, where $f_n^+=\max(f_n,0)$ and $f_n^-=\max(-f_n,0)$ are nonnegative functions. Hence, no generality is lost by assuming that f_n is nonnegative. Let us also assume that g_n is nondecreasing. Then

$$\int_0^t f_n(s)\,dg_n(s)-\int_0^t f(s)\,dg(s)=\int_0^t[f_n(s)-f(s)]\,dg_n(s)+\int_0^t f(s)\,d(g_n-g)(s). \tag{4.6}$$

The first term on the right-hand side of (4.6) is bounded in absolute value by

$$\int_0^t|f_n(s)-f(s)|\,dg_n(s)\le\sup_{0\le s\le\tau}|f_n(s)-f(s)|\,[g_n(t)-g_n(0)],$$

which converges to zero because $f_n\to f$ uniformly in t and $g_n(t)\to g(t)<\infty$.

The second Helly's theorem, as formulated in Serfling (1980) on page 352, states the following: If f is a continuous function and $g_n\to g$ on a finite interval $[0,\tau]$, where g_n are uniformly bounded and g is continuous at 0 and at τ, then $\int_0^\tau f\,dg_n\to\int_0^\tau f\,dg$. Thus, the Helly's theorem implies that the last term in (4.6) converges to zero, too.

It follows that

$$\int_0^t f_n(s)\,dg_n(s)\to\int_0^t f(s)\,dg(s)$$

and the convergence is uniform since $\int_0^t f_n(s)\,dg_n(s)$ is monotone in t. This concludes the proof. □

Lemma 4.3. *Let $B_n(t)$ be a sequence of processes converging weakly to a tight limit $B(t)$ with almost surely continuous sample paths. Assume that Conditions 2.1-2.4 hold and let $\{\boldsymbol{Z}(t), 0 \leq t \leq \tau\}$ be of bounded variation. Then*

$$\int_0^\tau [\boldsymbol{e}(t) - \overline{\boldsymbol{Z}}_H(t)] dB_n(t) \to_p 0.$$

Proof. By the Skorokhod strong embedding theorem (see Shorack and Wellner, 1986, p. 47), there exists another probability space on which $\overline{\boldsymbol{Z}}_H(t)$ and $B_n(t)$ can be defined so that $\{\overline{\boldsymbol{Z}}_H(t), B_n(t)\}$ converge to $\{\boldsymbol{e}(t), B(t)\}$ almost surely. Since $B(t)$ is almost surely continuous, the convergence of B_n to B also holds under the supremum norm, that is,

$$\sup_{0 \leq t \leq \tau} |B_n(t) - B(t)| \to_{as} 0.$$

Denote $\boldsymbol{S}_{1n}(t) = n^{-1} \sum \boldsymbol{Z}_i(t) Y_i(t)$ and $S_{0n}(t) = n^{-1} \sum Y_i(t)$. Then $\overline{\boldsymbol{Z}}_H(t) = \boldsymbol{S}_{1n}(t) S_{0n}^{-1}(t)$. Since $\boldsymbol{Z}_i = \boldsymbol{Z}_i^+ - \boldsymbol{Z}_i^-$, where $\boldsymbol{Z}_i^+ = \max(\boldsymbol{Z}_i, \boldsymbol{0})$ and $\boldsymbol{Z}_i^- = \max(-\boldsymbol{Z}_i, \boldsymbol{0})$ (the maxima are taken componentwise), we can assume without loss of generality that all components of $\boldsymbol{Z}_i(t)$ are positive for all i. Then, as functions of t, $\boldsymbol{S}_{1n}(t)$ is non-increasing and $S_{0n}^{-1}(t)$ is non-decreasing. It means that $\overline{\boldsymbol{Z}}_H(t)$ can be written as a product of two monotonic functions.

Using integration by parts, we get

$$\int_0^t S_{0n}^{-1}(s)\, dB_n(s) = S_{0n}^{-1}(t) B_n(t) - \int_0^t B_n(s)\, dS_{0n}^{-1}(s). \tag{4.7}$$

A sample path of the limiting process B is almost surely a bounded continuous function. A sample path of S_{0n}^{-1} is bounded with probability converging to 1, monotonic, and it converges to π_0^{-1} almost surely in the new probability space as $n \to \infty$. So, Lemma 4.2 is applicable, with $g_n := S_{0n}^{-1}$ and $f_n := B_n$. It follows that (4.7) converges almost surely to

$$\pi_0^{-1}(t) B(t) - \int_0^t B(s)\, d\pi_0^{-1}(s) = \int_0^t \pi_0^{-1}(s)\, dB(s).$$

where the last equality follows from integration by parts. The convergence is uniform in t.

Denote $B_n^*(t) = \int_0^t S_{0n}^{-1}(s)\, dB_n(s)$ and $B^*(t) = \int_0^t \pi_0^{-1}(s)\, dB(s)$. We have shown that $\sup_{0\le t\le\tau} |B_n^*(t) - B^*(t)| \to_{as} 0$. Obviously,

$$\int_0^\tau \overline{\boldsymbol{Z}}_H(t)\, dB_n(t) = \int_0^\tau \boldsymbol{S}_{1n}(t)\, B_n^*(t).$$

Since the sample paths of $\boldsymbol{S}_{1n}$ are monotonic and bounded and since they converge to $\boldsymbol{\pi}_1$ almost surely, we can apply integration by parts and Lemma 4.2 once more to get

$$\int_0^\tau \boldsymbol{S}_{1n}(t)\, B_n^*(t) \to_{as} \int_0^\tau \boldsymbol{\pi}_1(t)\, dB^*(t).$$

Thus, in the new probability space,

$$\int_0^\tau \overline{\boldsymbol{Z}}_H(t)\, dB_n(t) \to \int_0^\tau \boldsymbol{e}(t)\, dB(t)$$

almost surely and hence weakly in the original probability space.

Similarly, we can show that

$$\int_0^\tau \boldsymbol{e}(t)\, dB_n(t) \to \int_0^\tau \boldsymbol{e}(t)\, dB(t)$$

weakly in the original probability space. Since $\int_0^\tau \overline{\boldsymbol{Z}}_H(t)\, dB_n(t)$ and $\int_0^\tau \boldsymbol{e}(t)\, dB_n(t)$ converge in distribution to the same limit,

$$\int_0^\tau [\boldsymbol{e}(t) - \overline{\boldsymbol{Z}}_H(t)]\, dB_n(t)$$

converges to zero in distribution and hence in probability. □

The assertion of Lemma 4.2 would be trivial if $\overline{\boldsymbol{Z}}_H(t)$ were predictable. Since it is not, a more delicate argument is needed to prove the lemma. The assumption of bounded variation imposed on the covariates is very weak and we do not mention it among the conditions of Theorem 4.1.

Finally, we need a result on weak convergence of monotone processes. We state it as the third lemma, where we use the notation $\ell^\infty[0,\tau]$ for the space of bounded functions on the interval $[0,\tau]$.

Lemma 4.4. *Let A_i, $i = 1,\ldots,n$, be iid stochastic processes with nondecreasing sample paths, indexed by an interval $[0,\tau]$. If $\mathrm{E}\,A_i^2(0) < \infty$ and $\mathrm{E}\,A_i^2(\tau) < \infty$, then the sequence*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(A_i - \mathrm{E}\,A_i)$$

converges weakly in $\ell^\infty[0,\tau]$ to a tight Gaussian process.

Lemma 4.4 is proven in van der Vaart and Wellner (1996) as Example 2.11.16 on page 215. Its proof relies on the bracketing central limit theorem. Now we can proceed to the proof of Theorem 4.1.

Proof of Theorem 4.1. The case-cohort pseudoscore satisfies

$$\begin{aligned}
\frac{1}{\sqrt{n}}U_H(\boldsymbol{\beta}_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varrho_i\int_0^\tau [\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_H(t)]\,dM_i(t) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau [\boldsymbol{Z}_i - \boldsymbol{e}(t)]\,dM_i(t) \qquad (4.8)\\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau [\boldsymbol{e}(t) - \overline{\boldsymbol{Z}}_H(t)]\,dM_i(t) \qquad (4.9)\\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\varrho_i - 1)\int_0^\tau [\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_H(t)]\,dM_i(t). \qquad (4.10)
\end{aligned}$$

Clearly, (4.8) is the normalized full-data AH pseudoscore.

Denote $B_n(t) = n^{-1/2}\sum_{i=1}^{n} M_i(t)$. By the martingale central limit theorem (see Fleming and Harrington, 1991, p. 204), $B_n(t)$ converges weakly to a zero-mean Gaussian process $B(t)$ with almost all paths continuous and $B(0) = 0$. So, by Lemma 4.3, (4.9) converges to zero in probability.

Since $(1-\rho_i)\,dN_i = 0$, (4.10) is equal to

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\varrho_i - 1)\int_0^\tau [\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_H(t)]\,[Y_i(t)\,d\Lambda_0(t) + \boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{\beta}_0 Y_i(t)\,dt] \\
&\quad = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\varrho_i - 1)\boldsymbol{S}_i(\boldsymbol{\beta}_0) \\
&\qquad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\varrho_i - 1)\int_0^\tau [\boldsymbol{e}(t) - \overline{\boldsymbol{Z}}_H(t)]Y_i(t)[d\Lambda_0(t) + \boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{\beta}_0\,dt]. \qquad (4.11)
\end{aligned}$$

Expression (4.11) can be written as

$$\int_0^\tau [\boldsymbol{e}(t) - \overline{\boldsymbol{Z}}_H(t)]\,dB_n(t),$$

where $B_n(t)$ is now redefined as follows:

$$B_n(t) \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^t (\varrho_i - 1)Y_i(s)[d\Lambda_0(s) + \boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{\beta}_0\,ds].$$

To apply Lemma 4.3, we need to show that $B_n(t)$ converges weakly to a tight limit with continuous sample paths. But $B_n(t)$ can be expressed as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[A_i(t) - \mathrm{E}\,A_i(t)] - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[A_i^*(t) - \mathrm{E}\,A_i^*(t)],$$

where

$$A_i(t) = \varrho_i\int_0^t Y_i(s)[d\Lambda_0(s) + \boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{\beta}_0\,ds]$$

and

$$A_i^*(t) = \int_0^t Y_i(s)[d\Lambda_0(s) + \boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{\beta}_0\,ds]$$

are two nondecreasing processes satisfying $\mathrm{E}\,A_i(t) = \mathrm{E}\,A_i^*(t)$, $\mathrm{E}\,A_i^2(t) < \infty$ and $\mathrm{E}[A_i^*(t)]^2 < \infty$. By Lemma 4.4, $B_n(t)$ converges in $\ell^\infty[0,\tau]$ to a tight Gaussian limit $B(t)$. It follows from Lemma 1.5.9 and Example 1.5.10 of van der Vaart and Wellner (1996) that the limiting process is uniformly continuous with respect to the

semimetric $\rho_2(s,t) \equiv \mathrm{E}\,|B(s) - B(t)|^2$. Since we are in $\mathbf{R}$ and $\rho_2(s,t)$ is bounded, this is equivalent to uniform continuity of sample paths of B with respect to the Euclidean norm. Hence, Lemma 4.3 can be applied to show that (4.11) converges to zero in probability.

Since $\varrho_i - 1 = -(1-\Delta_i)(1-\xi_i/p_i)$, we have shown that

$$\frac{1}{\sqrt{n}}U_H(\boldsymbol{\beta}_0) = \frac{1}{\sqrt{n}}\sum_1^n \widetilde{\psi}_i^{(A)}(\boldsymbol{\beta}_0) - \frac{1}{\sqrt{n}}\sum_{i=1}^n (1-\Delta_i)(1-\xi_i/p_i)\boldsymbol{S}_i(\boldsymbol{\beta}_0) + o_P(1).$$

The expectation of $\widetilde{\psi}_i^{(A)}(\boldsymbol{\beta}_0)$ is zero (it is a martingale integral) and

$$\mathrm{E}\,[(1-\Delta_i)(1-\xi_i/p_i)\boldsymbol{S}_i(\boldsymbol{\beta}_0)] = \mathrm{E}\,\{(1-\Delta_i)\boldsymbol{S}_i(\boldsymbol{\beta}_0)\mathrm{E}\,[1-\xi_i/p_i \mid \Delta_i, X_i, \boldsymbol{Z}_i, \boldsymbol{V}_i\,]\} = 0.$$

Similarly,

$$\begin{aligned}\mathrm{E}\left[(1-\Delta_i)(1-\xi_i/p_i)\boldsymbol{S}_i(\boldsymbol{\beta}_0)\widetilde{\psi}_i^{(A)}(\boldsymbol{\beta}_0)\right] \\ = \mathrm{E}\left\{(1-\Delta_i)\boldsymbol{S}_i(\boldsymbol{\beta}_0)\widetilde{\psi}_i^{(A)}(\boldsymbol{\beta}_0)\mathrm{E}\,[1-\xi_i/p_i \mid \Delta_i, X_i, \boldsymbol{Z}_i, \boldsymbol{V}_i\,]\right\} = 0,\end{aligned}$$

and so, $\widetilde{\psi}_i^{(A)}$ and $(1-\Delta_i)(1-\xi_i/p_i)\boldsymbol{S}_i$ are uncorrelated. The theorem is proved. □

We are now in a position to prove weak convergence of the normalized case-cohort pseudoscore.

Theorem 4.2. *Let Conditions 2.1, 2.2, and 2.4 hold. Let Condition 2.3 be strengthened to*

$$\mathrm{E}\sup_{0\le t\le\tau}|Y(t)\boldsymbol{Z}^{\otimes 2}(\boldsymbol{\beta}_0^{\mathsf{T}}\boldsymbol{Z})^2| < \infty. \tag{4.12}$$

If

$$\frac{1}{n}\sum_{i=1}^n \frac{1-p_i}{p_i} \to \kappa \tag{4.13}$$

as $n \to \infty$, where κ is a positive real number, then

$$\frac{1}{\sqrt{n}}U_H(\boldsymbol{\beta}_0) \to_d \mathrm{N}_p(\mathbf{0}, \Sigma_A(\boldsymbol{\beta}_0) + \Sigma_H(\boldsymbol{\beta}_0)),$$

where, under fixed sampling probabilities p_i,

$$\Sigma_H(\boldsymbol{\beta}_0) = \kappa \, \mathrm{E}(1 - \Delta_i)\boldsymbol{S}_i^{\otimes 2}(\boldsymbol{\beta}_0)$$

and under random sampling probabilities p_i,

$$\Sigma_H(\boldsymbol{\beta}_0) = \mathrm{E}\,\frac{1 - p_i}{p_i}(1 - \Delta_i)\boldsymbol{S}_i^{\otimes 2}(\boldsymbol{\beta}_0).$$

Remark. Conditions (4.12) and (4.13) assure that $\Sigma_H(\boldsymbol{\beta}_0)$ exists. Obviously, Condition 2.3 is too weak to make $\mathrm{E}\,\boldsymbol{S}_i^{\otimes 2}(\boldsymbol{\beta}_0)$ finite. The matrix $\Sigma_H(\boldsymbol{\beta}_0)$ is positive semi-definite and so $\Sigma_A(\boldsymbol{\beta}_0) + \Sigma_H(\boldsymbol{\beta}_0)$ must be positive definite. Notice also that (4.13) is void when the sampling probabilities are iid random variables.

Proof. By Theorem 4.1, $n^{-1/2}U_H(\boldsymbol{\beta}_0)$ has the same limiting distribution as the right-hand side of (4.5). The limiting covariance matrix of $n^{-1/2}\sum \tilde{\psi}_i^{(A)}(\boldsymbol{\beta}_0)$ is $\Sigma_A(\boldsymbol{\beta}_0)$. Since $\mathrm{var}(1 - \xi_i/p_i) = (1 - p_i)/p_i$, we have

$$\begin{aligned}
&\mathrm{var}(1 - \xi_i/p_i)(1 - \Delta_i)\boldsymbol{S}_i(\boldsymbol{\beta}_0) \\
&\qquad = \mathrm{E}\,\mathrm{var}\,[(1 - \xi_i/p_i)(1 - \Delta_i)\boldsymbol{S}_i(\boldsymbol{\beta}_0) \mid \Delta_i, X_i, \boldsymbol{Z}_i] \\
&\qquad\quad + \mathrm{var}\,\mathrm{E}\,[(1 - \xi_i/p_i)(1 - \Delta_i)\boldsymbol{S}_i(\boldsymbol{\beta}_0) \mid \Delta_i, X_i, \boldsymbol{Z}_i] \\
&\qquad = \mathrm{E}\,\frac{1 - p_i}{p_i}(1 - \Delta_i)\boldsymbol{S}_i^{\otimes 2}(\boldsymbol{\beta}_0) + 0.
\end{aligned}$$

Hence, the limiting covariance matrix of $n^{-1/2}\sum(1 - \xi_i/p_i)(1 - \Delta_i)\boldsymbol{S}_i(\boldsymbol{\beta}_0)$ cannot be anything else than $\Sigma_H(\boldsymbol{\beta}_0)$, which is well defined.

If $n^{-1/2}\sum \tilde{\psi}_i^{(A)}(\boldsymbol{\beta}_0)$ and $n^{-1/2}\sum(1-\xi_i/p_i)(1-\Delta_i)\boldsymbol{S}_i(\boldsymbol{\beta}_0)$ jointly converge to a normal distribution, the limiting cobvariance matrix must be $\Sigma_A(\boldsymbol{\beta}_0) + \Sigma_H(\boldsymbol{\beta}_0)$ because they are uncorrelated by Theorem 4.1.

The joint weak convergence is trivial if the p_i's are iid. If they are fixed numbers, we must apply the Lindeberg central limit theorem together with the Cramér-Wold device to prove joint convergence to a normal distribution. To simplify notation, suppose that $\boldsymbol{Z}$ is a scalar and denote $\kappa_n = n^{-1}\sum(1 - p_i)/p_i$, $\sigma_1^2 = \Sigma_A$, $\sigma_2^2 =$

$\text{var}(1-\Delta_i)S_i^2$, and $V_i = c_1\widetilde{\psi}_i^{(A)} + c_2(1-\xi_i)/p_i(1-\Delta_i)S_i$, where c_1 and c_2 are real constants. The Lindeberg condition then takes the form

$$(c_1^2\sigma_1^2 + \kappa_n c_2^2\sigma_2^2)^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left\{V_i^2\mathbb{1}\left(|V_i| > \varepsilon\sqrt{n}\sqrt{c_1^2\sigma_1^2 + \kappa_n c_2^2\sigma_2^2}\right)\right\} \to 0$$

for any $\varepsilon > 0$. Since

$$|a+b|^2\,\mathbb{1}(|a+b| > \varepsilon) \le 4\,|a|^2\,\mathbb{1}(|a| > \varepsilon/2) + 4\,|b|^2\,\mathbb{1}(|b| > \varepsilon/2)$$

(see Andersen and Gill, 1982, proof of Theorem 3.2), and since the Lindeberg condition is satisfied for $\widetilde{\psi}_i^{(A)}$, it suffices to show that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\Bigg\{&\left|\left(1-\frac{\xi_i}{p_i}\right)(1-\Delta_i)S_i(\beta_0)\right|^2 \\ &\times\mathbb{1}\left(\left|\left(1-\frac{\xi_i}{p_i}\right)(1-\Delta_i)S_i(\beta_0)\right| > \varepsilon\sqrt{n}\sqrt{c_1^2\sigma_1^2+\kappa_n c_2^2\sigma_2^2}\right)\Bigg\}\end{aligned} \tag{4.14}$$

converges to zero for any $\varepsilon > 0$. But (4.14) is bounded by

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\Bigg\{\left|\left(1-\frac{\xi_i}{p_i}\right)(1-\Delta_i)S_i(\beta_0)\right|^2 \\ &\qquad\times\mathbb{1}\left(|(1-\Delta_i)S_i(\beta_0)| > \varepsilon\sqrt{n}\sqrt{c_1^2\sigma_1^2+\kappa_n c_2^2\sigma_2^2}\{\max_{i=1,\ldots,n}|1-\xi_i/p_i|\}^{-1}\right)\Bigg\} \\ \le&\left(\frac{1-p_0}{p_0}\right)^2\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\Bigg\{|(1-\Delta_i)S_i(\beta_0)|^2 \\ &\qquad\times\mathbb{1}\left(|(1-\Delta_i)S_i(\beta_0)| > \varepsilon\sqrt{n}\sqrt{c_1^2\sigma_1^2+\kappa_n c_2^2\sigma_2^2}\frac{p_0}{1-p_0}\right)\Bigg\}.\end{aligned}$$

Since $p_0 > 0$, $(1-\Delta_i)S_i(\beta_0)$ are iid, and $\sqrt{n}\sqrt{c_1^2\sigma_1^2+\kappa_n c_2^2\sigma_2^2} \to \infty$, (4.14) converges to 0 in probability. Hence the Lindeberg condition is satisfied. □

The limiting distribution of the case-cohort estimator is now easy to derive.

Theorem 4.3. *Under Conditions 2.1, 2.2 and 2.4, $\widehat{\boldsymbol{\beta}}_H$ is consistent and*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}_0) \to_d \mathrm{N}_p\left(\mathbf{0}, \mathbb{D}_A^{-1}[\Sigma_A(\boldsymbol{\beta}_0) + \Sigma_H(\boldsymbol{\beta}_0)]\mathbb{D}_A^{-1}\right),$$

where

$$\mathbb{D}_A = \mathrm{E} \int_0^\tau [\boldsymbol{Z}_i - \boldsymbol{e}(t)]^{\otimes 2} Y_i(t)\, dt.$$

Proof. Differentiating the observed case-cohort pseudoscore, we get

$$\begin{aligned}
-n^{-1}\frac{\partial U_H(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\mathsf{T}} =& \frac{1}{n}\sum_{i=1}^n \varrho_i \int_0^\tau [\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_H(t)]^{\otimes 2} Y_i(t)\, dt \\
=& \frac{1}{n}\sum_{i=1}^n \varrho_i \int_0^\tau [\boldsymbol{Z}_i - \boldsymbol{e}(t)]^{\otimes 2} Y_i(t)\, dt \\
&+ \frac{1}{n}\sum_{i=1}^n \varrho_i \int_0^\tau [\boldsymbol{Z}_i - \boldsymbol{e}(t)]\,[\boldsymbol{e}(t) - \overline{\boldsymbol{Z}}_H(t)]^\mathsf{T} Y_i(t)\, dt \\
&+ \frac{1}{n}\sum_{i=1}^n \varrho_i \int_0^\tau [\boldsymbol{e}(t) - \overline{\boldsymbol{Z}}_H(t)]\,[\boldsymbol{Z}_i - \boldsymbol{e}(t)]^\mathsf{T} Y_i(t)\, dt \\
&+ \frac{1}{n}\sum_{i=1}^n \varrho_i \int_0^\tau [\boldsymbol{e}(t) - \overline{\boldsymbol{Z}}_H(t)]^{\otimes 2} Y_i(t)\, dt.
\end{aligned}$$

The last three rows of the displayed equation converge to 0 in probability because $\sup_{0\le t\le\tau} |\boldsymbol{e}(t) - \overline{\boldsymbol{Z}}_H(t)| \to_p 0$ by Lemma 4.1. The remaining term converges in probability to

$$\mathrm{E}\, \varrho_i \int_0^\tau [\boldsymbol{Z}_i - \boldsymbol{e}(t)]^{\otimes 2} Y_i(t)\, dt,$$

which is equal to $\mathbb{D}_A$ because $\mathrm{E}[\varrho_i \mid \Delta_i, X_i, \boldsymbol{Z}_i] = 1$.

The rest of the argument is the same as in the proof of Lemma 2.3. By Taylor expansion,

$$U_H(\widehat{\boldsymbol{\beta}}_H) - U_H(\boldsymbol{\beta}_0) = \frac{\partial U_H(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^\mathsf{T}}(\widehat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}_0),$$

where $\boldsymbol{\beta}^*$ lies on the line segment between $\widehat{\boldsymbol{\beta}}_H$ and $\boldsymbol{\beta}_0$. Since $n^{-1}\partial U_H/\partial \boldsymbol{\beta}^\mathsf{T}$ is constant in $\boldsymbol{\beta}$ and converges in probability to $-\mathbb{D}_A$, we have

$$\frac{1}{\sqrt{n}} U_H(\boldsymbol{\beta}_0) = \mathbb{D}_A \sqrt{n}(\widehat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}_0) + o_P(1).$$

By Theorem 4.2, the left hand side converges in law to $\mathrm{N}_p(\boldsymbol{0}, \Sigma_A(\boldsymbol{\beta}_0) + \Sigma_H(\boldsymbol{\beta}_0))$. Hence, $\sqrt{n}(\widehat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}_0)$ converges in law to $\mathrm{N}_p(\boldsymbol{0}, \mathbb{D}_A^{-1}[\Sigma_A(\boldsymbol{\beta}_0) + \Sigma_H(\boldsymbol{\beta}_0)]\mathbb{D}_A^{-1})$. Again, consistency of $\widehat{\boldsymbol{\beta}}_H$ follows from asymptotic normality. □

Estimating the limiting covariance matrix

It remains to introduce a consistent estimator for the limiting covariance matrix of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}_0)$. We devote the rest of this section to this task. In the process, we assume that the subcohort selection probabilities p_i are independent and identically distributed random variables satisfying Condition 4.1 in the form $\mathrm{P}[p_i > p_0] = 1$ for some constant $p_0 > 0$. We also assume that Conditions 2.1, 2.2, 2.4 and 4.1 hold.

To estimate $\mathbb{D}_A^{-1}[\Sigma_A(\boldsymbol{\beta}_0) + \Sigma_H(\boldsymbol{\beta}_0)]\mathbb{D}_A^{-1}$, we need a consistent estimator for each of $\Sigma_A(\boldsymbol{\beta}_0)$, $\Sigma_H(\boldsymbol{\beta}_0)$ and $\mathbb{D}_A$. Estimating $\Sigma_A(\boldsymbol{\beta}_0)$ is easy: since the full-data estimator $\widehat{\Sigma}_A$ defined in Chapter 2 by (2.11) includes only the failures, it can be applied without a change in the case-cohort design. The negative expected pseudoscore derivative may be estimated by

$$\widehat{\mathbb{D}}_H = \frac{1}{n}\sum_{i=1}^{n} \varrho_i \int_0^\tau [\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}_H(t)]^{\otimes 2} Y_i(t)\, dt.$$

It has been already shown in the proof of Theorem 4.3 that $\widehat{\mathbb{D}}_H$ is consistent.

The most difficult task is estimating $\Sigma_H(\boldsymbol{\beta}_0)$, the extra pseudoscore covariance matrix. In order to estimate it consistently, we need to define a uniformly consistent estimator for the cumulative baseline hazard Λ_0. We propose using the estimator

$$\widehat{\Lambda}_0(t) = \int_0^t \frac{\sum_{i=1}^n dN_i(s)}{\sum_{j=1}^n Y_j(s)} - \int_0^t \widehat{\boldsymbol{\beta}}_H^{\mathsf{T}} \overline{\boldsymbol{Z}}_H(s)\, ds, \tag{4.15}$$

which mimics the full-data estimator proposed by Lin and Ying (1992). This estimator indeed possesses the desired uniform consistency:

Lemma 4.5. $\sup_{0\le t\le\tau} \left|\widehat{\Lambda}_0(t) - \Lambda_0(t)\right| \to_p 0.$

Proof. We can write $\widehat{\Lambda}_0(t) - \Lambda_0(t)$ as

$$\widehat{\Lambda}_0(t) - \Lambda_0(t) = \int_0^t \frac{\sum_i dN_i(s)}{\sum_j Y_j(s)} - \int_0^t \boldsymbol{\beta}_0^{\mathsf{T}} \overline{\boldsymbol{Z}}(s)\, ds - \int_0^t \frac{\sum_i Y_i(s) d\Lambda_0(s)}{\sum_j Y_j(s)} \tag{4.16}$$

$$+ \int_0^t \boldsymbol{\beta}_0^{\mathsf{T}} [\overline{\boldsymbol{Z}}(s) - \overline{\boldsymbol{Z}}_H(s)]\, ds \tag{4.17}$$

$$+ \int_0^t (\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_H)^{\mathsf{T}} \overline{\boldsymbol{Z}}_H(s)\, ds. \tag{4.18}$$

Now, (4.16) is equal to

$$\sum_{i=1}^{n} \int_0^t \left[\frac{dN_i(s)}{\sum_j Y_j(s)} - \frac{Y_i(s)\boldsymbol{\beta}_0^{\mathsf{T}} \boldsymbol{Z}_i \, ds}{\sum_j Y_j(s)} - \frac{Y_i(s)\, d\Lambda_0(s)}{\sum_j Y_j(s)} \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \int_0^t \frac{1}{n^{-1} \sum_j Y_j(s)} \, dM_i(s) \to_p 0$$

uniformly in t. Proceeding to (4.17),

$$\left| \int_0^t \boldsymbol{\beta}_0^{\mathsf{T}} [\overline{\boldsymbol{Z}}(s) - \overline{\boldsymbol{Z}}_H(s)] \, ds \right| \leq \tau \boldsymbol{\beta}_0^{\mathsf{T}} \left[\sup_{0 \leq t \leq \tau} |\overline{\boldsymbol{Z}}(t) - \boldsymbol{e}(t)| + \sup_{0 \leq t \leq \tau} |\overline{\boldsymbol{Z}}_H(t) - \boldsymbol{e}(t)| \right] \to_p 0.$$

Finally, the integral in (4.18) converges to a finite constant and $\widehat{\boldsymbol{\beta}}_H$ is consistent: hence (4.18) converges to zero in probability uniformly in t. □

We are in a position to introduce an estimator for $\Sigma_H(\boldsymbol{\beta}_0)$ and prove its consistency when the covariates are bounded. This is done in the last theorem.

