A SIMPLE LOCALLY EFFICIENT ESTIMATOR FOR RELATIVE RISK IN CASE-COHORT STUDIES

by

Emmanuel Sampene

BA in Mathematics, Caldwell College, 2006

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Emmanuel Sampene

It was defended on

11/14/2013

and approved by

Abdus S. Wahed, PhD.

Associate Professor of Biostatistics

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

Jong-Hyeon Jeong, PhD.

Associate Professor of Biostatistics

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

Vincent C. Arena, PhD.

Associate Professor of Biostatistics

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

Chung-Chou H. Chang, PhD.

Professor of Medicine,

Biostatistics, and

Translational Science

University of Pittsburgh

Raghavan Murugan, MD,MS,FRCP.

Assistant Professor of Critical Care Medicine,

Clinical, and

Translational Science Institute

University of Pittsburgh

Dissertation Director: Abdus S. Wahed, PhD.

Associate Professor of Biostatistics

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

A SIMPLE LOCALLY EFFICIENT ESTIMATOR FOR RELATIVE RISK IN CASE-COHORT STUDIES

Emmanuel Sampene, PhD

University of Pittsburgh, 2013

A case-cohort study is a two-phase study where at the first phase a representative sample, referred to as the study cohort, is selected from the target population. At the second phase, a subsample is selected from the cohort based on the case status. All cases are included in the subsample whereas only a random sample of controls is included. The endpoint of interest in such studies is usually the failure time. Several methods have been proposed to estimate the relative risk or hazard ratio from a case-cohort study. These methods almost always disregard the covariate information that is not included in the sampled study sub-cohort, and therefore, results in the loss of efficiency. While there have been attempts to derive the most efficient estimators, the resulting estimators were challenging from the data analysis point of view. We propose a locally efficient estimator (LEE) by restricting the estimator to a class of regular asymptotically linear estimators. The properties of this estimator are investigated through simulation and application to the Wilm's tumor study. The public health relevance of this dissertation is the use of innovative methodology to reduce cost associated with research.

Keywords: Case-cohort study; Influence Function; Martingale theory; Regular asymptotically linear estimator; Two-stage design

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to offer my sincere gratitude to my advisor Dr. Wahed. He has been a source of inspiration for me throughout my graduate education. He consistently provided me with guidance on statistical methodologies that have helped to build my career path. I am grateful for his mentorship and continuous support. Also, I would like to thank my committee members for their invaluable contributions to my education and the dissertation process. Thank you to Drs. Jong-Hyeon Jeong, Vincent C. Arena, Chung-Chou H. Chang, and Raghavan Murugan for your input. I would like to extend my gratitude to my colleagues, Olive, Semhar, Kidane, Yerke, Folefac, Abi, Akunna and Jesse, whose advice and support have helped me through graduate school. In addition, I would like to thank my parents and siblings for their constant prayers, love, and support throughout my education. Finally, I thank my wife Katie for her love, kindness and understanding. I thank her for all the encouragements and patience, and look forward to sharing many happy memories in the years to come as we build our lives together.

1.0 CASE-COHORT STUDY

In 1986, Prentice introduced the case-cohort design as a cost-reduction approach to studying large cohort designs. The case-cohort design is a two-phase study where at the first phase a representation sample, referred to as the study cohort, is selected from the target population. At the second phase, a subsample is selected from the cohort based on the case status.

Prentice's goal was to estimate the relative risk in a Cox proportional hazards (CPH) model without having to ascertain the covariate information of all cohort members. Since his initial work, many authors have developed variations of Prentice's method by proposing different estimating equations. Most of these methods, however, fail to utilize the covariate data collected outside the case-cohort sample, and thus incur the loss of efficiency. In particular, the Kalbfleisch & Lawless (1988) estimates simply ignore the first phase covariate data. To improve the efficiency of the case-cohort estimators, Barlow (1994) introduced an estimator that incorporates time-varying weights while Chen and Lo (1999) proposed an estimator that improves the efficiency only when the fully observed covariates are binary. In addition, Borgan et al. (2000) proposed an estimator that uses some of the the first phase covariate data. Also, Kulich & Lin (2004) introduced the combined doubly weighted estimator while Mark & Katki (2006) proposed the $\hat{\alpha}$- estimator.

The goal of the first part of this dissertation is to present a semiparametric efficient estimator for analyzing case-cohort studies. We derive the most efficient estimator along the lines of the semiparametric theory of Robins et al. (1994) by using the projection theory of Hilbert spaces and martingale probability theory. We restrict ourselves to the class of estimators which are regular and asymptotically linear (Newey, 1990). Hence, our proposed estimator is the best regular asymptotically linear estimator which is not only easily computable, but also the most efficient within this subclass of estimators.

In the second part of this dissertation, we extend the idea of our semiparametric efficient estimator to time-dependent covariates. The basic idea remains the same as that in part 1, except that the Cox proportional hazard model allows time-dependent covariates in the model. As a result, it is possible to assess the association between failure time and time-dependent covariates in a case-cohort design.

This dissertation is organized as follows. In Chapter 1, we introduce the case-cohort design along with the general framework of case-cohort estimation. In Chapter 2, we introduce the semiparametric efficient estimation, and also discuss topics such as Hilbert space, linear subspace and projection theorem for Hilbert spaces which are useful constructs for semiparametric theory. Chapters 3 and 4 are written as independent papers. In Chapter 3, we propose a semiparametric locally efficient estimator for analyzing case-cohort studies that do not require strong model assumptions, and yet contains quantities that are easy to compute. Chapter 4 extends the estimator described in Chapter 3 to time-dependent covariates. Therefore, there are some redundancies in the introductory sections. In the remaining part of this Chapter, we describe the model framework on which almost all existing estimators for analyzing case-cohort studies are based.

1.1 GENERAL FRAMEWORK OF CASE-COHORT ESTIMATION

The model framework of the case-cohort design is based on the popular Cox proportional hazard model. In this section, we describe a commonly used approach for obtaining an estimating equation for the regression parameters in a case-cohort design.

Let T be the failure time, C be a potential censoring time and Z be the vector of covariates. Suppose that T is conditionally independent of C given Z and that the conditional distribution of T given Z follows the Cox (1972) proportional hazards model:

$$\lambda(t|Z) = \lambda_o(t)exp(\beta^T Z)$$

where $\lambda(t|Z)$ is the conditional hazard for failure given the covariate history up to time t, β is a vector-valued parameter, and $\lambda_o(t)$ is an unspecified baseline hazard function. Let us define $U = min(T, C)$, $\Delta = I(T \leq C)$, $N(t) = I(U \leq t, \Delta = 1)$, and $Y(t) = I(U \geq t)$. A subject whose failure time is observed will have $\Delta = 1$ and will be treated as a case, and a censored subject with $\Delta = 0$ will be treated as a control. Suppose that the support of C is bounded above by $\tau > 0$ and that $Pr(Y(\tau) = 1) > 0$.

Under the case-cohort design, the complete observations $(U_i, \Delta_i, Z_i, \xi = 1)$ for all subsample members, and at least $(U_i, \Delta_i = 1, Z_i(U_i))$, are observed for the cases. With full data, one can estimate the parameter β by $\hat{\beta}$, the root of the partial likelihood Cox (1972) score function. This score function is defined as follows:

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \{Z_i - \bar{\mathbf{Z}}(\beta)\} dN_i(t) \tag{1.1.01}$$

where $\bar{\mathbf{Z}}(\beta) = \frac{S^1(\beta)}{S^o(\beta)}$, and these quantities are respectively defined as:

$$S^1(\beta) = \sum_{i=1}^{n} Z_i exp(\beta^T Z_i) Y_i(t),$$

$$S^o(\beta) = \sum_{i=1}^{n} exp(\beta^T Z_i) Y_i(t)$$

It should be noted that only the cases contribute to the summation in (1.1.01), while the controls affect $U(\cdot)$ only through the at-risk covariate average $\bar{\mathbf{Z}}$. In general, (1.1.01) cannot be calculated under the case-cohort design because $\bar{\mathbf{Z}}$ involves unobserved data. As a result, nearly all existing case-cohort estimators are based on estimating equations similar to (1.1.01), where we replace $\bar{\mathbf{Z}}$ with an approximate $\bar{\mathbf{Z}}_{\mathbf{C}}$,

$$U_C(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \{Z_i - \bar{\mathbf{Z}}_{\mathbf{C}}(\beta)\} dN_i(t), \tag{1.1.02}$$

where the at-risk average for (1.1.02) is defined as $\bar{\mathbf{Z}}_{\mathbf{C}}(\beta) = \frac{S_C^1(\beta)}{S_C^o(\beta)}$ with $S_C^j(\beta)$ defined as follows:

$$S_C^1(\beta) = \sum_{i=1}^{n} \frac{\xi_i}{\alpha_i} Z_i exp(\beta^T Z_i) Y_i(t),$$

$$S_C^o(\beta) = \sum_{i=1}^{n} \frac{\xi_i}{\alpha_i} exp(\beta^T Z_i) Y_i(t),$$

where ξ_i is a binary indicator for the controls and α_i refers to the sampling probability. The score function defined in Equation (1.1.02) eliminates subjects with incomplete data. As a result, all existing estimators that incorporate the score equation in (1.1.02) loss efficiency. Since our proposed estimator is based on the semiparametric efficiency theory, in the next section we discuss semiparametric efficient estimation and how it relates to case-cohort designs.

2.0 SEMIPARAMETRIC EFFICIENT ESTIMATION

A semiparametric model is one that has both a parametric component β, and nonparametric component η that describe the model. Consider data envisioned as realizations of random vectors $Z_1, \cdots, Z_n$, assumed iid. An estimator $\hat{\beta}$ of β is a q-dimensional measurable random function of $Z_1, \cdots, Z_n$. Most reasonable estimators for β are asymptotically linear. As a result, there exist a random vector $\varphi^{q\times 1}(Z)$, such that $E\varphi(Z) = 0^{q\times 1}$,

$$n^{1/2}(\hat{\beta}_n - \beta_o) = n^{-1/2}\sum_{i=1}^{n}\varphi(Z_i) + o_p(1), \tag{2.0.01}$$

where $o_p(1)$ is a term that converges in probability to zero as n goes to infinity and $E(\varphi\varphi^T)$ is finite and nonsingular. We refer to the random vector $\varphi(Z_i)$ defined in (2.0.01) as the influence function of the i-th observation of the estimator $\hat{\beta}_n$. In general semiparametric problems, one approach to construct estimators for β is to obtain some influence function $\varphi(H_i; \beta, \eta)$. This influence function is subsequently used to form estimating equations for β in the form of

$$\sum_{i=1}^{N}\varphi(H_i; \beta, \eta) = 0 \tag{2.0.02}$$

where H_i is the full data , β is the q-dimensional parameter of interest, and η is the infinite dimensional nuisance parameter. An advantage of using the influence

function in (2.0.01) to construct estimators for β is that there is less restriction on the probability constraint that our data might have. Hence, the solution in (2.0.01) is reasonable and robust. By solving for the estimating equation in (2.0.02), we obtain the solution $\hat{\beta}$, which is a semiparametric estimator and its variance has been shown to be equal to the variance of $n^{-1}\varphi(H_i; \beta, \eta)$. As a result, the optimal estimator among the class of all such estimators is the one whose influence function has the smallest variance. We refer to this as the semiparametric efficient estimator. Therefore, estimating the finite dimensional parameter β in the presence of an infinite dimensional nuisance parameter η, is a classical semiparametric problem (Tsiatis, 2010). Since the class of influence functions for the estimating equation defined in (2.0.01) belongs to the Hilbert space of all mean-zero q-dimensional random functions with finite variance, in the section that follows we review Hilbert spaces, the notion of orthogonality, minimum distances, and how it relates to efficient estimators (i.e. estimators with the smallest variance).

2.1 HILBERT SPACE

In this section, we introduce a Hilbert space without excessive technical details. Our focus will be on the Hilbert space whose elements are random vectors with mean zero and finite variance. A Hilbert space, denoted by $\mathcal{H}$, is a complete normed linear vector space equipped with an inner product. For example, consider the Hilbert space $\mathcal{H}$ of one-dimensional random functions, $h(Z)$, with mean zero and finite second moments. We can define the inner product for $h_1(Z), h_2(Z) \in \mathcal{H}$ as

$$\langle h_1, h_2 \rangle = E(h_1, h_2).$$

We refer to this inner product as the covariance inner product. Evidently, the space of all such h that consist of mean zero and finite variance is a linear space. By linear, it is implied that if h_1, h_2 are elements of the space, then for any real constants a and b, $ah_1 + bh_2$ also belongs to the space.

