A Z-estimation System for Two-phase Sampling with Applications to Additive Hazards Models and Epidemiologic Studies

Jie Hu

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

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Norman E. Breslow, Chair

Kwun C. Chan, Chair

Ying Q. Chen

Program Authorized to Offer Degree:
UW Biostatistics

University of Washington

Abstract

A Z-estimation System for Two-phase Sampling with
Applications to Additive Hazards Models and Epidemiologic Studies

Jie Hu

Co-Chairs of the Supervisory Committee:
Professor Norman E. Breslow
Biostatistics

Professor Kwun C. Chan
Biostatistics

An observational epidemiologic study often follows a large amount of participants for occurrence of diseases. If every covariate is measured for every participant, then the study can be highly expensive. Two-phase sampling reduces costs by oversampling more informative subjects from a large phase I sample into a small phase II subsample; only subjects in the subsample are measured for the expensive covariates. Analyzing this type of data is challenging, particularly for association study and risk prediction based on a semiparametric model. It requires new theoretical tools, methods, software and data analysis examples. This dissertation answers these timely challenges. We provide statisticians with a new theory to develop new tools for two-phase studies. This theory is general. It is not specific to a particular model or a two-phase study design. It can be used for association study via estimating regression parameters or for risk prediction via estimating the entire model, including both parametric and nonparametric parts of a model. It encompasses both likelihood and non-likelihood based inference. It provides correct inference in the presence or absence of model misspecification. Because a broad problem area is taken into account by this theory, the theory can be also considered as a framework to guide a researcher through a model development process for a two-phase study.

Next, we use our theoretical results to develop a semiparametric additive hazards model for general two-phase designs. We are able to obtain a collection of results systematically. These results include estimators for regression parameters, cumulative baseline hazards, and individual specific cumulative hazards from random sampling, two-phase sampling, two-phase sampling incorporating auxiliary information embedded in the phase I sample, as well as these estimators' model-based and robust asymptotic variances. Lastly, we apply our analyzing tools to an Atherosclerosis Risk In Community (ARIC) case-cohort study, we are able to use the biomarker information to create a new risk prediction function of coronary heart disease even though these biomarker information is only available for a selected sample. The individual risk profile calculated from this function can help physicians identify new patients who may not be discovered by traditional risk evaluation tools for prevention therapies. We then further improve prediction precision by incorporating additional information on standard risk factors in the main cohort. With these new tools for two-phase designs implemented in our software, researchers can use a new and expensive biomarker for risk prediction with substantially reduced costs.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am deeply grateful to many people who have helped me, encouraged me and supported me in graduate school. First and foremost, I would like to express my gratitude to Prof. Norman E. Breslow, who showed me how to find research questions, conduct research and write research results thoughtfully, who was always responsible and responsive as my teacher and my co-advisor despite circumstances in the past six years, and whose curiosity, courage and dedication influenced me. I would also like to express my sincere appreciation to my co-advisor, Prof. Kwun C. Chan, whose constructive advice, insight, and knowledge helped me whenever I ran into difficulties.

I wish to thank other members on my supervisory committee as well: Prof. Jon A. Wellner, for your attention to detail and insightful guidance; Prof. Ying Q. Chen, for your unceasing support and willingness to share with me your experiences as a researcher; Prof. Peter B. Gilbert, for your encouragement and approval during exams; and Prof. Loveday L. Conquest, my Graduate School Representative.

I wish to thank Sarah Chodakewitz from the Odegaard Writing and Research Center at the University of Washington. Discussions with Sarah made the writing process creative and fun. I wish to thank my fellow students in both Boston and Seattle, in particular, Victoria Ding, Laura Yee, Marlena Maziarz, Jessie Hsu, Elizabeth Ogburn, Brett Hanscom, and Shanshan Zhao. Their company and support kept me moving forward and enjoying graduate school. I wish to thank Prof. Andrea Rotnitzky. Without her passionate measure theory class and her encouragement, my research career would not have taken off, and I would not have chosen a theoretical dissertation direction. Finally, I wish to express my deep gratitude to my parents for their belief in me, patience and support.

DEDICATION

to my dear parents, Wei Hu and Qian Zhou

Chapter 1

INTRODUCTION

Epidemiologic cohort studies often follow a large amount of participants for a number of years for occurrence of diseases. If every covariate is measured for every participant, then the study can be highly expensive, especially when a disease is rare, the measurement of a covariate is labor-intensive or the measurement requires a new technology instrument. Two-phase sampling reduces costs substantially by arranging sampling in two stages so that expensive covariates are only assembled for a selected Phase II small sample, which is obtained by oversampling more informative subjects in the initial Phase I large cohort. However, a theoretical foundation and general methodology for analyzing this type of data are underdeveloped. My dissertation answers this timely challenge.