Theorem 4.4. *Let*

$$\widehat{\boldsymbol{S}}_i(\boldsymbol{\beta}) = \int_0^\tau [\boldsymbol{Z}_i - \boldsymbol{e}(t)] Y_i(t) [d\widehat{\Lambda}_0(t) + \boldsymbol{Z}_i^{\mathsf{T}} \boldsymbol{\beta} \, dt],$$

where $\widehat{\Lambda}_0$ *is defined by* (4.15). *Let*

$$\widehat{\Sigma}_H(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1 - p_i}{p_i^2} \xi_i (1 - \Delta_i) \widehat{\boldsymbol{S}}_i^{\otimes 2}(\boldsymbol{\beta}).$$

If there exists a constant M *such that* $\|\boldsymbol{Z}_i\| < M$ *for all* i, *then* $\widehat{\Sigma}_H(\widehat{\boldsymbol{\beta}}_H) \to_p \Sigma_H(\boldsymbol{\beta}_0)$.

To prove consistency of $\widehat{\Sigma}_H$, we need the following technical lemma which shows that stochastic integrals of certain form converge to zero in probability.

Lemma 4.6. *Let* $H(s,t)$ *be a bounded deterministic function of bounded variation in both arguments, defined on the set* $[0,\tau]^2$, $\tau < \infty$. *Let* $K_{1n}(t)$ *and* $K_{2n}(t)$, $0 \leq t \leq \tau$, *be stochastic processes with almost all paths of bounded variation for sufficiently large* n. *Suppose* $\sup_{0 \leq t \leq \tau} |K_{2n}(t)| \to_p 0$ *as* $n \to \infty$, $\sup_{0 \leq t \leq \tau} |K_{1n}(t)|$ *is bounded in probability, and* $\int_0^\tau H(s,x) \, dK_{1n}(s)$ *is bounded in probability for* $x = 0, \tau$. *Then*

$$\int_0^\tau \int_0^\tau H(s,t) \, dK_{1n}(s) dK_{2n}(t) \to_p 0.$$

Proof of Lemma 4.6. By Fleming and Harrington's (1991, p. 320) Theorem A.1.2 on integration by parts, for any right-continuous functions F, G of bounded variation on finite intervals,

$$F(t)G(t) - F(0)G(0) = \int_0^t F(x-)\,dG(x) + \int_0^t G(x)\,dF(x).$$

Suppressing the subscript n in K_{1n} and K_{2n} and using this theorem with $F(t) = K_2(t)$ and $G(t) = \int_0^\tau H(s,t)\,dK_1(s)$, we get

$$\int_0^\tau \int_0^\tau H(s,t)\,dK_1(s)\,dK_2(t) = K_2(\tau)\int_0^\tau H(s,\tau)\,dK_1(s) - K_2(0)\int_0^\tau H(s,0)\,dK_1(s) - \int_0^\tau \int_0^\tau K_2(t-)H(s,dt)\,dK_1(s).$$

The first two terms converge to zero in probability since both $K_2(\tau)$ and $K_2(0)$ converge to zero and the integrals involved are bounded. The third term can be decomposed by applying the integration by parts theorem once more, this time with $F(s) = K_1(s)$ and $G(s) = \int_0^\tau K_2(t-)H(s,dt)$. Thus we get

$$\int_0^\tau \int_0^\tau K_2(t-)H(s,dt)\,dK_1(s) = K_1(\tau)\int_0^\tau K_2(t-)H(\tau,dt) - K_1(0)\int_0^\tau K_2(t-)H(0,dt) - \int_0^\tau \int_0^\tau K_1(s-)K_2(t-)\,H(ds,dt).$$

When H is increasing in both arguments, this is bounded by

$$\sup_{0\le t\le\tau} |K_1(t)| \sup_{0\le t\le\tau} |K_2(t)|\,|2H(\tau,\tau) + H(0,\tau)|,$$

which converges to zero because K_2 converges to zero and K_1 is bounded in probability, both uniformly in time. Otherwise, H can be written as a difference of two increasing functions of s at each fixed t and vice versa. This can be used to obtain a similar upper bound that converges to zero. □

Proof of Theorem 4.4. To prove the theorem, we will assume without loss of generality that $\boldsymbol{Z}_i = Z_i$ is a scalar time-independent covariate. We have

$$\Sigma_H(\beta_0) - \widehat{\Sigma}_H(\widehat{\beta}_H) = \frac{1}{n}\sum_{i=1}^n \frac{1-p_i}{p_i}(1-\Delta_i)\left[S_i^2(\beta_0) - \frac{\xi_i}{p_i}\widehat{S}_i^2(\widehat{\beta}_H)\right] + o_P(1).$$

The approximation on the right-hand side can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1-p_i}{p_i}(1-\Delta_i)\frac{\xi_i}{p_i}\left[\widehat{S}_i^2(\widehat{\beta}_H)-S_i^2(\beta_0)\right]-\frac{1}{n}\sum_{i=1}^{n}\frac{1-p_i}{p_i}(1-\Delta_i)\left(1-\frac{\xi_i}{p_i}\right)S_i^2(\beta_0).$$

The second term is an average of independent and identically distributed zero-mean terms, whence it must converge to zero in probability.

Let us introduce some convenient notation. With $a_i = \sqrt{(1-\Delta_i)\xi_i(1-p_i)/p_i^2}$, let

$$G_i(t) = a_i[Z_i - e(t)]Y_i(t), \quad \text{and} \quad \widehat{G}_i(t) = a_i[Z_i - \overline{Z}_H(t)]Y_i(t).$$

Notice that $G_i(t)$, $i = 1,\ldots,n$, are iid random variables and that $G_i(t)$ is bounded. We have to prove that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\frac{1-p_i}{p_i}(1-\Delta_i)\frac{\xi_i}{p_i}\left[\widehat{S}_i^2(\widehat{\beta}_H)-S_i^2(\beta_0)\right]\\&\quad=\frac{1}{n}\sum_{i=1}^{n}\left(\left\{\int_0^\tau \widehat{G}_i(t)[d\widehat{\Lambda}_0(t)+Z_i\widehat{\beta}_H\,dt]\right\}^2-\left\{\int_0^\tau G_i(t)[d\Lambda_0(t)+Z_i\beta_0\,dt]\right\}^2\right)\end{aligned}$$

converges to zero in probability. Obviously,

$$\begin{aligned}&\left\{\int_0^\tau \widehat{G}_i(t)[d\widehat{\Lambda}_0(t)+Z_i\widehat{\beta}_H\,dt]\right\}^2\\&\qquad=\int_0^\tau\int_0^\tau \widehat{G}_i(s)\widehat{G}_i(t)[d\widehat{\Lambda}_0(s)+Z_i\widehat{\beta}_H\,ds][d\widehat{\Lambda}_0(t)+Z_i\widehat{\beta}_H\,dt]\\&\qquad=\int_0^\tau\int_0^\tau \widehat{G}_i(s)\widehat{G}_i(t)\,d\widehat{\Lambda}_0(s)d\widehat{\Lambda}_0(t)+2\int_0^\tau\int_0^\tau \widehat{G}_i(s)\widehat{G}_i(t)\,d\widehat{\Lambda}_0(s)Z_i\widehat{\beta}_H\,dt\\&\qquad\quad+\int_0^\tau\int_0^\tau \widehat{G}_i(s)\widehat{G}_i(t)Z_i^2\widehat{\beta}_H^2\,ds\,dt,\end{aligned}$$

and $\left\{\int_0^\tau G_i(t)[d\Lambda_0(t)+Z_i\beta_0\,dt]\right\}^2$ can be decomposed in the same way. So, we need

to show that the following three expressions,

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\int_0^\tau \widehat{G}_i(s)\widehat{G}_i(t)\,d\widehat{\Lambda}_0(s)d\widehat{\Lambda}_0(t) - \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\int_0^\tau G_i(s)G_i(t)\,d\Lambda_0(s)d\Lambda_0(t). \tag{4.19}$$

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\int_0^\tau \widehat{G}_i(s)\widehat{G}_i(t)Z_i\widehat{\beta}_H\,d\widehat{\Lambda}_0(s)\,dt - \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau\int_0^\tau G_i(s)G_i(t)Z_i\beta_0\,d\Lambda_0(s)\,dt. \tag{4.20}$$

$$\int_0^\tau\int_0^\tau \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{G}_i(s)\widehat{G}_i(t) - G_i(s)G_i(t)\right] Z_i^2\widehat{\beta}_H^2\,ds\,dt \tag{4.21}$$

all converge to zero in probability.

Let us start with (4.19). Denote $H(s,t) \equiv \mathrm{E}\,G_i(s)G_i(t)$. Then (4.19) is just

$$\int_0^\tau\int_0^\tau H(s,t)[d\widehat{\Lambda}_0(s)d\widehat{\Lambda}_0(t) - d\Lambda_0(s)d\Lambda_0(t)] \tag{4.22}$$

$$+\int_0^\tau\int_0^\tau \left[\frac{1}{n}\sum_{i=1}^{n}\widehat{G}_i(s)\widehat{G}_i(t) - H(s,t)\right] d\widehat{\Lambda}_0(s)d\widehat{\Lambda}_0(t) \tag{4.23}$$

$$-\int_0^\tau\int_0^\tau \left[\frac{1}{n}\sum_{i=1}^{n}G_i(s)G_i(t) - H(s,t)\right] d\Lambda_0(s)d\Lambda_0(t). \tag{4.24}$$

Let us first verify that

$$\sup_{0\le s,t\le\tau}\left|\frac{1}{n}\sum_{i=1}^{n}\widehat{G}_i(s)\widehat{G}_i(t) - H(s,t)\right| \to_p 0. \tag{4.25}$$

Using

$$\begin{aligned}&\widehat{G}_i(s)\widehat{G}_i(t) - G_i(s)G_i(t)\\ &\quad= (\widehat{G}_i - G_i)(s)\,(\widehat{G}_i - G_i)(t) + G_i(s)\,(\widehat{G}_i - G_i)(t) + G_i(t)\,(\widehat{G}_i - G_i)(s),\end{aligned} \tag{4.26}$$

and realizing that $(\widehat{G}_i - G_i)(t) = a_i[\overline{Z}_H(t) - e(t)]Y_i(t)$, we get

$$\begin{aligned}
&\sup_{0\le s,t\le\tau}\left|\frac{1}{n}\sum\left[\widehat{G}_i(s)\widehat{G}_i(t) - G_i(s)G_i(t)\right]\right| \\
&\qquad = \sup_{0\le s,t\le\tau}\Bigg|[\overline{Z}_H(s) - e(s)][\overline{Z}_H(t) - e(t)]\,\frac{1}{n}\sum a_i^2 Y_i(s)Y_i(t) \\
&\qquad\quad + [\overline{Z}_H(t) - e(t)]\,\frac{1}{n}\sum a_i G_i(s)Y_i(t) + [\overline{Z}_H(s) - e(s)]\,\frac{1}{n}\sum a_i G_i(t)Y_i(s)\Bigg| \\
&\qquad \le K_0^2 \sup_{0\le t\le\tau}|\overline{Z}_H(t) - e(t)|^2 + 2K_0 \sup_{0\le t\le\tau}|\overline{Z}_H(t) - e(t)| \sup_{0\le t\le\tau}\frac{1}{n}\sum|G_i(t)|.
\end{aligned}$$

where K_0 is the constant that bounds a_i. The right-hand side converges to zero by means of Lemma 4.1 and boundedness of $G_i(t)$.

For fixed s and t, $G_i(s)G_i(t)$ are iid random variables. Hence their sample average converges to $\mathrm{E}\,G_i(s)G_i(t)$. The convergence must be uniform in s and t because $G_i(s)G_i(t)$ is uniformly bounded over all $s, t \in [0, \tau]$. Thus, the convergence to zero of (4.25) is verified.

By (4.25) and Condition 2.1, both (4.23) and (4.24) converge to 0. Proving that (4.19) converges to 0 is now equivalent to showing that (4.22) converges to zero. Since H is symmetric in its arguments, a decomposition analogous to (4.26) applied to $d\widehat{\Lambda}_0(s)d\widehat{\Lambda}_0(t) - d\Lambda_0(s)d\Lambda_0(t)$ leads to an upper bound for (4.22) of the form

$$\begin{aligned}
&\left|\int_0^\tau\int_0^\tau H(s,t)[d\widehat{\Lambda}_0(s) - d\Lambda_0(s)]\,[d\widehat{\Lambda}_0(t) - d\Lambda_0(t)]\right| \\
&\qquad\qquad + 2\left|\int_0^\tau\int_0^\tau H(s,t)\,d\Lambda_0(s)[d\widehat{\Lambda}_0(t) - d\Lambda_0(t)]\right|.
\end{aligned}$$

Lemma 4.6 shows that both these terms converge to zero in probability. We take $K_2(t) = \widehat{\Lambda}_0(t) - \Lambda_0(t)$ and set $K_1(s)$ to either $\widehat{\Lambda}_0(s) - \Lambda_0(s)$ or $\Lambda_0(s)$. The condition that $\int_0^\tau H(s,x)\,dK_1(s)$ should be bounded in probability is fulfilled since $\Lambda_0(s)$ is finite and deterministic and $\widehat{\Lambda}_0(s) - \Lambda_0(s)$ itself converges to zero in probability. Lemma 4.6 is thus applicable. The convergence in probability to zero of (4.22), and hence of (4.19), is verified.

The remaining terms, (4.20) and (4.21), may be treated in exactly the same way. Since they both tend to zero, consistency of the extra pseudoscore variance estimator $\widehat{\Sigma}_H$ follows. □

4.4 *Selection of subcohort sampling probabilities*

Simple random sampling of the subcohort, which is defined by $p_i = \alpha$, $i = 1, \ldots, n$, is rarely the best choice for the case-cohort design. In this section we identify two cases where efficiency can be increased by oversampling certain subpopulations. This is made possible by letting the sampling probabilities p_i depend on a covariate observed during the first phase of the experiment, which is available for all subjects.

Let W be any such a first-phase covariate and let W_i be its individual observed values. W itself may or may not be a part of $\boldsymbol{Z}$. If it is not, we assume that W is a true surrogate, i.e., given $\boldsymbol{Z}$, W is conditionally independent of both survival and censoring times. Let the selection probabilities depend on the observed values of W by $\mathrm{P}[\xi_i = 1 \mid W_i] = p(W_i)$.

Suppose a cohort study is conducted to assess the effect of a certain binary exposure on survival. Suppose further that the death rate is low, so that the case-cohort design is feasible, and that the exposure is rare and expensive to ascertain precisely. A subcohort selected by simple random sampling would contain a relatively few exposed subjects. However, there sometimes exists another binary variable (a surrogate exposure) that is correlated with the true exposure and can be measured quite easily on all subjects. If the sampling probabilities are altered so that the subcohort is approximately balanced in terms of the surrogate exposure, it will be also more balanced in terms of the true exposure. Hence, the variance of the AH parameter estimator for the effect of the exposure will decrease.

How can be such a design implemented in practice? Let Z be a binary exposure covariate with $\mathrm{P}[Z = 1] = p_Z$ and let W be a binary surrogate for Z observable

during the first phase of the study. The association of W with Z can be described in terms of sensitivity $\eta = \mathrm{P}[W = 1 \mid Z = 1]$ and specificity $\nu = \mathrm{P}[W = 0 \mid Z = 0]$. If $\eta + \nu = 1$, W is independent of Z and does not carry any information about the exposure whatsoever. Therefore we assume, without loss of generality, that $\eta + \nu > 1$. The larger $\eta + \nu$ is, the "closer" is W to Z. Denote $p_W = \mathrm{P}[W_i = 1]$; we have $p_W = (1 - \nu)(1 - p_Z) + \eta p_Z$. To balance the subcohort in W, we apply the stratified sampling design described in the following paragraph.

Denote the subcohort sampling probabilities by $\alpha_0 = \mathrm{P}[\xi_i = 1 \mid W_i = 0]$ and $\alpha_1 = \mathrm{P}[\xi_i = 1 \mid W_i = 1]$. Let α be the desired overall proportion of the first-phase subjects selected into the subcohort and suppose that $\alpha < 0.5$. Then we define α_0 and α_1 as follows:

$$\alpha_0 = \begin{cases} (\alpha - p_W)/(1 - p_W) & \text{if } p_W < \alpha/2, \\ \alpha/(2 - 2p_W) & \text{if } \alpha/2 \le p_W \le 1 - \alpha/2, \\ 1 & \text{if } p_W > 1 - \alpha/2, \end{cases} \tag{4.27}$$

and

$$\alpha_1 = \begin{cases} 1 & \text{if } p_W < \alpha/2, \\ \alpha/(2p_W) & \text{if } \alpha/2 \le p_W \le 1 - \alpha/2, \\ (\alpha - 1 + p_W)/p_W & \text{if } p_W > 1 - \alpha/2. \end{cases} \tag{4.28}$$

These sampling probabilities define a stratified sampling design where the within-strata sample sizes are both random and the proportion of subjects sampled in each stratum converges to a fixed number.

It is easy to see that

$$\mathrm{P}[\xi_i = 1] = \alpha_0 \mathrm{P}[W_i = 0] + \alpha_1 \mathrm{P}[W_i = 1] = \alpha,$$

and hence the average subcohort proportion is indeed α. In addition, if $\alpha/2 \le p_W \le 1 - \alpha/2$, then the subcohort is balanced in W:

$$\mathrm{P}[W_i = 1 \mid \xi_i = 1] = \frac{\mathrm{P}[\xi_i = 1 \mid W_i = 1]\mathrm{P}[W_i = 1]}{\mathrm{P}[\xi_i = 1]} = \frac{\alpha/2}{\alpha} = \frac{1}{2}.$$

It follows that the subcohort is also more balanced in the true covariate Z: the proportion of subcohort subjects with $Z_i = 1$ exceeds p_Z and gets closer to 0.5. In the next sections, we demonstrate how large is the efficiency gain achieved by the stratified subcohort sampling design.

This idea can be also used with continuous exposures. Based on the information about the surrogate exposure, subjects whose true exposures are likely to fall into the extremes or into underrepresented ranges should be sampled with a higher probability.

Another interesting application arises when the expensive covariate is a confounder and there is a surrogate measure for it. Then the sampling probabilities may be set so that the distribution of the confounder within the subcohort is as similar as possible to that within the failures. This idea resembles matching: the parameter estimate for the confounder will be less precise but efficiency will be gained for the estimation of the effect of the main covariate. In the setting of the Cox model, a subcohort sampling design that makes use of this idea was proposed by Kim and DeGruttola (1996). These authors applied this approach to the analysis of an AIDS clinical trial to evaluate the effect of $CD4^+$ counts on the time to progression of AIDS.

4.5 *Asymptotic relative efficiency*

In this section, we address two important questions about the relative efficiency of the case-cohort estimator. First, compared to the full-data estimator, does the AH case-cohort estimator achieve an asymptotic relative efficiency which is comparable to the Cox case-cohort estimator? And second, what is the efficiency gain for the AH case-cohort estimator when the stratified sampling design described in the previous section is applied? We will show that the limiting variance of the AH case-cohort estimator can be calculated for the special case where there is only one binary covariate, the baseline hazard is constant and there is no censoring until the end of the study. We will calculate the limiting variance and use it to compare asymptotic relative

efficiencies.

4.5.1 Preliminaries

Let us return to the situation described in Section 4.4: Z is a binary covariate with $\mathrm{P}[Z = 1] = p_Z$ and W is a binary surrogate with sensitivity $\eta = \mathrm{P}[W = 1 \mid Z = 1]$ and specificity $\nu = \mathrm{P}[W = 0 \mid Z = 0]$. Let $\eta + \nu > 1$. Now, let the hazard for the failure time T be

$$\lambda(t \mid Z) = \begin{cases} \lambda & \text{if } Z = 0; \\ \lambda + \beta & \text{if } Z = 1. \end{cases}$$

The hazard is constant over time; hence T has an exponential distribution with the mean $\lambda^{-1}(t \mid Z)$. The model can be regarded either as the AH model, where β is estimated directly, or as the Cox model, where the estimated parameter is the log relative risk (the relative risk ϱ is given by $\varrho = \beta/\lambda + 1$). Hence this is a suitable example for comparing the AH case-cohort estimator with the Cox case-cohort estimator. In addition, the asymptotic variances of the estimators can be relatively easily calculated.

Let the time interval on which the data are observed be restricted to $[0, 1]$ and let $\mathrm{P}[C = 1] = 1$, i.e., let the censoring distribution be concentrated at 1. This means no censoring occurs until the end of the study. A calculation presented in Section 4.5.4 shows that, under these assumptions, the limiting variance of the full-data AH estimator $\widehat{\beta}_A$ is

$$\Sigma_A(\beta) = p_Z(1 - p_Z) \int_0^1 \frac{(\lambda + \beta)(1 - p_Z) + \lambda p_Z \, \mathrm{e}^{-\beta t}}{(p_Z \, \mathrm{e}^{-\beta t} + 1 - p_Z)^2} \, \mathrm{e}^{-(\lambda+\beta)t} \, dt \tag{4.29}$$

for $\beta \neq 0$, and

$$\Sigma_A(0) = p_Z(1 - p_Z) \left(1 - \mathrm{e}^{-\lambda}\right) \tag{4.30}$$

for $\beta = 0$. Given arbitrary subcohort selection probabilities $\alpha_0 = \mathrm{P}[\xi_i = 1 \mid W_i = 0]$ and $\alpha_1 = \mathrm{P}[\xi_i = 1 \mid W_i = 1]$, let us define two constants.

$$K_0 = \frac{1-\alpha_0}{\alpha_0}\nu + \frac{1-\alpha_1}{\alpha_1}(1-\nu)$$

and

$$K_1 = \frac{1-\alpha_0}{\alpha_0}(1-\eta) + \frac{1-\alpha_1}{\alpha_1}\eta.$$

Then the extra pseudoscore variance $\Sigma_H(\beta)$ equals

$$\Sigma_H(\beta) = K_0(1-p_Z)\lambda^2 \mathrm{e}^{-\lambda}\beta^{-2}\ln^2(p_Z \mathrm{e}^{-\beta} + 1 - p_Z) + K_1 p_Z(\lambda+\beta)^2 \mathrm{e}^{-(\lambda+\beta)}[1 + \beta^{-1}\ln(p_Z \mathrm{e}^{-\beta} + 1 - p_Z)]^2 \quad (4.31)$$

for $\beta \neq 0$, and

$$\Sigma_H(0) = p_Z(1-p_Z)[p_Z K_0 + (1-p_Z)K_1]\lambda^2 \mathrm{e}^{-\lambda} \quad (4.32)$$

for $\beta = 0$ (see Section 4.5.4).

4.5.2 Comparing the efficiency of the AH and Cox case-cohort estimators

In this section, we evaluate and compare the asymptotic relative efficiencies of the AH model and the Cox model case-cohort estimators. The ARE should be understood relative to the full-data estimators in the respective models. We work under the assumptions summarized in Section 4.5.1: a single binary covariate, constant hazard, and no censoring until the end of the study. For simplicity, we assume that the subcohort is selected by independent Bernoulli sampling. So, let $\Sigma_{H0}(\beta)$ be the extra AH pseudoscore variance for simple random sampling of the subcohort, i.e., $\Sigma_{H0}(\beta)$ is given by (4.31) and (4.32) where $\alpha_0 = \alpha_1 = \alpha$. This choice reduces both K_0 and K_1 to $(1-\alpha)/\alpha$. Then the asymptotic relative efficiency (ARE) of the AH case-cohort estimator with simple random sampling of the subcohort compared to the full-data

AH estimator is

$$ARE_{A0}(\beta) = \frac{\Sigma_A(\beta)}{\Sigma_A(\beta) + \Sigma_{H0}(\beta)}.$$

When $\beta = 0$,

$$ARE_{A0}(0) = \frac{1 - e^{-\lambda}}{1 - [1 - (1-\alpha)/\alpha\lambda^2]\, e^{-\lambda}}.$$

The ARE of the analogous Cox case-cohort estimator with independent Bernoulli sampling of the subcohort (that is, the estimator based on the score defined by (3.1) and (3.2) with $p_i = \alpha$) is given by

$$ARE_{C0}(\beta) = \frac{\Sigma_{PH}(\beta)}{\Sigma_{PH}(\beta) + \Sigma_{HC}(\beta)},$$

where Σ_{PH} is the variance of the full-data Cox partial likelihood score and Σ_{HC} is the extra score variance due to missing data. Denote the relative risk by ϱ, where $\varrho = \beta/\lambda + 1$. Then it can be shown that

$$\Sigma_{PH}(\beta) = \varrho p_Z(1 - p_Z) \int_0^\lambda \frac{e^{-(\varrho+1)x}}{\varrho p_Z\, e^{-\varrho x} + (1 - p_Z)\, e^{-x}}\, dx$$

for $\beta \neq 0$ ($\varrho \neq 1$), and $\Sigma_{PH}(0) = p_Z(1 - p_Z)(1 - e^{-\lambda})$ for $\beta = 0$ ($\varrho = 1$). The extra score variance can be written as

$$\Sigma_{HC}(\beta) = \varrho^2 p_Z \frac{1-\alpha}{\alpha} \left[p_Z(1 - p_Z)\, e^{-\lambda}\, I_1^2 + e^{-\varrho\lambda}(\lambda - \varrho p_Z I_1)^2\right]$$

for $\beta \neq 0$, where

$$I_1 = \int_0^\lambda \frac{e^{-\varrho x}}{\varrho p_Z\, e^{-\varrho x} + (1 - p_Z)\, e^{-x}}\, dx.$$

For $\beta = 0$,

$$\Sigma_{HC}(0) = p_Z(1 - p_Z) \frac{1-\alpha}{\alpha} \lambda^2\, e^{-\lambda}.$$

We can immediately notice that $\Sigma_{PH}(0) = \Sigma_A(0)$ and $\Sigma_{HC}(0) = \Sigma_H(0)$; so under the null hypothesis, the limiting variances of the AH and Cox case-cohort estimators

coincide. Hence, they have the same asymptotic relative efficiencies with respect to full-data estimators when $\beta = 0$.

The formulae for $\Sigma_{PH}(\beta)$ and $\Sigma_{HC}(\beta)$ can be derived in the same way as $\Sigma_A(\beta)$ and $\Sigma_H(\beta)$ for the AH model (see Section 4.5.4). Self and Prentice (1988) calculated the limiting variance of their Cox case-cohort estimator under the same assumptions. However, their formulae are different on two counts. First, their estimator is different, because they do not include the failures in estimating the weighted covariate average $\overline{Z}(t, \beta)$. Second, their formula for Σ_{HC} is correct only for $\beta = 0$ or $\beta = \ln 2$ and has to be recalculated in order to give the correct extra score variance for a general β.

Tables 4.1 and 4.2 show the asymptotic relative efficiencies ARE_{C0} and ARE_{A0} of the Cox and AH case-cohort estimators relatively to their respective full-data counterparts. The efficiencies are calculated for independent Bernoulli sampling of the subcohort. The proportion p_Z of exposed subjects is set to 0.1 or 0.5. The baseline hazard λ is calculated so that the overall proportion of deaths p_d is 0.01, 0.10 or 0.20. The death rates p_d span the range where conducting a case-cohort study may lead to substantial savings in covariate assessment cost. The subcohort consists of p_d, $2p_d$ or $4p_d$ subjects, so that the approximate case-control ratios are 1:1, 1:2 and 1:4, respectively. The covariate effects correspond to relative risk values of 1, 2 and 3.

When the exposure has no effect on failure time ($\varrho = 1$), the ARE's are the same for the AH model as for the Cox model, as indicated earlier. However, with increasing magnitude of covariate effect, differences between the ARE's start to emerge. When the exposure is common (Table 4.1 with $p_Z = 0.5$), ARE for the Cox model increases with increasing magnitude of the covariate effect, while ARE for the AH model decreases. For $\rho = 2$, the differences are still small; for $\rho = 3$ they may get as large as 0.13. The picture changes when the exposure is relatively rare (Table 4.2 with $p_Z = 0.1$). There, both ARE of the Cox model and ARE of the AH model decrease as the covariate effect gets larger. Moreover, the differences between the two seem to be relatively small throughout the table.

Table 4.1: ARE of the Cox and AH case-cohort estimators with respect to their full-data counterparts under $p_Z = 0.5$.