Definition 1. *Let h_1,h_2 belonging to a linear vector space $\mathcal{H}$, an inner product, defined by $\langle h_1, h_2 \rangle$, is a function that maps to the real line. That is $\langle h_1, h_2 \rangle$ is a scalar that satisfies:*

(i) $\langle h_1, h_2 \rangle = \langle h_2, h_1 \rangle$,

(ii) $\langle h_1 + h_2, h_3 \rangle = \langle h_1, h_3 \rangle + \langle h_2, h_3 \rangle$, *where h_1,h_2, and h_3 belong to $\mathcal{H}$,*

(iii) $\langle \lambda h_1, h_2 \rangle = \lambda \langle h_1, h_2 \rangle$ *for any scalar constant $\lambda \in \Re$,*

(iv) $\langle h_1, h_1 \rangle \geq 0$ *with equality if and only if $h_1 = 0$*

With this definition of inner product, we can proceed to define the norm of any vector (i.e., element of $\mathcal{H}$). Furthermore, we denote the distance from any point $h \in \mathcal{H}$ to the origin as $\| h \| = \langle h, h \rangle^{1/2}$. Since the properties of Hilbert spaces allow us to define orthogonality, we can state that $h_1, h_2 \in \mathcal{H}$ are orthogonal if $\langle h_1, h_2 \rangle = 0$ [Luenberger, 1969].

2.2 LINEAR SUBSPACE AND PROJECTION THEOREM FOR HILBERT SPACES

In this section, we explain the importance of using the projection theorem in Hilbert space ($\mathcal{H}$) geometry. Consider a space $\mathcal{U} \in \mathcal{H}$ as a linear subspace if $v_1, v_2 \in \mathcal{U}$. This means that $av_1 + bv_2 \in \mathcal{U}$ for all scalar constants a, b. A linear subspace must contain the origin. We achieve this by setting $a = b = 0$. A simple example of

a linear subspace is the space $(a_1h_1 + \cdots + a_kh_k)$, where $h_1, \cdots, h_k$ are arbitrary elements of $\mathcal{H}$. One might reasonably conjecture that in n-dimensional Euclidean space, the shortest distance from a point to a subspace is orthogonal to the subspace. In fact, this optimization principle is called the projection theorem.

It should be noted that the inner product defined in Section 2.1 corresponds to a covariance. This means that we can use the projection theorem to find the minimum variance estimate. Consequently, finding the projection of the q-dimensional vector h onto the subspace $\mathcal{U}$ is equivalent to taking each element of h and projecting it individually to the subspace spanned by $(v_1, \cdots, v_r)$ for the Hilbert space of one-dimensional random functions (Tsiatis, 2010). An advantage of using the projection theorem approach is that once we have a well defined inner product, there is no need to minimize the variance estimate. Therefore, we define a space $\mathcal{U} \in \mathcal{H}$ as a linear subspace if v_1, $v_2 \in \mathcal{U}$. This implies that $av_1 + bv_2 \in \mathcal{U}$ for all a,$b \in \Re$.

Theorem 2.2.1. *Let $\mathcal{H}$ be a Hilbert space and $\mathcal{U}$ a linear subspace that is closed (i.e., contains all its limit points). Corresponding to any $h \in \mathcal{H}$, there exists a unique $v_o \in \mathcal{U}$ that is closest to h; that is, $\|h - v_o\| \leq \|h - v\| \ \forall \ v \in \mathcal{U}$. Furthermore, $h - v_o$ is orthogonal to $\mathcal{U}$; that is, $\langle h - v_o, v \rangle = 0 \ \forall \ v \in \mathcal{U}$.*

We refer to v_o as the projection of h onto the space $\mathcal{U}$, and this is denoted as $\prod(h|\mathcal{U})$. Moreover, v_o is the only element in $\mathcal{U}$ such that $h - v$ is orthogonal to $\mathcal{U}$. For example, let $u_1(Z), \cdots, u_k(Z)$ be arbitrary elements of this space and $\mathcal{U}$ be a linear subspace spanned by $u_1, \cdots, u_k$. That is,

$$\mathcal{U} = a^T u;$$

for $a \in \Re^k$, where $u = (u_1, \cdots, u_k)^T$. Let h be an arbitrary element of $\mathcal{H}$. Then the projection of h onto the linear subspace $\mathcal{U}$ is given by the unique element $a_o^T u$ that

satisfies

$$\left\langle h - a_o^T u, a^T u \right\rangle = 0,$$

for all $a = (a_1, \cdots, a_k)^T \in \Re^k$. Also, a Hilbert space must satisfy the condition of completeness to guarantee the existence of the projection. By completeness we mean that every Cauchy sequence has a limit point that belongs to the space. A graphical representation of the projection theorem is found below:

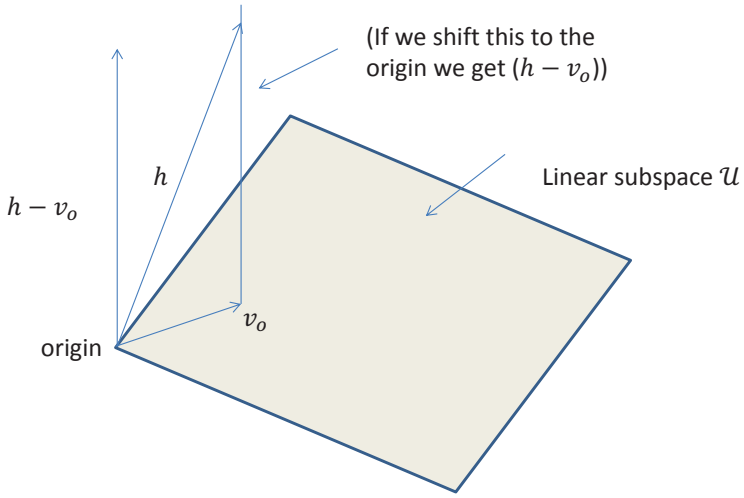

Figure 1: Geometrical Interpretation of Projection Theorem (adapted from Tsiatis (2010)).

3.0 A SIMPLE LOCALLY EFFICIENT ESTIMATOR FOR RELATIVE RISK IN CASE-COHORT STUDIES

3.1 INTRODUCTION

The standard cohort design requires assembly of all covariate (exposure) histories. In such studies, participants are followed prospectively, and subsequent status evaluations with respect to a disease or outcome are ascertained for exposure characteristics. Although subjects in a cohort design can be matched, which limits the influence of confounding variables, the outcome of interest could take time to occur. As a result, this type of studies are very costly to conduct. Cohort studies typically measure exposure in many controls. This measurement of exposures in controls are wasteful and thus inefficient. Not only that but also, cohort designs require extended follow-up to observe the development of the condition of interest, say, death due to lung cancer.

Prentice (1986) introduced the case-cohort design as an economical way of studying large cohort studies. This type of design is widely used in epidemiological studies with time-to-event data. The case-cohort study is a two-phase study where at the first phase a representative sample, referred to as the study cohort, is selected from the target population. In practice, certain covariates such as treatment allocations,

gender, age, and surrogate measurements of expensive covariates are obtained on all subjects in the cohort. Considering that evaluation of all covariates on this cohort could be expensive, this list of covariates is usually kept to a minimum. At the second phase, a subsample is selected from the cohort based on the case status, and the most expensive covariates that were not measured in the first phase are evaluated for the cases and the subsample. As a result, all cases are included in the subsample whereas only random samples of controls are included. The endpoint of interest in such studies is usually the failure time.

Rothman (2002) described the advantages of using the case-cohort design. This design is very efficient, since controls can be used in all risk sets for which they qualify. Furthermore, this type of design is very flexible, and allows testing hypotheses that were not anticipated when the cohort was drawn from the subsample. That is, the subsample can be used to study multiple outcomes. Also, the case-cohort analysis is less sensitive to missing covariate information. In addition, the case-cohort design reduces selection and information bias, because cases and non-cases are sampled from the same population.

Analysis of case-cohort studies is very similar to the usual Cox regression approach with a few modifications. It is assumed that, if we had full data, then the standard Cox proportional hazard model would suffice. Prentice (1986) showed how to estimate the relative risk from a Cox proportional hazard model without necessarily obtaining the covariate information of all subjects in the cohort. His method used an estimating equation for the regression parameters through a pseudolikelihood approach that weighted the contributions from the cases and subsample members using the inverses of their true or estimated sampling probabilities.

Since Prentice's initial work, many authors have derived variations of his method by proposing different estimating equations. Most of these methods, however, fail

to account for the covariate data collected outside the case-cohort sample, and thus incur the loss of efficiency. In particular, the Kalbfleisch & Lawless (1988) estimator, which is based on the modifications of the full data partial likelihood score function, weights the contributions from the cases and subsample members with inverses of their true or estimated probabilities, called sampled fraction. However, it ignores the first phase covariate data. Similarly, the estimator proposed by Self & Prentice (1988) ignores all the first phase information. Only Borgan & Goldstein et al. (1995) method utilizes some of the first phase information by stratifying on the first phase covariates (Kulich & Lin, 2004).

To improve the efficiency of the case-cohort estimators, several authors have introduced different estimators. For example, Barlow (1994) introduced an estimator that incorporates time-varying weights by proposing different weighting schemes. This weighting scheme assigned a value of one for all cases inside or outside the subsample. On the contrary, the cases in the subsample prior to failure, and the subsample controls are weighted by the inverses of the sampling fraction. Also, Chen & Lo (1999) proposed an estimator based on the partial likelihood score functions that improves the efficiency only when the fully observed covariates are binary (Kulich & Lin, 2004).

The efficiency of the relative risk estimation in analyzing case-cohort studies was further improved by Borgan & Langholz et al. (2000) where they proposed an exposure stratified case-cohort estimator, whereby complete covariate information is assembled for all failures (cases) and a stratified random subsample of the non-failures (controls). The stratification was based on an inexpensive easily observable covariate that is measured on all members in the cohort. By this approach, Borgan & Langholz et al. (2000) showed that stratified sampling designs can lead to substantial efficiency gain in case-cohort studies by employing the weighted versions of

the pseudolikelihood methods. In addition, Kulich & Lin (2004) introduced the combined doubly weighted estimator (CDW), which is a linear combination of the class of augmented estimators described in Robins et al. (1994) and Borgan & Goldstein (1995) estimators. Thus, the CDW estimator is protected against a deterioration of efficiency below that of the Borgan et. al. estimators due to an incorrectly specified model (Kulich & Lin, 2004). Mark & Katki (2006) proposed a simple estimator based on inverse-probability weighting (Horvitz, 1952), which they refer to as the $\hat{\alpha}$-estimator, reflecting that the probability of being included in the sample is estimated. In addition, Nan (2004) proposed an estimator for case-cohort designs with discrete covariates by solving the efficient score equations using a one-step Newton-Raphson approximation.

All the aforementioned estimators seek to efficiently estimate the association parameters from a case-cohort design by using the Cox proportional hazard model. Although the CDW estimator has appealing asymptotic properties, it may not always perform well in finite samples. Furthermore, it is computationally complex, and often infeasible to implement, and the estimator becomes unstable when the subsample control becomes small (Mark & Katki, 2006). In addition, the efficiency gain of the combined doubly weighted estimator depends on whether a fully observed continuous or binary covariate is observed at baseline. That is, fully observed continuous covariates achieve more efficiency compared to fully observed binary covariates.

The $\hat{\alpha}$-estimator proposed by Mark & Katki (2006) relaxes the requirement that the selection of subjects into the subsample be independent with known probabilities. It is similar to the usual inverse-probability weighted Horvitz-Thompson estimator (Breslow et al., 2009). The authors replaced the sampling probability α with its maximum likelihood estimates $\hat{\alpha}$. The efficiency gain of the $\hat{\alpha}$-estimator depends on the assumed correctness of the logistic model used to estimate α. The estimation

procedure ignores the first phase information, and hence there is further room for efficiency gain.

In this chapter, we present a semiparametric locally efficient estimator for analyzing case-cohort studies that do not require strong model assumptions and yet achieve efficiency gain over $\hat{\alpha}$-estimator. We derive the most efficient estimator along the lines of the semiparametric theory of Robins et. al (1994) by restricting ourselves to a subclass of estimators which are regular and asymptotically linear (Newey, 1990). Our proposed estimator is an asymptotically linear estimator which is not only easily computable, but also the most efficient within this subclass of estimators, and enjoys nice asymptotic properties.

In the next section, we present the model, notation, and assumptions used throughout this chapter. In Section 3.2, we derive a class of estimating equations for analyzing case-cohort studies. These class of estimators are regular and asymptotically linear. Section 3.3 presents our proposed locally efficient estimator. We demonstrate how to draw inference on the regression parameters on our proposed locally efficient estimator (LEE). Also, we show that LEE has the smallest asymptotic variance among all the class of restricted asymptotic linear (RAL) estimators. Section 3.4 shows our simulation experiment comparing our proposed estimator to the $\hat{\alpha}$-estimator. Finally, Section 3.5 discusses the analysis of the Wilm's tumor data.