For example, the Atherosclerosis Risk In Community (ARIC) study as a prospective cohort study followed approximately 16,000 middle aged men and women in four US communities to investigate the etiology and natural history of atherosclerosis. In one substudy, Ballantyne et al. [2004] was interested in investigating whether elevated levels of high sensitivity C-reactive protein (hs-CRP) was associated with the increased risk of Coronary Heart Disease (CHD), particularly among patients without elevated low density lipoprotein cholesterol (LDL-C). The motivation behind this investigation is that many CHD events occur among people without elevated LDL-C. Risk evaluation based on traditional risk factors can not identify patients with low LDL-C but at a high risk of CHD. It is believed hs-CRP measurements may identify these patients for subsequent preventive therapies. However, the hs-CRP assay was a newly developed biochemical technology, which is usually expensive. Performing hs-CRP assays on plasma samples from every participant in such a large cohort was prohibitive. To save costs, two-phase sampling was implemented. hs-CRPs were only

measured for 608 subjects who subsequently developed CHD or stroke and a cohort random sample (CRS) of 785 subjects. To analyze this data, Ballantyne et al. [2004] considered fitting the semiparametric Cox's regression model due to its simplicity and flexibility. The parametric part of the semiparametric model offers a simple summary on association between hs-CRP levels and CHD events, while the nonparametric part of the model allows the baseline hazard function to be flexible.

Ballantyne et al. [2004]'s relatively new study designs, combined with their interest in semi-parametric models and using hs-CRPs to refine risk assessment, pose many new and practical questions beyond their association analysis. These include:

(1) In order to use hs-CRPs for individual risk prediction, all the parameters in the semiparametric model become our interest. This will involve joint estimation of the coefficients of increased CHD risk associated with each risk factor as well as the baseline hazards curve. No such risk prediction has been applied to an actual two-phase data.

(2) Cox's proportional hazards model assumes a multiplicative effect of hs-CRP levels on the increased risk of CHD. However, Kang et al. [2013] demonstrated that increased levels of hs-CRP acted additively rather than proportionally on the baseline risk of CHD. No statistical software has been developed for fitting the simplest additive hazards model— Lin and Ying's additive hazards (AH) model —to two-phase sampling data for applied statisticians to use, for either association analysis or risk prediction.

(3) When the assumed model, for instance, Cox's regression model in Ballantyne et al. [2004], does not correctly specify the true underlying distribution for observed data, how do we describe the target of a proposed estimator and this estimator's precision in a two-phase sampling scenario?

(4) If we only use the selected small sample of 1373 subjects [Ballantyne et al., 2004] in ARIC for association analysis or building risk prediction function, a wealth of information in auxiliary variables embedded in the main cohort is ignored, such as traditional

risk factors of CHD and demographic variables that were measured for every participant in the main cohort of approximately 16,000 subjects. How can we incorporate these pieces of information into our estimation procedure to further improve the precision of coefficients and individual risk estimates?

(5) Although a case-cohort design was implemented in Ballantyne et al. [2004], hs-CRP measurements were missing among a fraction of CHD cases and these CHD cases were removed from the sample for association analysis. Ballantyne et al. [2004] did not take into account the problem of sampling among cases and used the analyzing method [Barlow, 1994, Barlow et al., 1999] that was proposed for case-cohort designs, which assumed all the cases are selected. By contrast, Cox's regression model developed from weighted likelihood methods for two-phase stratified samples [Breslow and Wellner, 2007] allows more general two-phase designs including the complex sampling scenario in Ballantyne et al. [2004]. However, if we used the weighted likelihood approach to develop the AH model for two-phase designs, then the score equations for regression parameters and cumulative hazards would not be as simple as Cox's regression model. Developing weighted likelihood estimators for the AH model would be complex and difficult. For random sampling, AH models are usually fitted by methods of estimating equation rather than maximum likelihood estimation. The asymptotic behaviors of resulting estimators are then studied by martingale central limit theorem. However, since Ballantyne et al. [2004]'s study selected their sample for association analysis based on outcomes, one crucial condition for applying martingale theory does not hold. Hence new theoretical tools are desired for developing the AH model with two-phase sampling.