$\frac{\alpha}{p_d}$	Model	$p_d = 0.01$			$p_d = 0.10$			$p_d = 0.20$		
		$\varrho = 1$	$\varrho = 2$	$\varrho = 3$	$\varrho = 1$	$\varrho = 2$	$\varrho = 3$	$\varrho = 1$	$\varrho = 2$	$\varrho = 3$
1	Cox	0.503	0.532	0.573	0.527	0.553	0.591	0.557	0.580	0.613
	AH	0.503	0.476	0.447	0.527	0.500	0.471	0.557	0.531	0.502
2	Cox	0.671	0.696	0.731	0.714	0.736	0.765	0.770	0.786	0.809
	AH	0.671	0.647	0.620	0.714	0.693	0.667	0.770	0.751	0.729
4	Cox	0.806	0.824	0.847	0.870	0.881	0.897	0.953	0.957	0.962
	AH	0.806	0.789	0.769	0.870	0.857	0.842	0.953	0.948	0.942

Notes: p_d is the overall probability of death; ϱ is the relative risk; α/p_d is the ratio of subcohort size to number of deaths.

Table 4.2: ARE of the Cox and AH case-cohort estimators with respect to their full-data counterparts under $p_Z = 0.1$.

$\frac{\alpha}{p_d}$	Model	$p_d = 0.01$			$p_d = 0.10$			$p_d = 0.20$		
		$\varrho = 1$	$\varrho = 2$	$\varrho = 3$	$\varrho = 1$	$\varrho = 2$	$\varrho = 3$	$\varrho = 1$	$\varrho = 2$	$\varrho = 3$
1	Cox	0.503	0.380	0.327	0.527	0.410	0.359	0.557	0.449	0.403
	AH	0.503	0.364	0.294	0.527	0.395	0.328	0.557	0.435	0.374
2	Cox	0.671	0.553	0.496	0.714	0.610	0.558	0.770	0.685	0.642
	AH	0.671	0.536	0.457	0.714	0.595	0.523	0.770	0.673	0.615
4	Cox	0.806	0.717	0.668	0.870	0.806	0.771	0.953	0.929	0.915
	AH	0.806	0.703	0.632	0.870	0.797	0.745	0.953	0.925	0.905

Notes: p_d is the overall probability of death; ϱ is the relative risk; α/p_d is the ratio of subcohort size to number of deaths.

Overall, it is surprising how close the ARE of the AH case-cohort estimator is to the ARE of the Cox case-cohort estimator, given that, unlike the Cox case-cohort score, the AH case-cohort pseudoscore loses all the contributions of non-cohort non-failures. There is virtually no difference when the covariate effect is small: the difference is largest when the covariate effect is large, the subcohort is small and the exposure is common. It is also interesting to notice how well the case-cohort design performs when the failure rate is small. The ARE's in both tables barely change when the death rate increases from 0.01 to 0.10; however, the estimator uses ten times as many subjects under $p_d = 0.1$ as under $p_d = 0.01$.

4.5.3 Efficiency of the AH case-cohort estimator under stratified subcohort sampling

In Section 4.4, we defined sampling probabilities that balance the subcohort in terms of a binary surrogate covariate. We claimed that such a stratified sampling design improves the efficiency of the case-cohort estimator, especially when the exposure rate is low. In the current section we evaluate the efficiency gain under the assumptions of Section 4.5.1.

Let $\Sigma_{H1}(\beta)$ be the extra pseudoscore variance given by (4.31) and (4.32), where α_0 and α_1 are defined in Section 4.4 by Equations (4.27) and (4.28). We continue to write $\Sigma_{H0}(\beta)$ for the extra pseudoscore variance under independent Bernoulli sampling of the subcohort. The asymptotic relative efficiency of the stratified case-cohort estimator relative to the case-cohort estimator with independent Bernoulli sampling is

$$ARE_{A1}(\beta) = \frac{\Sigma_A(\beta) + \Sigma_{H0}(\beta)}{\Sigma_A(\beta) + \Sigma_{H1}(\beta)}.$$

We evaluated $ARE_{A1}(\beta)$ in various settings and plotted them. The plots are shown in Figures 4.1–4.3 on pages 73–75. In Figure 4.1, the deaths-subcohort ratio is approximately 1:1; in Figure 4.2, it is 1:2; and in Figure 4.3 it is 1:4. The overal probability of death varies between 0.01, 0.1 and 0.2 within each column of the figures. The

relative risk varies between 1, 2 and 3 within each row. Each plot shows how ARE_{A1} changes with the proportion of exposed subjects p_Z when the sensitivity η and specificity ν of the surrogate W are high ($\eta = \nu = 0.9$), medium ($\eta = 0.7$, $\nu = 0.9$) or low ($\eta = \nu = 0.7$). The abrupt changes in slopes that can be observed in the lower rows of Figures 4.2 and 4.3 occur when the subcohort is so large and the exposure so rare that there are not enough subjects with $W_i = 1$ to make the subcohort balanced in the surrogate exposure (see the definitions of α_0 and α_1 in (4.27) and (4.28)).

The plots suggest that the efficiency gain is highest when the subcohort is small, the exposure is rare, the surrogate is precise, and the covariate effect is large. In some cases, the ARE may be as high as 1.9. But in all cases, some efficiency is gained; this is true even if the surrogate is imprecise, the exposure is only moderately rare and the covariate effect is weak. This is important: since the application of stratified subcohort sampling does not increase the complexity of either the study or its analysis, any gain in efficiency is essentially a free windfall. So, even a small gain is quite satisfactory; the better if the gain is substantial in some cases.

4.5.4 Calculation of asymptotic variances

In this section, we explain how the asymptotic variance formulae (4.29)–(4.32) can be calculated. We work under the assumptions introduced in Section 4.5.1 to evaluate

$$\Sigma_A(\beta_0) = \mathrm{E} \int_0^1 [Z_i - e(t)]^2 Y_i(t)[d\Lambda_0(t) + \beta_0 Z_i \, dt]$$

and

$$\Sigma_H(\beta_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{1 - p_i}{p_i} \mathrm{E}(1 - \Delta_i) S_i^2(\beta_0),$$

where

$$S_i(\beta_0) = \int_0^1 [Z_i - e(t)] Y_i(t)[d\Lambda_0(t) + \beta_0 Z_i \, dt].$$

Realizing that $Z_i^2 = Z_i$ and writing $\lambda_0(t)\,dt$ for $d\Lambda_o(t)$, we get

$$\begin{aligned}\Sigma_A(\beta_0) &= \int_0^1 \mathrm{E}[Z_i - 2Z_i e(t) + e^2(t)]Y_i(t)\lambda_0(t)\,dt \\ &\quad + \beta_0 \int_0^1 \mathrm{E}[Z_i - 2Z_i e(t) + e^2(t)]Z_i Y_i(t)\,dt \\ &= \int_0^1 e(t)[1-e(t)]\pi_0(t)\lambda_0(t)\,dt + \beta_0 \int_0^1 [1-e(t)]^2\pi_1(t)\,dt \\ &= \int_0^1 [1-e(t)]\pi_1(t)\{\beta_0[1-e(t)] + \lambda_0(t)\}.\end{aligned}$$

Since both Z_i and W_i are binary, we have

$$\begin{aligned}\Sigma_H(\beta_0) = &\sum_{k,l\in\{0,1\}} \frac{1-\alpha_l}{\alpha_l}\mathrm{E}\left[(1-\Delta_i)S_i^2(\beta_0)\,\middle|\, Z_i = k, W_i = l\right] \\ &\times \mathrm{P}\left[W_i = l \,|\, Z_i = k\right]\mathrm{P}[Z_i = k] \\ =&K_0(1-p_Z)\mathrm{E}\left[(1-\Delta_i)S_i^2(\beta_0)\,\middle|\, Z_i = 0\right] \\ &+ K_1 p_Z \mathrm{E}\left[(1-\Delta_i)S_i^2(\beta_0)\,\middle|\, Z_i = 1\right],\end{aligned} \tag{4.33}$$

where

$$\begin{aligned}K_0 =&\frac{1-\alpha_0}{\alpha_0}\mathrm{P}\left[W_i = 0 \,|\, Z_i = 0\right] + \frac{1-\alpha_1}{\alpha_1}\mathrm{P}\left[W_i = 1 \,|\, Z_i = 0\right] \\ =&\frac{1-\alpha_0}{\alpha_0}\nu + \frac{1-\alpha_1}{\alpha_1}(1-\nu),\end{aligned}$$

and

$$\begin{aligned}K_1 =&\frac{1-\alpha_0}{\alpha_0}\mathrm{P}\left[W_i = 0 \,|\, Z_i = 1\right] + \frac{1-\alpha_1}{\alpha_1}\mathrm{P}\left[W_i = 1 \,|\, Z_i = 1\right] \\ =&\frac{1-\alpha_0}{\alpha_0}(1-\eta) + \frac{1-\alpha_1}{\alpha_1}\eta.\end{aligned}$$

The hazard for failure at time t is $\lambda(t \mid Z) = \lambda + \beta_0 Z$. Hence, $\mathrm{P}\left[T > t \mid Z = 0\right] = \exp\{-\lambda t\}$ and $\mathrm{P}\left[T > t \mid Z = 1\right] = \exp\{-(\lambda+\beta_0)t\}$. The censoring variable attains the value 1 with probability 1. Thus,

$$\pi_0(t) = \mathrm{E}\,Y(t) = \mathrm{P}[T > t] = \mathrm{e}^{-\lambda t}(p_Z\,\mathrm{e}^{-\beta_0 t} + 1 - p_Z),$$

$$\pi_1(t) = \mathrm{E}\,ZY_i(t) = \mathrm{P}[T > t, Z = 1] = p_Z\,\mathrm{e}^{-(\lambda+\beta_0)t},$$

$$e(t) = \frac{\pi_1(t)}{\pi_0(t)} = \frac{p_Z\,\mathrm{e}^{-\beta_0 t}}{p_Z\,\mathrm{e}^{-\beta_0 t} + 1 - p_Z}, \quad \text{and} \quad 1 - e(t) = \frac{1-p_Z}{p_Z\,\mathrm{e}^{-\beta_0 t} + 1 - p_Z}.$$

It follows from the preceding calculations that

$$\Sigma_A(\beta_0) = \int_0^1 \frac{1-pz}{pz\,e^{-\beta_0 t}+1-pz}\left[\beta_0 pz\,e^{-(\lambda+\beta_0)t}\frac{1-pz}{pz\,e^{-\beta_0 t}+1-pz}+\lambda pz\,e^{-(\lambda+\beta_0)t}\right]dt$$
$$= pz(1-pz)\int_0^1 \frac{(\lambda+\beta_0)(1-pz)+\lambda pz\,e^{-\beta_0 t}}{(pz\,e^{-\beta_0 t}+1-pz)^2}\,e^{-(\lambda+\beta_0)t}\,dt.$$

When $\beta_0 = 0$, the integral above is easy to evaluate:

$$\Sigma_A(0) = pz(1-pz)\int_0^1 \frac{\lambda(1-pz)+\lambda pz}{1}\,e^{-\lambda t}\,dt = pz(1-pz)(1-e^{-\lambda}).$$

This verifies (4.29) and (4.30). We proceed to calculating the extra score variance $\Sigma_H(\beta_0)$. By (4.33), it remains to evaluate $\mathrm{E}\left[(1-\Delta_i)S_i^2(\beta_0)\mid Z_i\right]$. We do it in two steps:

$$\begin{aligned}
&\mathrm{E}\left[(1-\Delta_i)S_i^2(\beta_0)\mid Z_i=0\right]\\
&\quad=\mathrm{E}\left[(1-\Delta_i)\left\{\int_0^1[Z_i-e(t)]Y_i(t)(\lambda+\beta_0 Z_i)\,dt\right\}^2\,\middle|\,Z_i=0\right]\\
&\quad=\mathrm{E}\left[(1-\Delta_i)\left\{-\lambda\int_0^1 e(t)Y_i(t)\,dt\right\}^2\,\middle|\,Z_i=0\right]\\
&\quad=\lambda^2\int_0^1\int_0^1 e(s)e(t)\mathrm{E}\left[(1-\Delta_i)Y_i(s)Y_i(t)\mid Z_i=0\right]ds\,dt.
\end{aligned}$$

Since censoring can occur only at $t=1$, we have

$$\mathrm{E}\left[(1-\Delta_i)Y_i(s)Y_i(t)\mid Z_i\right] = \mathrm{P}\left[T>1\mid Z_i\right] = e^{-(\lambda+\beta_0 Z_i)},$$

and hence

$$\mathrm{E}\left[(1-\Delta_i)S_i^2(\beta_0)\mid Z_i=0\right] = \lambda^2 e^{-\lambda}\left[\int_0^1 e(t)\,dt\right]^2.$$

The integral can be calculated explicitly:

$$\begin{aligned}
\int_0^1 e(t)\,dt &= pz\int_0^1 \frac{e^{-\beta_0 t}dt}{pz\,e^{-\beta_0 t}+1-pz} = \frac{pz}{\beta_0}\int_{e^{-\beta_0}}^1 \frac{dx}{pzx+1-pz}\\
&= -\frac{1}{\beta_0}\ln(pz\,e^{-\beta_0}+1-pz)
\end{aligned}$$

for $\beta_0 \neq 0$ and $\int_0^1 e(t)\,dt = p_Z$ for $\beta_0 = 0$. Thus,

$$\mathrm{E}\left[(1-\Delta_i)S_i^2(\beta_0) \mid Z_i = 0\right] = \frac{\lambda^2 \mathrm{e}^{-\lambda}}{\beta_0^2} \ln^2(p_Z \mathrm{e}^{-\beta_0} + 1 - p_Z)$$

if $\beta_0 \neq 0$, and

$$\mathrm{E}\left[(1-\Delta_i)S_i^2(\beta_0) \mid Z_i = 0\right] = \lambda^2 \mathrm{e}^{-\lambda} p_Z^2$$

if $\beta_0 = 0$. Conditioning on $Z_i = 1$ instead on $Z_i = 0$, an analogous calculation reveals that

$$\mathrm{E}\left[(1-\Delta_i)S_i^2(\beta_0) \mid Z_i = 1\right] = (\lambda+\beta_0)^2 \mathrm{e}^{-(\lambda+\beta_0)} \left[1 + \frac{1}{\beta_0} \ln(p_Z \mathrm{e}^{-\beta_0} + 1 - p_Z)\right]^2$$

if $\beta_0 \neq 0$, and

$$\mathrm{E}\left[(1-\Delta_i)S_i^2(\beta_0) \mid Z_i = 1\right] = \lambda^2 \mathrm{e}^{-\lambda}(1-p_Z)^2$$

if $\beta_0 = 0$. Thus, Equations (4.31) and (4.32) have been verified.

4.6 Simulation study

We conducted a simulation study to illustrate how the AH case-cohort estimator works in settings that mimic the circumstances where the case-cohort estimator might be used in practice. We generated samples of 5,000 subjects with failure times distributed according to the model $\lambda_0(t \mid Z) = \lambda_0 + \beta_0 Z$, where Z was a binary exposure. So, the failure time distribution was exponential. The censoring distribution was uniform over the interval [0,2]; such a censoring is typical for studies with a sequential entry and no loss to follow-up. We chose the baseline hazard λ_0 so that the overall probability of failure was 0.05 or 0.1. The proportion of exposed subjects was varied between 0.05 and 0.2. The subcohort consisted of either the same proportion of subjects as have died (case-control ratio 1:1), or twice as many (case-control ratio 1:2). We also generated two surrogate exposures, one with sensitivity and specificity for

the true exposure $\eta = \nu = 0.7$ and one with $\eta = \nu = 0.9$. We used the surrogate exposures to apply the stratified subcohort sampling design, as proposed in Section 4.4. For each of the 1,000 samples we generated, four estimates were calculated: the full-data AH estimate (F); the case-cohort estimate with subcohort selected by independent Bernoulli sampling (CC); and two stratified case-cohort estimates with subcohort sampling probabilities defined by (4.27) and (4.28), based on the surrogates with $\eta = \nu = 0.7$ (SC-1) and $\eta = \nu = 0.9$ (SC-2). Like all simulation studies reported in this work, this study was programmed in Fortran 77 and run on a Unix SPARC workstation.

The results are given in three tables presented on pages 76–78: Table 4.3 for 5% deaths and the parameter to be estimated $\beta_0 = 0.1$; Tables 4.4 and 4.5 for 10% deaths and β_0 equal to 0.2 and 0, respectively. The means of the 1,000 simulated parameter estimates shown in the three tables suggest that the case-cohort estimate is usually very close to the true parameter value, although it may be slightly biased upwards. The estimated standard errors are close to the sample standard errors based on the simulated parameter estimates. The fact that the limiting variance estimator works quite well is also reflected in the coverage probabilities of the 95% confidence intervals based on the estimated parameters and estimated standard errors. The coverage probabilities range from 0.93 to 0.95 when the subcohort proportion is small (0.05), and are virtually identical to 0.95 when the subcohort proportion rises to 0.2. Comparing the three case-cohort estimators, the one stratifying on the more precise surrogate exposure (SC-2) has almost always much smaller standard errors than the simple CC estimator, so the efficiency gain suggested by the theoretical calculations presented in Section 4.5.3 is confirmed by the simulations. Even when the surrogate exposure is less precise (estimator SC-1), both the sample and estimated standard errors are smaller than those for the simple CC estimator, unless the subcohort is large and the exposure relatively common. All these results are consistent regardless of the magnitude of the covariate effect.

4.7 *Example: Analysis of NWTSG data*

We introduced the National Wilms Tumor Study as an illustrating example in Chapter 1. We noted there that unfavorable histology was an important prognostic factor for survival in Wilms tumor patients and that histology evaluated by the central pathologist was the true covariate, while the evaluation done by the institutions themselves could be regarded as a misclassified surrogate measure. We also mentioned that stage (I–IV) was another risk factor acting independently of histology. In this section, we apply the case-cohort design to analyze the NWTSG data. Since the actual data contain nearly complete covariate information on all subjects, we simulate a case-cohort study by drawing subcohorts at random from the complete data set.

We use data on 4335 subjects (1915 in NWTSG-3 and 2420 in NWTSG-4) with complete survival information. The median follow-up time for this sample was 5.6 years and the death rate was 0.099. The tumor stage was distributed as follows: 39.1% subjects were Stage I, 25.6% were Stage II, 23.4% were Stage III, and 11.9% were Stage IV. Unfavorable central histology was found on 11.4% of the subjects. The sensitivity of unfavorable institutional histology for unfavorable central histology was 0.72, the specificity was 0.98. This means that the local pathologists who performed the institutional evaluations missed almost 30% of tumors with truly unfavorable histology.

We first fitted the full-data AH model to the whole cohort of 4335 subjects. We included five covariates in the model: an indicator of unfavorable central histology, indicators of Stages II–IV (as opposed to Stage I), and an indicator of NWTSG-4 (as opposed to NWTSG-3). The inclusion of the last covariate is motivated by the fact that the treatment of Wilms tumor improved in time and so NWTSG-4 patients had a lower death rate than NWTSG-3 patients. To generate the case-cohort design, we drew 500 random subcohorts from the total sample by independent Bernoulli sampling, and 500 subcohorts by stratified random sampling, balanced on

the institutional histology (the surrogate exposure). The subcohorts consisted of 10%, 20% and 30% of the total sample; that corresponds to approximate case-control ratios of 1:1, 1:2 and 1:3, respectively.

Table 4.6 on page 79 summarizes the results. The first row gives the full-data AH estimates of the five parameters and their estimated standard errors. All of them are significant at the 5% level. The estimate for unfavorable central histology was 0.0677, which means that the patients with unfavorable histology had on average 6.8 more deaths per 100 person-years of follow-up than the patients with favorable histology. Similarly, the remaining four parameters estimate the excess deaths due to Stages II–IV and deaths prevented thanks to improved treatment in NWTSG-4. The rest of the table shows the averages of the case-cohort estimates over the 500 simulated subcohorts and their averaged standard error estimates; the size of the subcohort varies as indicated above and the method alternates between Bernoulli and stratified sampling. All the averaged case-cohort parameter estimates are close to the full-data estimates. The standard errors are generally larger, but they get smaller as the subcohort size increases. When stratified sampling is used instead of Bernoulli sampling, the standard errors for unfavorable histology get much smaller and closer to the full-data standard errors. Stratification has little or no effect on the standard errors of the other parameter estimates.

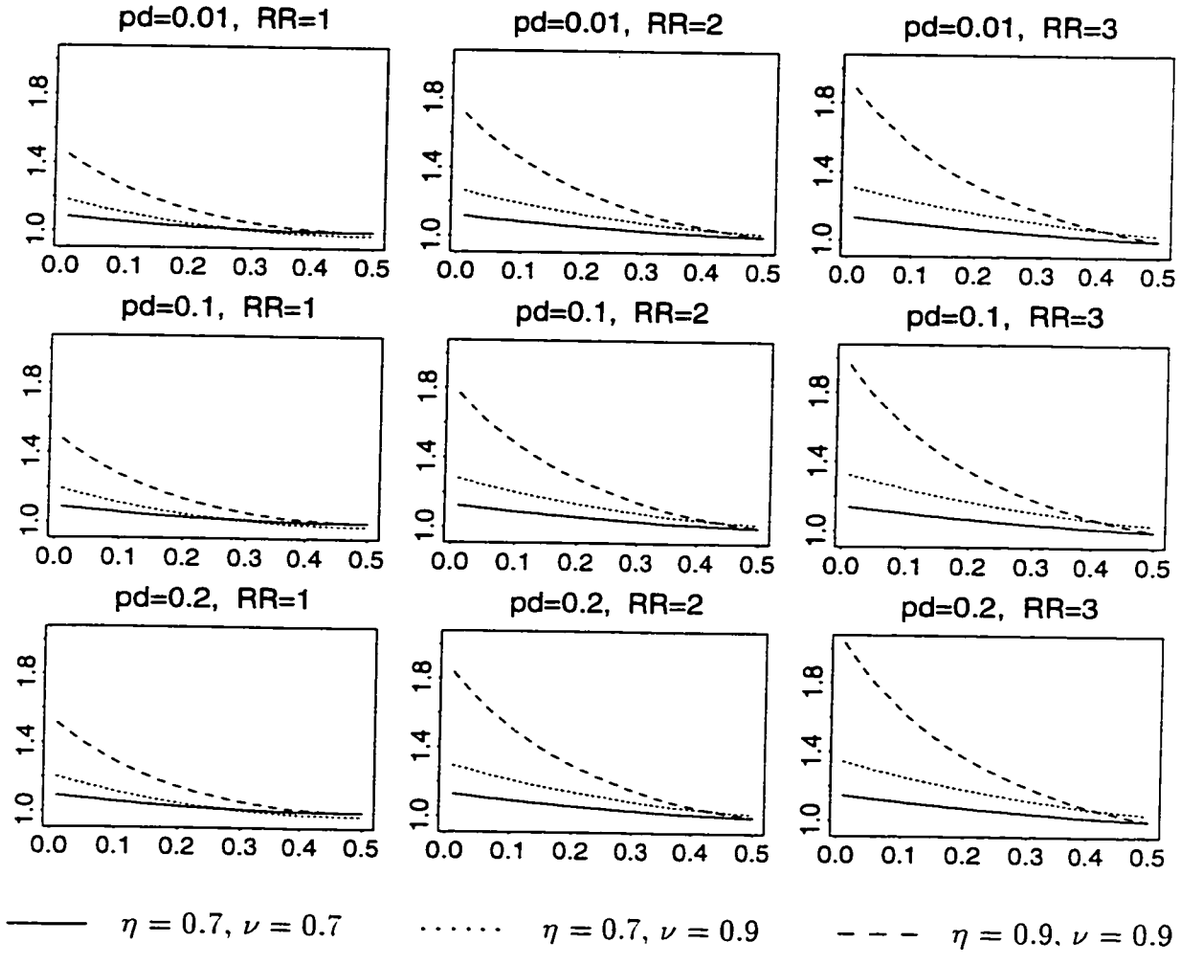


Figure 4.1: Case-cohort estimator: ARE of stratified sample versus Bernoulli sample. Subcohort size is equal to the expected number of deaths.

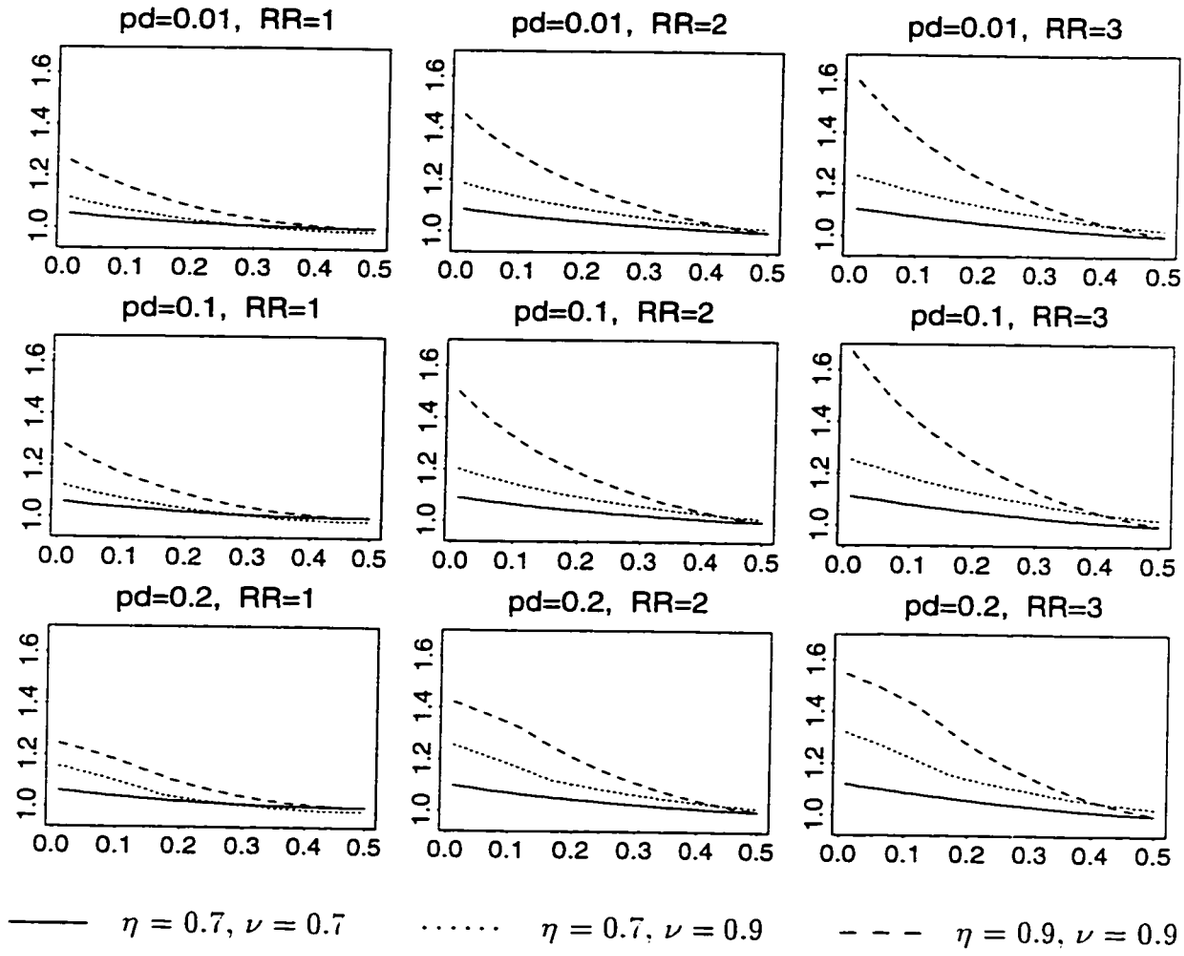


Figure 4.2: Case-cohort estimator: ARE of stratified sample versus Bernoulli sample. Subcohort size is two times the expected number of deaths.