3.2 MODEL, NOTATION, AND ASSUMPTIONS

Let T be the failure time, C be the potential censoring time and Z be the vector of covariates. Suppose that T is conditionally independent of C given Z and that the

conditional distribution of T given Z follows the Cox (1972) hazards model

$$\lambda(t|Z) = \lambda_o(t)exp(\beta^T Z) \tag{3.2.01}$$

where $\lambda(t|Z)$ is the conditional hazard for failure given the covariate history up to time t, β is a vector-valued parameter, and $\lambda_o(t)$ is an unspecified baseline hazard function. We will often write $\lambda(t|Z)$ as $\lambda(t)$ for brevity. Our goal is to draw inference about β from data observed through a case-cohort sampling scheme.

The case-cohort design evaluates some covariates on the overall cohort and measures expensive covariates on the subcohort of controls and all cases. Let the observed data be denoted by

$$\{\Delta_i, U_i, (1-\Delta_i)\xi_i, \{\Delta_i + (1-\Delta_i)\xi_i\}Z_i, W_i\}$$

$i = 1, \cdots, n$, where $\Delta_i = I(T_i \leq C_i)$, the indicator for cases, $U_i = min(T_i, C_i)$ the observed time, ξ is the indicator for the controls ($\Delta_i = 0$) who are in the subcohort, Z_i be the covariate of interest that is observed for all individuals in the subcohort ($\Delta_i = 1$) or $(1-\Delta_i)\xi_i = 1$, or equivalently, $\Delta_i + (1-\Delta_i)\xi_i = 1$, and W_i is the other covariates observed for the i-th individual. We assume that the cohort study involves n individuals.

Inference for Cox model with simple random sampling follows a partial likelihood approach. Since our method primarily relies on the influence function from this process, we briefly describe the procedure in this section. Suppose the exposure Z_i is observed on all individuals in the sample. The partial likelihood score equation for estimating β in such case is given by

$$\sum_{i=1}^{n} \int_0^{T_i} (Z_i - \mathbf{E}(\mathbf{u}, \beta))dN_i(u) = 0 \tag{3.2.02}$$

where

$$\mathbf{E}(\mathbf{u},\beta) = \frac{\sum_{i=1}^{n}[Y_i(u)Z_i exp(\beta^T Z_i)]}{\sum_{i=1}^{n}[Y_i(u)exp(\beta^T Z_i)]}$$

is the weighted average of the exposure vector Z_i among those who are at risk at time u, $Y_i(u) = I(U_i \geq u)$ be the at risk indicator at time u, and $N_i(u) = \Delta_i I(U_i \leq u)$, $i = 1, 2, \cdots, n$; τ is a fixed time chosen to limit the analysis to a follow-up time beyond which there is still a reasonable number at risk. The solution of β, $\hat{\beta}_{PL}$ from Equation (3.2.02) is known to follow an asymptotic normal distribution. Moreover, one can write

$$n^{1/2}(\hat{\beta}_{PL} - \beta_o) = n^{-1/2}\sum_{i=1}^{n}\varphi_i + o_p(1),$$

where

$$\varphi_i = I_i^{-1}(\beta_o)\int_0^{\tau}\{Z_i - \mathcal{E}(u,\beta_o)\}dM_i(u), \tag{3.2.03}$$

is the influence function of $\hat{\beta}_{PL}$, with

$$\mathcal{E}(u,\beta) = \frac{s^1(u,\beta)}{s^o(u,\beta)} = \frac{E[Y_i(u)Z_i exp(\beta^T Z_i)]}{E[Y_i(u)exp(\beta^T Z_i)]},$$

$$I_i(\beta) = E\left[\left\{\int_0^{\tau}(Z_i - \mathcal{E}(u,\beta))dM_i(u)\right\} \times \left\{\int_0^{\tau}(Z_i - \mathcal{E}(u,\beta))dM_i(u)\right\}^T\right],$$

where $M_i(u) = N_i(u) - \int_0^u \lambda(u|Z_i)Y_i(u)du$ is the martingale process corresponding to the hazard function $\lambda(u|Z_i)$, and the death and at-risk processes $N_i(u)$ and $Y_i(u)$ respectively; $o_p(1)$ is a term that converges to zero in probability as n approaches infinity.

In the case-cohort sampling, however, not all the exposure variables are measured on all individuals. Therefore, it is not possible to estimate β from Equation (3.2.02).

To account for the variable probability of being included in the subsample, Equation (3.2.02) is usually weighted by the inverse of the probability of inclusion. In what follows, we first describe common approaches to estimating β from the case-cohort sampling, and their limitations, and then we describe our proposed estimator.

Almost all existing estimators for analyzing case-cohort studies are based on score equations similar to Equation (3.2.02). In a typical case-cohort setting, it takes the form

$$\sum_{i=1}^{n} \int_0^{\tau} (Z_i - \mathbf{E_c}(\mathbf{u}, \beta)) dN_i(u) = 0 \tag{3.2.04}$$

where

$$\mathbf{E_c}(\mathbf{u}, \beta) = \frac{\sum_{i=1}^{n} \frac{\xi_i}{\alpha_i} Y_i(u) Z_i exp(\beta^T Z_i)}{\sum_{i=1}^{n} \frac{\xi_i}{\alpha_i} Y_i(u) exp(\beta^T Z_i)}$$

is the at-risk average of the exposure vector Z_i, and α_i is the sampling fraction which is estimated from true or estimated sampling probability. Various proposals for obtaining the sampling weight, α_i, have been published, yielding different case-cohort estimators (Kulich & Lin, 2004). In particular, the $\hat{\alpha}$-estimator is obtained by solving the equation

$$\sum_{i=1}^{n} \left\{ \Delta_i + \frac{(1 - \Delta_i)}{\hat{\alpha}_i} \xi_i \right\} \int_0^{\tau} (Z_i - \mathbf{E_c}(\mathbf{u}, \beta)) dN_i(u) = 0 \tag{3.2.05}$$

This estimator replaces the sampling probability α_i by its maximum likelihood estimate $\hat{\alpha}_i$ in a correctly specified model. Since the estimation procedure in Equation (3.2.04) eliminates subjects with incomplete data, all existing estimators that incorporate the pseudoscore in Equation (3.2.02) for analyzing case-cohort study including the $\hat{\alpha}$-estimator, use only the second phase subjects, while ignoring the first phase information. This leads to loss of efficiency.

To account for the missing covariates which result from ignoring the first phase information in the estimation in Equation (3.2.04), Robins et al. (1994), Laan & Robins (2003), and Nan (2004) have proposed estimators that augment the full data influence function, φ_i, by projecting it onto the orthogonal complement of the nuisance tangent space. In particular, the estimator proposed by Nan (2004), is obtained by a one-step Newton-Raphson approximation, solves the efficient score equation with initial values from existing estimators. In our notation, this estimator has influence function

$$\left\{\Delta_i + \frac{(1-\Delta_i)}{\alpha_i}\xi_i\right\}\varphi_i + \left\{1 - \Delta_i - \frac{(1-\Delta_i)}{\alpha_i}\xi_i\right\}E\{\varphi_i|data\}, \tag{3.2.06}$$

where $E\{\varphi_i|data\}$ indicates the expectation of the full data influence function given the observed data. The expectation in Equation (3.2.06) contains population quantities which are often intractable, and without additional assumptions on the full data model and censoring mechanism, can not be be reasonably estimated with finite samples (Van der Laan et al., 2003 p.35). As a result, in the section that follows, we propose an estimator that restricts the influence function to a class of estimators that are regular and asymptotically normally distributed. Our proposed locally efficient estimator (LEE) is built on the semiparametric efficiency theory [Tsiatis, 2006] contains quantities that are easy to calculate, and is more efficient than the $\hat{\alpha}$-estimator.

3.3 PROPOSED LOCALLY EFFICIENT ESTIMATOR

In this section, we describe the procedure in deriving our proposed estimator. We restrict ourselves to the class of estimators that are regular and asymptotically linear

(RAL) (Newey, 1990). Following the theory of inverse-weighting and the semiparametrics, all influence functions for case-cohort estimator of β can be written as

$$\left\{\Delta_i + \frac{(1-\Delta_i)}{\alpha_i}\xi_i\right\}\varphi_i + \left\{1 - \Delta_i - \frac{(1-\Delta_i)}{\alpha_i}\xi_i\right\}g[\mathcal{F}(T_i)], \tag{3.3.01}$$

where φ_i is as defined in Equation (3.2.03), and $\mathcal{F}(t)$ is the history of covariates up to time t. The optimal influence function that gives rise to semiparametric efficient estimator is given by Equation (3.2.06). However, it is not easy to implement in practice due to its complex structure with intractable expectations. Alternatively, we restrict the class by setting $g(\cdot)$ as a linear function of the data indexed by a vector parameter γ, namely,

$$g[\mathcal{F}(T_i)] = \gamma^T H(T_i),$$

where $H(T_i)$ is a vector function of the covariates. In other words, we start with the influence function

$$\Psi_i = \left\{\Delta_i + \frac{(1-\Delta_i)}{\alpha_i}\xi_i\right\}\varphi_i + (1-\Delta_i)\left(\frac{\alpha_i - \xi_i}{\alpha_i}\right)\gamma^T H(T_i), \tag{3.3.02}$$

The influence function (3.3.02) is indexed by the $q-$dimensional vector parameter γ. Choice of this vector parameter γ will determine how efficient the corresponding estimator will be. Therefore, the problem of finding the estimator with the minimum variance is equivalent to finding optimal γ for which the variance of Ψ_i in (3.3.02) is minimum. Also, the influence function for the $\hat{\alpha}$-estimator belongs to this class with $\gamma = 0$. Therefore, optimal influence function in this class will be more efficient than the influence function for $\hat{\alpha}$-estimator.

Let us define $K(u) = Pr(C_i > u)$ and $M_i^c(u) = N_i^c(u) - \int_0^u \lambda^c(t)Y_i(t)dt$ be the martingale associated with the censoring process, where $N_i^c(u) = I(U_i \leq u, \Delta = 0)$,

and $\lambda^c(u)$ is the hazard rate for the censoring distribution. Then plugging in φ_i from Equation (3.2.03) and using Gill (1980) equality

$$\frac{\Delta_i}{K(T_i)} = 1 - \int_0^{T_i} \frac{dM_i^c(u)}{K(u)},$$

we can express Equation (3.3.02) as

$$\begin{aligned}
&\left\{K(T_i) - K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)} + \frac{\xi_i}{\alpha_i}\left[1 - K(T_i) + K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)}\right]\right\} \\
&\times I_i^{-1}(\beta)\int_0^{\tau}(Z_i - \mathcal{E}(u,\beta))dM_i(u) + \left\{1 - K(T_i) + K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)} - \right. \\
&\left. \frac{\xi_i}{\alpha_i}\left[1 - K(T_i) + K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)}\right]\right\} \times \gamma^T H(T_i)
\end{aligned} \tag{3.3.03}$$

To find the optimal influence function, as in Robins et al. (1994), we consider the Hilbert space $\mathcal{H}$ consisting of all zero-mean random functions of the observed data with finite variance equipped with the covariance inner product. Within this space we define the closed linear subspace $\mathcal{U}$ consisting of random functions

$$\mathcal{U}_i = \left\{1 - K(T_i) + K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)} - \frac{\xi_i}{\alpha_i}\left[1 - K(T_i) + K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)}\right]\right\} \times \gamma^T H(T_i). \tag{3.3.04}$$

Our aim is to find the γ which minimizes the variance of Equation (3.3.03), or equivalently, to find the element in $\mathcal{U}$ which is at the minimum distance from

$$\begin{aligned}
\mathcal{V}_i = &\left\{K(T_i) - K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)} + \frac{\xi_i}{\alpha_i}\left[1 - K(T_i) + K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)}\right]\right\} \\
&\times I_i^{-1}(\beta)\int_0^{\tau}(Z_i - \mathcal{E}(u,\beta))dM_i(u).
\end{aligned} \tag{3.3.05}$$

By the projection theorem for Hilbert spaces (Tsiatis, 2006, pp.13-19), and the results presented in the Appendix, we deduce that the optimal γ is given by

$$\gamma^{opt} = E\big[\zeta_i H(T_i) H^T(T_i)\big]^{-1} E\big[H(T_i)\eta_i\big] \tag{3.3.06}$$

where

$$\zeta_i = -\left(1 - \frac{\xi_i}{\alpha_i} + \frac{\xi_i}{\alpha_i^2}\right)\left\{(1 - K(T_i))^2 - K^2(T_i)\int_0^{T_i} \frac{\lambda^c(u)S(u)}{K(u)}du\right\}$$

and

$$\begin{aligned}\eta_i = K(T_i) I_i^{-1}\Bigg[&\left\{(1 - K(T_i))\left(1 - \frac{\xi_i}{\alpha_i} + \frac{\xi_i}{\alpha_i^2}\right)\right.\\ &\left. -K(T_i)\left(1 - \frac{\xi_i}{\alpha_i}\right)\right\}\int_0^{T_i}(Z_i - \mathcal{E}(u,\beta))\lambda^c(u)\lambda(u)S(u)du\\ &+K(T_i)\left(1 - \frac{\xi_i}{\alpha_i} + \frac{\xi_i}{\alpha_i^2}\right)\int_0^{T_i}\frac{(Z_i - \mathcal{E}(u,\beta))}{K(u)}\lambda^c(u)\lambda(u)S(u)du\Bigg].\end{aligned}$$