From the ARIC study example we see that utilizing two-phase sampling for epidemiologic studies requires new tools from every aspect of statistics. Not only does this real-life problem demand extensions of numerous statistical methods used for random sampling to two-phase sampling, but also alternative theoretical tools that are as general as weighted likelihood methods [Breslow and Wellner, 2007] for methods development and extension. In addition,

in order to effectively assist epidemiologists to use the cost-effect two-phase sampling designs, software development is in great need and data analysis examples that illustrate the method of analyzing this type of data will be useful.

Therefore, this dissertation aims to provide a new collection of theoretical results, a series of statistical tools, new software, and a data analysis example for general two-phase sampling problems. This dissertation aims to offer solutions to the five aforementioned questions that arise from implementing a two-phase sampling design to an ARIC study but not limited to ARIC. These solutions are designed as simple as possible.

In Chapter 2, I build a Z-estimation system for semiparametric inference with two-phase sampling. This system is built upon a Z-estimation theorem in van der Vaart [1998] for infinite-dimensional parameters and modern empirical process theory from van der Vaart and Wellner [1996].

The data analysis methods in this system are easy to implement because I draw simple techniques from surveys, which are commonly used across disciplines. I connect problems from two-phase epidemiologic studies to survey sampling. I adapt inverse probability weighted estimating equation (IPW-EE) method Binder [1983] and the calibration technique Deville and Särndal [1992] created for analyzing complex surveys to epidemiologic studies. Since complex surveys are widely used in broad areas of research such as psychology, political science, and economics, tools from surveys are often straightforward, general, and easy to follow. Therefore, my tools inherit these advantages in the ease of implementation.

The theoretical results in this system are general and can be used for a variety of scenarios. Using this system, I can estimate both Euclidean and infinite-dimensional parameters, and these two types of parameters jointly. In the context of epidemiology, it means I can use semiparametric models to study risk factors associated with a disease as well as build a risk prediction function to predict future incidence of a disease for each individual based on his/her measurements of risk factors. This system encompasses both likelihood [Breslow and Wellner, 2007] and non-likelihood based inference. In addition, it allows for model misspecification. It is not restricted to a particular model. It can be applied for various

two-phase sampling designs, not only for the most familiar case-cohort study.

The development of my analysis methods is systematic. I am able to make it systematic because 1) Z-estimation is applicable to solving problems across different circumstances; 2) the Z-estimation theorem due to [Huber, 1967, Pollard, 1985, van der Vaart, 1998], on which my theoretical tools are based, has a systematic asymptotic analysis of estimators— a large asymptotic study is decomposed to small components, and these components are connected to each other working together as a complex whole; 3) the preservation theorems [van der Vaart and Wellner, 2000, 1996] in modern empirical results allow existing results in some of these components to be easily extended to complex scenarios; 4) I connect estimation problems in two-phase sampling to Z-estimation. Then I adapt Z-estimation to new two-phase sampling problems for the development of new theoretical tools. As a result, I am able to use this Z-estimation system to obtain a collection of results systematically, rapidly, and ready to use for two-phase sampling.

Although this system is general, easy to implement, and systematic, it does bear some limitations. I limit myself to the estimating equations that are in the form of or could be transformed to an average of i.i.d. functions; I only consider the scenarios that both types of parameters have root-N convergence rates; I assume data are collected by two-phase i.i.d. variable probability sampling (Bernoulli sampling) [Lawless et al., 1999]. However, by these restrictions, I am able to simplify the application of modern empirical process theory. I sacrifice some generality in return for accessibility .

In Chapters 3 & 4, I apply my Z-estimation system to Lin & Ying (1994)'s additive hazards (AH) model. Due to the systematic nature and generalizability of this system, I am able to develop nine estimators based on the AH model and their model based and robust standard errors. On the other hand, Chapters 3 & 4 are examples to illustrate the theoretical framework and results from Chapter 2.

Additive hazards model is an important tool to study survival outcomes, and particularly useful if epidemiologists would like to study the risk difference attributable to an exposure. The attributable risk tells the amount of absolute risk that could be reduced if an exposure

were removed from the population. This interpretation makes a statistical result from the AH model well communicated among public health practitioners for policy making. However, due to the shortage of software for fitting Lin & Ying's additive hazards (AH) model and the lack of knowledge of it in the applied statistics research community, this important model was seldom used. Therefore, I also implement my results for the AH model in computer software R, so that this model can be easily implemented for scientific and clinical investigations. With these tools available in R, public health research community may start to recognize the importance and usefulness of this model.