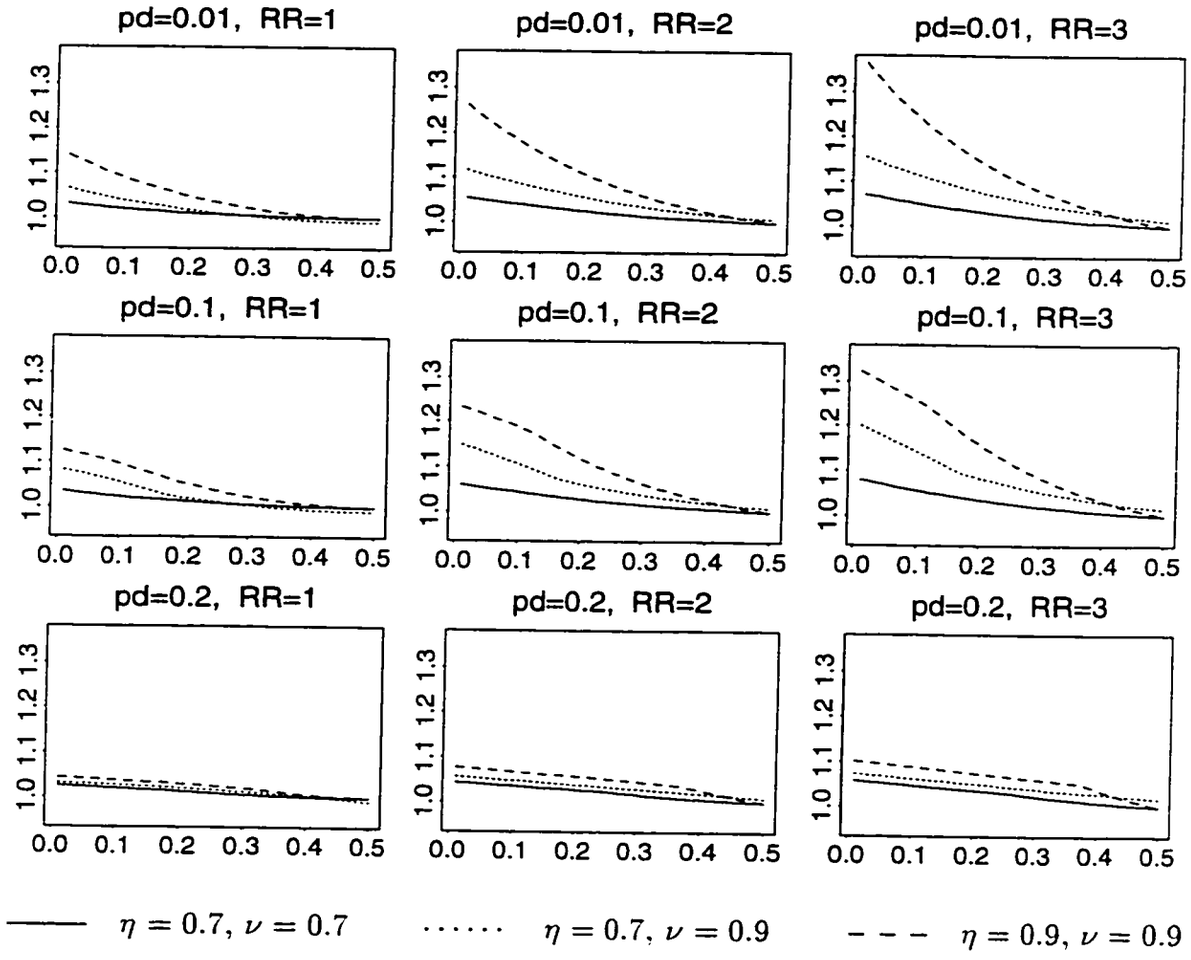


Figure 4.3: Case-cohort estimator: ARE of stratified sample versus Bernoulli sample. Subcohort size is four times the expected number of deaths.

Table 4.3: Simulation study of AH case-cohort design with rare binary exposure. Overall probability of death = 0.05. Model: $\lambda(t \mid Z) = 0.0465 + \beta_0 Z$ where $\beta_0 = 0.1$.

pz	α	Est.	Mean Est.	Sample SE	Av. estim. SE	95% cover. probab.	ARE_F
0.05	0.05	F	0.099	0.0260	0.0256	0.936	
		CC	0.118	0.0764	0.0696	0.931	0.116
		SC-1	0.114	0.0616	0.0601	0.927	0.177
		SC-2	0.108	0.0433	0.0406	0.943	0.360
0.05	0.10	F	0.101	0.0257	0.0258	0.946	
		CC	0.107	0.0468	0.0458	0.937	0.300
		SC-1	0.107	0.0418	0.0423	0.946	0.378
		SC-2	0.103	0.0333	0.0327	0.941	0.595
0.20	0.05	F	0.100	0.0124	0.0124	0.944	
		CC	0.103	0.0260	0.0258	0.947	0.231
		SC-1	0.103	0.0236	0.0244	0.951	0.277
		SC-2	0.102	0.0211	0.0212	0.955	0.347
0.20	0.10	F	0.100	0.0126	0.0124	0.946	
		CC	0.102	0.0193	0.0196	0.957	0.427
		SC-1	0.101	0.0195	0.0189	0.943	0.419
		SC-2	0.101	0.0171	0.0169	0.949	0.545

Notes:

Estimators: F = full data, CC = subcohort selected by Bernoulli sampling, SC-1 = stratified subcohort sampling ($\eta = \nu = 0.7$), SC-2 = stratified subcohort sampling ($\eta = \nu = 0.9$).

pz is the proportion of exposed subjects, α is the subcohort proportion, ARE_F is estimated ARE with respect to the full model.

Table 4.4: Simulation study of AH case-cohort design with rare binary exposure. Overall probability of death = 0.10. Model: $\lambda(t \mid Z) = 0.0963 + \beta_0 Z$ where $\beta_0 = 0.2$.

p_Z	α	Est.	Mean Est.	Sample SE	Av. estim. SE	95% cover. probab.	ARE_F
0.05	0.10	F	0.201	0.0393	0.0383	0.936	
		CC	0.217	0.0855	0.0831	0.945	0.211
		SC-1	0.211	0.0780	0.0745	0.924	0.254
		SC-2	0.204	0.0534	0.0544	0.944	0.541
0.05	0.20	F	0.202	0.0395	0.0384	0.940	
		CC	0.208	0.0589	0.0596	0.947	0.449
		SC-1	0.208	0.0572	0.0560	0.936	0.476
		SC-2	0.204	0.0452	0.0446	0.943	0.763
0.20	0.10	F	0.200	0.0180	0.0184	0.965	
		CC	0.204	0.0363	0.0354	0.945	0.244
		SC-1	0.203	0.0337	0.0338	0.949	0.285
		SC-2	0.202	0.0292	0.0292	0.951	0.379
0.20	0.20	F	0.199	0.0182	0.0183	0.952	
		CC	0.201	0.0260	0.0270	0.955	0.489
		SC-1	0.201	0.0260	0.0260	0.953	0.490
		SC-2	0.200	0.0231	0.0233	0.945	0.618

Notes:

Estimators: F = full data, CC = subcohort selected by Bernoulli sampling, SC-1 = stratified subcohort sampling ($\eta = \nu = 0.7$), SC-2 = stratified subcohort sampling ($\eta = \nu = 0.9$).

p_Z is the proportion of exposed subjects, α is the subcohort proportion, ARE_F is estimated ARE with respect to the full model.

Table 4.5: Simulation study of AH case-cohort design with rare binary exposure. Overall probability of death = 0.10. Model: $\lambda(t \mid Z) = 0.1054 + \beta_0 Z$ where $\beta_0 = 0$ (no covariate effect).

p_Z	α	Est.	Mean Est.	Sample SE	Av. estim. SE	95% cover. probab.	ARE_F
0.05	0.10	F	0.000	0.0222	0.0220	0.954	
		CC	0.005	0.0378	0.0365	0.948	0.345
		SC-1	0.005	0.0327	0.0336	0.954	0.460
		SC-2	0.001	0.0266	0.0272	0.951	0.698
0.05	0.20	F	0.000	0.0220	0.0219	0.938	
		CC	0.002	0.0284	0.0281	0.936	0.604
		SC-1	0.003	0.0264	0.0272	0.948	0.697
		SC-2	0.001	0.0246	0.0239	0.931	0.802
0.20	0.10	F	-0.000	0.0119	0.0119	0.934	
		CC	0.001	0.0171	0.0178	0.958	0.486
		SC-1	0.001	0.0178	0.0175	0.945	0.450
		SC-2	0.001	0.0166	0.0165	0.952	0.518
0.20	0.20	F	-0.001	0.0119	0.0119	0.948	
		CC	0.000	0.0150	0.0147	0.952	0.630
		SC-1	0.001	0.0152	0.0146	0.942	0.611
		SC-2	-0.001	0.0142	0.0139	0.949	0.709

Notes:

Estimators: F = full data, CC = subcohort selected by Bernoulli sampling, SC-1 = stratified subcohort sampling ($\eta = \nu = 0.7$), SC-2 = stratified subcohort sampling ($\eta = \nu = 0.9$).

p_Z is the proportion of exposed subjects, α is the subcohort proportion, ARE_F is estimated ARE with respect to the full model.

Table 4.6: Results of AH model fit to NWTS data: Full data estimates (estimated standard errors in parentheses) and case-cohort estimates, averaged over 500 simulated subcohorts (averaged estimated standard errors in parentheses).

	Parameters:				
	UH (central)	Stage II	Stage III	Stage IV	NWTS-4
Full	0.0667	0.0067	0.0152	0.0373	-0.0056
	(0.00571)	(0.00154)	(0.00196)	(0.00386)	(0.00195)
Subcohort proportion $\alpha = 0.1$*:*					
Bern.	0.0700	0.0067	0.0155	0.0387	-0.0059
	(0.01513)	(0.00351)	(0.00397)	(0.00823)	(0.00382)
Stratif.	0.0673	0.0067	0.0154	0.0390	-0.0058
	(0.01112)	(0.00325)	(0.00370)	(0.00967)	(0.00399)
Subcohort proportion $\alpha = 0.2$*:*					
Bern.	0.0678	0.0067	0.0154	0.0381	-0.0057
	(0.01057)	(0.00257)	(0.00300)	(0.00614)	(0.00291)
Stratif.	0.0676	0.0067	0.0153	0.0380	-0.0056
	(0.00843)	(0.00242)	(0.00281)	(0.00682)	(0.00303)
Subcohort proportion $\alpha = 0.3$*:*					
Bern.	0.0673	0.0067	0.0152	0.0378	-0.0057
	(0.00881)	(0.00220)	(0.00261)	(0.00530)	(0.00255)
Stratif.	0.0669	0.0066	0.0153	0.0374	-0.0056
	(0.00713)	(0.00200)	(0.00238)	(0.00533)	(0.00249)

Chapter 5

THE ADDITIVE HAZARDS MODEL WITH SURROGATE COVARIATES

5.1 Introduction

This chapter addresses the problem of AH regression parameter estimation with covariates subject to measurement error. We treat the problem as a special case of a two-phase design where the true covariate is observed on a random subsample of the study subjects and a surrogate covariate is available for the total first-phase sample. Initially, it is assumed that the underlying additive hazards model includes only a single covariate. This assumption is removed in one of the final sections. We apply the corrected score method to derive an asymptotically unbiased pseudoscore and use error calibration to specify the association between the true and surrogate covariates (see Section 3.3.1). The true covariate values are assumed to be independent and identically distributed, but no conditions are imposed on the distribution of the true covariate.

We first show that if the standard AH pseudoscore is used in a naive way, the resulting estimator is biased. We proceed to define a corrected pseudoscore where the bias term is estimated and subtracted to remove the bias. The estimator based on the corrected pseudoscore, the CS estimator, is shown to be consistent and asymptotically normal.

In Section 5.2, the study design is defined, and working assumptions about the relationship between the true and the surrogate covariates are formulated. In Section 5.3 it is assumed that this relationship is completely known. Under these cir-

cumstances, a corrected pseudoscore is derived. Regularity conditions are introduced to prove consistency and asymptotic normality of the CS estimator. A consistent estimator for its asymptotic variance is proposed. Section 5.4 relaxes the assumption of known relationship between the true and surrogate covariates. It is shown there how the error parameters describing the relationship can be estimated from the data and how the asymptotic variance of the CS estimator is affected. The last part of that section suggests a way to minimize the variance of the CS estimator by an optimal weight selection. Several concrete applications of the corrected pseudoscore method for different structures of covariate measurement errors are discussed in Section 5.5. Section 5.6 deals with generalization of the method to multiple covariates. Results of several simulation studies and an analysis of the NWTSG data set are given in Sections 5.7 and 5.8.

5.2 *Basic assumptions and preliminary considerations*

Before we start deriving the corrected pseudoscore estimator it is necessary to specify the sampling design we will work with and to introduce working assumptions on the surrogate covariate. This is the goal of the current section.

Suppose that the hazard for the uncensored failure time T follows the additive hazards model

$$\lambda(t|Z) = \lambda_0(t) + \beta_0 Z, \qquad 0 \le t \le \tau < \infty.$$

For the moment we assume that the true covariate Z is a time-independent scalar. Let the validation subsample be defined as the set of subjects $\mathcal{V} = \{i : \xi_i = 1\}$ where the selection indicators $\xi_1, \ldots, \xi_n$ are independent binary variables with $\mathrm{P}[\xi_i = 1] = 1 - \mathrm{P}[\xi_i = 0] = \alpha$. The average proportion α of the first-phase sample selected to the validation set is a known constant. We will usually write $\overline{\xi}_i$ for $1 - \xi_i$ and $\overline{\alpha}$ for $1 - \alpha$. Suppose each ξ_i is independent of the failure indicator Δ_i, the censored failure time X_i, the true covariate Z_i, and the surrogate covariate W_i. The validation set $\mathcal{V}$ is

thus selected by simple random sampling with random total sample size, that is, by iid Bernoulli sampling. When $i \in \mathcal{V}$, we observe the full data $(X_i, \Delta_i, Z_i, W_i, \xi_i = 1)$. Otherwise we observe the incomplete data $(X_i, \Delta_i, W_i, \xi_i = 0)$ and the true covariate Z_i is unknown.

We impose the following conditions on the surrogate covariate W_i:

Condition 5.1. $\mathrm{E}[W_i \mid Z_i] = \gamma_0 + \gamma_1 Z_i$, where $\gamma_1 \neq 0$.

Condition 5.2. $\mathrm{var}[W_i \mid Z_i] = V(Z_i) = v_0 + v_1 Z_i + v_2 Z_i^2$.

Condition 5.3. W_i and W_j are independent given Z_i and Z_j.

Condition 5.4. Given Z_i, W_i is independent of X_i and Δ_i.

Conditions 5.1 and 5.2 specify the first two conditional moments of the surrogate covariate given the true covariate. These are the only assumptions on the distribution of W. The variance function $V(\cdot)$ in Condition 5.2 is an arbitrary constant, linear, or quadratic function. Later we explain why more general forms of the variance function are not admissible. Condition 5.3 assures errors are independent. Condition 5.4 means that W must be a true surrogate, i.e., that all the effect of W on survival and censoring is mediated through Z.

The first two conditions explain the relationship between the true and surrogate covariates through two mean parameters γ_0 and γ_1, and three variance parameters v_0, v_1, and v_2. These are the *error parameters* and they will be jointly denoted by $\boldsymbol{\theta}$. In many practical situations some of the error parameters will be fixed at certain values (e.g., $\gamma_0 = 0$, $\gamma_1 = 1$, $v_1 = v_2 = 0$). So, the number of components in $\boldsymbol{\theta}$ may be anywhere from 1 to 5. This structure is flexible enough to accommodate various biases in the surrogate covariate as well as many patterns of nonconstant error variance. Unfortunately, the methodology used to derive the corrected score estimator prohibits using a non-linear link function relating the surrogate covariate mean to the true covariate.

5.3 Known error parameters

Let us first assume that the error parameters $\boldsymbol{\theta}$ are all known. We will define the corrected pseudoscore under this working assumption and derive the asymptotic distribution of the CS estimator. We will also show how to estimate its asymptotic variance.

5.3.1 *Definition of the corrected pseudoscore*

Under Condition 5.1, it is easy to estimate the unobserved Z_i. Define the bias-adjusted surrogate covariate W_i^* by

$$W_i^* = \frac{W_i - \gamma_0}{\gamma_1}.$$

Clearly, $\mathrm{E}[W_i^* \mid Z_i] = Z_i$ and $\mathrm{var}[W_i^* \mid Z_i] = \gamma_1^{-2} V(Z_i)$. Since Z_i can be estimated unbiasedly by W_i^*, it is interesting to see what happens when we consider the full-data pseudoscore

$$U_A = \sum_{i=1}^{n} \int_0^\tau [Z_i - \bar{Z}(t)]\,[dN_i(t) - Z_i \beta Y_i(t)\,dt]$$

and simply replace Z_i by W_i^* whenever Z_i is unobserved, that is, whenever $i \notin \mathcal{V}$. This idea leads us to the naive estimator $\widehat{\beta}_N$, defined as the solution of $U_N(\beta) \equiv \sum_i \psi_i^{(N)}(\beta) = 0$, where

$$\psi_i^{(N)}(\beta) = \int_0^\tau [R_i - \bar{R}(t)]\,[dN_i(t) - R_i \beta Y_i(t)\,dt], \tag{5.1}$$

$R_i = \xi_i Z_i + \bar{\xi}_i W_i^*$, and $\bar{R}(t) = \sum Y_i(t) R_i / \sum Y_i(t)$.

Although W_i^* is unbiased for Z_i, the naive estimator is not unbiased for β_0 because, as we will see, $\mathrm{E}\,U_N(\beta_0) \neq 0$. This is what happens with the naive estimator in most regression settings with errors-in-variables. However, we may try to apply the corrected score method described in Section 3.3.2 to derive an asymptotically unbiased corrected pseudoscore. A necessary condition for the corrected score method

to work well is a relatively simple functional form of the pseudoscore. The AH pseudoscore is quadratic in the true covariate, which facilitates the evaluation of the naive pseudoscore bias.

So, to derive an asymptotically unbiased pseudoscore, we calculate $\mathrm{E}\, U_N(\beta_0)$ conditionally on the true covariate, survival and censoring, and subtract the bias from the naive pseudoscore. Denote $\mathcal{F}_\tau$ the σ-algebra generated by $(Z_i, N_i(t), Y_i(t) : i = 1, \ldots, n; 0 \le t \le \tau)$. Under Conditions 5.1, 5.3 and 5.4, we get

$$\mathrm{E}\left[\psi_i^{(N)} \,\middle|\, \mathcal{F}_\tau, \xi_i = 0\right] = \int [Z_i - \bar{Z}(t)]\, dN_i(t) - \int \mathrm{E}\left[[W_i^* - \overline{W^*}(t)]W_i^* \,\middle|\, \mathcal{F}_\tau, \xi_i = 0\right] \beta Y_i(t)\, dt$$

and

$$\mathrm{E}\left[[W_i^* - \overline{W^*}(t)]W_i^* \,\middle|\, \mathcal{F}_\tau, \xi_i = 0\right] = [Z_i - \bar{Z}(t)]Z_i + \bar{\xi}_i \gamma_1^{-2}\left[V(Z_i) - \frac{Y_i(t)V(Z_i)}{\sum Y_j(t)}\right].$$

Summing over i, we get from the previous two equations,

$$\mathrm{E}\left[\sum_{i=1}^n \psi_i^{(N)}(\beta) \,\middle|\, \mathcal{F}_\tau\right] = U_A(\beta) - \gamma_1^{-2}\beta \sum_{i=1}^n \bar{\xi}_i V(Z_i) X_i + \gamma_1^{-2}\beta \int \frac{\sum \bar{\xi}_i Y_i(t) V(Z_i)}{\sum Y_j(t)}\, dt.$$

The last term is $o_P(n^{1/2})$ and therefore asymptotically negligible. At β_0, the expectation of U_A is zero, but the expectation of the middle term is not. That must be the term causing the bias in the naive pseudoscore. Hence, an asymptotically unbiased pseudoscore should be obtained by adding $\gamma_1^{-2}\beta \sum_{i=1}^n \bar{\xi}_i V(Z_i) X_i$ to $U_N(\beta)$.

To evaluate this term, $V(Z_i)$ is needed for the nonvalidation subjects. Unfortunately, it depends on the unobserved Z_i. Here is the point where a restriction on V brought by Condition 5.2 is necessary. Indeed, if $V(x) = v_0 + v_1 x + v_2 x^2$, then it is easy to show that $\mathrm{E}\,[V(W_i^*) \mid Z_i] = (1 + \gamma_1^{-2} v_2) V(Z_i)$. This means that $V(Z_i)$ can be estimated unbiasedly by $\gamma_1^2(\gamma_1^2 + v_2)^{-1} V(W_i^*)$. Thus, the bias correction to the naive pseudoscore becomes

$$\gamma_1^{-2}\beta \sum_{i=1}^n \bar{\xi}_i \frac{\gamma_1^2}{\gamma_1^2 + v_2} V(W_i^*) X_i = \sum_{i=1}^n \bar{\xi}_i \int_0^\tau \frac{V(W_i^*)}{\gamma_1^2 + v_2} \beta Y_i(t)\, dt. \tag{5.2}$$

The previous two paragraphs suggest how to obtain a corrected pseudoscore. Let us now proceed to a formal definition. We will specify the contributions of the validation set subjects and of the nonvalidation set subjects and combine them into a single corrected pseudoscore. But before we do so, let us consider the at-risk covariate average. To estimate it using all observed data, we define

$$\bar{R}(t,w) = \frac{\sum(\xi_i Z_i + w\bar{\xi}_i W_i^*)Y_i(t)}{\sum(\xi_i + w\bar{\xi}_i)Y_i(t)}. \tag{5.3}$$

The estimated at-risk average $\bar{R}(t,w)$ is a weighted average of the true (where available) and adjusted surrogate covariate values over the subjects at risk at time t. The surrogate values are multiplied by a fixed weight w satisfying $0 \le w \le 1$. By down-weighting the surrogates we hope to obtain a better estimator since they have a larger variance. To simplify the notation, however, we usually drop the argument w from $\bar{R}(t,w)$ and write $\bar{R}(t)$.

The true covariates of the validation set subjects are known. Thus, we suggest that the contribution to the corrected pseudoscore of a validation set subject be obtained by replacing $\overline{Z}(t)$ with $\bar{R}(t)$ in the original full-data AH pseudoscore:

$$\psi_i^{(V)}(\beta) = \int_0^\tau [Z_i - \bar{R}(t)]\, dN_i(t) - \int_0^\tau [Z_i - \bar{R}(t)] Z_i \beta Y_i(t)\, dt.$$

The true covariates of the nonvalidation set members have to be estimated by the bias-adjusted surrogates. To account for the resulting bias, we subtract the correction term, as suggested by (5.2). Hence, the corrected pseudoscore contribution of a nonvalidation subject is

$$\psi_i^{(NV)}(\beta) = \int_0^\tau [W_i^* - \bar{R}(t)]\, dN_i(t) - \int_0^\tau \left\{ [W_i^* - \bar{R}(t)] W_i^* - \frac{V(W_i^*)}{\gamma_1^2 + v_2} \right\} \beta Y_i(t)\, dt.$$

The two kinds of contributions are combined to form the corrected pseudoscore as follows:

$$U_C(\beta, w) = \sum_{i=1}^{n} \left[\xi_i \psi_i^{(V)}(\beta) + w\bar{\xi}_i \psi_i^{(NV)}(\beta) \right].$$

The rationale for including the weight w at this place again is the same as for using it in the at-risk average: downweighting the nonvalidation set contributions. Notice that setting $w = 0$ eliminates all nonvalidation subjects from the pseudoscore and the resulting estimator is the complete-case AH estimator based on the validation set.

The corrected pseudoscore U_C can be written in a different and sometimes more convenient form if we define $\varrho_i = \xi_i + \overline{\xi}_i w$ and $R_i = \xi_i Z_i + \overline{\xi}_i W_i^*$. Then $\bar{R}(t, w) = \sum \varrho_i R_i Y_i(t) / \sum \varrho_i Y_i(t)$ and

$$U_C(\beta, w) = \sum_{i=1}^{n} \left\{ \varrho_i \int_0^\tau [R_i - \bar{R}(t)] \, [dN_i(t) - R_i \beta Y_i(t)\, dt] + \overline{\xi}_i w \frac{V(W_i^*)}{\gamma_1^2 + v_2} \beta X_i \right\}.$$

Since $\sum_{i=1}^n \varrho_i [R_i - \bar{R}(t)] Y_i(t) = 0$, $U_C(\beta, w)$ is equal to

$$\sum_{i=1}^{n} \left\{ \varrho_i \int_0^\tau [R_i - \bar{R}(t)] \, [dN_i(t) - Y_i(t)\, d\Lambda_0(t) - R_i \beta Y_i(t)\, dt] + \overline{\xi}_i w \frac{V(W_i^*)}{\gamma_1^2 + v_2} \beta X_i \right\} \tag{5.4}$$

and also to

$$\sum_{i=1}^{n} \varrho_i \left(\int_0^\tau [R_i - \bar{R}(t)]\, dN_i(t) - \int_0^\tau \left\{ [R_i - \bar{R}(t)]^2 - \overline{\xi}_i \frac{V(W_i^*)}{\gamma_1^2 + v_2} \right\} \beta Y_i(t)\, dt \right).$$

The last expression shows that the corrected pseudoscore is invariant with respect to linear transformations of the covariate. It also implies that the corrected pseudoscore (CS) estimator $\widehat{\beta}_C(w)$ defined as the solution of $U_C(\beta, w) = 0$ possesses a closed form.

$$\widehat{\beta}_C(w) = \left(\sum_{i=1}^{n} \varrho_i \int_0^\tau \left\{ [R_i - \bar{R}(t)]^2 - \overline{\xi}_i \frac{V(W_i^*)}{\gamma_1^2 + v_2} \right\} Y_i(t)\, dt \right)^{-1} \sum_{i=1}^{n} \varrho_i \int_0^\tau [R_i - \bar{R}(t)]\, dN_i(t). \tag{5.5}$$

5.3.2 Limiting distributions of the corrected pseudoscore and the CS estimator

In order to derive the limiting distribution of $\widehat{\beta}_C(w)$, we approximate $U_C(\beta_0, w)$ by a sum of asymptotically independent contributions. For U_C itself the independence does

not hold due to the presence of $\bar{R}(t)$ in the estimating equation. We assume that the data are observed on the time interval $[0, \tau]$, $0 < \tau < \infty$ and that $(X_i, \Delta_i, Z_i, W_i, \xi_i)$, $i = 1, \ldots, n$, are independent and identically distributed replicates of (X, Δ, Z, W, ξ). Let us define

$$\pi_k(t) \equiv \mathrm{E}\, Z^k Y(t), \qquad k = 0, 1, 2,$$

and denote $e(t) = \pi_1(t)/\pi_0(t)$. As mentioned in Section 2.3, $\pi_k(t)$ is also the limit in probability of $n^{-1} \sum_{i=1}^{n} Z_i^k Y_i(t)$.

Recall the regularity conditions introduced in Section 2.3. Here are their univariate versions for time-independent covariates:

Condition 2.1. $\Lambda_0(\tau) < \infty$.

Condition 2.2. $\mathrm{P}[Y(\tau) = 1] > 0$.

Condition 2.3. $\mathrm{E}\,|Z|^3 < \infty$.

Condition 2.4. $\Sigma_A(\beta_0) \equiv \mathrm{E} \int_0^\tau [Z - e(t)]^2 dN(t) > 0$.

The conditions are assumed to be fulfilled and will be used together with Conditions 5.1–5.4 to investigate the limiting distribution of the CS estimator.

By Lemma 2.1(i),

$$\sup_{0 \le t \le \tau} \left| \frac{1}{n} \sum_{i=1}^{n} Z_i^k Y_i(t) - \pi_k(t) \right| \to_p 0, \qquad k = 0, 1, 2,$$

provided Conditions 2.2 and 2.3 hold. Let us generalize Lemma 2.1(ii) to a covariate measured with error:

Lemma 5.1. *Under Conditions 2.2 and 2.3, the estimated at-risk average converges in probability to the limiting at-risk average uniformly in time, that is*

$$\sup_{0 \le t \le \tau} |\bar{R}(t, w) - e(t)| \to_p 0 \qquad \textit{for any } w.$$

Proof. By the weak law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n}(\xi_i + w\bar{\xi}_i)Y_i(t) \to_p \mathrm{E}(\xi_i + w\bar{\xi}_i)Y_i(t) = (\alpha + w\bar{\alpha})\pi_0(t) \tag{5.6}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}(\xi_i Z_i + w\bar{\xi}_i W_i^*)Y_i(t) \to_p \mathrm{E}(\xi_i Z_i + w\bar{\xi}_i W_i^*)Y_i(t) = (\alpha + w\bar{\alpha})\pi_1(t). \tag{5.7}$$

We have to show that these two convergences are uniform in t. However, the uniformity follows immediately from Corollary III.2 of Andersen and Gill (1982). The sufficient condition, uniformly bounded means in t, is trivially fulfilled for $(\xi_i Z_i + w\bar{\xi}_i W_i^*)Y_i(t)$ as well as for $(\xi_i + w\bar{\xi}_i)Y_i(t)$.

The largest error in the at-risk average, $\sup_{0\le t\le\tau}|\bar{R}(t) - e(t)|$, is bounded by the sum of

$$(\alpha + w\bar{\alpha})^{-1}\pi_0^{-1}(\tau)\sup_{0\le t\le\tau}\left|\frac{1}{n}\sum(\xi_i Z_i + w\bar{\xi}_i W_i^*)Y_i(t) - (\alpha + w\bar{\alpha})\pi_1(t)\right| \tag{5.8}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}|\xi_i Z_i + w\bar{\xi}_i W_i^*|\sup_{0\le t\le\tau}\left|\left[\frac{1}{n}\sum(\xi_i + w\bar{\xi}_i)Y_i(t)\right]^{-1} - (\alpha + w\bar{\alpha})^{-1}\pi_0^{-1}(t)\right|. \tag{5.9}$$

By Condition 2.2, $\pi_0^{-1}(\tau) < \infty$ and the absolute value in (5.8) tends to zero in probability uniformly in t by means of the uniform convergence in (5.7). In (5.9), the average of absolute covariates converges to its expectation and the supremum of the absolute value converges to zero by means of (5.6), Condition 2.2, and the fact that $\mathrm{P}[\sum_1^n Y_i(\tau) = 0]$ converges to zero. This finishes the proof. □

Asymptotic linearity of the corrected pseudoscore

Here we show that the corrected pseudoscore evaluated at the true parameter can be approximated by a sum of independent and identically distributed zero-mean random variables.