Thus the optimal influence function is given by (3.3.02) with γ replaced by γ^{opt}. Consequently, we can obtain the optimal estimator of β, $\hat{\beta}_{LE}$, by solving

$$\sum_{i=1}^{n}\hat{\Psi}_i = \sum_{i=1}^{n}\left[\left\{\Delta_i + \frac{(1-\Delta_i)}{\alpha_i}\xi_i\right\}\hat{\varphi}_i + \left\{1 - \Delta_i - \frac{(1-\Delta_i)}{\alpha_i}\xi_i\right\}\hat{\gamma}^{opt^T} H(T_i)\right] = 0, \tag{3.3.07}$$

where we estimate the following quantities as follows:

$$\hat{\varphi}_i = \hat{I}_i^{-1}(\hat{\beta}_{LE})\int_0^{\tau}\{Z_i - \mathbf{E_c}(\mathbf{u}, \hat{\beta}_{\mathbf{LE}})\}\{dN_i(u) - Y_i(u)\hat{\lambda}^c(u)du\}, \tag{3.3.08}$$

$$\hat{I}_i(\beta) = \Bigg[\Bigg\{\int_0^{T_i}(Z_i - \mathbf{E_c}(\mathbf{u},\beta)\{dN_i(u) - Y_i(u)\hat{\lambda}^c(u)du\}\Bigg\} \times\Bigg\{\int_0^{T_i}(Z_i - \mathbf{E_c}(\mathbf{u},\beta)\{dN_i(u) - Y_i(u)\hat{\lambda}^c(u)du\}\Bigg\}^T\Bigg], \tag{3.3.09}$$

$$\hat{\gamma}^{opt} = E\big[\hat{\zeta}_i H(T_i)H^T(T_i)\big]^{-1}E\big[H(T_i)\hat{\eta}_i\big], \tag{3.3.010}$$

$$\hat{\zeta}_i = -\left(1 - \frac{\xi_i}{\alpha_i} + \frac{\xi_i}{\alpha_i^2}\right)\left\{(1-\hat{K}(T_i))^2 - \hat{K}^2(T_i)\int_0^{T_i}\frac{\hat{\lambda}^c(u)\hat{S}(u)}{\hat{K}(u)}du\right\}, \tag{3.3.011}$$

$$\begin{aligned}\hat{\eta}_i = \hat{K}(T_i)I_i^{-1}(\beta)\Bigg[\Bigg\{&(1-\hat{K}(T_i))\left(1 - \frac{\xi_i}{\alpha_i} + \frac{\xi_i}{\alpha_i^2}\right)\\ &-K(T_i)\left(1 - \frac{\xi_i}{\alpha_i}\right)\Bigg\}\int_0^{\tau}(Z_i - \mathbf{E_c}(\mathbf{u},\hat{\beta}_{\mathbf{LE}}))\hat{\lambda}^c(u)\hat{\lambda}(u)\hat{S}(u)du\\ &+\hat{K}(T_i)\left(1 - \frac{\xi_i}{\alpha_i} + \frac{\xi_i}{\alpha_i^2}\right)\int_0^{\tau}\frac{(Z_i - \mathbf{E_c}(\mathbf{u},\hat{\beta}_{\mathbf{LE}}))}{\hat{K}(u)}\hat{\lambda}^c(u)\hat{\lambda}(u)\hat{S(}u)du\Bigg],\end{aligned} \tag{3.3.012}$$

$$\hat{\lambda}^c(u) = \frac{dN_i^c(u)}{Y_i(u)}, \tag{3.3.013}$$

$$\hat{\lambda}(u) = \frac{dN_i(u)}{Y_i(u)}, \tag{3.3.014}$$

and $\hat{S}(u)$ is the survival function estimated by the product limit estimator. Note that even though the hazard $\lambda(u|Z_i)$ and the survival $S(u|Z_i)$ are dependent on covariates, for simplicity, in estimating ζ_i and η_i, we ignored the covariates and

simply used Nelson-Aalen and product limit estimator to estimate them. This leads to an estimator for β that is consistent and asymptotically normal. The variance of $\hat{\beta}_{LE}$ can be estimated by

$$\text{var}(\hat{\beta}_{LE}) = \frac{\sum_{i=1}^{n} \hat{\Psi}_i^2}{n^2}. \tag{3.3.015}$$

3.4 SIMULATION STUDY

Simulation experiments have been carried out to evaluate the large sample properties of our proposed efficient estimator LEE. For comparisons, we also assessed the $\hat{\alpha}$-estimator, and then we used the full data Cox estimator as a reference estimator. The simulation set up is very similar to that described by Kulich & Lin (2004). Our simulation study involves three covariates, a binary covariate Z_1 with $Pr(Z_1 = 1) = p$, and two continuous covariates $Z_2 \sim N(0, 0.5^2)$, and $\log(Z_3) \sim N(cz_2, 0.5^2)$ conditional on $Z_2 = z_2$, and we set $p = 0.5$ and $c = 0.2$. Thus, our model contains three parameters: one for binary covariate Z_1, and two for continuous covariates $\tilde{Z}_2$ and Z_3. Also, the failure times are generated from the exponential distribution, and the censoring times are generated from a uniform distribution independent of the survival data. We choose 3,000 subjects in the study cohort and the subsample are drawn from the entire cohort.

Simulation results presented are for $exp(\beta_1) = 1.3, exp(\beta_2) = 1.2$, and $exp(\beta_3) = 1.2$ corresponding to Z_1, $\tilde{Z}_2$ and Z_3 respectively. We assumed that Z_1 and Z_3 were observed at phase one, while Z_2 was only observed at phase two. We generated a surrogate variable $\tilde{Z}_2 \equiv Z_2 + \epsilon$ for every subject, where ϵ is normal with mean zero independent of Z_2. We note that Z_2 and $\tilde{Z}_2$ have correlation equal to either

0.71 or 0.93. We designated the vector function of covariates, $H(T_i)$, as $\tilde{Z}_2$ and Z_3. Simulation results are based on 1,000 replications.

Table 1: Estimator, Monte Carlo standard deviations (MCSE) and relative efficiencies (RE) of LEE and $\hat{\alpha}$-estimators with $Corr(Z_2, \tilde{Z}_2) = 0.71$. True Relative Risks are $exp(\beta_1) = 1.3, exp(\beta_2) = 1.2$, and $exp(\beta_3) = 1.2$.

		Full data Cox	LE estimator		$\hat{\alpha}$-estimator	
Event Rate:(n)	Variable	Estimate (MCSE)	Estimate (MCSE)	RE	Estimate (MCSE)	RE
10% : 1275	Z_1	1.331(0.140)	1.339(0.145)	0.93	1.320(0.150)	0.87
	$\tilde{Z}_2$	1.201(0.099)	1.206(0.102)	0.94	1.197(0.109)	0.82
	Z_3	1.155(0.083)	1.158(0.085)	0.95	1.151(0.089)	0.87
20% : 1537	$Z1$	1.868(0.096)	1.876(0.100)	0.92	1.887(0.104)	0.85
	$\tilde{Z}_2$	1.486(0.076)	1.483(0.078)	0.95	1.477(0.084)	0.82
	$Z3$	1.107(0.083)	1.111(0.086)	0.93	1.112(0.088)	0.89

Table 1 presents the estimators and relative efficiencies for the different estimators. The first row shows the results of 10% event rate for all three estimators. This means that after including all the cases, we obtained 1275 samples for analysis. We set the full data Cox as the reference category, and compared both our proposed locally efficient estimator (LEE), and the $\hat{\alpha}$-estimator to the full data Cox model. The relative efficiencies have been calculated by using the ratio of the Monte Carlo means-squared errors. For instance, the entry 1.320 (0.150), RE = 0.87 for $\hat{\alpha}$-estimator in row one of Table 1 refers to the case where with a sample size of 1275, Monte Carlo mean of relative risk estimates for the covariate Z_1 is 1.32, showing a bias of 0.03. The ratio of the Monte Carlo mean-squared error of the full data Cox model to that of the $\hat{\alpha}$-estimator is 0.87. Thus, the $\hat{\alpha}$-estimator in this case is 13% less efficient

compared to the full data Cox estimator. In contrast, for the same scenario, the LE estimator has Monte Carlo mean of 1.339 and a relative efficiency of 0.93, showing a bias of 0.01. This suggests that the LE estimator is only 7% less efficient compared to the full data Cox estimator. Furthermore, comparing the estimators show that the efficiency gain of the LE estimator over the $\hat{\alpha}$-estimator is approximately 95%. In addition, for the variable $\tilde{Z}_2$ the LE estimator has Monte Carlo mean estimate of 1.206 and a relative efficiency of 0.94 compared to the $\hat{\alpha}$-estimator, which has Monte Carlo mean estimate of 1.197 and relative efficiency of 0.82. In terms of efficiency gain for the $\tilde{Z}_2$ variable, the LE estimator gains 87% efficiency over the $\hat{\alpha}$-estimator. For the variable Z_3, the LE estimator has Monte Carlo mean estimate of 1.158 and relative efficiency of 0.95, compared to the $\hat{\alpha}$-estimator which has Monte Carlo mean estimate of 1.151 and a relative efficiency of 0.87. This suggests an efficiency gain of 92% of the LE estimator over the $\hat{\alpha}$-estimator. From the relative efficiencies of the two methods, it is evident that our proposed estimator always performs better. Also, when we increased the event rate by sampling 20% from the controls (n= 1537), a similar trend is observed. The LE estimator performs better and, therefore, more efficient than the $\hat{\alpha}$-estimator.

Table 2: Estimator, Monte Carlo standard deviations (MCSE) and relative efficiencies (RE) of LEE and $\hat{\alpha}$-estimators with $Corr(Z_2, \tilde{Z}_2) = 0.93$. True Relative Risks are $exp(\beta_1) = 1.3, exp(\beta_2) = 1.2$, and $exp(\beta_3) = 1.2$.

		Full data Cox	LE estimator		$\hat{\alpha}$-estimator	
Event Rate:(n)	Variable	Estimate (MCSE)	Estimate (MCSE)	RE	Estimate (MCSE)	RE
10% : 1201	$Z1$	1.328(0.147)	1.336(0.149)	0.97	1.318(0.151)	0.95
	$\tilde{Z}_2$	1.196(0.101)	1.198(0.104)	0.94	1.191(0.108)	0.87
	$Z3$	1.151(0.081)	1.154(0.084)	0.93	1.149(0.089)	0.83
20% : 1425	$Z1$	1.866(0.095)	1.878(0.100)	0.90	1.891(0.103)	0.85
	$\tilde{Z}_2$	1.484(0.076)	1.490(0.078)	0.95	1.487(0.082)	0.86
	$Z3$	1.106(0.083)	1.112(0.085)	0.95	1.111(0.088)	0.89

Likewise, Table 2 shows the relative efficiency of our proposed estimator compared with other estimators, while varying the correlation coefficient from 0.71 to 0.93. We notice that the differences in correlation has little effect on the overall results. This holds true for all three estimators. However, the higher the event rate the (i.e 10% to 20%), the higher the relative efficiencies of case-cohort estimators compared to full-cohort estimator. The MCSEs are certainly smaller for higher event rate as the effective total sample size is larger compared to lower event rate.

3.5 ANALYSIS OF WILM'S TUMOR DATA

Wilm's tumor is a rare type of kidney cancer that occurs in children. Many factors contribute to the survival or relapse of this tumor. Some of the factors include: age at diagnosis, stage of diagnosis of disease (usually 4 stages), histological type of tumor, and the tumor diameter. We dichotomized the stage variable by combining stages I and II into one group and stages III and IV into another group. There are six parameters in our model excluding the intercept term. We have two interaction variables: histology and the continuous age variable, and the stage and tumor diameter variables. A total of 3915 patients were enrolled into the Wilm's Tumor Study [D'Angio et al, 1989]. We assume that all covariates except histological group are observed at phase one. Also, we assign the vector function of covariates, $H(T_i)$ as age and stage variables. We illustrate the proposed LE estimator by analyzing data from the Wilm's tumor study. Thus, after the second phase sampling we obtain 660 control subjects and 669 cases. As a result, a total of 1329 observations are analyzed.