In Chapter 5, I conduct simulation studies to validate numerically various results I derived in Chapters 3 & 4 for the AH model with two-phase sampling. I also demonstrate the two-phase sampling and the calibration technique's ability to substantially improve estimation efficiency compared to random sampling by simulations in various settings. Then I study the performance of these methods under model misspecification.

In Chapter 6, I apply the AH model I developed in Chapters 3 & 4 for two-phase sampling to re-analyze the ARIC study I introduced at the beginning of this chapter. With these new tools, I demonstrate I am able to use additional information in the whole cohort to improve both association analysis and risk prediction precision. I am able to use the ARIC case-cohort data to make risk profiles for individuals with different levels of hs-CRPs at a reduced cost—which has never been done before. These new risk profiles including the new biomarker information may help physicians to discover new patients at high risk of the coronary heart disease. These new patients, who may be missed by traditional tools of risk evaluations but now can be identified, will benefit from therapies in the prevention of coronary heart diseases. The new risk prediction tool is not restricted to a particular disease such as the coronary heart disease or a particular two-phase study such as ARIC.

Chapter 2

A Z-ESTIMATION SYSTEM FOR TWO-PHASE SAMPLING DATA

Two-phase sampling reduces costs substantially in prospective epidemiological studies. It uses outcomes and/or inexpensive covariates obtained for each individual to determine their following sub-sampling probabilities, so that certain expensive and difficult to be obtained covariates are only measured for the most informative individuals in the subsamples. Epidemiology usually concerns with the incidents and determinants of diseases. To study them, previously parametric model and now semiparametric models are often used for analyzing data collected by random sampling. However, when the data are from a two-phase design, how do we fit these models? Both general methods and the theory for developing these methods are lacking. Thus, in this chapter, we provide statisticians a new theory to develop new tools for two-phase study. This theory is very general. It is not specific to a particular model or a two-phase study design. It can be used for association study via estimating regression parameters or for risk prediction via estimating the entire model, including both parametric and nonparametric parts of a model. It encompasses both likelihood and non-likelihood based inference. It provides correct inference in the presence or absence of model misspecification. Because a broad problem area is taken into account by this theory, the theory can be also considered as a framework to guide a researcher through a model development process for a two-phase study.

2.1 Background, Motivation, Contribution

This dissertation carefully selects and consolidates results from survey sampling, Z-estimation and the modern empirical process theory, so that it provides a *systematic* and *easy-to-access* theoretical foundation for analyzing two-phase studies.

From Survey Sampling

Neyman [1938] proposed two-phase sampling, also called double sampling, for field surveys. A field survey is usually conducted at the local level to estimate an average of a population quantity, for instance, the household food expenditure, and it often requires interviews by experienced surveyors. The variable collected through a field survey is usually expensive. Thus, within budget the sample size can be small for obtaining an estimate of a desired accuracy. Neyman [1938]'s idea is to use a second inexpensive variable that is correlated with the expensive variable to select the subjects for expensive variable measurements. In the first phase, a relatively large random sample is collected to ascertain the inexpensive variable. In the second phase, based on the phase I variable, the sample is stratified and subsamples are randomly drawn from each stratum for the expensive variable of interest. Because the variation of expensive variable within each stratum is less than it is for the whole population, a more accurate estimate will be produced. This two-phase design was later further generalized by Horvitz and Thompson [1952] through the use of unequal probabilities to select a sample.

Similar ideas on the two-phase design appeared in epidemiology separately. Walker [1982] and White [1982] first considered it for study of association between dichotomized diseases and exposures. Association analysis often requires collecting covariates to adjust for confounders and effect modifiers. To reduce costs in the collection of these covariates, both authors proposed to divide the sample into four groups based on the dichotomized diseases and exposures first and then use unequal sampling fractions for different exposure-disease

categories to select subsamples, so that smaller categories will be oversampled for covariate measurements. Their intuition were that additional observations from a smaller group should add more information than additional observations from a larger group. Driven by the same cost concern, Prentice [1986] proposed a case-cohort design for follow-up studies, in which covariate histories are only assembled for a random subcohort and all cases. Since then various two-phase designs were proposed such as exposure stratified case-cohort designs [Borgan et al., 2000], generalized case-cohort designs that do not require sampling all the cases [Cai and Zeng, 2004, Kang and Cai, 2009], and cohort sampling that allows sub-sampling for both cases and controls [Gray, 2009]. Each of these designs was proposed for a specific sampling situation with a particular data analysis method, but each can be also considered as a special case of the two-phase survey sampling we described at the end of the last section. Connecting two-phase epidemiologic designs to the original two-phase survey sampling [Breslow and Wellner, 2008] provides us a systematic outlook on these epidemiologic designs — they all involve drawing random subsamples for expensive variables measurements within each stratum from a stratified large random sample like a two-phase survey.