Theorem 5.1. *Under Conditions 2.1–2.4, the corrected pseudoscore is asymptotically linear, that is*

$$\frac{1}{\sqrt{n}}U_C(\beta_0, w) = \frac{1}{\sqrt{n}}\widetilde{U}_C(\beta_0, w) + o_P(1),$$

where

$$\widetilde{U}_C(\beta_0, w) \equiv \sum_{i=1}^{n}\left[\xi_i\widetilde{\psi}_i^{(V)}(\beta_0) + w\overline{\xi}_i\widetilde{\psi}_i^{(NV)}(\beta_0)\right],$$

$$\widetilde{\psi}_i^{(V)}(\beta_0) \equiv \int_0^\tau [Z_i - e(t)]\, dM_i(t) = \widetilde{\psi}_i^{(A)}(\beta_0),$$

and

$$\widetilde{\psi}_i^{(NV)}(\beta_0) \equiv \int_0^\tau [W_i^* - e(t)]\,[dN_i(t) - Y_i(t)\, d\Lambda_0(t) - W_i^*\beta_0 Y_i(t)\, dt] + \beta_0\frac{V(W_i^*)}{\gamma_1^2 + \upsilon_2}X_i.$$

Corollary. If $\mathrm{E}\, W_i^4 < \infty$, then $n^{-1/2}U_C(\beta_0, w)$ converges in law to normal distribution with zero mean and variance

$$\Sigma_C(\beta_0, w) = \alpha\Sigma_A(\beta_0) + \overline{\alpha}w^2\,\mathrm{E}\left[\widetilde{\psi}_i^{(NV)}(\beta_0)\right]^2,$$

where $\Sigma_A(\beta_0)$ is the limiting variance of the full-data AH pseudoscore contribution.

Proof of Corollary. Theorem 5.1 implies that $n^{-1/2}U_C(\beta_0, w)$ has the same limiting distribution as $n^{-1/2}\widetilde{U}_C(\beta_0, w)$. Since $\widetilde{\psi}_i^{(V)}(\beta_0)$ is a martingale integral, $\mathrm{E}\,[W_i^* \mid Z_i] = Z_i$, and $\mathrm{E}\,[(W_i^*)^2 \mid Z_i] = Z_i^2 + V(Z_i)/\gamma_1^2$, it follows that

$$\mathrm{E}\,\widetilde{\psi}_i^{(NV)}(\beta_0) = \mathrm{E}\,\mathrm{E}\left[\widetilde{\psi}_i^{(NV)}(\beta_0)\,\middle|\, Z_i\right] = \mathrm{E}\,\widetilde{\psi}_i^{(V)}(\beta_0) = 0.$$

Obviously, $\xi_i\widetilde{\psi}_i^{(V)}(\beta_0) + w\overline{\xi}_i\widetilde{\psi}_i^{(NV)}(\beta_0)$ are independent and identically distributed random variables with zero mean and variance $\Sigma_C(\beta_0, w)$. The condition $\mathrm{E}\, W_i^4 < \infty$ is necessary to make sure $\mathrm{E}\left[\widetilde{\psi}_i^{(NV)}(\beta_0)\right]^2$ is finite. The variance is positive by Condition 2.4. So, the central limit theorem for iid random variables can be applied to get

$$n^{-1/2}\widetilde{U}_C(\beta_0, w) \to_d \mathrm{N}(0, \Sigma_C(\beta_0, w)).$$

□

Proof of Theorem 5.1. By Equation (5.4), $U_C(\beta_0)$ can be written as $U_1 + wU_2$, where

$$U_1 = \sum_{i=1}^{n} \xi_i \int_0^\tau [Z_i - \bar{R}(t)][dN_i(t) - Y_i(t)d\Lambda_0(t) - Z_i\beta_0 Y_i(t)\,dt]$$

and

$$U_2 = \sum_{i=1}^{n} \bar{\xi}_i \left\{ \int_0^\tau [W_i^* - \bar{R}(t)][dN_i(t) - Y_i(t)d\Lambda_0(t) - W_i^*\beta_0 Y_i(t)\,dt] + \frac{V(W_i^*)}{\gamma_1^2 + v_2}\beta_0 X_i \right\}.$$

The normalized validation part $n^{-1/2}U_1$ can be decomposed as follows:

$$n^{-1/2}U_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \int_0^\tau [Z_i - e(t)]\,[dN_i(t) - Y_i(t)\,d\Lambda_0(t) - Z_i\beta_0 Y_i(t)\,dt] \tag{5.10}$$

$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \int_0^\tau [e(t) - \bar{R}(t)]\,[dN_i(t) - Y_i(t)\,d\Lambda_0(t) - Z_i\beta_0 Y_i(t)\,dt]. \tag{5.11}$$

The first term, (5.10), is a martingale integral. It is, in fact, the sum of contributions of validation set members to the full-data pseudoscore U_A. The second term, (5.11), is a martingale integral with predictable variance process

$$\frac{1}{n} \sum_{i=1}^{n} \xi_i \int_0^\tau [e(t) - \bar{R}(t)]^2 dN_i(t) \le \sup_{0 \le t \le \tau} [e(t) - \bar{R}(t)]^2 \frac{1}{n} \sum_{i=1}^{n} \xi_i \Delta_i. \tag{5.12}$$

By Lemma 5.1, $\sup_t [e(t) - \bar{R}(t)]^2 \to_p 0$. The average of $\xi_i\Delta_i$ converges in probability to its expectation. So, (5.12) converges to zero in probability and consequently (5.11) converges in probability to its expectation, which is zero. Hence, we have

$$n^{-1/2}U_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \int_0^\tau [Z_i - e(t)]\,dM_i(t) + o_P(1).$$

The normalized nonvalidation part $n^{-1/2}U_2$ can be decomposed similarly:

$$n^{-1/2}U_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \bar{\xi}_i \int_0^\tau [W_i^* - e(t)]\,[dN_i(t) - Y_i(t)\,d\Lambda_0(t) - W_i^*\beta_0 Y_i(t)\,dt] \tag{5.13}$$

$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \bar{\xi}_i \beta_0 \frac{V(W_i^*)}{\gamma_1^2 + v_2} X_i \tag{5.14}$$

$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \bar{\xi}_i \int_0^\tau [e(t) - \bar{R}(t)]\,[dN_i(t) - Y_i(t)\,d\Lambda_0(t) - Z_i\beta_0 Y_i(t)\,dt] \tag{5.15}$$

$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \bar{\xi}_i \int_0^\tau [e(t) - \bar{R}(t)](Z_i - W_i^*)\beta_0 Y_i(t)\,dt. \tag{5.16}$$

Here, (5.15) tends to zero by the same argument as (5.11). Expression (5.16) can be written as

$$\beta_0 \int_0^\tau [e(t) - \bar{R}(t)]\, dB_n(t),$$

where

$$B_n(t) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n \overline{\xi}_i (Z_i - W_i^*) \int_0^t Y_i(s)\, ds.$$

Now it is easy to apply Lemma 4.3 and Lemma 4.4 in the same way as in the proof of Theorem 4.1 to show that (5.16) converges to zero in probability.

Thus, the nonvalidation part can be written as

$$n^{-1/2} U_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^n \overline{\xi}_i \Big\{ \int_0^\tau [W_i^* - e(t)]\, [dN_i(t) - Y_i(t)\, d\Lambda_0(t) - W_i^* \beta_0 Y_i(t)\, dt] + \beta_0 \frac{V(W_i^*)}{\gamma_1^2 + v_2} X_i \Big\} + o_P(1).$$

If follows that $n^{-1/2} U_C(\beta_0) = n^{-1/2} \widetilde{U}_C(\beta_0) + o_P(1)$, where

$$\widetilde{U}_C(\beta_0, w) = \sum_{i=1}^n \xi_i \int_0^\tau [Z_i - e(t)]\, dM_i(t)$$

$$+ w \sum_{i=1}^n \overline{\xi}_i \Big\{ \int_0^\tau [W_i^* - e(t)]\, [dN_i(t) - Y_i(t)\, d\Lambda_0(t) - W_i^* \beta_0 Y_i(t)\, dt] + \beta_0 \frac{V(W_i^*)}{\gamma_1^2 + v_2} X_i \Big\}.$$

This proves Theorem 5.1. □

Limiting distribution of the CS estimator

Consistency and asymptotic normality of $\widehat{\beta}_C$ are stated in the following theorem:

Theorem 5.2. *Assume that Conditions 2.1–2.4 hold and* $\mathrm{E}\, W_i^4 < \infty$. *Then* $\widehat{\beta}_C \to_p \beta_0$ *and* $\sqrt{n}(\widehat{\beta}_C - \beta_0)$ *is asymptotically normal with zero mean and variance* $\Sigma_C(\beta_0)/D_C^2$ *where*

$$D_C \equiv (\alpha + \overline{\alpha} w) D_A = (\alpha + \overline{\alpha} w)\, \mathrm{E} \int_0^\tau [Z_i - e(t)]^2 Y_i(t)\, dt$$

is the negative expected derivative of $n^{-1} U_C(\beta)$.

Proof. The proof is essentially the same as the proof of Lemma 2.3. Using the techniques employed in the proof of Theorem 5.1 it is easy to show that $n^{-1}[\partial \widetilde{U}_C(\beta_0)/\partial\beta - \partial\widetilde{U}_C(\beta_0)/\partial\beta] \to_p 0$. Both partial derivatives are constant in β and therefore it does not matter at which point they are calculated.

By Taylor expansion,

$$U_C(\widehat{\beta}_C) - U_C(\beta_0) = \frac{\partial U_C(\beta^*)}{\partial\beta}(\widehat{\beta}_C - \beta_0)$$

for some β^* between $\widehat{\beta}_C$ and β_0, whose precise value is irrelevant. It follows that

$$\frac{1}{\sqrt{n}}U_C(\beta_0) = -\frac{1}{n}\frac{\partial\widetilde{U}_C(\beta)}{\partial\beta}\sqrt{n}(\widehat{\beta}_C - \beta_0) + o_P(1).$$

By the Corollary to Theorem 5.1, the left-hand side converges in distribution to $N(0, \Sigma_C)$, where the variance Σ_C is positive. Since $\widetilde{U}_C(\beta_0)$ is a sum of iid terms, the normalized negative partial derivative of $\widetilde{U}_C(\beta_0)$ converges in probability to D_C, which can be expressed as

$$\begin{aligned} D_C &= -\,\mathrm{E}\,\frac{\partial}{\partial\beta}\left[\xi_i\widetilde{\psi}_i^{(V)}(\beta_0) + w\overline{\xi}_i\widetilde{\psi}_i^{(NV)}(\beta_0)\right] \\ &= \alpha\,\mathrm{E}\int_0^\tau [Z_i - e(t)]Z_iY_i(t)\,dt + w\overline{\alpha}\,\mathrm{E}\left\{\int_0^\tau [W_i^* - e(t)]W_i^*Y_i(t)\,dt + \frac{V(W_i^*)}{\gamma_1^2 + v_2}\right\} \\ &= (\alpha + w\overline{\alpha})\,\mathrm{E}\int_0^\tau [Z_i - e(t)]^2Y_i(t)\,dt = (\alpha + w\overline{\alpha})D_A. \end{aligned}$$

In the last equation above we conditioned on (Z_i, Y_i) and used the identity

$$\mathrm{E}\int_0^\tau [Z_i - e(t)]e(t)Y_i(t)\,dt = \int_0^\tau e(t)[\pi_1(t) - e(t)\pi_0(t)]\,dt = 0.$$

It follows that $\sqrt{n}(\widehat{\beta}_C - \beta_0)$ converges in distribution to $N(0, \Sigma_C/D_C^2)$. The consistency of $\widehat{\beta}_C$ is a consequence of this fact. □

5.3.3 Estimation of the limiting variance

The goal of this section is to derive a consistent estimator of the asymptotic variance of the CS estimator. We assume that Conditions 2.1–2.4 hold, with Condition 2.3

strengthened to $\mathrm{E}\, W_i^4 < \infty$. The negative partial derivative D_C can be estimated by differentiating the observed corrected pseudoscore. Thus we obtain the estimator

$$\widehat{D}_C = \frac{1}{n}\sum_{i=1}^{n}\left(\xi_i \int_0^\tau [Z_i - \bar{R}(t)]^2 Y_i(t)\, dt \right.$$
$$\left. + \overline{\xi}_i w \int_0^\tau \left\{[W_i^* - \bar{R}(t)]^2 - \frac{V(W_i^*)}{\gamma_1^2 + v_2}\right\} Y_i(t)\, dt\right).$$

To find a consistent estimator for $\Sigma_C(\beta_0, w)$, we need to estimate the cumulative baseline hazard using the observed data. We propose using the estimator

$$d\widehat{\Lambda}_0(t) = \frac{\sum_{i=1}^n dN_i(t)}{\sum_{i=1}^n Y_i(t)} - \bar{R}(t)\widehat{\beta}_C\, dt.$$

This is just a simple generalization of Lin and Ying's full data estimator (2.12). Now,

$$Y_i(t)\, d\widehat{\Lambda}_0(t) + W_i^*\widehat{\beta}_C Y_i(t)\, dt = Y_i(t)\frac{\sum_{j=1}^n dN_j(t)}{\sum_{j=1}^n Y_j(t)} + [W_i^* - \bar{R}(t)]\widehat{\beta}_C Y_i(t)\, dt$$

and $\Sigma_C(\beta_0, w)$ can be estimated by

$$\widehat{\Sigma}_C(\beta, w) = \frac{1}{n}\sum_{i=1}^{n}\left\{\xi_i \int_0^\tau [Z_i - \bar{R}(t)]^2\, dN_i(t) + w^2\overline{\xi}_i\left[\widehat{\psi}_i^{(NV)}(\beta)\right]^2\right\}$$

evaluated at $\beta = \widehat{\beta}_C$, where

$$\widehat{\psi}_i^{(NV)}(\beta) =$$
$$\int_0^\tau [W_i^* - \bar{R}(t)]\left[dN_i(t) - Y_i(t)\frac{\sum_{j=1}^n dN_j(t)}{\sum_{j=1}^n Y_j(t)} - [W_i^* - \bar{R}(t)]\beta Y_i(t)\, dt\right] + \beta\frac{V(W_i^*)}{\gamma_1^2 + v_2}X_i.$$

The following two theorems show that $\widehat{D}_C$, $\widehat{\Lambda}_0(t)$ and $\widehat{\Sigma}_C$ are consistent.

Theorem 5.3. *Under the regularity conditions summarized at the beginning of this section,*

1. $\widehat{D}_C \to_p D_C$,

2. $\sup_{0\le t\le\tau} |\widehat{\Lambda}_0(t) - \Lambda_0(t)| \to_p 0$.

Proof. **1.** Let us first prove the consistency of $\widehat{D}_C$. Its validation part can be written as

$$\begin{aligned}\frac{1}{n}\sum \xi_i \int_0^\tau [Z_i - \bar{R}(t)]^2 Y_i(t)\,dt =& \frac{1}{n}\sum \xi_i \int_0^\tau [Z_i - e(t)]^2 Y_i(t)\,dt \\ &+ \frac{2}{n}\sum \xi_i \int_0^\tau [Z_i - e(t)][e(t) - \bar{R}(t)] Y_i(t)\,dt \quad (5.17) \\ &+ \frac{1}{n}\sum \xi_i \int_0^\tau [e(t) - \bar{R}(t)]^2 Y_i(t)\,dt.\end{aligned}$$

The absolute value of the second term on the right hand side of (5.17) is bounded by

$$2 \sup_t |e(t) - \bar{R}(t)| \left[\frac{1}{n}\sum \xi_i |Z_i| X_i + \frac{1}{n}\sum \xi_i \int_0^\tau |e(t)| Y_i(t)\,dt\right] \to_p 0$$

since the supremum converges in probability to zero and the square bracket converges in probability to $2\alpha\, \mathrm{E}\,|Z_i| X_i$, which is finite. Similarly, the third term tends to zero because $\sup_t |e(t) - \bar{R}(t)|^2 \to_p 0$. The leading term converges in probability to its expectation αD_A.

By the same argument, $\bar{R}(t)$ in the nonvalidation part of $\widehat{D}_C$ may be replaced by $e(t)$ with an error of $o_P(1)$. It follows that the nonvalidation part tends to

$$w\overline{\alpha} \int_0^\tau \mathrm{E}\left\{[W_i^* - e(t)]^2 - \frac{V(W_i^*)}{\gamma_1^2 + v_2}\right\} Y_i(t)\,dt = w\overline{\alpha} \int_0^\tau \mathrm{E}[Z_i - e(t)]^2 Y_i(t)\,dt = w\overline{\alpha} D_A.$$

Hence $\widehat{D}_C \to_p (\alpha + w\overline{\alpha}) D_A = D_C$.

2. To show the uniform consistency of the baseline hazard estimator, let us first fix time t. We express $\widehat{\Lambda}_0(t) - \Lambda_0(t)$ as follows:

$$\begin{aligned}\widehat{\Lambda}_0(t) - \Lambda_0(t) =& \int_0^t \frac{\sum_i dN_i(s)}{\sum_j Y_j(s)} - \int_0^t \frac{\sum_i (\xi_i + w\bar{\xi}_i) dN_i(s)}{\sum_j (\xi_j + w\bar{\xi}_j) Y_j(s)} + \int_0^t \frac{\sum_i (\xi_i + w\bar{\xi}_i) dN_i(s)}{\sum_j (\xi_j + w\bar{\xi}_j) Y_j(s)} \\ &- \int_0^t \frac{\sum_i (\xi_i + w\bar{\xi}_i)(\xi_i Z_i + \bar{\xi}_i W_i^*) Y_i(s)}{\sum_j (\xi_j + w\bar{\xi}_j) Y_j(s)} \widehat{\beta}_C\, ds - \int_0^t \frac{\sum_i (\xi_i + w\bar{\xi}_i) Y_i(s)}{\sum_j (\xi_j + w\bar{\xi}_j) Y_j(s)}\, d\Lambda_0(s) \\ =& \int_0^t \frac{\sum_i (\xi_i + w\bar{\xi}_i)}{\sum_j (\xi_j + w\bar{\xi}_j) Y_j(s)} [dN_i(s) - Y_i(s)\, d\Lambda_0(s) - (\xi_i Z_i + \bar{\xi}_i W_i^*) \widehat{\beta}_C Y_i(s)\, ds] \\ &+ \sum_i \int_0^t \left[\frac{1}{\sum_j Y_j(s)} - \frac{\xi_i + w\bar{\xi}_i}{\sum_j (\xi_j + w\bar{\xi}_j) Y_j(s)}\right] dN_i(s).\end{aligned}$$

This can be written as a sum of four terms.

$$\int_0^t \frac{1}{\sum_j(\xi_j + w\bar{\xi}_j)Y_j(s)} \sum_i (\xi_i + w\bar{\xi}_i)[dN_i(s) - Y_i(s)\,d\Lambda_0(s) - Z_i\beta_0 Y_i(s)\,ds]$$
$$+ \int_0^t \frac{\beta_0 - \widehat{\beta}_C}{\sum_j(\xi_j + w\bar{\xi}_j)Y_j(s)} \sum_i (\xi_i + w\bar{\xi}_i)(\xi_i Z_i + \bar{\xi}_i W_i^*)Y_i(s)\,ds$$
$$+ \int_0^t \frac{w\beta_0}{\sum_j(\xi_j + w\bar{\xi}_j)Y_j(s)} \sum_i \bar{\xi}_i(Z_i - W_i^*)Y_i(s)\,ds$$
$$+ \sum_i \int_0^t \left[\frac{1}{\sum_j Y_j(s)} - \frac{\xi_i + w\bar{\xi}_i}{\sum_j(\xi_j + w\bar{\xi}_j)Y_j(s)}\right] dN_i(s).$$

By Condition 2.2, the probability that $Y_i(t) = 0$ for all i converges to zero as $n \to \infty$. Thus the expressions above are finite with probability converging to 1. Since $\sup_s |n^{-1}\sum_j Y_j(s) - \pi_0(s)| \to_p 0$ and $\pi_0(s) \geq \pi_0(\tau) > 0$, only a negligible error is made by replacing $\sum_j Y_j(s)$ with $n\pi_0(s)$ and $\sum_j(\xi_j + w\bar{\xi}_j)Y_j(s)$ with $n(\alpha + w\bar{\alpha})\pi_0(s)$. In addition, $n\pi_0(s)$ is bounded below by $n\pi_0(\tau)$ and $n(\alpha + w\bar{\alpha})\pi_0(s)$ by $nw\pi_0(\tau)$. So we are free to make such replacements whenever it is convenient.

The first term is a martingale integral with predictable variance function

$$\frac{1}{n^2}\sum_i \int_0^t \frac{(\xi_i + w\bar{\xi}_i)^2}{[n^{-1}\sum_j(\xi_j + w\bar{\xi}_j)Y_j(s)]^2}[Y_i(s)\,d\Lambda_0(s) + Z_i\beta_0 Y_i(s)\,ds],$$

which converges to 0 in probability uniformly in t. The second term is equal to $(\beta_0 - \widehat{\beta}_C)\int_0^t \bar{R}(s)\,ds$. It converges to zero uniformly in t because $\widehat{\beta}_C$ is consistent and $\int_0^t \bar{R}(s)\,ds$ converges to $\int_0^t e(s)\,ds$, which is bounded. The third term is asymptotically equivalent to a term bounded in absolute value by

$$\frac{\beta_0}{(\alpha + w\bar{\alpha})\pi_0(\tau)}\left|\frac{1}{n}\sum_i \bar{\xi}_i(Z_i - W_i^*)\min(X_i, t)\right|.$$

This converges to 0 pointwise and also uniformly in t. Hence the third term also converges uniformly to zero. For any t, the fourth term is asymptotically equivalent to

$$\sum_i \int_0^t \left[\frac{1}{\pi_0(s)} - \frac{\xi_i + w\bar{\xi}_i}{\sum_j(\alpha + w\bar{\alpha})\pi_0(s)}\right] dN_i(s).$$

This is bounded in absolute value by

$$\frac{1}{(\alpha + w\overline{\alpha})\pi_0(\tau)} \left| \frac{1}{n} \sum_i \Delta_i(\alpha + w\overline{\alpha} - \xi_i - w\overline{\xi}_i) \right|.$$

which converges to zero by the weak law of large numbers and independence of ξ_i and Δ_i. This finishes the proof of uniform consistency of $\widehat{\Lambda}_0$. □

Theorem 5.4. *Let the conditions of the previous theorem hold and let Z_i be bounded. Then $\widehat{\Sigma}_C(\widehat{\beta}_C) \to_p \Sigma_C(\beta_0)$.*

Proof. The validation part of the pseudoscore variance estimator $\widehat{\Sigma}_C$ obviously converges to $\alpha\Sigma_A$. It remains to be shown that the nonvalidation part of the estimator converges to $w^2\overline{\alpha}\,\mathrm{E}\left[\widetilde{\psi}_i^{(NV)}\right]^2$. Without loss of generality, we assume that $\gamma_0 = 0$, $\gamma_1 = 1$ and $v_2 = 0$ so that we may write W_i instead of W_i^* and the pseudoscore bias correction becomes simply $V(W_i)$. A nonvalidation subject's contribution to $\widehat{\Sigma}_C$ can be written as

$$\begin{aligned}
&\left\{ \int_0^\tau [W_i - \bar{R}(t)]\,[dN_i(t) - Y_i(t)\,d\widehat{\Lambda}_0(t) - W_i\widehat{\beta}_C Y_i(t)\,dt] + \widehat{\beta}_C X_i V(W_i) \right\}^2 \\
&\quad = [W_i - \bar{R}(X_i)]^2 && (5.18)\\
&\qquad + \int_0^\tau \int_0^\tau [W_i - \bar{R}(s)]Y_i(s)[W_i - \bar{R}(t)]Y_i(t)\,d\widehat{\Lambda}_0(s)\,d\widehat{\Lambda}_0(t) && (5.19)\\
&\qquad + \int_0^\tau \int_0^\tau \{[W_i - \bar{R}(s)]W_i - V(W_i)\}Y_i(s) \\
&\qquad\quad \times \{[W_i - \bar{R}(t)]W_i - V(W_i)\}Y_i(t)\widehat{\beta}_C^2\,ds\,dt && (5.20)\\
&\qquad - [W_i - \bar{R}(X_i)] \int_0^\tau [W_i - \bar{R}(t)]Y_i(t)\,d\widehat{\Lambda}_0(t) && (5.21)\\
&\qquad - [W_i - \bar{R}(X_i)] \int_0^\tau \{[W_i - \bar{R}(t)]W_i - V(W_i)\}Y_i(t)\widehat{\beta}_C\,dt && (5.22)\\
&\qquad + \int_0^\tau \int_0^\tau [W_i - \bar{R}(s)]Y_i(s)\{[W_i - \bar{R}(t)]W_i - V(W_i)\}Y_i(t)\widehat{\beta}_C\,d\widehat{\Lambda}_0(s)\,dt. && (5.23)
\end{aligned}$$

It is easy to see that $\mathrm{E}\left[\widetilde{\psi}_i^{(NV)}\right]^2$ can be decomposed into six parts corresponding to (5.18)–(5.23) except that $\widehat{\beta}_C$ is replaced by β_0, $\widehat{\Lambda}_0$ by Λ_0, $\bar{R}(t)$ by $e(t)$ and expectation is taken over each of the terms. We will show that sample averages of

terms (5.18)–(5.23) converge to the corresponding terms in the decomposition of $\mathrm{E}\left[\widetilde{\psi}_i^{(NV)}\right]^2$.

Let us start with (5.18). Here,

$$\frac{1}{n}\sum[W_i - \bar{R}(X_i)]^2 = \frac{1}{n}\sum\{[W_i - e(X_i)] + [e(X_i) - \bar{R}(X_i)]\}^2 \to_p \mathrm{E}[W_i - e(X_i)]^2$$

since $\sup_t |e(t) - \bar{R}(t)| \to_p 0$.

The most difficult term is (5.19). Denote $\widehat{G}_i(t) \equiv [W_i - \bar{R}(t)]Y_i(t)$ and $G_i(t) \equiv [W_i - e(t)]Y_i(t)$. Then $G_i(t)$, $i = 1, \ldots, n$ are iid random variables and $\mathrm{E}\, G_i(s)G_i(t)$ is bounded by a constant. We have to show that

$$\left| \int_0^\tau \int_0^\tau \frac{1}{n}\sum \widehat{G}_i(s)\widehat{G}_i(t)\, d\widehat{\Lambda}_0(s) d\widehat{\Lambda}_0(t) - \int_0^\tau \int_0^\tau \mathrm{E}\, G_i(s)G_i(t)\, d\Lambda_0(s) d\Lambda_0(t) \right| \to_p 0. \tag{5.24}$$

But this is exactly what was shown in the proof of Theorem 4.4, only with slightly different $G_i(t)$ and $\widehat{G}_i(t)$. We can apply the same approach here, that is, first proving that

$$\sup_{0 \le s,t \le \tau} \left| \frac{1}{n}\sum_{i=1}^n \widehat{G}_i(s)\widehat{G}_i(t) - \mathrm{E}\, G_i(s)G_i(t) \right| \to_p 0$$

and then using Lemma 4.6 in the same way. For details, see the proof of Theorem 4.4.

The remaining terms (5.20)–(5.23) may be treated similarly. The idea is to approximate the sample average of each of these terms by an average of independent and identically distributed terms which converge to their expectation by the weak law of large numbers. In the process, $\bar{R}(t)$ has to be replaced by $e(t)$, $\widehat{\beta}_C$ by β_0 and $\widehat{\Lambda}_0(t)$ by $\Lambda_0(t)$. All the steps combined show that the estimator $\widehat{\Sigma}_C$ is consistent for Σ_C. □

5.4 *Unknown error parameters*

We finally relax the assumption that the error parameters describing the conditional distribution of W given Z are known. We defined five of them: the mean parameters

γ_0 and γ_1 and the variance parameters v_0, v_1 and v_2. Let $\boldsymbol{\theta}$ denote all the error parameters taken as a column vector: let $\boldsymbol{\theta}_0$ be the true value of $\boldsymbol{\theta}$: and let $1 \leq q \leq 5$ be the number of components in $\boldsymbol{\theta}$. With the current experimental design, it is easy to estimate $\boldsymbol{\theta}_0$ from the validation set. The estimated $\boldsymbol{\theta}$ can be substituted into the corrected pseudoscore U_C. We show that the estimator $\widehat{\beta}_C$ obtained from such a pseudoscore is consistent but its variance is larger than what it would have been should $\boldsymbol{\theta}_0$ be known. Therefore, an adjusted variance formula is derived and a consistent variance estimator is proposed.

5.4.1 Corrected pseudoscore with estimated error parameters

Since the validation set is selected by Bernoulli sampling and both the true and surrogate covariates are available for the validation set subjects, it is possible to estimate $\boldsymbol{\theta}_0$ consistently by applying any standard estimation procedure to the validation set data. A quasi-likelihood approach is convenient to estimate the error parameters even in the most general case. We will sketch that approach later on. But no matter how the estimator for $\boldsymbol{\theta}_0$ is obtained, we will assume that the following condition always holds:

Condition 5.5. The q-vector of error parameters $\boldsymbol{\theta}_0$ is estimated by $\widehat{\boldsymbol{\theta}}$, a solution of $\sum_{i=1}^{n} \xi_i \boldsymbol{\phi}_i(\boldsymbol{\theta}) = 0$ that satisfies

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = [\alpha \mathbb{D}_T(\boldsymbol{\theta}_0)]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \boldsymbol{\phi}_i(\boldsymbol{\theta}_0) + o_P(1), \tag{5.25}$$

where

$$\mathbb{D}_T(\boldsymbol{\theta}_0) \equiv - \mathrm{E} \frac{\partial \boldsymbol{\phi}_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^{\mathsf{T}}}$$

is a $q \times q$ matrix and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \boldsymbol{\phi}_i(\boldsymbol{\theta}_0) \rightarrow_d \mathrm{N}(\mathbf{0}, \alpha \Sigma_T(\boldsymbol{\theta}_0)).$$

Condition 5.5 merely assures that the estimator $\widehat{\boldsymbol{\theta}}$ is asymptotically linear, and hence consistent and asymptotically normal.