Results in Table 3 show that the for each centimeter increase in tumor diameter, the relative risk of death is 1.06 for patients in stage I-II. This represents a 6% increase in risk of death per 1 unit increase in tumor diameter for individuals in this group. However, individuals in stage III-IV have a slightly lower risk (RR=0.98). Thus, an increase in tumor diameter for persons in stage I-II can be fatal but not for persons in stage III-IV. This is may seem counter-intuitive, however it is because individuals in stage III-IV are already at higher risk of dying from cancer and so an additional increase in diameter of the tumor would not necessarily determine the survival or otherwise death of the patient. The age effect is amplified among patients with unfavorable histology compared to those with favorable histology. For every year increase in age, the risk for a person with unfavorable histology increases

Table 3: Analysis of Wilm's Tumor Data Using LE Estimator.

Variable	Relative Risk	SE	CI
Tumor Diameter (per cm):			
For Stage I-II patients	1.06	0.015	(1.03, 1.09)
For Stage III-IV patients	0.98	0.013	(0.95, 1.01)
Stage III-IV vs Stage I-II:			
At tumor diameter = 5cm	0.71	0.065	(0.58, 0.84)
At tumor diameter = 11.5cm	0.43	0.096	(0.24, 0.61)
At tumor diameter = 20cm	0.22	0.086	(0.05, 0.39)
Age effect (per year):			
With Favorable Histology	0.95	0.034	(0.89, 1.02)
With Unfavorable Histology	1.10	0.018	(1.07, 1.13)
Unfavorable vs Favorable Histology:			
At age = 1 year	0.95	0.001	(0.95, 0.96)
At age = 3.5 years	0.66	0.086	(0.49, 0.83)
At age = 10 years	0.26	0.037	(0.19, 0.32)

by 10% (RR =1.10, CI =(1.07,1.13)) whereas for a person with favorable histology, age is not statistically significant. We show the interaction of age and histology, and stage and tumor diameter side by side in Figure 2.

As patients become older, the risk of death becomes higher for patients in the unfavorable histology group compared to patients in the favorable histology group. Relative risk of death in Stage III-IV compared to that in Stage I-II depends on the

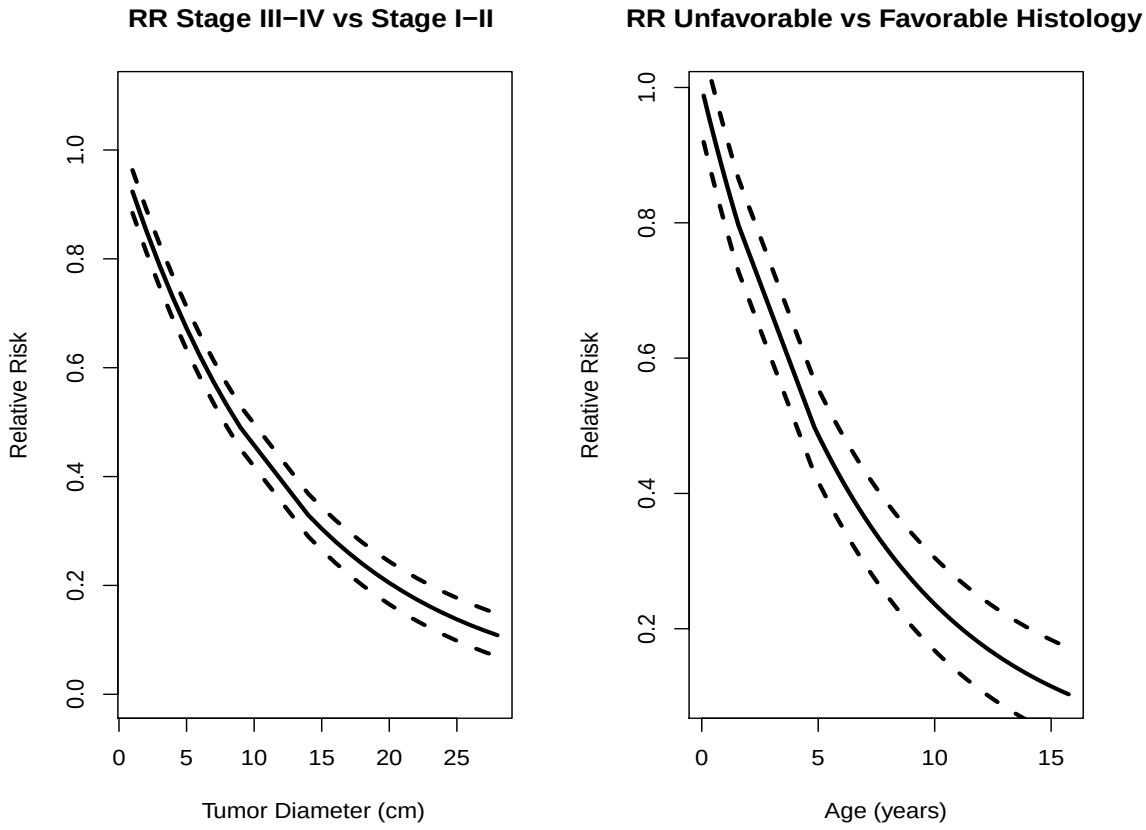

Figure 2: Relative risk of stage III-IV vs stage I-II as a function of tumor diameter (left panel), and relative risk of unfavorable vs favorable histology as a function of age (right panel).

diameter of tumor such that for smaller tumor diameter the relative risk is close to one whereas for larger tumor size, the Stage I-II patients are at greater risk of death compared to Stage III-IV patients (Figure 2).

3.6 DISCUSSION

In this chapter, we have derived an expression for the most efficient estimator for the restricted asymptotically linear estimators for analyzing case-cohort designs. The proposed LE estimator works well for both binary and continuous covariates. The LE estimator is guaranteed to gain efficiency over other available estimators such as CDW estimators or the estimators proposed by Mark & Katki (2006). In particular, the LE estimator is more efficient than the $\hat{\alpha}$-estimator. Our proposed estimator uses all the first phase information, and therefore, for inferential procedures, our estimator will give more accurate results than existing ones. Furthermore, our proposed estimator contains quantities that are easy to calculate, and has nice asymptotic properties.

On the contrary, the efficiency of the $\hat{\alpha}$-estimator depends on whether we have completely observed continuous covariates versus completely observed binary covariates. Not only that but also, the efficiency of the $\hat{\alpha}$-estimator depends on the correctness of the assumed logistic model, while the efficiency of LEE does not depend on whether the observed covariates are binary or continuous. Through simulations and data analysis, we have shown that the efficiency gain of the LE estimator is substantial. Also, the LE estimator is consistent and asymptotically normal. In Chapter 4, we implement the LE estimator that handles time-dependent covariates.

4.0 A SIMPLE LOCALLY EFFICIENT ESTIMATOR FOR RELATIVE RISK FOR TIME-DEPENDENT VARIABLES IN CASE-COHORT STUDIES

In survival analysis, a frequently used method of associating covariates with the time of failure is Cox regression (Cox, 1972; Andersen & Gill, 1982). The Cox proportional hazard model can be written as

$$\lambda_i(t) = \lambda_o(t) exp(\beta^T Z_i) \tag{4.0.01}$$

where $\lambda_i(t)$ is the hazard at time t of the i-th individual, $\lambda_o(t)$ is the baseline hazard at time t, Z_i is a vector of covariate values corresponding to the i-th individuals, and β is a vector of coefficients. The Cox model defined in Equation (4.0.01) is very popular because it is robust to distributional assumption of the survival time and can be utilized in many situations (Liu et al., 2010). Although the baseline hazard, $\lambda_o(t)$, is left unspecified, we can estimate β and thus compute the hazard ratio. This is because the baseline hazard does not depend on Z, but only on t. In addition, the baseline hazard can be regarded as infinite dimensional, while our parameter of interest, β, is finite dimensional. As a result, the Cox model is a classic semiparametric model. Hence, it requires few model assumptions.

To make inference on the regression parameters, Cox (1972) developed a non-

parametric method called partial likelihood. Estimation of the parameter values is obtained by use of maximum partial likelihood estimation. The partial likelihood for Cox model is given by

$$L_p = \prod_{i=1}^{n} \left[\frac{exp(\beta^T Z_i)}{\sum_{j \in R(t_i)} exp(\beta^T Z_j)} \right]^{\Delta_i}, \tag{4.0.02}$$

where t_i denotes the failure time for the i-th individual, $R(t_i)$ is the number of people at risk of experiencing an event at time t_i, that is, the risk set, and Δ_i is an indicator for a case ($\Delta_i = 1$) or censored ($\Delta_i = 0$) individuals. It should be noted that the partial likelihood estimation described in (4.0.02) is valid only when there are no ties in the data set. That is, time is assumed to be continuous and no two subjects have the same event time (Klein & Moeschberger, 2005). Several methods have been proposed to handle situations where there are ties in the data set. For example, Breslow (1974) suggested replacing the term $\sum_{j \in R(t_i)} exp(\beta^T Z_j)$ in the denominator in Equation (4.0.02) by

$$\left(\sum_{j \in R(t_i)} exp(\beta^T Z_j) \right)^{d_i},$$

where d_i is the number of events occurring at time t_i. The denominator corresponds to the sum of all hazards at t_i for all individuals who were at risk at time t_i, while the numerator is the hazard for individual i at time t_i. This approach is known as the Breslow approximation. Also, Efron (1977) extended the Breslow approximation to situations where there are large number of ties. Like the Breslow approximation, Efron's method will yield estimates of β which are biased toward zero when there are many ties (Pugh, 1993). In fact, the Efron method gives much more closer estimates than the Breslow approximation. In addition, Kalbfleisch & Prentice (1978) derived a likelihood involving only β and Z, excluding $\lambda_o(t)$, based on the marginal distribution

of the ranks of the observed failure time. By this approach, all possible orderings of the tied events are calculated, and prababilities of each are summed. Usually, this approach is referred to as the exact method. It is the most complex, but also the most accurate (Pugh, 1993).

Several methods have been proposed to extend the Cox proportional hazard model in (4.0.01) to time-dependent covariates. The Cox regression model specifies the intensity function for the observed time counting process N_i for a time-dependent covariate vector $Z(t)$ evaluated at time t as

$$\lambda(t|Z(t)) = \lambda_o(t)exp(\beta^T Z(t)) \tag{4.0.03}$$

where $\lambda_o(t)$ is an unspecified baseline hazard function, $\lambda(t|Z(t))$ is the conditional hazard for failure given the covariate history up to time t, and β is a vector-valued parameter. Evidently, Equation (4.0.03) generalizes the Cox regression model defined in (4.0.01). A key feature in (4.0.01) is proportionality of hazard functions for individuals with different covariates is lost in (4.0.03). Different authors have considered variable-influence covariates in models similar to Equation (4.0.03), but only under rather stringent assumptions on the functional form of β.

Fisher & Lin (1999) described the advantages of using the Cox proportional hazards regression model with time-dependent covariate. A time-dependent covariate is defined as a variable whose value for a given individual may change over time. The modeling of time-dependent covariates involve the choice of a functional form and this may require some deep biological insights (Liu, et al. 2010). Kalbfleisch & Prentice (2002) distinguished between external and internal time-dependent covariates. An external time-dependent covariate is one that is not directly related to the failure mechanism. An example would be an individual's age in a long-term follow-up study. On the contrary, an internal time-dependent covariate is a value over time generated

by the individual under study. An example would be blood pressure measured over the course of a study.

The estimation and inference for the Cox proportional hazard model for time-dependent covariates defined in Equation (4.0.03) have been studied by several authors. For example, Zucker & Karr (1990) proposed estimating the regression parameter function β, by maximizing a penalized version of the partial likelihood. Using a penalized likelihood technique, their estimator allowed β to be infinite dimensional. Thus, they outlined a computational approach appropriate for the maximization of the partial likelihood. Their technique, however, applies to only large sample sizes (Murphy, 1993). Also, Hastie & Tibshirani (1990) described a general framework of varying-coefficient models for survival data. They proposed an estimator that allowed the regression coefficients to change smoothly with the value of other variables. Their estimator is based on a penalized least squares criterion that imposes restrictions on the coefficient functions. In addition, Murphy & Sen (1991) proposed an estimator similar to that described by Zucker & Karr (1990), which also allowed the regression parameter β to be infinite dimensional. Their estimator, however, relies on maximization of the likelihood estimator based on simple histogram sieves. Other methods that have been proposed based on Cox proportional hazard model with time-dependent covariates involve assessing the model adequacy and goodness-of-fit measures. These methods seek to handle the violation of the proportional hazard model assumption, which occurs in a Cox proportional hazard model with time-dependent covariates. For example, Grambsch & Therneau (1994) visualized the parameter function as a smooth function, and fitted a weighted least squares line to the residual plots. The estimator proposed by Marzec & Marzec (1997) used a chi-squared type goodness-of-fit test to handle the violation of the proportional hazard assumption, while Cai & Sun (2003) developed an estimator that transformed

the regression parameter into a smooth coefficient function. Thus, the estimator proposed by Cai & Sun (2003) can be used as a diagnostic tool to handle the departures from the proportional hazard model.