This connection of two-phase epidemiologic studies to surveys suggests that analyzing methods for complex surveys could be adapted to many two-phase epidemiology studies and a common analysis approach will exist.

Surveys are primarily designed to estimate the mean of some quantity from a finite population. The standard analyzing technique is inverse probability weighting [Horvitz and Thompson, 1952]: the finite population quantity was estimated by weighting each observation with its inverse probability of inclusion in the sample. Later regression models started to be used for surveys. How to make inference about model parameters from surveys became a new issue, especially when a survey was multi-staged and stratified. Binder [1983] considered to define this type of parameters as an implicit function on a finite population. This implicit function was motivated but not defined by the model. When the model was misspecified, Binder's parameters were still well-defined. Subsequently, Binder [1992] proposed a weighted

estimating equations method to estimate these parameters when a sample was drawn with a complex design. The resulting estimators were still consistent when the assumed model was wrong. In two-phase surveys, a large amount of information on auxiliary variables is often obtained at the initial phase. Survey statisticians have developed various methods to use these variables for efficiency improvement. Calibration [Deville and Särndal, 1992], for instance, provides a systematic approach to take into account auxiliary information by adjusting the sampling weights. In both survey theory and practice, calibration has established itself as an important instrument because it is transparent, easy to understand and not specific to a specialized situation [Särndal, 2007].

I am interested in using all these techniques created for complex surveys for analyzing two-phase epidemiologic studies. From a classical survey perspective, the population is fixed and finite; the parameter is an explicit or implicit function of this finite population. The modern population-based studies such as epidemiologic studies, however, adopt a superpopulation perspective. The phase I population is considered to be a random sample from a superpopulation defined by a probability distribution; the parameter is an explicit or implicit function of this probability distribution instead. Therefore, adaptation of inference methods for the finite population to the superposition is required. Godambe and Thompson [1986] related superpopulation model parameters to survey population parameters. Lin [2000] extended Binder's method of weighted estimating equation for the finite population to the superpopulation and claimed the robustness of Binder's estimators remained for the estimators obtained in the superpopulation inference. Breslow et al. [2009a,b] adapted calibration techniques to further improving precision of semiparametric inference, which is often desired for epidemiologic studies, by adjusting the weights in the inverse probability weighted likelihood equations. They used these new tools to analyze dataset obtained or simulated by two-phase designs and found the precision of estimators was dramatically improved.

Inspired by these works, in this dissertation I will follow Binder's definition of parameters for survey populations and will consider my parameters as an implicit function on probability models. The function is motivated by an assumed model but does not reply on the correct-

ness of the model. This construction is in line with Newey [1994]'s treatment of a parameter as a functional allowing for general misspecification. I will adopt the method of weighted estimating equation as my inference method due to its robustness and ease of implementation for complex sampling designs. In the presence of auxiliary variables, I will incorporate calibration into my inference procedure for further efficiency improvement due to its ease of implementation and transparency. In summary, by connecting to statistical methods for surveys I find a general approach to analyze the two-phase sampling epidemiologic data.

From Z-estimation and modern empirical process theory

To fully develop this general approach for analyzing two-phase sampling studies, I also need theoretical tools to study the asymptotic properties of my estimators obtained from weighted estimating equations and calibration methods. I choose to build my theoretical tools upon a Z-estimation theorem from van der Vaart [1998] and results from modern empirical process theory. I make this choice because semiparametric inference with two-phase sampling data creates new problems for which the traditional theoretical tools can not solve. It is also because the Z-estimation theorem from van der Vaart [1998] is an extension of Huber [1967]'s Z-estimation, and therefore it preserves the robustness and systematic approach of Huber's Z-estimation.