Let us demonstrate how $\widehat{\boldsymbol{\theta}}$ can be obtained from the quasi-likelihood approach. Suppose there are five error parameters to be estimated. When the variance parameters are known, the mean parameters can be estimated by solving the estimating equation $\sum \xi_i \phi_i^{(\gamma)}(\boldsymbol{\theta}) = 0$, where

$$\phi_i^{(\gamma)}(\boldsymbol{\theta}) = \frac{W_i - \gamma_0 - \gamma_1 Z_i}{V(Z_i)} \begin{pmatrix} 1 \\ Z_i \end{pmatrix}.$$

In turn, the solution of $\sum \xi_i \phi_i^{(v)}(\boldsymbol{\theta}) = 0$, where

$$\phi_i^{(v)}(\boldsymbol{\theta}) = \frac{(W_i - \gamma_0 - \gamma_1 Z_i)^2 - V(Z_i)}{V^2(Z_i)} \begin{pmatrix} 1 \\ Z_i \\ Z_i^2 \end{pmatrix},$$

provides a consistent estimator for v_0, v_1, v_2 when γ_0, γ_1 are known (Carroll and Ruppert, 1988). When neither the mean nor the variance parameters are known, a joint estimating equation is formed by solving the two concurrently: $\sum \xi_i \phi_i(\boldsymbol{\theta}) = 0$ where ϕ_i is a column vector of $\phi_i^{(\gamma)}$ and $\phi_i^{(v)}$.

The quasi-likelihood estimating equations provide consistent estimates of the error parameters in every case to which the corrected pseudoscore method can be applied. They satisfy Condition 5.5 under mild regularity conditions. However, whenever possible, they may be replaced by any other set of consistent estimating equations, e.g., those that define the least squares estimators for homoscedastic linear regression.

Expressing the dependence of the corrected pseudoscore on $\boldsymbol{\theta}$ as $U_C(\beta, \boldsymbol{\theta})$, the pseudoscore with known error parameters is obtained as $U_C(\beta, \boldsymbol{\theta}_0)$ and with estimated error parameters as $U_C(\beta, \widehat{\boldsymbol{\theta}})$. The asymptotic linearity of the former has been established earlier in Theorem 5.1; to prove the asymptotic linearity of the latter, we show that the difference between the two is asymptotically linear.

Lemma 5.2. *If Condition 5.5 holds then*

$$\frac{1}{\sqrt{n}}[U_C(\beta_0,\widehat{\boldsymbol{\theta}}) - U_C(\beta_0,\boldsymbol{\theta}_0)] = -w\frac{\overline{\alpha}}{\alpha}\boldsymbol{\Gamma}(\beta_0,\boldsymbol{\theta}_0)^{\mathsf{T}}\frac{1}{\sqrt{n}}\sum \xi_i o_i(\boldsymbol{\theta}_0) + o_P(1).$$

where $\boldsymbol{\Gamma}(\beta,\boldsymbol{\theta})^{\mathsf{T}} \equiv \boldsymbol{D}_X(\beta,\boldsymbol{\theta})\mathbb{D}_T^{-1}(\boldsymbol{\theta})$ *and* $\boldsymbol{D}_X(\beta,\boldsymbol{\theta}) \equiv -\,\mathrm{E}\,\partial\widetilde{\psi}_i^{(NV)}(\beta,\boldsymbol{\theta})/\partial\boldsymbol{\theta}^{\mathsf{T}}$.

Proof. By Taylor expansion,

$$\frac{1}{\sqrt{n}}[U_C(\beta_0,\widehat{\boldsymbol{\theta}}) - U_C(\beta_0,\boldsymbol{\theta}_0)] = \frac{1}{n}\frac{\partial U_C(\beta_0,\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}^{\mathsf{T}}}\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

where $\boldsymbol{\theta}^*$ lies on the line segment between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. Obviously,

$$\frac{1}{n}\frac{\partial U_C(\beta_0,\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}^{\mathsf{T}}} = \frac{1}{n}\frac{\partial\widetilde{U}_C(\beta_0,\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}^{\mathsf{T}}} + o_P(1) \to_p w\overline{\alpha}\,\mathrm{E}\frac{\partial\widetilde{\psi}_i^{(NV)}(\beta_0,\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}^{\mathsf{T}}} = w\overline{\alpha}\boldsymbol{D}_X(\beta_0,\boldsymbol{\theta}_0).$$

The desired result follows from the last two displayed equations and the equality (5.25). □

Corollary.

$$\begin{aligned}&\frac{1}{\sqrt{n}}U_C(\beta_0,\widehat{\boldsymbol{\theta}})\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\xi_i\widetilde{\psi}_i^{(V)}(\beta_0) + w\overline{\xi}_i\widetilde{\psi}_i^{(NV)}(\beta_0,\boldsymbol{\theta}_0) - w\xi_i\frac{\overline{\alpha}}{\alpha}\boldsymbol{\Gamma}(\beta_0,\boldsymbol{\theta}_0)^{\mathsf{T}}\phi_i(\boldsymbol{\theta}_0)\right] + o_P(1) \quad (5.26)\end{aligned}$$

and the limiting variance of $\sqrt{n}(\widehat{\beta}_C - \beta_0)$ is Σ_{CN}/D_C^2, where

$$\Sigma_{CN}(\beta_0,\boldsymbol{\theta}_0) = \alpha\Sigma_A(\beta_0) + w^2\overline{\alpha}\left\{\mathrm{E}\left[\widetilde{\psi}_i^{(NV)}(\beta_0,\boldsymbol{\theta}_0)\right]^2 + \frac{\overline{\alpha}}{\alpha}\boldsymbol{\Gamma}(\beta_0,\boldsymbol{\theta}_0)^{\mathsf{T}}\Sigma_T(\boldsymbol{\theta}_0)\boldsymbol{\Gamma}(\beta_0,\boldsymbol{\theta}_0)\right\} \quad (5.27)$$

and $\Sigma_{CN}(\beta_0,\boldsymbol{\theta}_0) \geq \Sigma_C(\beta_0,\boldsymbol{\theta}_0)$.

Proof. It is only necessary to prove (5.27). By (5.26),

$$\Sigma_{CN}(\beta_0,\boldsymbol{\theta}_0) = \mathrm{E}\left[\xi_i\widetilde{\psi}_i^{(V)}(\beta_0) + w\overline{\xi}_i\widetilde{\psi}_i^{(NV)}(\beta_0,\boldsymbol{\theta}_0) - w\xi_i\frac{\overline{\alpha}}{\alpha}\boldsymbol{\Gamma}(\beta_0,\boldsymbol{\theta}_0)^{\mathsf{T}}\phi_i(\boldsymbol{\theta}_0)\right]^2.$$

Suppressing all arguments, this can be written as

$$\begin{aligned}\Sigma_{CN} &= \text{var}\left[\xi_i\left(\widetilde{\psi}_i^{(V)} - \frac{\overline{\alpha}}{\alpha} w\boldsymbol{\Gamma}^{\mathsf{T}}\phi_i\right) + w\overline{\xi}_i\widetilde{\psi}_i^{(NV)}\right] \\ &= \alpha\,\text{var}\left(\widetilde{\psi}_i^{(V)} - \frac{\overline{\alpha}}{\alpha} w\boldsymbol{\Gamma}^{\mathsf{T}}\phi_i\right) + \overline{\alpha}w^2\,\text{var}\,\widetilde{\psi}_i^{(NV)}.\end{aligned}$$

The estimating function ϕ_i depends only on the pair of true and surrogate covariates Z_i and W_i^* while the validation contribution $\widetilde{\psi}_i^{(V)}$ involves Z_i, X_i and Δ_i, but not W_i^*. By Condition 5.4, W_i^* is independent of X_i and Δ_i given Z_i. It follows that $\widetilde{\psi}_i^{(V)}$ and ϕ_i are conditionally independent given Z_i and hence

$$\text{cov}\left(\widetilde{\psi}_i^{(V)}, \phi_i\right) = \text{E}\,\text{E}\left[\widetilde{\psi}_i^{(V)}\phi_i \,\middle|\, Z_i\right] = \text{E}\,\text{E}\left[\widetilde{\psi}_i^{(V)} \,\middle|\, Z_i\right]\text{E}\left[\phi_i \mid Z_i\right] = 0.$$

Thus we can write Σ_{CN} as

$$\alpha\,\text{var}\,\widetilde{\psi}_i^{(V)} + \alpha\left(\frac{\overline{\alpha}}{\alpha}\right)^2 w^2\boldsymbol{\Gamma}^{\mathsf{T}}\,\text{var}\,\phi_i\boldsymbol{\Gamma} + \overline{\alpha}w^2\,\text{var}\,\widetilde{\psi}_i^{(NV)}$$

which is equal to (5.27). The fact that $\Sigma_{CN} \geq \Sigma_C$ is obvious. □

The q-dimensional vector $\boldsymbol{\Gamma}(\beta_0, \boldsymbol{\theta}_0)$ will be called *the correction vector* and $\boldsymbol{\Gamma}^{\mathsf{T}}\phi_i$ will be referred to as *the correction factor* for the ith subject. The correction vector can be consistently estimated by $\widehat{\boldsymbol{\Gamma}}^{\mathsf{T}} = \widehat{\boldsymbol{D}}_X(\widehat{\beta}_C, \widehat{\boldsymbol{\theta}})\widehat{\mathbb{D}}_T^{-1}(\widehat{\boldsymbol{\theta}})$. There are two ways to obtain $\widehat{\boldsymbol{D}}_X$ and $\widehat{\mathbb{D}}_T$. First, we can take

$$\widehat{\mathbb{D}}_T(\boldsymbol{\theta}) = -\frac{1}{n\alpha}\sum_{i=1}^{n}\xi_i\frac{\partial\phi_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^{\mathsf{T}}}$$

and find $\widehat{\boldsymbol{D}}_X(\beta, \boldsymbol{\theta})$ in each special case by differentiating $\widetilde{\psi}_i^{(NV)}(\beta, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, replacing $\Lambda_0(t)$ by $\widehat{\Lambda}_0(t)$ and $e(t)$ by $\bar{R}(t)$, and averaging these terms over the nonvalidation set. Alternatively, we can calculate the expectations theoretically and estimate the resulting matrices term by term. Examples of the second approach will be discussed in Section 5.5.

With a consistent estimator of the correction vector, the adjusted pseudoscore variance Σ_{CN} can be consistently estimated by

$$\widehat{\Sigma}_{CN}(\beta,\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\xi_i\int_0^\tau [Z_i - \bar{R}(t)]^2\, dN_i(t) + w^2\left[\overline{\xi}_i\widehat{\psi}_i^{(NV)}(\beta) - \xi_i\frac{\overline{\alpha}}{\alpha}\widehat{\boldsymbol{\Gamma}}^{\mathsf{T}}\phi_i(\boldsymbol{\theta})\right]^2\right\} \tag{5.28}$$

evaluated at $\beta = \widehat{\beta}_C$ and $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$.

5.4.2 Selection of optimal weight

The previous sections have ignored the fact that the CS estimator we have proposed depends on a fixed but so far arbitrary weight w. Indeed, the CS estimator is actually a whole family of estimators indexed by w. All of them are consistent but the variance of any particular estimator is driven by the choice of w. In this section we propose a method for an adaptive selection of the weight in order to maximize efficiency. The resulting estimator is always more efficient than the estimator based only on the validation data. The following lemma shows how to calculate the weight w that minimizes the asymptotic variance of the CS estimator and how to estimate it from the data.

Lemma 5.3. *Let* $\sigma_1^2 = \operatorname{var}\widetilde{\psi}_i^{(V)}(\beta_0)$, $\sigma_2^2 = \operatorname{var}\widetilde{\psi}_i^{(NV)}(\beta_0,\boldsymbol{\theta}_0)$ *and*

$$\sigma_3^2 = \boldsymbol{\Gamma}(\beta_0,\boldsymbol{\theta}_0)^{\mathsf{T}}\operatorname{var}\phi_i(\boldsymbol{\theta}_0)\boldsymbol{\Gamma}(\beta_0,\boldsymbol{\theta}_0).$$

Then the estimator $\widehat{\beta}_C(w_{\text{opt}})$, *where*

$$w_{\text{opt}} = \frac{\sigma_1^2}{\sigma_2^2 + \overline{\alpha}/\alpha\sigma_3^2} \le 1,$$

possesses the minimal asymptotic variance within the class of estimators $\{\widehat{\beta}_C(w), 0 \le w \le 1\}$. *The optimal weight can be estimated consistently by replacing* σ_1^2, σ_2^2 *and* σ_3^2

by their consistent estimators

$$\widehat{\sigma}_1^2 = \frac{1}{\alpha} \sum \xi_i \int_0^\tau \{Z_i - \bar{R}(t)\}^2 \, dN_i(t).$$
$$\widehat{\sigma}_2^2 = \frac{1}{\overline{\alpha}} \sum \overline{\xi}_i \widehat{\psi}_i^{(NV)}(\widehat{\beta}_C)^2$$
$$\widehat{\sigma}_3^2 = \widehat{\mathbf{\Gamma}}^\mathsf{T} \frac{1}{\alpha n} \sum \xi_i \boldsymbol{\phi}_i^{\otimes 2}(\widehat{\boldsymbol{\theta}}) \widehat{\mathbf{\Gamma}}.$$

Corollary. The estimator $\widehat{\beta}_C(w_{\text{opt}})$ always has a smaller asymptotic variance than the complete-case estimator $\widehat{\beta}_C(0)$.

Proof. By the corollary to Lemma 5.2, the pseudoscore variance with unknown error parameters can be written as

$$\Sigma_{CN}(w) = \alpha\sigma_1^2 + \overline{\alpha} w^2 \left(\sigma_2^2 + \frac{\overline{\alpha}}{\alpha} \sigma_3^2 \right).$$

The variance of $\widehat{\beta}_C(w)$, taken as a function of w, has the form

$$f(w) = \text{const} \times \frac{\alpha\sigma_1^2 + \overline{\alpha} w^2 (\sigma_2^2 + \overline{\alpha}/\alpha\sigma_3^2)}{(\alpha + \overline{\alpha} w)^2}.$$

The derivative of $f(w)$,

$$f'(w) = \frac{2\overline{\alpha}\alpha}{(\alpha + \overline{\alpha} w)^3} \left[w \left(\sigma_2^2 - \frac{\overline{\alpha}}{\alpha} \sigma_3^2 \right) - \sigma_1^2 \right],$$

attains zero at w_{opt}. Since

$$\sigma_2^2 = \text{var}\, \widetilde{\psi}_i^{(NV)} = \text{E}\, \text{var} \left[\widetilde{\psi}_i^{(NV)} \,\middle|\, Z_i, X_i, \Delta_i \right] + \text{var} \underbrace{\text{E} \left[\widetilde{\psi}_i^{(NV)} \,\middle|\, Z_i, X_i, \Delta_i \right]}_{\widetilde{\psi}_i^{(V)}}$$

$$\geq \text{var}\, \widetilde{\psi}_i^{(V)} = \sigma_1^2,$$

the optimal weight never exceeds 1. The consistency of $\widehat{\sigma}_1^2$, $\widehat{\sigma}_2^2$ and $\widehat{\sigma}_3^2$ can be proved similarly to the consistency of the variance estimator in the proof of Theorem 5.4. □

Remark. When the error parameters are known, the optimal weight is simply $w_{\text{opt}} = \sigma_1^2/\sigma_2^2$.

Since $\widehat{w}_{\text{opt}}$ depends on $\widehat{\beta}_C$, both can be estimated simultaneously by a simple iterative procedure or, perhaps preferably, a working estimator of β_0 can be calculated with $w = 0$ and plugged into the expression for $\widehat{w}_{\text{opt}}$, which in turn is used to obtain the final $\widehat{\beta}_C$.

5.5 *Special cases*

In this section we show how the corrected score estimator is applied in different measurement error settings. For each case we identify the pseudoscore bias correction and calculate the variance adjustment for unknown error parameters.

5.5.1 Continuous covariate measured with error of constant variance

This is the simplest practical application of the corrected pseudoscore method. Let $W_i = Z_i + \varepsilon_i$, where $\varepsilon_1, \ldots, \varepsilon_n$ are zero-mean random variables, mutually independent, and independent of Z_i, N_i and Y_i. Let $\operatorname{var} \varepsilon_i = \sigma^2$, where σ^2 is unknown. The nonvalidation part of the corrected pseudoscore is given by

$$\psi_i^{(NV)}(\beta) = \int_0^\tau [W_i - \bar{R}(t)]\, dN_i(t) - \int_0^\tau \left\{[W_i - \bar{R}(t)]W_i - \widehat{\sigma}^2\right\} \beta Y_i(t)\, dt,$$

where

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^n \xi_i (W_i - Z_i)^2}{\sum_{i=1}^n \xi_i}.$$

It follows that $\widehat{\sigma}^2$ is the solution of $\sum \xi_i \phi_i = 0$ with $\phi_i = (W_i - Z_i)^2 - \sigma^2$. Denote $\varrho_i = \xi_i + \bar{\xi}_i w$ and $R_i = \xi_i Z_i + \bar{\xi}_i W_i$. Then the corrected pseudoscore estimator is

$$\widehat{\beta}_C = \frac{\sum \int_0^\tau \varrho_i [R_i - \bar{R}(t)]\, dN_i(t)}{\sum \int_0^\tau \varrho_i [R_i - \bar{R}(t)]^2 Y_i(t)\, dt + w\widehat{\sigma}^2 \sum \bar{\xi}_i X_i}.$$

The variance correction factor Γ equals $-\beta_0 \operatorname{E} X_i$. Hence, the asymptotic variance of $n^{-1/2}(\widehat{\beta}_C - \beta_0)$ is

$$\operatorname{E}\left\{\xi_i \widetilde{\psi}_i^{(V)} + w\bar{\xi}_i \widetilde{\psi}_i^{(NV)} + \frac{1-\alpha}{\alpha} \xi_i w \beta_0 [(W_i - Z_i)^2 - \sigma^2](\operatorname{E} X_i)\right\}^2 \Big/ D_C^2.$$

5.5.2 Linear regression with constant variance

Assume that the relationship between Z and W possesses the linear regression form

$$W_i = \gamma_0 + \gamma_1 Z_i + \varepsilon_i, \qquad \operatorname{var} \varepsilon_i = \sigma^2,$$

where γ_0, γ_1 and σ^2 are unknown and $\varepsilon_1, \ldots, \varepsilon_n$ are zero-mean random variables, mutually independent, and independent of Z_i, N_i and Y_i. The nonvalidation part of the corrected pseudoscore is now given by

$$\psi_i^{(NV)}(\beta) = \int_0^\tau [W_i^* - \bar{R}(t)]\, dN_i(t) - \int_0^\tau \left\{[W_i^* - \bar{R}(t)]W_i^* - \widehat{\gamma}_1^{-2}\widehat{\sigma}^2\right\} \beta Y_i(t)\, dt,$$

where $W_i^* = \widehat{\gamma}_1^{-1}(W_i - \widehat{\gamma}_0)$. The error parameters $\boldsymbol{\theta} = (\gamma_0, \gamma_1, \sigma^2)^\mathsf{T}$ have been replaced by $\widehat{\gamma}_0$ and $\widehat{\gamma}_1$, the least squares estimators of the regression parameters, and the maximum likelihood estimator $\widehat{\sigma}^2$ of the error variance given by

$$\widehat{\sigma}^2 \equiv \frac{1}{\sum_{i=1}^n \xi_i} \sum_{i=1}^n \xi_i (W_i - \widehat{\gamma}_0 - \widehat{\gamma}_1 Z_i)^2.$$

Hence, the error parameter estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\gamma}_0, \widehat{\gamma}_1, \widehat{\sigma}^2)^\mathsf{T}$ solves the estimating equation $\sum \xi_i \phi_i(\boldsymbol{\theta}) = 0$, where

$$\phi_i(\boldsymbol{\theta}) = \begin{pmatrix} \gamma_0 + \gamma_1 Z_i - W_i \\ \gamma_0 Z_i + \gamma_1 Z_i^2 - W_i Z_i \\ (W_i - \gamma_0 - \gamma_1 Z_i)^2 - \sigma^2 \end{pmatrix}.$$

At the end of this section we show that the correction factor is given by

$$\boldsymbol{\Gamma}(\boldsymbol{\theta})^\mathsf{T} \phi_i(\boldsymbol{\theta}) = -\frac{\beta_0}{\gamma_1}\Big\{ D_A(Z_i - \mathrm{E}\, Z_i)(W_i - \gamma_0 - \gamma_1 Z_i) + \frac{1}{\gamma_1} \mathrm{E}\, X_i \,\mathrm{var}\, Z_i \left[(W_i - \gamma_0 - \gamma_1 Z_i)^2 - \sigma^2\right] \Big\},$$

where $D_A = \mathrm{E} \int_0^\tau [Z_i - e(t)]^2 Y_i(t)\, dt$. It can be estimated by replacing $\mathrm{E}\, Z_i$ and $\mathrm{var}\, Z_i$ with their empirical counterparts and D_A with $\widehat{D}_A$. The pseudoscore variance is given by Equation (5.27) and its consistent estimator by (5.28).

Derivation of the correction factor. Since $W_i^*(\gamma_0, \gamma_1) = \gamma_1^{-1}(W_i - \gamma_0)$,

$$\frac{\partial W_i^*}{\partial \gamma_0} = -\frac{1}{\gamma_1}, \qquad \frac{\partial W_i^*}{\partial \gamma_1} = -\frac{1}{\gamma_1} W_i^*, \qquad \frac{\partial W_i^*}{\partial \sigma^2} = 0. \tag{5.29}$$

The pseudoscore bias correction, $V(W_i^*)/\gamma_1^2 = \sigma^2/\gamma_1^2$, has partial derivatives

$$\frac{\partial}{\partial\gamma_0}\frac{\sigma^2}{\gamma_1^2} = 0, \qquad \frac{\partial}{\partial\gamma_1}\frac{\sigma^2}{\gamma_1^2} = -2\frac{\sigma^2}{\gamma_1^3}, \qquad \frac{\partial}{\partial\sigma^2}\frac{\sigma^2}{\gamma_1^2} = \frac{1}{\gamma_1^2}.$$

Denote

$$Q_i(W_i^*) = \int_0^\tau [W_i^* - e(t)]\,[dN_i(t) - Y_i(t)\,d\Lambda_o(t) - W_i^*\beta_0 Y_i(t)\,dt].$$

Then

$$\frac{\partial Q_i(W_i^*)}{\partial W_i^*} = \int_0^\tau [dN_i(t) - Y_i(t)\,d\Lambda_o(t) - W_i^*\beta_0 Y_i(t)\,dt] - \int_0^\tau [W_i^* - e(t)]\beta_0 Y_i(t)\,dt \tag{5.30}$$

and, since $\mathrm{E}\,[W_i^* \mid Z_i] = Z_i$ and $\mathrm{E}\,[W_i^{*2} \mid Z_i] = Z_i^2 + \sigma^2/\gamma_1^2$,

$$\begin{aligned}
\mathrm{E}\,\frac{\partial Q_i(W_i^*)}{\partial W_i^*} &= \mathrm{E}\left\{M_i(\tau) - \int_0^\tau [Z_i - e(t)]\beta_0 Y_i(t)\,dt\right\} \\
&= -\beta_0 \int_0^\tau [\mathrm{E}\,Z_i Y_i(t) - e(t)\,\mathrm{E}\,Y_i(t)]\,dt \\
&= -\beta_0 \int_0^\tau [\pi_1(t) - e(t)\pi_0(t)]\,dt = 0, \\
\mathrm{E}\,W_i^*\frac{\partial Q_i(W_i^*)}{\partial W_i^*} &= \mathrm{E}\left\{Z_i M_i(\tau) - \frac{\sigma^2}{\gamma_1^2}\beta_0 X_i - \int_0^\tau Z_i[Z_i - e(t)]\beta_0 Y_i(t)\,dt - \frac{\sigma^2}{\gamma_1^2}\beta_0 X_i\right\} \\
&= -\beta_0\,\mathrm{E}\int_0^\tau Z_i[Z_i - e(t)]\,dt - 2\frac{\sigma^2}{\gamma_1^2}\beta_0\,\mathrm{E}\,X_i \\
&= -\beta_0\,\mathrm{E}\int_0^\tau [Z_i - e(t)]^2\,dt - 2\frac{\sigma^2}{\gamma_1^2}\beta_0\,\mathrm{E}\,X_i.
\end{aligned}$$

It follows that

$$\begin{aligned}
-\mathrm{E}\,\frac{\partial\tilde{\psi}_i^{(NV)}(\beta_0,\boldsymbol{\theta})}{\partial\gamma_0} &= \frac{1}{\gamma_1}\mathrm{E}\,\frac{\partial Q_i(W_i^*)}{\partial W_i^*} = 0, \\
-\mathrm{E}\,\frac{\partial\tilde{\psi}_i^{(NV)}(\beta_0,\boldsymbol{\theta})}{\partial\gamma_1} &= \frac{1}{\gamma_1}\mathrm{E}\,W_i^*\frac{\partial Q_i(W_i^*)}{\partial W_i^*} + 2\frac{\sigma^2}{\gamma_1^3}\beta_0\,\mathrm{E}\,X_i = -\frac{\beta_0}{\gamma_1}D_A \\
-\mathrm{E}\,\frac{\partial\tilde{\psi}_i^{(NV)}(\beta_0,\boldsymbol{\theta})}{\partial\sigma^2} &= -\frac{1}{\gamma_1^2}\beta_0\,\mathrm{E}\,X_i.
\end{aligned}$$

These are the components of $\boldsymbol{D}_X$, which is a row vector in this case. Let us now calculate $\mathbb{D}_T$. Denoting $u_i = W_i - \gamma_0 - \gamma_1 Z_i$, we have

$$\frac{\partial \phi_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\mathsf{T}} = \begin{pmatrix} 1 & Z_i & 0 \\ Z_i & Z_i^2 & 0 \\ -2u_i & -2Z_i u_i & -1 \end{pmatrix} \quad \text{and} \quad \mathrm{E}\,\frac{\partial \phi_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\mathsf{T}} = \begin{pmatrix} 1 & \mathrm{E}\,Z_i & 0 \\ \mathrm{E}\,Z_i & \mathrm{E}\,Z_i^2 & 0 \\ 0 & 0 & -1 \end{pmatrix} = -\mathbb{D}_T$$

and hence

$$\mathbb{D}_T^{-1} = \frac{1}{\operatorname{var} Z_i} \begin{pmatrix} -\mathrm{E}\,Z_i^2 & \mathrm{E}\,Z_i & 0 \\ \mathrm{E}\,Z_i & -1 & 0 \\ 0 & 0 & \operatorname{var} Z_i \end{pmatrix}.$$

Thus, $\boldsymbol{\Gamma}^\mathsf{T} = \boldsymbol{D}_X \mathbb{D}_T^{-1} = \beta_0 \gamma_1^{-1}(-D_A\,\mathrm{E}\,Z_i, D_A, \mathrm{E}\,X_i \operatorname{var} Z_i/\gamma_1)$ and

$$\boldsymbol{\Gamma}^\mathsf{T}\phi_i = \frac{\beta_0}{\gamma_1}\left[D_A u_i\,\mathrm{E}\,Z_i - D_A u_i Z_i - \mathrm{E}\,X_i \operatorname{var} Z_i(u_i^2 - \sigma^2)/\gamma_1\right].$$

This finishes the derivation. □

5.5.3 *Misclassified binary covariate*

Suppose that Z is a binary covariate with $\mathrm{P}[Z = 1] = p_Z$ and W is a misclassified surrogate for Z. Denote the sensitivity of W for Z by $\eta \equiv \mathrm{P}[W = 1 \mid Z = 1]$ and the specificity by $\nu \equiv \mathrm{P}[W = 0 \mid Z = 0]$. Then the expectation of W given Z is

$$\mathrm{E}[W \mid Z] = \eta Z + (1 - \nu)(1 - Z) = 1 - \nu + (\eta + \nu - 1)Z.$$

Thus, the conditional expectation possesses the required linear form with $\gamma_0 = 1 - \nu$ and $\gamma_1 = \eta + \nu - 1$. We assume that $\gamma_1 \neq 0$. The conditional variance of W is

$$V(Z) = \eta(1 - \eta)Z + \nu(1 - \nu)(1 - Z) = \nu(1 - \nu) + [\eta(1 - \eta) - \nu(1 - \nu)]Z,$$

which is also of a linear form $v_0 + v_1 Z$ with $v_0 = \nu(1-\nu)$ and $v_1 = \eta(1-\eta) - \nu(1-\nu)$. Because all the error parameters are functions of sensitivity and specificity, we can

parametrize the problem either in η and ν or in γ_0 and γ_1. We will switch from one parametrization to the other as convenient.