In this chapter, we present a semiparametric locally efficient estimator for analyzing case-cohort studies involving time-dependent covariates that do not require strong model assumptions. The basic idea is the same as that described in Chapter 3, except that we introduce time-dependent covariate vectors in the model. In the next section, we introduce the model, notation, and assumptions used throughout this chapter, and we derive a class of estimating equations that allow time-dependent covariates in a case-cohort design. Section 4.2 discusses our proposed locally efficient estimator (LEE). In Section 4.3, we test our proposed estimator by running two simulation studies. The first simulation study incorporates a binary time-dependent covariate, while the second simulation study uses a continuous time-dependent covariate.

4.1 MODEL, NOTATION, AND ASSUMPTIONS

Let T be the failure time, C be the potential censoring time, and Z be a vector of covariates, possibly time dependent. Suppose that T is conditionally independent of C given Z, and that the conditional distribution of T given $Z(t)$ follows Cox (1972) proportional hazards model

$$\lambda(t|Z_i(t)) = \lambda_o(t)exp(\beta^T Z_i(t)),$$

Suppose that we have n individuals in the study, such that the data consists of

$$\{U_i, \Delta_i, Z_i(t_j), j = 1, \cdots, m_i i = 1, \cdots, n, \}$$

where $m_i \leq U_i$, Δ_i indicates the status of the observed event, taking a value of 1 for observing a real failure event and 0 otherwise. At a specific time t, let $R(t) = \{i : U_i \geq t\}$ denote the risk set. As described in Chapter 3, the case-cohort design evaluates some covariates on the overall cohort and measures expensive covariates on the subcohort of controls and all cases. In a simple random sampling design, where $Z_i(t)$ is observed for all individuals in the sample, the estimation of β can be done by maximizing the partial likelihood

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{exp(\beta^T Z_i(t_i))}{\sum_{j \in R_i} exp(\beta^T Z_j(t_i))} \right]^{\Delta_i}, \tag{4.1.01}$$

where $R_i \equiv R(t_i)$ is the set of individuals who are at risk at time t_i. The partial likelihood score function incorporating time-dependent covariates is given by

$$U(\beta) = \sum_{i=1}^{n} \Delta_i \left\{ Z_i(t_i) - \frac{\sum_{j \in R_i} exp(\beta^T Z_j(t_i)) Z_j(t_i)}{\sum_{j \in R_i} exp(\beta^T Z_j(t_i))} \right\} \tag{4.1.02}$$

where

$$\frac{\sum_{j \in R_i} exp(\beta^T Z_j(t_i)) Z_j(t_i)}{\sum_{j \in R_i} exp(\beta^T Z_j(t_i))}$$

is the population weighted average of the covariates of individuals at risk at time t_i. The maximum partial likelihood estimator $\hat{\beta}$ is the solution of $U(\beta) = 0$

Our goal is to adapt the above expression in Equation (4.1.02) to a case-cohort sampling design. When the exposure $Z_i(t)$ is observed on all individuals in the sample, the expression in (4.1.02) can be re-written as stochastic integrals with respect to the counting process as follows:

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\tau} [Z_i(u) - \mathbf{E}(u, \beta)] dN_i(u), \tag{4.1.03}$$

where

$$\mathbf{E}(u,\beta) = \frac{\sum_{i=1}^{n} Y_i(u) Z_i(u) exp(\beta^T Z_i(u))}{\sum_{i=1}^{n} Y_i(u) exp(\beta^T Z_i(u))}$$

is the population weighted average of the exposure vector $Z_i(u)$ among those who are at risk at time u, $Y_i(u) = I(U_i \geq u)$ be the at risk indicator at time u, and $N_i(u) = \Delta_i I(U_i \leq u)$, $i = 1, 2, \cdots, n$; τ is a fixed time chosen to limit the analysis to a follow-up time beyond which there is still a reasonable number at risk. The solution of β, $\hat{\beta}_{PL}$ from Equation (4.1.03) has been shown to follow an asymptotic normal distribution (Fisher & Lin, 1999). In fact, $\hat{\beta}_{PL}$ is a consistent estimator of β. The expectation of the terms inside the integral from Equation (4.1.03) equals zero. Furthermore, one can write

$$n^{1/2}(\hat{\beta}_{PL} - \beta_o) = n^{-1/2} \sum_{i=1}^{n} \chi_i + o_p(1),$$

where

$$\chi_i = I_i^{-1}(\beta_o) \int_0^{\tau} \{Z_i(u) - \mathcal{E}(u, \beta_o)\} dM_i(u), \tag{4.1.04}$$

is the influence function of $\hat{\beta}_{PL}$, with

$$\mathcal{E}(u,\beta) = \frac{s^1(u,\beta)}{s^o(u,\beta)} = \frac{E[Y_i(u) Z_i(u) exp(\beta^T Z_i(u))]}{E[exp(\beta^T Z_i(u))]},$$

$$I_i(\beta) = \frac{-\partial^2 log L(\beta)}{\partial\beta\partial\beta^T} = \sum_{i=1}^{n} \int_0^{\tau} \mathcal{V}(u,\beta) dN_i(u), \tag{4.1.05}$$

where

$$\mathcal{V}(u,\beta) = [Z_i(u) - \mathcal{E}(u,\beta)]^{\otimes 2}$$

with $b^{\otimes 2} = bb^T$ for a vector b, and $M_i(u) = N_i(u) - \int_0^u \lambda(u|Z_i(u))Y_i(u)du$ is the martingale process corresponding to the hazard function $\lambda(u|Z_i(u))$, and the death and at-risk processes $N_i(u)$ and $Y_i(u)$ respectively; $o_p(1)$ is a term that converges to zero in probability as n approaches infinity.

In the case-cohort sampling design, it is not possible to estimate β from Equation (4.1.03) because exposure measurements are not obtained on all individuals. Usually, the score Equation (4.1.03) is modified as

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\tau} [Z_i(u) - \mathbf{E_{tc}}(\mathbf{u}, \beta)]dN_i(u), \tag{4.1.06}$$

to make it applicable to the case-cohort sampling where

$$\mathbf{E_{tc}}(\mathbf{u}, \beta) = \frac{\sum_{i=1}^{n} \frac{\xi_i}{\alpha_i} Y_i(u)Z_i(u)exp(\beta^T Z_i(u))}{\sum_{i=1}^{n} \frac{\xi_i}{\alpha_i} Y_i(u)exp(\beta^T Z_i(u))}$$

is the weighted at-risk average of the exposure vector $Z_i(u)$ over the subcohort sample weighted by the sampling fraction α_i, and ξ_i is a binary indicator for the controls. Unfortunately, the estimation procedure in (4.1.06) eliminates all subjects with incomplete data. As a result, estimation procedures based on (4.1.06) lead to the loss of efficiency. In the next section, we propose a locally efficient estimator (LEE). The LE estimator is asymptotically normally distributed, and it is built on the semiparametric efficiency theory (Wahed, 2006). Also, the LE estimator uses all the first phase covariate information under a case-cohort sampling design. Therefore, it is guaranteed to be more efficient than any estimator built on the pseudoscore in Equation (4.1.03).

4.2 LOCALLY EFFICIENT ESTIMATOR

In this section, we explain how our proposed estimator is derived. We restrict ourselves to the class of estimators that are regular and asymptotically linear (RAL) (Newey, 1990). Applying the theory of inverse weighting and semiparametrics, all influence functions for case-cohort estimator of β involving time-dependent covariates can be written as

$$\left\{\Delta_i + \frac{(1-\Delta_i)}{\alpha_i}\xi_i\right\}\chi_i + \left\{1 - \Delta_i - \frac{(1-\Delta_i)}{\alpha_i}\xi_i\right\}g[\mathcal{F}(T_i)], \tag{4.2.01}$$

where χ_i is as defined in Equation (4.1.04), and $\mathcal{F}(t)$ is the history of covariates up to time t. The influence function defined in (4.1.04) contains quantities that are not easy to implement. Therefore, we set $g(\cdot)$ as a linear function of the data indexed by a q-dimensional parameter γ, by restricting the class, leading to

$$g[\mathcal{F}(T_i)] = \gamma^T H(T_i),$$

where $H(T_i)$ is a vector function of the covariates. In other words, we start with the influence function below

$$\Psi_i = \left\{\Delta_i + \frac{(1-\Delta_i)}{\alpha_i}\xi_i\right\}\chi_i + (1-\Delta_i)\left(\frac{\alpha_i - \xi_i}{\alpha_i}\right)\gamma^T H(T_i). \tag{4.2.02}$$

Suppose $K(u) = Pr(C_i > u)$ and $M_i^c(u) = N_i^c(u) - \int_0^u \lambda^c(t)Y_i(t)dt$ is the martingale associated with the censoring process, where $N_i^c(u) = I(U_i \leq u, \Delta = 0)$, and $\lambda^c(u)$ is the hazard rate for the censoring distribution. When we plug in φ_i from Equation (4.1.04) and using Gill (1980) equality

$$\frac{\Delta_i}{K(T_i)} = 1 - \int_0^{T_i} \frac{dM_i^c(u)}{K(u)},$$

we are able to write Equation (4.2.02) as

$$\left\{K(T_i) - K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)} + \frac{\xi_i}{\alpha_i}\left[1 - K(T_i) + K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)}\right]\right\}$$
$$\times I_i^{-1}(\beta)\int_0^{\tau}(Z_i(u) - \mathcal{E}(u,\beta))dM_i(u) + \left\{1 - K(T_i) + K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)} - \right. \quad (4.2.03)$$
$$\left.\frac{\xi_i}{\alpha_i}\left[1 - K(T_i) + K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)}\right]\right\} \times \gamma^T H(T_i)$$

As described in Chapter 3, we consider the Hilbert space $\mathcal{H}$ consisting of all mean zero random functions of the observed data with finite variance equipped with the covariance inner product. We define a closed linear subspace $\mathcal{U}$, within this space, consisting of random functions

$$\left\{1 - K(T_i) + K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)} - \frac{\xi_i}{\alpha_i}\left[1 - K(T_i) + K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)}\right]\right\} \times \gamma^T H(T_i). \quad (4.2.04)$$

Our goal, again, is to find the γ which minimizes the variance of Equation (4.2.03). Specifically, we find the element in $\mathcal{U}$ which is at the minimum distance from

$$\left\{K(T_i) - K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)} + \frac{\xi_i}{\alpha_i}\left[1 - K(T_i) + K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)}\right]\right\}$$
$$\times I_i^{-1}(\beta)\int_0^{\tau}(Z_i(u) - \mathcal{E}(u,\beta))dM_i(u).$$

By the projection theorem for Hilbert spaces (Tsiatis, 2006, pp.13-19), we derive that the optimal γ is given by

$$\gamma^{opt} = E\big[\zeta_i H(T_i)H^T(T_i)\big]^{-1}E\big[H(T_i)\eta_i\big] \quad (4.2.05)$$

where

$$\zeta_i = -\left(1 - \frac{\xi_i}{\alpha_i} + \frac{\xi_i}{\alpha_i^2}\right)\left\{(1 - K(T_i))^2 - K^2(T_i)\int_0^{T_i} \frac{\lambda^c(u)S(u)}{K(u)}du\right\}$$

and

$$\begin{aligned}\eta_i = K(T_i)I_i^{-1}(\beta)\Bigg[&\left\{(1 - K(T_i))\left(1 - \frac{\xi_i}{\alpha_i} + \frac{\xi_i}{\alpha_i^2}\right)\right.\\ &\left.-K(T_i)\left(1 - \frac{\xi_i}{\alpha_i}\right)\right\}\int_0^{T_i}(Z_i(u) - \mathcal{E}(u,\beta))\lambda^c(u)\lambda(u)S(u)du\\ &+K(T_i)\left(1 - \frac{\xi_i}{\alpha_i} + \frac{\xi_i}{\alpha_i^2}\right)\int_0^{T_i}\frac{(Z_i(u) - \mathcal{E}(u,\beta))}{K(u)}\lambda^c(u)\lambda(u)S(u)du\Bigg].\end{aligned}$$