It follows that the bias-corrected version of W is $W^* = \gamma_1^{-1}(W - 1 + \nu)$. An unbiased estimator for $V(Z)$ is obtained by $V(W^*) = \nu(1-\nu)+[\eta(1-\eta)-\nu(1-\nu)]W^*$. The nonvalidation part of the pseudoscore is

$$\psi_i^{(NV)}(\beta) = \int_0^\tau [W_i^* - \bar{R}(t)] \{dN_i(t) - [W_i^* - \bar{R}(t)]\beta Y_i(t)\, dt\} + \beta\gamma_1^{-2} X_i V(W_i^*).$$

The error parameters $\boldsymbol{\theta} = (\eta, \nu)^\mathsf{T}$ can be estimated by $\widehat{\eta} = n_{11}/(n_{10} + n_{11})$ and $\widehat{\nu} = n_{00}/(n_{00} + n_{01})$, where $n_{ij} = \sum_k \xi_k \mathbb{1}(Z_k = i, W_k = j)$. This means that $\widehat{\boldsymbol{\theta}}$ is the solution of $\sum \xi_i \phi_i(\boldsymbol{\theta}) = 0$ with

$$\phi_i(\boldsymbol{\theta}) = \begin{pmatrix} \mathbb{1}(Z_i = 1, W_i = 1) - \eta \mathbb{1}(Z_i = 1) \\ \mathbb{1}(Z_i = 0, W_i = 0) - \nu \mathbb{1}(Z_i = 0) \end{pmatrix}.$$

After some clumsy but simple algebra (see the end of this section), it can be shown that the correction factor is

$$\boldsymbol{\Gamma}^\mathsf{T} \phi_i(\boldsymbol{\theta}) = -\frac{\beta_0}{\gamma_1} \left\{ p_Z^{-1} \int_0^\tau e^2(t)\pi(t)\, dt Z_i(W_i - \eta) \right. \\ \left. + (1 - p_Z)^{-1} \int_0^\tau [1 - e(t)]^2 \pi(t)\, dt (1 - Z_i)(1 - W_i - \nu) \right\}.$$

Derivation of the correction factor. To derive the correction factor, we will work with γ_0 and γ_1 and express the result in terms of η and ν at the very end. Since $\nu = 1 - \gamma_0$ and $\eta = \gamma_0 + \gamma_1$, the variance function is

$$\begin{aligned} V(\gamma_0, \gamma_1, W_i^*) &= \gamma_0(1 - \gamma_0)(1 - W_i^*) + (\gamma_0 + \gamma_1)(1 - \gamma_0 - \gamma_1) W_i^* \\ &= \gamma_0(1 - \gamma_0) + \gamma_1(1 - 2\gamma_0 - \gamma_1) W_i^*. \end{aligned}$$

The partial derivatives of the pseudoscore bias correction, $V(W_i^*)/\gamma_1^2$, are

$$\frac{\partial}{\partial \gamma_0} \frac{V(W_i^*)}{\gamma_1^2} = \frac{1}{\gamma_1^2} \left[1 - 2\gamma_0 - 2\gamma_1 W_i^* - \gamma_1(1 - 2\gamma_0 - \gamma_1) \frac{1}{\gamma_1} \right] = \frac{1}{\gamma_1}(1 - 2W_i^*)$$

and

$$\frac{\partial}{\partial\gamma_1}\frac{V(W_i^*)}{\gamma_1^2} = -2\frac{V(W_i^*)}{\gamma_1^3} + \frac{1}{\gamma_1^2}\left[(1-2\gamma_0-2\gamma_1)W_i^* - \gamma_1(1-2\gamma_0-\gamma_1)\frac{1}{\gamma_1}W_i^*\right]$$
$$= -\frac{2}{\gamma_1}\frac{V(W_i^*)}{\gamma_1^2} - \frac{W_i^*}{\gamma_1}.$$

Using the first two equalities in (5.29), Equation (5.30), $\mathrm{E}[W_i^* \mid Z_i] = Z_i$ and $\mathrm{E}[W_i^{*2} \mid Z_i] = Z_i^2 + V(Z_i)/\gamma_1^2$, and realizing that $Z_i^2 = Z_i$, we get

$$-\mathrm{E}\frac{\partial\tilde{\psi}_i^{(NV)}(\beta_0,\boldsymbol{\theta})}{\partial\gamma_0} = \frac{1}{\gamma_1}\mathrm{E}\frac{\partial Q_i(W_i^*)}{\partial W_i^*} - \mathrm{E}\left[\beta_0 X_i\frac{\partial}{\partial\gamma_0}\frac{V(W_i^*)}{\gamma_1^2}\right]$$
$$= -\frac{\beta_0}{\gamma_1}\mathrm{E}\int_0^\tau (1-2Z_i)Y_i(t)\,dt = -\frac{\beta_0}{\gamma_1}\mathrm{E}\int_0^\tau [1-2e(t)]\pi_0(t)\,dt,$$

$$-\mathrm{E}\frac{\partial\tilde{\psi}_i^{(NV)}(\beta_0,\boldsymbol{\theta})}{\partial\gamma_1} = \frac{1}{\gamma_1}\mathrm{E}\,W_i^*\frac{\partial Q_i(W_i^*)}{\partial W_i^*} - \mathrm{E}\left[\beta_0 X_i\frac{\partial}{\partial\gamma_1}\frac{V(W_i^*)}{\gamma_1^2}\right]$$
$$= -\frac{\beta_0}{\gamma_1}\mathrm{E}\int_0^\tau Z_i[Z_i - e(t)]Y_i(t)\,dt$$
$$-2\frac{\beta_0}{\gamma_1}\mathrm{E}\frac{V(Z_i)}{\gamma_1^2}\int_0^\tau Y_i(t)\,dt + \beta_0\,\mathrm{E}\left[\frac{2}{\gamma_1}\frac{V(Z_i)}{\gamma_1^2} + \frac{Z_i}{\gamma_1}\right]\int_0^\tau Y_i(t)\,dt$$
$$= \frac{\beta_0}{\gamma_1}\int_0^\tau e^2(t)\pi_0(t)\,dt.$$

We have got $\boldsymbol{D}_X$. The estimating function for γ_0 and γ_1 can be written as

$$\phi_i(\boldsymbol{\theta}) = \begin{pmatrix} (1-Z_i)\gamma_0 - (1-Z_i)W_i \\ Z_i\gamma_0 + Z_i\gamma_1 - Z_iW_i \end{pmatrix}.$$

Thus, $\mathbb{D}_T = -E\partial\phi_i(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^\mathsf{T}$ and its inverse turn out to be

$$\mathbb{D}_T = -\begin{pmatrix} 1-p_Z & 0 \\ p_Z & p_Z \end{pmatrix} \quad \text{and} \quad \mathbb{D}_T^{-1} = \begin{pmatrix} -(1-p_Z)^{-1} & 0 \\ (1-p_Z)^{-1} & -p_Z^{-1} \end{pmatrix},$$

which implies that

$$\boldsymbol{\Gamma}^\mathsf{T} = \boldsymbol{D}_X\mathbb{D}_T^{-1} = \left(\frac{1}{1-p_Z}\mathrm{E}\left[\frac{\partial\tilde{\psi}_i^{(NV)}}{\partial\gamma_0} - \frac{\partial\tilde{\psi}_i^{(NV)}}{\partial\gamma_1}\right], -\frac{1}{p_Z}\mathrm{E}\frac{\partial\tilde{\psi}_i^{(NV)}}{\partial\gamma_1}\right)$$
$$= \frac{\beta_0}{\gamma_1}\left(\frac{1}{1-p_Z}\int_0^\tau [e^2(t)+1-2e(t)]\pi_0(t)\,dt, \frac{1}{p_Z}\int_0^\tau e^2(t)\pi_0(t)\,dt\right)$$

and hence

$$\begin{aligned}\boldsymbol{\Gamma}^{\mathsf{T}}\phi_i =& \frac{\beta_0}{\gamma_1}\frac{1}{1-p_Z}\int_0^\tau [1-e(t)]^2\pi_0(t)\,dt(1-Z_i)(\gamma_0-W_i) \\ &+ \frac{\beta_0}{\gamma_1}\frac{1}{p_Z}\int_0^\tau e^2(t)\pi_0(t)\,dt Z_i(\gamma_0+\gamma_1-W_i) \\ =& -\frac{\beta_0}{\gamma_1}\left\{\frac{Z_i}{p_Z}(W_i-\eta)\int_0^\tau e^2(t)\pi_0(t)\,dt\right. \\ &\left.+\frac{1-Z_i}{1-p_Z}(1-W_i-\nu)\int_0^\tau [1-e(t)]^2\pi_0(t)\,dt\right\}\end{aligned}$$

□

5.6 *Multiple covariates*

In this section, we define the corrected score estimator for the situation where one covariate is measured with an error and the remaining covariates are known exactly. We assume that the true values of the first covariate are observed on a validation set selected by Bernoulli sampling. A surrogate covariate is available for all the study subjects, as in the univariate case described previously. We will work under the same assumptions on the surrogate covariate and its relationship to the true covariate and other observed data as introduced for the univariate case.

5.6.1 Definition of the multivariate corrected pseudoscore

Let the true covariate vector for the ith subject be $\boldsymbol{Z}_i(t) = \begin{pmatrix} Z_{1i} \\ \boldsymbol{Z}_{2i}(t) \end{pmatrix}$ where $\boldsymbol{Z}_i(t)$ is a time-dependent p-vector, Z_{1i} is the incompletely observed time-independent covariate, and $\boldsymbol{Z}_{2i}(t)$ is a $(p-1)$-vector of time-dependent covariates that are observed for all the subjects. Let W_i be the surrogate for Z_{1i} subject to Conditions 5.1–5.4, where all the conditioning is made on Z_{1i} only. We continue to write W_i^* for the bias-adjusted surrogate. The true value of Z_{1i} is known only if the subject is in the validation set, i.e., if $\xi_i = 1$. So, define the observed first covariate as $R_{1i} = \xi_i Z_{1i} + \bar{\xi}_i W_i^*$

and the vector of observed covariates as $\boldsymbol{R}_i(t) = (R_{1i}, \boldsymbol{Z}_{2i}(t))^\mathsf{T}$. Other p-vectors, like β_0, $\tilde{\psi}_i^{(V)}$, etc., are similarly partitioned into the first component and the remaining $(p-1)$-vector.

To define the corrected pseudoscore, we have to generalize the downweighting constant w to the multivariate case. We propose to weight the pseudoscore contributions of the nonvalidation subjects by a general $p \times p$ invertible matrix Ω. Intuitively, one would rather use a diagonal matrix; however the definition of the pseudoscore works fine with general weight matrices and, as we will show, the optimal weight matrix is actually non-diagonal.

So, the contribution of the ith subject is weighted by the matrix $\mathbb{A}_i = \xi_i \mathbb{I}_p + \overline{\xi}_i \Omega$. The expectation of $\mathbb{A}_i$ is $\mathbb{A} = \alpha \mathbb{I}_p + \overline{\alpha}\Omega$. The weighted at-risk covariate average is defined by

$$\overline{\boldsymbol{R}}(t) = \left[\sum_{i=1}^{n} \mathbb{A}_i Y_i(t)\right]^{-1} \left[\sum_{i=1}^{n} \mathbb{A}_i \boldsymbol{R}_i(t) Y_i(t)\right].$$

The multivariate corrected pseudoscore then attains the form

$$U_C(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbb{A}_i \int_0^\tau [\boldsymbol{R}_i(t) - \overline{\boldsymbol{R}}(t)]\,[dN_i(t) - \boldsymbol{R}_i(t)^\mathsf{T}\boldsymbol{\beta} Y_i(t)\,dt]$$
$$+ \sum_{i=1}^{n} \overline{\xi}_i \mathbb{A}_i \boldsymbol{j}^{\otimes 2}\boldsymbol{\beta} \frac{V(W_i^*)}{\gamma_1^2 + v_2} X_i,$$

where $\boldsymbol{j}$ is the p-vector with 1 as the first component and zeros elsewhere. The whole construction $\boldsymbol{j}^{\otimes 2}\boldsymbol{\beta}$ just creates a vector with β_1 as the first component and zeros elsewhere. It is easy to see that

$$\sum \mathbb{A}_i [\boldsymbol{R}_i(t) - \overline{\boldsymbol{R}}(t)] Y_i(t) = 0.$$

This is an important property because it implies that U_C can be written as

$$U_C(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbb{A}_i \int_0^\tau [\boldsymbol{R}_i(t) - \overline{\boldsymbol{R}}(t)]\,[dN_i(t) - Y_i(t)\,d\Lambda_0(t) - \boldsymbol{R}_i(t)^\mathsf{T}\boldsymbol{\beta} Y_i(t)\,dt]$$
$$+ \sum_{i=1}^{n} \overline{\xi}_i \mathbb{A}_i \boldsymbol{j}^{\otimes 2}\boldsymbol{\beta} \frac{V(W_i^*)}{\gamma_1^2 + v_2} X_i$$

and also as

$$U_C(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbb{A}_i \int_0^\tau [\boldsymbol{R}_i(t) - \overline{\boldsymbol{R}}(t)]\, dN_i(t) \\ - \sum_{i=1}^{n} \mathbb{A}_i \int_0^\tau \left\{ [\boldsymbol{R}_i(t) - \overline{\boldsymbol{R}}(t)]^{\otimes 2} - \overline{\xi}_i \frac{V(W_i^*)}{\gamma_1^2 + v_2} \boldsymbol{j}^{\otimes 2} \right\} \boldsymbol{\beta} Y_i(t)\, dt.$$

The estimator defined by $U_C(\widehat{\boldsymbol{\beta}}_C) = 0$ is

$$\widehat{\boldsymbol{\beta}}_C = \left(\sum_{i=1}^{n} \mathbb{A}_i \int_0^\tau \left\{ [\boldsymbol{R}_i(t) - \overline{\boldsymbol{R}}(t)]^{\otimes 2} - \overline{\xi}_i \frac{V(W_i^*)}{\gamma_1^2 + v_2} \boldsymbol{j}^{\otimes 2} \right\} Y_i(t)\, dt \right)^{-1} \\ \times \left(\sum_{i=1}^{n} \mathbb{A}_i \int_0^\tau [\boldsymbol{R}_i(t) - \overline{\boldsymbol{R}}(t)]\, dN_i(t) \right).$$

5.6.2 Limiting distribution of the multivariate CS estimator

The normalized pseudoscore is asymptotically linear, that is

$$\frac{1}{\sqrt{n}} U_C(\boldsymbol{\beta}_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{A}_i \begin{pmatrix} \xi_i \widetilde{\psi}_{1i}^{(V)}(\boldsymbol{\beta}_0) + \overline{\xi}_i \widetilde{\psi}_{1i}^{(NV)}(\boldsymbol{\beta}_0) \\ \xi_i \widetilde{\psi}_{2i}^{(V)}(\boldsymbol{\beta}_0) + \overline{\xi}_i \widetilde{\psi}_{2i}^{(NV)}(\boldsymbol{\beta}_0) \end{pmatrix} + o_P(1) \\ = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\xi_i \widetilde{\psi}_i^{(V)}(\boldsymbol{\beta}_0) + \overline{\xi}_i \Omega \widetilde{\psi}_i^{(NV)}(\boldsymbol{\beta}_0) \right] + o_P(1),$$

where

$$\widetilde{\psi}_{1i}^{(V)}(\boldsymbol{\beta}_0) = \int_0^\tau [Z_{1i} - e_1(t)]\, dM_i(t),$$

$$\widetilde{\psi}_{1i}^{(NV)}(\boldsymbol{\beta}_0) = \int_0^\tau [W_i^* - e_1(t)]\, [dN_i(t) - Y_i(t)\, d\Lambda_0(t) - \boldsymbol{R}_i(t)^{\mathsf{T}} \boldsymbol{\beta}_0 Y_i(t)\, dt] \\ + \frac{V(W_i^*)}{\gamma_1^2 + v_2} \beta_{01} X_i,$$

$$\widetilde{\psi}_{2i}^{(V)}(\boldsymbol{\beta}_0) = \int_0^\tau [\boldsymbol{Z}_{2i}(t) - \boldsymbol{e}_2(t)]\, dM_i(t),$$

$$\widetilde{\psi}_{2i}^{(NV)}(\boldsymbol{\beta}_0) = \int_0^\tau [\boldsymbol{Z}_{2i}(t) - \boldsymbol{e}_2(t)]\, dM_i(t) + \int_0^\tau [\boldsymbol{Z}_{2i}(t) - \boldsymbol{e}_2(t)](Z_{1i} - W_i^*) \beta_{01} Y_i(t)\, dt$$

are mean zero independent and identically distributed terms. Thus, $n^{-1/2} U_C(\boldsymbol{\beta}_0)$ converges in law to p-variate normal distribution with zero mean and covariance

matrix

$$\Sigma_C(\boldsymbol{\beta}_0) = \mathrm{E}\left[\xi_i\widetilde{\psi}_i^{(V)}(\boldsymbol{\beta}_0) + \overline{\xi}_i\Omega\widetilde{\psi}_i^{(NV)}(\boldsymbol{\beta}_0)\right]^{\otimes 2} = \alpha\Sigma_A(\boldsymbol{\beta}_0) + \overline{\alpha}\Omega\Sigma_B(\boldsymbol{\beta}_0)\Omega^{\mathsf{T}}.$$

where Σ_A is the covariance of the full-data pseudoscore and Σ_B is the covariance of $\widetilde{\psi}_i^{(NV)}$. The negative expected derivative matrix of U_C is

$$-\mathrm{E}\,\frac{\partial}{\partial\boldsymbol{\beta}^{\mathsf{T}}}U_C(\boldsymbol{\beta}) = (\alpha\mathbb{I} + \overline{\alpha}\Omega)\mathbb{D}_A,$$

where $\mathbb{D}_A$ is the negative expected derivative of the full-data pseudoscore U_A.

It follows that the asymptotic covariance matrix of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}_0)$ is

$$\mathbb{D}_A^{-1}(\alpha\mathbb{I} + \overline{\alpha}\Omega)^{-1}(\alpha\Sigma_A + \overline{\alpha}\Omega\Sigma_B\Omega^{\mathsf{T}})(\alpha\mathbb{I} + \overline{\alpha}\Omega^{\mathsf{T}})^{-1}\mathbb{D}_A^{-\mathsf{T}}.$$

Denote the product of the inner three matrices by $\Phi(\Omega)$. The goal is to find Ω_0 such that $\Phi(\Omega) - \Phi(\Omega_0) \geq 0$ for any invertible weight matrix Ω. By analogy with the univariate case one can guess that the optimal matrix satisfies $\Omega_0\Sigma_A = \Omega_0\Sigma_B\Omega_0^{\mathsf{T}}$, which implies that $\Omega_0 = \Sigma_A\Sigma_B^{-1}$. With this Ω_0, we get

$$\begin{aligned}\Phi(\Omega_0) &= (\alpha\mathbb{I} + \overline{\alpha}\Sigma_A\Sigma_B^{-1})^{-1}(\alpha\Sigma_A + \overline{\alpha}\Sigma_A\Sigma_B^{-1}\Sigma_B\Sigma_B^{-1}\Sigma_A)(\alpha\mathbb{I} + \overline{\alpha}\Sigma_B^{-1}\Sigma_A)^{-1}\\ &= (\alpha\mathbb{I} + \overline{\alpha}\Sigma_A\Sigma_B^{-1})^{-1}(\alpha\mathbb{I} + \overline{\alpha}\Sigma_A\Sigma_B^{-1})\Sigma_A(\alpha\mathbb{I} + \overline{\alpha}\Sigma_B^{-1}\Sigma_A)^{-1}\\ &= \Sigma_A(\alpha\mathbb{I} + \overline{\alpha}\Sigma_B^{-1}\Sigma_A)^{-1} = (\alpha\Sigma_A^{-1} + \overline{\alpha}\Sigma_B^{-1})^{-1}.\end{aligned}$$

The following lemma shows that Ω_0 is indeed optimal.

Lemma 5.4.

$$(\alpha\mathbb{I} + \overline{\alpha}\Omega)^{-1}(\alpha\Sigma_A + \overline{\alpha}\Omega\Sigma_B\Omega^{\mathsf{T}})(\alpha\mathbb{I} + \overline{\alpha}\Omega^{\mathsf{T}})^{-1} - (\alpha\Sigma_A^{-1} + \overline{\alpha}\Sigma_B^{-1})^{-1} \geq 0,$$

for any Ω invertible.

Proof. Consider the linear model $\boldsymbol{y} = \mathbb{X}\boldsymbol{b} + \boldsymbol{\varepsilon}$, where $\mathbb{X} = \begin{pmatrix}\sqrt{\alpha}\mathbb{I}_p\\ \sqrt{1-\alpha}\mathbb{I}_p\end{pmatrix}$ is a $2p \times p$ design matrix, $\boldsymbol{y}$ is a $2p$-vector of responses, $\boldsymbol{b}$ is a p-vector of parameters, and $\boldsymbol{\varepsilon} \sim$

$N_{2p}(0, \mathbb{W})$ is the random error. Let the error covariance matrix be $\mathbb{W} = \begin{pmatrix} \Sigma_A & 0 \\ 0 & \Sigma_B \end{pmatrix}$. All unbiased linear estimators in such a model have the form $\hat{\boldsymbol{b}} = (\mathbb{X}^\mathsf{T}\mathbb{C}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\mathbb{C}\boldsymbol{y}$ where $\mathbb{C}$ is a $2p \times 2p$ matrix of rank $2p$. The covariance of $\hat{\boldsymbol{b}}$ is

$$\operatorname{var}\hat{\boldsymbol{b}} = (\mathbb{X}^\mathsf{T}\mathbb{C}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\mathbb{C}\mathbb{W}\mathbb{C}^\mathsf{T}\mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{C}^\mathsf{T}\mathbb{X})^{-1}.$$

Notice that the choice $\mathbb{C}_0 = \begin{pmatrix} \mathbb{I}_p & 0 \\ 0 & \Omega \end{pmatrix}$ leads to

$$\operatorname{var}\hat{\boldsymbol{b}} = (\alpha\mathbb{I} + \overline{\alpha}\Omega)^{-1}(\alpha\Sigma_A + \overline{\alpha}\Omega\Sigma_B\Omega^\mathsf{T})(\alpha\mathbb{I} + \overline{\alpha}\Omega^\mathsf{T})^{-1}. \tag{5.31}$$

By the Gauss-Markov Theorem, the best linear unbiased estimator is obtained by setting $\mathbb{C} = \mathbb{W}^{-1} = \begin{pmatrix} \Sigma_A^{-1} & 0 \\ 0 & \Sigma_B^{-1} \end{pmatrix}$ and its covariance matix equals

$$(\mathbb{X}^\mathsf{T}\mathbb{W}^{-1}\mathbb{X})^{-1} = (\alpha\Sigma_A^{-1} + \overline{\alpha}\Sigma_B^{-1})^{-1}. \tag{5.32}$$

Thus, the difference between (5.31) and (5.32) must be a positive semi-definite matrix. □

The optimal weight matrix Ω_0 is non-symmetric. Its diagonal elements should range from 0 to 1, with the last $p-1$ of them being much closer to 1 than the first one. The off-diagonal elements should be much smaller than the diagonal ones. This is just a conjecture based on the meaning of Σ_A and Σ_B and on the fact that Ω_0 must be close to the identity matrix when the measurement error is small.

The pseudoscore variance Σ_C can be estimated by

$$\widehat{\Sigma}_C(\widehat{\boldsymbol{\beta}}_C) = \frac{1}{n}\sum_{i=1}^{n}\left\{\xi_i\int_0^\tau [\boldsymbol{Z}_i(t) - \overline{\boldsymbol{R}}(t)]^{\otimes 2}\, dN_i(t) + \overline{\xi}_i\Omega\left[\widehat{\psi}_i^{(NV)}(\widehat{\boldsymbol{\beta}}_C)\right]^{\otimes 2}\Omega^\mathsf{T}\right\},$$

where $\widehat{\psi}_i^{(NV)}(\boldsymbol{\beta})$ is a column vector having

$$\int_0^\tau [W_i^* - \bar{R}_1(t)]\left[dN_i(t) - Y_i(t)\, d\widehat{\Lambda}_0(t) - \boldsymbol{R}_i(t)^\mathsf{T}\boldsymbol{\beta}Y_i(t)\, dt\right] + \boldsymbol{j}^\mathsf{T}\boldsymbol{\beta}\frac{V(W_i^*)}{\gamma_1^2 + v_2}X_i$$

as the first component and

$$\int_0^\tau [\boldsymbol{Z}_{2i}(t) - \overline{\boldsymbol{R}}_2(t)] \left[dN_i(t) - Y_i(t)\, d\widehat{\Lambda}_0(t) - \boldsymbol{R}_i(t)^\mathsf{T} \boldsymbol{\beta} Y_i(t)\, dt \right]$$

as the last $p-1$ components. The baseline hazard estimator has an easy extension to multiple covariates,

$$d\widehat{\Lambda}_0(t) = \frac{\sum_{i=1}^n dN_i(t)}{\sum_{j=1}^n Y_j(t)} - \overline{\boldsymbol{R}}(t)^\mathsf{T} \widehat{\boldsymbol{\beta}}_C\, dt.$$

When the q-vector of error parameters $\boldsymbol{\theta}$ is estimated from the validation set, the pseudoscore variance includes an extra term:

$$\begin{aligned} \Sigma_C(\boldsymbol{\beta}_0) &= \mathrm{E} \left[\xi_i \widetilde{\psi}_i^{(V)}(\boldsymbol{\beta}_0) + \overline{\xi}_i \Omega \widetilde{\psi}_i^{(NV)}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) - \xi_i \frac{\overline{\alpha}}{\alpha} \Omega \boldsymbol{\Gamma}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) \phi_i(\boldsymbol{\theta}_0) \right]^2 \\ &= \alpha \Sigma_A + \overline{\alpha} \Omega \Sigma_B \Omega^\mathsf{T} + \frac{\overline{\alpha}^2}{\alpha} \Omega \boldsymbol{\Gamma} \Sigma_T \boldsymbol{\Gamma}^\mathsf{T} \Omega^\mathsf{T}, \end{aligned}$$

where $\Sigma_T(\boldsymbol{\theta}_0) = \operatorname{var} \phi_i(\boldsymbol{\theta}_0)$ and

$$\boldsymbol{\Gamma}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) = \mathrm{E}\, \frac{\partial \widetilde{\psi}_i^{(NV)}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^\mathsf{T}} \left[\mathrm{E}\, \frac{\partial \phi_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^\mathsf{T}} \right]^{-1}$$

is a $p \times q$ adjustment matrix. The optimal weighting matrix Ω_0 is given by $\Omega_0 = \Sigma_A(\Sigma_B + \overline{\alpha}/\alpha \boldsymbol{\Gamma} \Sigma_T \boldsymbol{\Gamma}^\mathsf{T})^{-1}$.

5.7 *Simulations*

5.7.1 Continuous covariate with error of constant variance

A simulation study was performed to investigate the performance of the corrected pseudoscore estimator in the setting of Section 5.5.1. The population distribution of the covariate Z was standard normal truncated at ± 1.96. The random error ε was zero-mean normal with standard deviation σ. The baseline hazard was constant and censoring was uniform over the interval $[0, 0.4]$. The total number of observations was 1000. One thousand data sets were generated and analyzed for each setting. Four

estimators were calculated: the full-data estimator that assumes all the true covariate values are known; the naive estimator defined by Equation (5.1) that ignores the error in nonvalidation set covariates; the complete-case estimator $\widehat{\beta}_C(0)$ that uses only the validation set; and finally, the corrected pseudoscore (CS) estimator (5.5) that uses the complete-case estimate to select the optimal weight $\widehat{w}_{\mathbf{opt}}$ and estimates the error variance σ^2 from the validation data. The optimal weight was calculated according to the formula given in Lemma 5.3. Simulation results are summarized in Table 5.1 on page 120.

The table shows the behavior of the four estimators under different validation set sizes and magnitudes of measurement error. The case of $\sigma = 0.5$ represents a small measurement error (error variance is a quarter of the population variance of the covariate), $\sigma = 1$ is a moderate error (error variance equal to covariate variance), and $\sigma = 2$ results in a relatively large error (error variance four times as large as covariate variance). The validation set consisted of either a fifth or a half of the observations. As expected, there is a significant bias in the naive estimator. The smaller the validation set and the larger the measurement error the more the naive estimator underestimates the true parameter. The CS estimator removed the bias quite well. The standard error of the CS estimates is always smaller than that of the complete-case estimates, even when the error variance is large. Moreover, the estimated variance of the CS estimator is on average close to the sample variance of the simulated CS estimates. The coverage probability of the 95% confidence intervals based on the CS estimator and its estimated standard error ranges from 0.945 to 0.961, which is quite satisfactory. The table also shows the empirical power for testing the hypothesis $H_0 : \beta_0 = 0$. When the error variance is small, the power of CS estimator is quite close to the power of the full-data estimator.