Thus the optimal influence function is given by (4.2.02) with γ replaced by γ^{opt}. As a result, we can obtain the optimal estimator of β, $\hat{\beta}_{LE}$, by solving

$$\sum_{i=1}^{n}\hat{\Psi}_i = \sum_{i=1}^{n}\left[\left\{\Delta_i + \frac{(1-\Delta_i)}{\alpha_i}\xi_i\right\}\hat{\chi}_i + \left\{1 - \Delta_i - \frac{(1-\Delta_i)}{\alpha_i}\xi_i\right\}\hat{\gamma}^{opt^T}H(T_i)\right] = 0, \tag{4.2.06}$$

where the above quantities are estimated as:

$$\hat{\chi}_i = \hat{I}_i^{-1}(\hat{\beta}_{LE})\int_0^{\tau}\{Z_i(u) - \mathbf{E_{tc}}(\mathbf{u}, \hat{\beta}_{\mathbf{LE}}))\}\{dN_i(u) - Y_i(u)\hat{\lambda}^c(u)du\}, \tag{4.2.07}$$

$$\hat{I}_i(\beta) = \sum_{i=1}^{n}\Delta_i\left[\frac{G_{2i} - G_{1i}G_{1i}^T}{G_{oi}}\right] \tag{4.2.08}$$

where $G_{2i} = exp(\hat{\beta}^T Z_j(t_i))Z_j(t_i)Z_j(t_i)^T$, $G_{1i} = exp(\hat{\beta}^T Z_j(t_i))Z_j(t_i)$, and $G_{oi} = exp(\hat{\beta}^T Z_j(t_i))$

$$\hat{\gamma}^{opt} = E\big[\hat{\zeta}_i H(T_i)H^T(T_i)\big]^{-1}E\big[H(T_i)\hat{\eta}_i\big], \tag{4.2.09}$$

$$\hat{\lambda}^c(u) = \frac{dN_i^c(u)}{Y_i(u)}, \tag{4.2.010}$$

$$\hat{\lambda}(u) = \frac{dN_i(u)}{Y_i(u)}, \tag{4.2.011}$$

$$\hat{\zeta}_i = -\left(1 - \frac{\xi_i}{\alpha_i} + \frac{\xi_i}{\alpha_i^2}\right)\left\{(1 - \hat{K}(T_i))^2 - \hat{K}^2(T_i)\int_0^{T_i} \frac{\hat{\lambda}^c(u)\hat{S}(u)}{\hat{K}(u)}du\right\}, \tag{4.2.012}$$

$$\begin{aligned}\hat{\eta}_i = \hat{K}(T_i)I_i^{-1}(\beta)\Bigg[&\left\{(1 - \hat{K}(T_i))\left(1 - \frac{\xi_i}{\alpha_i} + \frac{\xi_i}{\alpha_i^2}\right)\right. \\ &\left. -K(T_i)\left(1 - \frac{\xi_i}{\alpha_i}\right)\right\}\int_0^{T_i}(Z_i(u) - \mathbf{E_{tc}}(\mathbf{u}, \hat{\beta}_{\mathbf{LE}}))\hat{\lambda}^c(u)\hat{\lambda}(u)\hat{S}(u)du \\ &+\hat{K}(T_i)\left(1 - \frac{\xi_i}{\alpha_i} + \frac{\xi_i}{\alpha_i^2}\right)\int_0^{T_i}\frac{(Z_i(u) - \mathbf{E_{tc}}(\mathbf{u}, \hat{\beta}_{\mathbf{LE}}))}{\hat{K}(u)}\hat{\lambda}^c(u)\hat{\lambda}(u)\hat{S}(u)du\Bigg],\end{aligned} \tag{4.2.013}$$

and $\hat{S}(u)$ is the survival function estimated by the product limit estimator. This leads to an estimator for β that is consistent and asymptotically normal. The variance of $\hat{\beta}_{LE}$ can be estimated by

$$\text{var}(\hat{\beta}_{LE}) = \frac{\sum_{i=1}^{n}\hat{\Psi}_i^2}{n^2}. \tag{4.2.014}$$

4.3 SIMULATION STUDY

We adapt the algorithm described by Hendry (2013) to generate survival times that follow a Cox proportional hazards model with time-dependent covariates. To assess the large sample properties of our proposed locally efficient estimator (LEE), we conduct two simulation studies. In the first simulation, we generate a binary time-dependent covariate Z_1. Suppose that Z_1 represents whether a person had surgery or no surgery. This means that once an individual has the event (i.e. surgery), their event status remains constant for the rest of the study. Time to surgery, S_o, follow a uniform, U(0,20), distribution. The failure time T is generated from a piecewise exponential distribution, while the observed time U_i is obtained by using rejection sampling to randomly accept each value such that $TS_o \leq 20$. We set $Z_1 = 1$ if U_i is greater or equal to S_o, and zero otherwise to indicate whether there was a surgery or not before failure or censoring. We have 1000 observations. Table 4 presents a sample simulated data set that shows the data structure for the binary time-dependent covariate. Also, we generate continuous covariate $Z_2 \sim U(0, 1)$ independent of Z_1. Therefore, our model in the first simulation study contains two parameters: one binary time-dependent covariate Z_1 and a continuous covariate Z_2. The first row of Table 4 represents the time a subject enters the study. For example, subject 1 initially enters the study at time 0, and after the first interval, the binary-dependent covariate Z_1 is zero. This means that subject 1 does not have an event (i.e. surgery) at this time interval. However, on the second row the interval shows that subject 1 has an event, and once an event occurs, it stays the same for the rest of the follow-up. Also, notice that the covariate Z_2 remains the same at each time point since it is not time-dependent. On the contrary, subject 2 does not have an event in the first two intervals, but an event is recorded at the third interval. Once an

Table 4: Sample Simulated Data Set For Binary Time-dependent Covariate Z_1 and Continuous Covariate Z_2.

Subject	U_i	Δ_i	Z_1	Z_2
1	9.03	0	0	0.19
1	9.03	1	1	0.19
1	9.03	1	1	0.19
1	9.03	1	1	0.19
1	9.03	1	1	0.19
1	9.03	1	1	0.19
1	9.03	1	1	0.19
1	9.03	1	1	0.19
2	5.32	0	0	0.13
2	5.32	0	0	0.13
2	5.32	1	1	0.13
2	5.32	1	1	0.13
2	5.32	1	1	0.13
2	5.32	1	1	0.13

event is recorded for subject 2, it remains the same for the rest of the follow-up study. Similarly, the Z_2 covariate value for subject 2 remains constant over time since it is not time-dependent. In the second simulation setup, we generate a continuous time-dependent covariate $Z_3 \sim U(-0.5, 0.5)$ and a binary covariate Z_4 with $Pr(Z_4 = 1) = 0.5$. Likewise, Table 5 shows a sample simulated data set for the

Table 5: Sample Simulated Data Set For Continuous Time-dependent Covariate Z_3 and Binary Covariate Z_4.

Subject	U_i	Δ_i	Z_3	Z_4
1	8.65	1	-0.01	0
1	8.65	1	-0.08	0
1	8.65	1	-0.47	0
1	8.65	1	-0.30	0
1	8.65	1	0.05	0
1	8.65	1	0.01	0
1	8.65	1	-0.14	0
1	8.65	1	-0.17	0
1	8.65	1	0.16	0
1	8.65	1	-0.19	0
1	8.65	1	0.27	0
1	8.65	1	-0.24	0
2	5.18	1	-0.01	1
2	5.18	1	0.39	1
2	5.18	1	-0.20	1
2	5.18	1	-0.45	1
2	5.18	1	0.06	1
2	5.18	1	-0.46	1
2	5.18	1	0.12	1

continuous time-dependent covariate Z_3 and a binary covariate Z_4. From Table 5 we notice that the continuous time-dependent covariate Z_3 changes for every interval. Thus, our model in simulation study 2 contains two parameters: one continuous time-dependent covariate Z_3 and a binary covariate Z_4.

Results for simulation study 1 (Table 6) are presented for $exp(\beta_1) = 0.9$ and $exp(\beta_2) = 1.0$ corresponding to Z_1 and Z_2 respectively. We assume that Z_2 is observed at phase one, while Z_1 is only observed at phase two. In addition, we designate the vector value function of covariates, $H(T_i)$, as both Z_1 and Z_2. Simulation results are based on 500 replications. Table 6 presents the estimated coefficients and Monte Carlo mean-squared errors (MCSE) for the full cohort by running a Cox model with time-dependent covariates compared to our proposed estimator. From Table 6 we see that the full-cohort estimator reports a relative risk of death after surgery compared to the relative risk of death before surgery as 1.23. Similarly, the LE estimator reports a relative risk of death for the binary time-dependent covariate Z_1 as 1.27, with Monte Carlo mean-squared error 0.106. A similar trend is observed in Table 7 when we treat the time-dependent covariate as continuous. The full-cohort estimator shows a relative risk of 1.28 and MCSE of 0.158 for the continuous time-dependent covariate Z_3, while the LE estimator shows a relative risk of 1.36 and its corresponding MCSE as 0.162.

Also, in simulation 2 (Table 7) the binary covariate Z_4 reports MCSE of 0.094 when the time-dependent covariate Z_3 is treated as continuous compared to an MCSE of 0.150 when the time-dependent covariate Z_1 is treated as binary. Overall, the values are quite close considering the fact that our proposed estimator does not use the full-cohort unlike the Cox model.

Results for simulation 2 are presented for $\beta_3 = 1.3$ and $\beta_4 = 0.9$ corresponding to Z_3 and Z_4 respectively. In the case-cohort sampling design for this simulation setup,

Table 6: Simulation 1 Results Showing The Estimator, And Monte Carlo Standard Deviations (MCSE) For The Full-cohort And Our Proposed LE Estimator. True Relative Risks Are $exp(\beta_1) = 1.2$, and $exp(\beta_2) = 1.1$.

	Full-Cohort		LEE	
Variable	Estimate	MCSE	Estimate	MCSE
Z_1	1.23	0.095	1.17	0.101
Z_2	1.12	0.150	1.06	0.156

we assume that Z_4 is observed at phase one, while Z_3 is only observed at phase two. Also, the vector function of covariates, $H(T_i)$, are presented by Z_3 and Z_4. Table 7 shows the estimated coefficients and Monte Carlo mean-squared errors (MCSE) for the full data cohort compared with our proposed LE estimator.

Table 7: Simulation 2 Results Showing The Estimator, And Monte Carlo Standard Deviations (MCSE) For The Full-cohort And Our Proposed LE Estimator. True Relative Risks Are $exp(\beta_3) = 1.3$, and $exp(\beta_4) = 0.9$.

	Full-Cohort		LEE	
Variable	Estimate	MCSE	Estimate	MCSE
Z_3	1.28	0.158	1.32	0.161
Z_4	0.92	0.094	0.94	0.099

4.4 DISCUSSION

In this chapter, we show how to derive the most efficient estimator for analyzing a case-cohort design with time-dependent covariates. Our proposed estimator is built on the semiparametric efficiency theory, and restricts to the class of asymptotically linear estimators. We show that our estimator is consistent and asymptotically normally distributed. The proposed LE estimator contains quantities that are easy to calculate, and works well for both binary time-dependent and continuous time-dependent covariates.

The major difficulty with time-dependent covariates in a case cohort model framework is computing. At each event (i.e. surgery) time, we need to know the exact value of the covariate at that event time for all individuals at risk. Evidently, this complicates the management, collection, and storage of such data.

4.5 PUBLIC HEALTH SIGNIFICANCE

The usual cohort design is very expensive to conduct and also inefficient. As the solicitation of funding to conduct research becomes increasingly challenging, it is imperative for researchers to develop sophisticated but highly efficient methods to reduce the cost of research. This dissertation demonstrates the importance of using innovative methodology to reduce cost associated with research.

APPENDIX

DERIVATION OF γ^{OPT}

Define the filtration $\mathcal{F}(u)$ as the increasing sequence of sub$-\sigma-$algebras

$$\mathcal{F}(u) = \sigma\{N_i(s), N_i^c(s), Z_i, 0 \leq s \leq u, i = 1, \cdots, n\},$$

and $H_{1i}(\cdot)$ and $H_{2i}(\cdot)$ are predictable functions with respect to $\mathcal{F}(u)$. Unless otherwise stated, this is the filtration with respect to which all the martingales are defined in this dissertation. In order to derive the optimal gamma we use the following results which follow from theorem 2.4.4 in Fleming and Harrington (1991).