We also calculated the average estimated optimal weights. When the measurement error standard deviation σ was 0.5, the weight was around 0.75–0.80. So, a lot of the nonvalidation information was used for estimation of β_0. When σ was 1, the optimal

weight decreased to 0.45–0.50. This means that the nonvalidation contributions were downweighted to less than a half. With large measurement error ($\sigma = 2$), the optimal weight dropped below 0.2. In that case, the unweighted CS estimator with $w = 1$ had much larger standard error than the complete-case estimator. So, the introduction of optimal weight significantly improved the precision of the CS estimator.

5.7.2 Misclassified binary covariate

The finite sample performance of the corrected pseudoscore method for misclassified binary covariates (see Section 5.5.3) was the topic of another simulation study. Data sets of 1000 subjects were generated. The baseline hazard was constant and censoring uniform so that failure was observed for about 420 subjects. The probability of being exposed was $p_Z = 0.5$. The validation set consisted of either 20% or 50% of the subjects. One thousand data sets were generated and for each four estimates were calculated: the full-data estimate, the naive estimate, the complete-case estimate, and the corrected pseudoscore (CS) estimate with estimated optimal weight and estimated error parameters. Results are shown in Table 5.2 on page 121.

Table 5.2 includes the results for different misclassification rates in the binary covariate: the sensitivity and specificity pairs of (0.7,0.7), (0.9,0.7) and (0.9,0.9) give misclassification rates 30%, 20% and 10%, respectively. The table indicates that the corrected score method successfully removes the bias which is evident in the naive estimator. The variance of the CS estimator is always lower than the variance of the complete-case estimator, even when the misclassification rate is 30% and the validation set is small. The 95% coverage probabilities for the CS estimator are in the range 0.934 to 0.957, which suggests that both the CS estimator and its variance estimator work reasonably well with binary covariates. The empirical power of the CS estimator for testing the hypothesis $H_0 : \beta_0 = 0$ approaches the power of the full-data estimator when the misclassification rate is small.

5.8 *Example: NWTSG data set*

To provide an example of an application of the corrected pseudoscore method, we present an analysis of relapse rate for patients enrolled in two studies conducted by the National Wilms Tumor Study Group (NWTSG-3 and NWTSG-4). Wilms tumor is a rare kidney cancer occurring only in children, with a relatively good prognosis when treated with combination chemotherapy. The results of NWTSG-3 were previously published by D'Angio *et al.* (1989) and by Breslow *et al.* (1991); a paper on NWTSG-4 will appear in the near future (Green *et al.*, 1996).

One of the most important prognostic factors for relapse in Wilms tumor patients is histologic type of the tumor, which can be roughly classified into favorable (FH) and unfavorable (UH). Histologic type was first evaluated by a pathologist in the institution that treated a particular patient. The NWTS Pathology Center then re-evaluated histologic type for patients whose tissue samples had been submitted. The institutional and central histology evaluations agreed in most, but not all, cases. Central histology can be regarded as more precise and reliable than institutional histology. For estimating the excess risk of relapse associated with UH, the central and institutional histology assessments can be treated as the true covariate and the surrogate, respectively. In NWTSG studies, central histology was evaluated for nearly all patients. However, the cost and complexity of the studies would decrease if the Pathology Center reviewed only a random subsample of the entire cohort.

The data set consisted of 4119 subjects (1911 in NWTSG-3 and 2208 in NWTSG-4) with known relapse status, follow-up time, and both histology evaluations. Of them, 11.5% had unfavorable central histology (11.6% in NWTSG-3 versus 11.5% in NWTSG-4) and 14.5% relapsed (15.6% in NWTSG-3 and 13.5% in NWTSG-4). The sensitivity and specificity of institutional histology for central histology were 0.72 and 0.98, respectively. Preliminary analyses of relapse rates confirmed that, given central histology, institutional histology had no effect on relapse.

We fitted the standard AH model to the whole data, using either central or institutional histology as the covariate. Then validation sets were drawn at random and the CS and complete-case estimators were calculated assuming that central histology was available only for the validation set subjects. The results are shown in Table 5.3 on page 122. The CS and CC estimators and their estimated standard errors are averaged over 500 different validation sets. The parameter estimate based on all subjects with central histology known was 0.0744. Since time was measured in years, this number estimates how many extra relapses per one person-year of follow-up are expected in the UH group compared to the FH group. Using institutional histology instead of central, we estimated the risk difference as 0.0557. This underestimates the truth by about 30%. When central histology was available for a validation set of 20% or 50% of the original sample, the CS estimator was on average 0.0768 and 0.0753, respectively. The standard errors were 21% to 55% larger compared to the full data estimator but with a 50% validation set, the Pathology Center would have had over 2000 less tissue samples to analyze. The complete-case estimator was on average close to the full-data estimate but its standard errors were much larger than those of the CS estimator.

Table 5.1: Simulation study of the AH corrected score estimator with a continuous covariate subject to random error. Model: $\lambda(t \mid Z) = 3.4 + \beta_0 Z$ where $\beta_0 = 0.3$.

Valid. set prop.	Error SD σ	Est.	Mean Est.	Sample SE	Av. estim. SE	95% cover. probab.	Power
0.2	0.5	F	0.298	0.154	0.162	0.968	0.474
		N	0.245	0.142	0.148	0.939	0.378
		CC	0.303	0.365	0.371	0.958	0.131
		CS	0.299	0.173	0.178	0.961	0.398
0.2	1	F	0.314	0.155	0.162	0.968	0.490
		N	0.167	0.115	0.120	0.809	0.269
		CC	0.316	0.355	0.370	0.962	0.125
		CS	0.312	0.212	0.217	0.955	0.300
0.2	2	F	0.303	0.155	0.161	0.962	0.462
		N	0.071	0.075	0.078	0.166	0.143
		CC	0.282	0.363	0.369	0.948	0.117
		CS	0.315	0.291	0.314	0.961	0.148
0.5	0.5	F	0.298	0.159	0.162	0.961	0.454
		N	0.262	0.152	0.153	0.953	0.416
		CC	0.295	0.228	0.230	0.948	0.252
		CS	0.298	0.171	0.172	0.953	0.424
0.5	1	F	0.290	0.160	0.162	0.954	0.449
		N	0.190	0.131	0.132	0.864	0.293
		CC	0.292	0.239	0.230	0.935	0.253
		CS	0.290	0.194	0.189	0.946	0.348
0.5	2	F	0.300	0.161	0.162	0.950	0.478
		N	0.104	0.094	0.093	0.442	0.194
		CC	0.301	0.229	0.230	0.947	0.262
		CS	0.312	0.213	0.214	0.945	0.296

Notes:

Estimators: F = full data, N = naive, CC = complete-case, CS = corrected score.

Population variance of the covariate was 1.

Table 5.2: Simulation study of the AH corrected score estimator with a misclassified binary covariate. Model: $\lambda(t \mid Z) = 2.6 + \beta_0 Z$ where $\beta_0 = 0.9$.

Valid. set prop.	Sens. η	Spec. ν	Est.	Mean Est.	Sample SE	Av. estim. SE	95% cover. probab.	Power
0.2	0.7	0.7	F	0.908	0.310	0.299	0.939	0.848
			N	0.178	0.131	0.130	0.000	0.288
			CC	0.943	0.701	0.680	0.947	0.285
			CS	0.940	0.554	0.537	0.944	0.411
0.2	0.9	0.7	F	0.888	0.318	0.299	0.935	0.835
			N	0.390	0.195	0.190	0.241	0.539
			CC	0.909	0.683	0.676	0.949	0.271
			CS	0.932	0.442	0.435	0.957	0.583
0.2	0.9	0.9	F	0.899	0.309	0.299	0.941	0.853
			N	0.616	0.253	0.247	0.780	0.707
			CC	0.882	0.683	0.680	0.955	0.259
			CS	0.900	0.371	0.360	0.946	0.711
0.5	0.7	0.7	F	0.888	0.313	0.299	0.939	0.833
			N	0.245	0.161	0.156	0.015	0.366
			CC	0.919	0.437	0.426	0.950	0.578
			CS	0.912	0.408	0.395	0.938	0.646
0.5	0.9	0.7	F	0.893	0.327	0.299	0.925	0.833
			N	0.481	0.231	0.216	0.512	0.598
			CC	0.911	0.448	0.424	0.941	0.573
			CS	0.902	0.387	0.362	0.934	0.693
0.5	0.9	0.9	F	0.905	0.308	0.299	0.945	0.848
			N	0.708	0.268	0.263	0.868	0.764
			CC	0.902	0.446	0.424	0.940	0.578
			CS	0.908	0.343	0.331	0.946	0.775

Notes:

Estimators: F = full data, N = naive, CC = complete-case, CS = corrected score.
Population exposure rate was 0.5.

Table 5.3: Analysis of NWTSG data: estimates of excess risk for relapse associated with unfavorable central histology.

Estimator	Valid. set proportion	Param. est.	SE
F	1	0.0744	0.00683
F	0[a]	0.0557	0.00609
CC	0.2	0.0755[b]	0.01555[b]
CC	0.5	0.0746[b]	0.00969[b]
CS	0.2	0.0768[b]	0.01058[b]
CS	0.5	0.0753[b]	0.00829[b]

Notes:

Estimators: F = full data, CC = complete-case, CS = corrected score.

[a] Institutional UH used as covariate.

[b] Average over 500 validation sets.

Chapter 6

DISCUSSION AND FURTHER RESEARCH

In this dissertation, we proposed consistent and asymptoticaly normal estimators of additive hazards regression parameters for two special cases of two-phase designs. We derived their asymptotic distributions, developed consistent estimators for the limiting variances, demonstrated the behavior of the estimators in moderate sample sizes by numerical simulation studies, and illustrated the usefulness of the new methods on a real-life data set.

In Chapter 4, we dealt with the case-cohort design. To this end, we derived an unbiased pseudoscore which works with general subcohort sampling schemes. The pseudoscore relied on Horvitz-Thompson type weights and utilized information on all second-phase subjects at each failure time. In particular, covariates of the failures were included in the at-risk covariate average no matter if the failure occured before, at, or after the time the at-risk average was evaluated. Because of that we could not use the standard martingale theory to study the asymptotic distribution of the pseudoscore but on the other hand, the estimator gained some efficiency. Because missing covariates in the AH model cause a total elimination of some contributions from the pseudoscore, which was not the case in the Cox model, we were concerned about the eventually large loss in efficiency of the AH case-cohort estimator. However, as we demonstrated, the efficiency loss due to missing data for the case-cohort estimator in the AH model was only slightly larger than in the Cox model. The AH case-cohort estimator was consistent and asymptotically normal under mild regularity conditions we introduced. The estimator of its limiting variance that we proposed was shown to be consistent and worked well in moderately large samples.

Chapter 5 was devoted to the errors-in-variables design. We have introduced a corrected score (CS) estimator for the additive hazards regression parameter that is consistent when covariates are subject to measurement error and a validation set is available. The conditions on the measurement error mechanism were weak: we worked under a linear error calibration model and allowed for some types of nonconstant error variance. We did not make any parametric assumptions about the distributions of either the true covariate or the measurement error. It was demonstrated on examples and simulation studies that the method works well for both discrete and continuous covariates. We proposed a consistent estimator for the limiting variance of the CS estimator and showed how the variance is affected when the error model parameters have to be estimated from the validation set. By introducing optimal weights into the corrected pseudoscore, we assured that the CS estimator is always more efficient than the complete-case estimator even when the measurement error is large. Even though we initially worked with a single covariate subject to error, we have shown how the corrected score estimator may be generalized to multiple covariates, one of which is measured with error. Similarly, the CS estimator can be also used with more covariates subject to error.

In Chapter 3, the current methods for estimating the Cox model parameters in the presence of covariate measurement error were reviewed. Compared to them, the corrected score estimator for the additive hazards model is much simpler to calculate and works under much weaker conditions. Thus, the additive hazards model is much more convenient for analyzing data in the presence of covariate measurement error. Because of this, it could be used as the primary method of statistical analysis in the studies that deal with serious covariate measurement error and/or covariate assessment problem. The presence of the validation set is not crucial for the CS estimator, as long as the error parameters are known from other sources.

It should not be difficult to construct score-type statistics to test hypotheses about various subsets of $\boldsymbol{\beta}$ for either case-cohort or corrected score estimators. Some hints

how to do that can be found in Lin and Wei (1989). Such tests would be useful for model building as well as for testing hypotheses in the final model. Experience with score-type tests in other models suggests that they might better keep the level in small samples than the Wald test based on estimated standard deviations of individual parameter estimates.

The case-cohort design suffers from not being able to use any first-phase covariate data, except for modifying the subcohort selection probabilities. The errors-in-variables design uses all the observed first-phase covariates; however, that does not mean the errors-in-variables design is always preferable to the case-cohort design. In the case-cohort design, the covariates of the failures are all known, and most of the information about the AH regression parameter $\boldsymbol{\beta}_0$ is contained in the failures. So, the greatest weakness of the errors-in-variables design is the selection of the validation set by simple random sampling only. A generalization of the errors-in-variables design that makes possible selecting validation set members on the basis of their failure status and covariates would therefore be most welcome.

To this end, let us sketch a general two-phase design that combines the strengths of the case-cohort and errors-in-variables design and contains both as special cases. Let the second-phase sample be $\mathcal{V} = \{i : \xi_i = 1\}$, where $\mathrm{P}\,[\xi_i = 1 \mid \Delta_i = 1\,] = p_i$ and $\mathrm{P}\,[\xi_i = 1 \mid \Delta_i = 0\,] = q_i$. So, a failure is selected to the validation set with probability p_i and a censored observation with probability q_i. The sampling probabilities p_i and q_i are fixed constants or functions of first-phase covariates. Now, suppose again that Z_i is the true covariate observed when $i \in \mathcal{V}$, and that W_i^* is a bias-adjusted surrogate available for all subjects. Define the observed covariate R_i by $R_i = \xi_i Z_i + (1 - \xi_i) W_i^*$ and consider the weighted availability indicator

$$\varrho_i = \xi_i \left(\frac{\Delta_i}{p_i} + \frac{1 - \Delta_i}{q_i} \right) + w(1 - \xi_i) \left(\frac{\Delta_i}{1 - p_i} + \frac{1 - \Delta_i}{1 - q_i} \right),$$

where $w \in [0, 1]$ is a downweighting constant. If ξ_i and W_i^* are independent given Z_i, and W_i^* is independent of Δ_i and $Y_i(t)$ given Z_i, then it is easy to see that $\mathrm{E}\,\varrho_i = 1 + w$.

With

$$\bar{R}(t) = \frac{\sum_{i=1}^{n} \varrho_i R_i Y_i(t)}{\sum_{i=1}^{n} \varrho_i Y_i(t)}.$$

it should be easy to show that $\sup_{0\le t\le\tau} \left|\bar{R}(t) - e(t)\right| \to_p 0$ and hence $\bar{R}(t)$ is a uniformly consistent estimator for the at-risk covariate average. A generalized corrected pseudoscore can be defined by

$$\begin{aligned} U_G(\beta) = &\sum_{i=1}^{n} \varrho_i \int_0^\tau [R_i - \bar{R}(t)]\,[dN_i(t) - R_i \beta Y_i(t)\,dt] \\ &+ \sum_{i=1}^{n} w(1-\xi_i) \left(\frac{\Delta_i}{1-p_i} + \frac{1-\Delta_i}{1-q_i} \right) \frac{V(W_i^*)}{v_2 + \gamma_1^2} \beta X_i. \end{aligned}$$

We can immediately see that for $p_i = 1$ and $w = 0$, we get the case-cohort pseudoscore. With $p_i = q_i = \alpha$, this is the corrected pseudoscore for the errors-in-variables design described in Chapter 5. But U_G may be used to estimate AH parameters for many other designs that combine the features of the case-cohort and errors-in-variables design and use general sampling schemes for the selection of the validation set.

Asymptotic theory for the general corrected pseudoscore U_G needs to be worked out and regularity conditions have to be determined. This is the main topic for further research originating in this work.

As for the error model we used, it would certainly be useful to model the measurement errors more flexibly. We might extend the error calibration model by including first phase covariates other than just the surrogate. That could apply to both mean and variance parts of the error model. We could then estimate the AH parameters correctly in all the situations where the mean and/or the variance of the measurement error depend on other observed covariates.

BIBLIOGRAPHY

Andersen, P.K., Borgan, Ø., Gill, R.D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* New York: Springer-Verlag.

Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.*, **10**, 1100–1120.

Begun, J.M., Hall, W.J., Huang, W.M., and Wellner, J.A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.*, **11**, 432–452.

Borgan, Ø., Goldstein, L., and Langholz, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *Ann. Statist.*, **23**, 1749–1778.

Breslow, N.E. (1972). Contribution to the discussion on the paper by D.R. Cox. Regression and life tables. *J. R. Statist. Soc.* B, **34**, 216–217.

Breslow, N.E. and Day, N.E. (1987). *Statistical Models in Cancer Research.* **2**. *The Design and Analysis of Cohort Studies.* Lyon: IARC.

Breslow, N., Sharples, K., Beckwith, J.B., Takashima, J., Kelalis, P.P., Green, D.M., D'Angio, G.J. (1991). Prognostic factors in nonmetastatic, favorable histology Wilms tumor. *Cancer*, **68**, 2345–2353.

Carroll, R.J. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* London: Chapman & Hall.

Carroll, R.J., Ruppert, D., and Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models.* London: Chapman & Hall.

Chung, K.L. (1974). *A Course in Probability Theory: 2nd Edition.* San Diego: Academic Press.

Cox, D.R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc.* B, **34**, 187–220.

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

Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data.* London: Chapman & Hall.

D'Angio, G.J., Breslow, N., Beckwith, J.B. *et al.* (1989). Treatment of Wilms tumor: Results of the third National Wilms Tumor Study. *Cancer*, **64**, 349–360.

Fleming, T.R. and Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. New York: Wiley.

Goldstein. L.. and Langholz. B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model. *Ann. Statist.*, **20**. 1903–1928.

Green, D.M.. Beckwith, J.B., Breslow, N.E. *et al.* (1994a). Treatment of children with stages II to IV anaplastic Wilms tumor: A report from the National Wilms Tumor Study Group. *J. Clin. Oncology*, **12**, 2126–2131.

Green, D.M., Breslow, N.E., Beckwith, J.B. *et al.* (1994b). Treatment of children with clear-cell sarcoma of the kidney: A report from the National Wilms Tumor Study Group. *J. Clin. Oncology*, **12**, 2132–2137.

Green, D.M., Breslow, N.E., Beckwith, J.B. *et al.* (1996). A comparison between single dose and divided dose administration of dactinomycin and doxorubicin. A report from the National Wilms Tumor Study Group. *J. Clin. Oncology*. submitted.

Horvitz, D.G. and Thompson, D.J. (1951). A generalization of sampling without replacement from a finite universe. *Journal Am. Stat. Assoc.*, **47**, 663–685.

Hughes, M.D. (1993). Regression dilution in the proportional hazards model. *Biometrics*, **49**, 1056–1066.

Kalbfleisch, J.D., and Lawless, J.F. (1988). Likelihood analysis of multi-state models for disease incidence and mortality. *Statistics in Medicine*, **7**, 149–160.

Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.

Kim, S. and DeGruttola, V. (1996). Strategies for cohort sampling under the Cox proportional hazards model: Application to an AIDS clinical trial. Technical report. Harvard School of Public Health, Department of Biostatistics.

Klaassen, C.A.J. (1989). Efficient estimation in the Cox model for survival data. In: Mandl, P. and Hušková, M., editors, *Proceedings of the Fourth Prague Symposium on Asymptotic Statistics*. Praha: Charles University.

Lin, D.Y. and Wei, L.J. (1989). The robust inference for the Cox proportional hazards model.*Journal Am. Stat. Assoc.*, **84**, 1074–1078.

Lin, D.Y. and Ying, Z. (1993). Cox regression with incomplete covariate measurements. *Journal Am. Stat. Assoc.*, **88**, 1341–1349.

Lin, D.Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, **81**, 61–71.

Nakamura, T. (1990). Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika*, **77**, 127–137.

Nakamura, T. (1992). Proportional hazards model with covariates subject to measurement error. *Biometrics*, **48**, 829–838.

Pepe, M.S., Self, S.G., and Prentice, R.L. (1989). Further results on covariate measurement errors in cohort studies with time to response data. *Statistics in Medicine*, **8**, 1167–1178.

Prentice, R.L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, **69**, 331–342.

Prentice, R.L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, **73**, 1–11.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

Samuelsen, S.O. (1989). *Two incomplete data problems in life-history analysis: Double censoring and the case-cohort design.* Dr. Scient. Dissertation. Oslo: University of Oslo.

Self, S.G. and Prentice, R.L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann. Statist.*, **16**, 64–81.

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics.* New York: Wiley.

Shorack, G.R. and Wellner, J.A. (1986). *Empirical Processes with Applications to Statistics.* New York: Wiley.

Thomas, D.C. (1977). Addendum to: Methods of cohort analysis: Appraisal by application to asbestos mining. By F.D.K. Lidell, J.C. McDonald, and D.C. Thomas. *J. Roy. Statist. Soc. A*, **140**, 469–491.

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes.* New York: Springer-Verlag.

Wang, C.Y., Hsu, L., Feng, Z.D., Prentice, R.L. (1997). Regression calibration in failure time regression. *Biometrics*, **53**, 131–145.

Zhou, H. and Pepe, M.S. (1995). Auxiliary covariate data in failure time regression. *Biometrika*, **82**, 139–149.

Zhou, H. and Wang, C.Y. (1995) *Failure time regression analysis with measurement error in covariates.* Technical report, Fred Hutchinson Cancer Research Center, Seattle.

Appendix A

OVERVIEW OF NOTATION

We routinely use boldface letters for vectors ($\boldsymbol{e}, \boldsymbol{\beta}$) and blackboard capitals or capital greek letters for matrices ($\mathbb{A}, \mathbb{D}, \Sigma$). The exception is the (pseudo-) score vector and its individual terms, which we do not boldface. Matrix transposition is denoted by $\mathbb{A}^\mathsf{T}$ and scalar product by $\boldsymbol{a}^\mathsf{T}\boldsymbol{b}$. The notation $\mathbb{A}^{-\mathsf{T}}$ means just $(\mathbb{A}^{-1})^\mathsf{T}$. Sometimes we use the vector power notation $\boldsymbol{a}^{\otimes k}$, which stands for scalar 1 if $k = 0$, for $\boldsymbol{a}$ when $k = 1$, and for $\boldsymbol{a}\boldsymbol{a}^\mathsf{T}$ when $k = 2$. For a right-continuous function $N(t)$, we define $\Delta N(t)$ by $N(t) - N(t-)$. We denote the indicator function of an event A by $\mathbb{1}(A)$. It equals one if the event occurs and zero otherwise. The Euler's constant is typeset in upright font (e). The exponential function is sometimes written as e^x and sometimes as $\exp\{x\}$. It's inverse, the natural logarithm, is denoted by $\ln x$. Sometimes we use the shorthand notation $\overline{\alpha}$ for $1 - \alpha$, $\overline{\xi}_i$ for $1 - \xi_i$, etc.

Probability is written as $\mathrm{P}[\ldots]$ and expectation as E. If the argument of the expectation is a product, we usually do not enclose it in parentheses; so $\mathrm{E}\,XYZ$ is to be interpreted as $\mathrm{E}(XYZ)$. The distribution of a random variable X is denoted by $\mathcal{L}(X)$; normal distribution is written as $\mathrm{N}(\mu, \sigma^2)$, p-variate normal as $\mathrm{N}_p(\boldsymbol{\mu}, \Sigma)$. Convergence in probability is denoted by $\to_p$ and convergence in distribution by $\to_d$.

For easy reference, Tables A.1 and A.2 list the symbols encountered frequently throughout the dissertation.

Table A.1: Most important symbols used throughout the dissertation.

Symbol	Meaning
$\boldsymbol{\beta}_0$	true regression parameter
$\boldsymbol{Z}_i$	covariate vector
W_i	surrogate covariate
T_i	failure time
C_i	censoring time
X_i	censored failure time
Δ_i	indicator of failure
$\lambda_0(t)$	baseline hazard
$\Lambda_0(t)$	cumulative baseline hazard
τ	end of observation time

Table A.2: List of other symbols.

Symbol	Section(s) where defined	Meaning
α	3.2.2, 4.1, 5.2	proportion of validation set (subcohort)
$\widehat{\boldsymbol{\beta}}_A$	2.3	full-data AH estimator
$\widehat{\boldsymbol{\beta}}_H$	4.2	case-cohort AH estimator
$\widehat{\beta}_C(w)$	5.3.1	CS estimator
γ_0, γ_1	5.2	mean error parameters
$\boldsymbol{\Gamma}(\beta, \boldsymbol{\theta})$	5.4.1	variance correction vector, $\boldsymbol{D}_X(\beta, \boldsymbol{\theta})\mathbb{D}_T^{-1}(\boldsymbol{\theta})$
η	4.5, 5.5.3	specificity of binary surrogate
$\boldsymbol{\theta}$	5.2	error parameters
ν	4.5, 5.5.3	sensitivity of binary surrogate
ξ_i	3.2.2, 4.1, 5.2	indicator of subcohort (validation set) selection
$\pi_k(t)$	2.3, 4.3, 5.3.2	$\mathrm{E}\,\boldsymbol{Z}^{\otimes k}Y(t)$
ϱ_i	3.2.2, 4.2, 5.3.1	weighted availability indicator
$\Sigma_A(\boldsymbol{\beta})$	2.3	limiting variance of $U_A(\boldsymbol{\beta})$
$\Sigma_H(\boldsymbol{\beta})$	4.3	extra variance of case-cohort pseudo-score
$\Sigma_C(\beta, w)$	5.3.2	limiting variance of $U_C(\beta, w)$
$\phi_i(\boldsymbol{\theta})$	5.4.1	estimating function for error parameters
$\widetilde{\psi}_i^{(A)}(\boldsymbol{\beta})$	2.3	asymptotically iid contribution to $U_A(\boldsymbol{\beta})$
$\psi_i^{(V)}(\beta)$	5.3.1	validation set contribution to corrected pseudo-score
$\psi_i^{(NV)}(\beta)$	5.3.1	nonvalidation set contribution to corrected pseudo-score
$\widetilde{\psi}_i^{(V)}(\beta)$	5.3.2	asymptotically iid approximation to $\psi_i^{(V)}$
$\widetilde{\psi}_i^{(NV)}(\beta)$	5.3.2	asymptotically iid approximation to $\psi_i^{(NV)}$
Ω	5.6.1	weighting matrix for multivariate corrected pseudo-score

Table A.2: (continued)

Symbol	Section(s) where defined	Meaning
$\mathbb{A}_i$	5.6.1	weighted selection matrix for multivariate CS
$\mathbb{D}_A$	2.3	expected negative partial derivative of $U_A(\boldsymbol{\beta})$
$D_C(w)$	5.3.2	expected negative partial derivative of $U_C(\beta, w)$
$\mathbb{D}_T(\boldsymbol{\theta})$	5.4.1	expected negative partial derivative of $\phi_i(\boldsymbol{\theta})$
$\boldsymbol{D}_X(\beta, \boldsymbol{\theta})$	5.4.1	$-\mathrm{E}\,\partial\widetilde{\psi}_i^{(NV)}(\beta, \boldsymbol{\theta})/\partial\boldsymbol{\theta}^{\mathsf{T}}$
$\boldsymbol{e}(t)$	2.3, 4.3, 5.3.2	limiting at-risk covariate average
$M_i(t)$	2.3	counting process martingale for the AH model
$N_i(t)$	2.3	counting process
p_i	3.2.2, 4.2	probability of selection
$\bar{R}(t, w)$	5.3.1	mean covariate at-risk for corrected pseudo-score
$U_A(\boldsymbol{\beta})$	2.3	full-data additive hazards pseudo-score
$U_H(\boldsymbol{\beta})$	4.2	case-cohort AH pseudo-score
$U_C(\beta, w)$	5.3.1	corrected pseudo-score
v_0, v_1, v_2	5.2	error variance parameters
$V(Z_i)$	5.2	error variance function
w	5.3.1	weighting constant in corrected pseudo-score
w_{opt}	5.4.2	optimal weight
$Y_i(t)$	2.3	at-risk process
$\overline{\boldsymbol{Z}}(t)$	2.3	mean covariate at-risk
$\overline{\boldsymbol{Z}}_H(t)$	4.2	mean covariate at-risk for case-cohort design

VITA

Michal Kulich was born on August 16, 1967, in Prague, Czech Republic. He is the eldest son of Světlana Kulichová and Pavel Kulich. He graduated from high school in 1985 and enrolled in the Charles University in Prague, where he received a MS degree in Statistics in 1991. With the help of a grant from the TEMPUS agency, he entered a one-year master's program in Biostatistics at the Limburgs Universitair Centrum in Diepenbeek, Belgium. During his study in Diepenbeek, he spent three months on an internship with the European Organisation for Research and Treatment of Cancer in Brussels. After his graduation from the LUC in 1992, he stayed another year in Diepenbeek as a fellow in the Biostat Program. Since the fall 1993, he has been a graduate student at the Dept. of Biostatistics, University of Washington. He earned a MS degree in Biostatistics from the UW in 1995 and continued towards a PhD degree. Upon his graduation, he will return to his home country to join the faculty at the Dept. of Probability and Statistics at the Charles University in Prague.