Under the assumption of independent censoring

$$\begin{aligned} E\Big[\int_0^{T_i} H_{1i}(u)dM_i(u) \times \int_0^{T_i} H_{2i}(u)dM_i^c(u)\Big] &= -E\Big[\int_0^{T_i} H_{1i}(u)H_{2i}(u)Y_i(u)\lambda(u)\lambda^c(u)du\Big] \\ &= -E\Big[\int_0^{T_i} H_{1i}(u)H_{2i}(u)\lambda^c(u)\lambda(u)S(u)du\Big], \end{aligned} \tag{A.0.01}$$

$$E\left[H_{1i}(u)\int_0^{T_i} H_{2i}(u)dM_i^c(u)\right] = E\left[H_{2i}(u)\int_0^{T_i} H_{1i}(u)dM_i(u)\right] = 0, \qquad \text{(A.0.02)}$$

and

$$E\left[\int_0^{T_i} H_{1i}(u)dM_i^c(u) \times \int_0^{T_i} H_{2i}(u)dM_i^c(u)\right] = E\left[\int_0^{T_i} H_{1i}(u)H_{2i}(u)\lambda^c(u)K(u)S(u)du\right]. \qquad \text{(A.0.03)}$$

From the discussion in Section 3.3, the optimal influence functions, or equivalently, the optimal γ is obtained by projecting $\mathcal{V}_i$ in Equation (3.3.05) on $\mathcal{U}$ defined by the elements $\mathcal{U}_i$ in Equation (4.1.03). By the projection theorem, the optimal gamma (γ^{opt}) must satisfy

$$\begin{aligned} E\Bigg[\Bigg[\Bigg\{K(T_i) - K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)} + \frac{\xi_i}{\alpha_i}\left[1 - K(T_i) + K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)}\right]\Bigg\} \\ \times I_i^{-1}(\beta)\int_0^{\tau}(Z_i - \mathcal{E}(u,\beta))dM_i(u) + \Bigg\{1 - K(T_i) + K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)} - \frac{\xi_i}{\alpha_i}\Bigg[1 - K(T_i) \\ + K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)}\Bigg]\Bigg\} \times \gamma^{opt^T}H(T_i)\Bigg] \times \Bigg[\Bigg\{1 - K(T_i) + K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)} \\ - \frac{\xi_i}{\alpha_i}\left[1 - K(T_i) + K(T_i)\int_0^{T_i}\frac{dM_i^c(u)}{K(u)}\right]\Bigg\} \times \gamma^T H(T_i)\Bigg]\Bigg] = 0 \end{aligned} \qquad \text{(A.0.04)}$$

Since Equation (A.0.04) has to be true for any γ, we set $\gamma = 0$ and solve Equation (A.0.04) for γ^{opt} using iterative conditional expectation. Consequently, we obtain

$$\gamma^{opt}E\Bigg[-\left(1-\frac{\xi_i}{\alpha_i}+\frac{\xi_i}{\alpha_i^2}\right)\left\{(1-K(T_i))^2-K^2(T_i)\int_0^{T_i}\frac{\lambda^c(u)S(u)}{K(u)}du\right\}\Bigg]$$
$$=E\Bigg[K(T_i)I_i^{-1}(\beta)\Bigg[\left\{(1-K(T_i))\left(1-\frac{\xi_i}{\alpha_i}+\frac{\xi_i}{\alpha_i^2}\right)\right.$$
$$\left.-K(T_i)\left(1-\frac{\xi_i}{\alpha_i}\right)\right\}\int_0^{\tau}(Z_i-\mathcal{E}(u,\beta))\lambda^c(u)\lambda(u)S(u)du$$
$$+K(T_i)\left(1-\frac{\xi_i}{\alpha_i}+\frac{\xi_i}{\alpha_i^2}\right)\int_0^{\tau}\frac{(Z_i-\mathcal{E}(u,\beta))}{K(u)}\lambda^c(u)\lambda(u)S(u)du\Bigg] \qquad \text{(A.0.05)}$$

Therefore, by solving Equation (A.0.05) for γ^{opt} we obtain

$$\gamma^{opt}=E\big[\zeta_i H(T_i)H^T(T_i)\big]^{-1}E\big[H(T_i)\eta_i\big] \qquad \text{(A.0.06)}$$

where

$$\zeta_i=-\left(1-\frac{\xi_i}{\alpha_i}+\frac{\xi_i}{\alpha_i^2}\right)\left\{(1-K(T_i))^2-K^2(T_i)\int_0^{T_i}\frac{\lambda^c(u)S(u)}{K(u)}du\right\}$$

and

$$\eta_i=K(T_i)I_i^{-1}\Bigg[\left\{(1-K(T_i))\left(1-\frac{\xi_i}{\alpha_i}+\frac{\xi_i}{\alpha_i^2}\right)\right.$$
$$\left.-K(T_i)\left(1-\frac{\xi_i}{\alpha_i}\right)\right\}\int_0^{\tau}(Z_i-\mathcal{E}(u,\beta))\lambda^c(u)\lambda(u)S(u)du$$
$$+K(T_i)\left(1-\frac{\xi_i}{\alpha_i}+\frac{\xi_i}{\alpha_i^2}\right)\int_0^{\tau}\frac{(Z_i-\mathcal{E}(u,\beta))}{K(u)}\lambda^c(u)\lambda(u)S(u)du\Bigg].$$

BIBLIOGRAPHY

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

[2] P. C. Austin. Generating survival times to simulate cox proportional hazards models with time-varying covariates. *Statistics In Medicine*, 31:3946–3958, 2012.

[3] W.E. Barlow. Robust variance estimation for the case-cohort design. *Biometrics*, 50:1062–1072, 1994.

[4] P. J. Bickel and J. A. Wellner. *Efficient and Adaptive Inference in Semiparametric Models*. Johns Hopkins University Press, Baltimore, 1993.

[5] O. Borgan and B. Langholz et al. Exposure stratified case-cohort designs. *Lifetime Data Analysis*, 6:39–58, 2000.

[6] O. Borgan and L. Goldstein et al. Methods for the analysis of sampled cohort data in the cox proportional hazards model. *The Annals of Statistics*, 23:1749–1778, 1995.

[7] N. E. Breslow. Improved horvitz-thompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology. *Stat Biosci*, 1:1–19, 2009.

[8] N. E. Breslow and J. A. Wellner et al. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression. *Scan J Stat*, 34:86–102, 2007.

[9] N. E. Breslow and R. Holubkov et al. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *J. R. Stat. Soc.*, 59:447–461, 1997.

[10] N. E. Breslow and T. Lumley et al. Using the whole cohort in the analysis of case-cohort data. *Am J Epidem*, 169:1–8, 2009.

[11] J. Cai and J. Fan et al. Marginal hazard models with varying-coefficient for multivariate failure time data. *Ann. Stats*, 35:324–354, 2007.

[12] Z. Cai and Y. Sun. Local linear estimation for time-dependent coefficients in cox's regression models. *Scand. J. Stats*, 30:93–111, 2003.

[13] H. Y. Chen. Double-semiparametric method for missing covariates in cox regression models. *J. Am Stat Assoc*, 97:565–576, 2002.

[14] H. Y. Chen and R. J. A. Little. Proportional hazards regression with missing covariates. *J. Am Stat Assoc*, 94:896–908, 1999.

[15] K. Chen and S. Lo. Case-cohort and case-control analysis with cox's model. *Biometrika*, 86:755–764, 1999.

[16] D. R. Cox. Regression models and life tables with discussion. *J. R. Stat. Soc*, 34:187–220, 1972.

[17] G. J. D'Angio and N. Breslow. Treatment of wilm's tumor:results of the third national wilm's tumor study. *Cancer*, 64:349–360, 1989.

[18] M. J. Van der Laan and J. M. Robins. *Unified Methods for Censoring Longitudinal Data and Causality*. Springer Series in Statistics, New York, 2003.

[19] Z. Donglin and D. Y. Lin. Semiparametric transformation models with random effects for recurrent events. *J Am Stat Assoc*, 102:167–180, 2007.

[20] B. Efron. The efficiency of cox's likelihood function for censored data. *J Am Stat Assoc*, 72:557–565, 1977.

[21] B. Nan, Emond, M. and J. A. Wellner. Information bounds for cox regression models with missing data. *The Annals of Statistics*, 32:723–753, 2004.

[22] L. D. Fisher and D. Y. Lin. Time-dependent covariates in the cox proportional-hazards regression model. *Annu. Rev. Public Health*, 20:145–157, 1999.

[23] T. R. Fleming and D. P. Harrington. *Counting Processes and Survival Analysis*. Wiley Series, New York, 1991.

[24] D. Gamerman. Markov chain monte carlo for dynamic generalized linear models. *Biometrika*, 85:215–227, 1991.

[25] R. D. Gill. *Censoring and Stochastic Integrals*. Mathematical Centre tracts No. 124, Amsterdam: Mathematisch Centrum, 1980.

[26] P. M. Grambsch and T. M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81:515–526, 1994.

[27] T. J. Hastie and R. J. Tibshirani. Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, 46:1005–1016, 1990.

[28] D. J. Hendry. Data generation for the cox proportional hazards model with time-dependent covariates: A method for medical researchers. *Statistics in Medicine*, 4:1–35, 12013.

[29] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc*, 47:663–685, 1952.

[30] L. Kong J. Cai and P. K. Sen. Weighted estimating equations for semiparametric transformation models with censored data from a case-cohort design. *Biometrika*, 91:305–319, 2004.

[31] S. Kim J. Cai and W. Lu. More efficient estimators for case-cohort studies. *Biometrika*, 100:695–708, 2013.

[32] J. D. Kalbfleisch and J. F. Lawless. Likelihood analysis of multi-state models for disease incidence and mortality. *Statistics in Medicine*, 7:149–160, 1988.

[33] J. D. Kalbfleisch and R. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, New Jersey, 2002.

[34] J. D. Kalbfleisch and R. L. Prentice. Analysis of failure times in the presence of competing risks. *Biometrics*, 34:541–554, 1978.

[35] J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, 2005.

[36] D. G. Kleinbaum and M. Klein. *Survival Analysis*. Springer, New York, 2005.

[37] L. Kong and J. Cai. Case-cohort analysis with accelerated failure time. *Biometrics*, 65:135–142, 2009.

[38] M. Kulich and D. Y. Lin. Improving the efficiency of relative-risk in case-cohort studies. *J Am Stat Assoc*, 99:832–844, 2004.

[39] M. Pugh, Lipsitz, S. and J. Robins. Proportional hazards model with missing covariate data. *Department of Biostatistics, Harvard School of Public Health*, 1:1–23, 1993.

[40] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley and Sons, Inc, New Jersey, 2002.

[41] D. G. Luenberger. *Optimization by Vector Space Methods*. John Wiley and Sons, Inc, New York, 1969.

[42] W. Lu M. Liu and R. E. Shore. Cox regression model with time-varying coefficients in nested case-control studies. *Biostatistics*, 11:693–706, 2010.

[43] S. D. Mark and H. A. Katki. Specifying and implementing nonparametric and semiparametric survival estimators in two-stage(nested)cohort studies with missing case data. *Journal of the American Statistical Association*, 101:460–471, 2006.

[44] L. Marzec and P. Marzec. On fitting cox's regression model with time-dependent coefficients. *Biometrika*, 84:901–908, 1997.

[45] E. C. Norton, Morris, C. N. and X. H. Zhou. *Parametric duration analysis of nursing home usage*. Wiley Series, New York, 1994.

[46] S. A. Murphy. Testing for a time dependent coefficient in cox's regression model. *Scand. j. Stats*, 20:35–50, 1993.

[47] S. A. Murphy and P. K. Sen. Time-dependent coefficients in a cox-type regression model. *Stochastic Process Application*, 39:153–180, 1991.

[48] B. Nan. Efficient estimation for case-cohort studies. *The Canadian Journal of Statistics*, 32:403–419, 2004.

[49] W. K. Newey. Semiparametric efficiency bounds. *J. Appl. Econometric*, 5:99–135, 1990.

[50] R. L. Prentice. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73:1–11, 1986.

[51] K. J. Rothman. *Epidemiology: An Introduction*. Oxford University Press, New York, 2002.

[52] J. M. Robins, Rotnitzky, A. and L.P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of The American Statistical Association*, 89:846–866, 1994.

[53] S. G. Self and R. L. Prentice. Asymptotic distribution theory and efficiency results for case-cohort studies. *The Annals of Statistics*, 16:64–81, 1988.

[54] T. H. Scheike T. Martinussen and I. M. Skovgaard. Efficient estimation of fixed and time-varying covariates effects in multiplicative intensity models. *Scand. J. Stats*, 29:57–74, 2000.

[55] D. Zucker, Tian, L. and L. J. Wei. On the cox model with time-varying regression coeffcients. *J. Am. Stats. Association*, 100:172–183, 2005.

[56] A. A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer, New York, 2010.

[57] A. S. Wahed and A. A. Tsiatis. Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomized designs in clinical trials. *Biometrics*, 60:124–133, 2004.

[58] A. S. Wahed and A. A. Tsiatis. Semiparametric efficient estimation of survival distributions in two-stage randomized designs in clinical trials with censored data. *Biometrika*, 93(1):167–177, 2006.

[59] C. Y. Wang and H. Y. Chen. Augmented inverse probability weighted estimator for cox missing covariate regression. *Biometrics*, 57:414–419, 2001.

[60] H. Zhao and A. A. Tsiatis. A consistent estimator for the distribution of quality adjusted survival time. *Biometrika*, 84(2):339–348, 1997.

[61] D. M. Zucker and A. F. Karr. Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *Ann. Stats*, 18:329–353, 1990.