Z-estimation, together with modern empirical process results, can solve many new problems traditional tools can not solve. Because epidemiologic studies often collect censored survival data, semiparametric models are frequently used for data analysis. The standard way to make semiparametric inference for censored survival data with random sampling is to use martingale theory. To show asymptotic normality of estimators, this approach formulates the model in the framework of counting processes and then applies the martingale central limit theorem to a martingale integral [Andersen and Gill, 1982]. However, when the model is misspecified, the martingale integral does not exist. When outcome based sampling occurs, weights, which adjust the bias introduced by the sampling, depend on a complete history of the study. One crucial condition for the martingale central limit theorem that the integrand of the martingale integral needs to be predictable is violated. Therefore, martingale theory does not apply. Self and Prentice [1988] and Lin [2000] considered the Taylor expansion technique for proportional hazards models with case-cohort sampling. However, an important tightness condition was left without further exploration or treated as an assumption because tightness was very difficult to prove algebraically. Z-estimation combined with modern empirical process theory, on the other hand, can solve the problem of semiparametric inference with outcome-based sampling.

In addition, these two theoretical tools enable us to develop risk prediction function based on semiparametric models. In epidemiological and clinical studies, model based prediction of an individual's outcome is of great interest according to the measurements of this individual's risk factors and the time since these measurements. If this model is semiparametric, then for prediction purposes the entire model needs to be estimated. This means joint estimation of both Euclidean (finite-dimension) and non-Euclidean (infinite-dimension) parameters are required (see Bickel et al. [1998, p.1-2] and van der Vaart [1998, p.358] for definitions). Estimating these two types of estimators jointly has not been seen when martingale theory or Taylor expansion techniques were used for developing estimators.

Z-estimation is an estimation method for Z-estimators, which are defined as solutions to estimating equations [van der Vaart and Wellner, 1996]. Before van der Vaart and Wellner [1996], they were called M-estimators. An M-estimator refers to an estimator obtained by maximizing a criterion functions and the "M" means maximum-likelihood-type [Huber, 1981]. In many situations, finding an M-estimator $\hat{\alpha}$ for a parameter α is sought by solving a zero valued estimating equation:

$$\Psi_N(\alpha) = \frac{1}{N} \sum \psi_\alpha(x_i) = 0 \tag{2.1}$$

where ψ_α is the derivative of a criterion function with respect to α. In some cases, solving an equation for an estimator does not correspond to a maximization problem. To distinguish such estimator from the estimator obtained by maximizing a criterion function, the name Z-estimators was invented in van der Vaart and Wellner [1996, chap. 3.2], where the "Z" means zero. However, the name of M-estimators has been widespread in literature.

Huber's M-estimators and therefore Z-estimators were originally motivated by robustness. In Huber [1964], robustness meant insensitivity to small deviation from assumptions. In 1967, Huber proved consistency and asymptotic normality of maximum likelihood (ML) estimators without assuming that the true distribution underlying the observations belonged to the parametric family defining ML estimators. The meaning of robustness was generalized with respect to all misspecified models thereof. In 1986, Royall [1986] showed the asymptotic

variance in Huber [1967] could be easily modified so that the asymptotic variance was also consistently estimated under model misspecification. This paper completed the meaning of " robustness" for today's use: validity of estimators under model misspecification in term of their consistency and limiting distributions.

Suppose the expectation of $\Psi_N(\alpha)$ in (2.1) is $\Psi(\alpha)$. The old classic approach studied asymptotics of Z-estimators by the Taylor expansion on $\Psi_N(\alpha)$. In contrast, Huber's 1967 landmark paper proposed a new approach for proof by separating the contribution to $\Psi_N(\alpha)$ into a deterministic part $\Psi(\alpha)$ and a stochastic remainder (see Pollard [1984, section VII] and Pollard [1985] for details). As a result, the conditions for asymptotics were decomposed into two components: analytical conditions on Ψ and stochastic conditions on the deviation of Ψ from Ψ_N. By using this systematic approach, Huber was able to relax conditions on the second and higher order derivatives of likelihood function for ML estimators. This new approach was not widely appreciated until Pollard [1984, 1985] made the connection of Huber's work with empirical process theory, which provides tools to check the stochastic conditions.

Modern empirical process theory studies empirical process indexed by classes of sets or functions. It extends classical results for empirical distributions to multidimensional space and abstract infinite-dimensional space. The connection of these results to asymptotic statistics was well phrased in Pollard [1989]: "much asymptotic effort has been devoted to bounding error terms in Taylor expansions; empirical process theory provides some effective new tools for doing this", in short, "asymptotics via empirical processes". Due to the rapid developments in abstract empirical process theory in 1970s and 1980s, stochastic conditions in Z-estimation can be established even when the estimating equation becomes infinite-dimensional and abstract. This extension is very useful for semiparametric inference since the non-Eulidean parameters range over an infinite-dimensional space.

Later Huber [1967] and Pollard [1985]'s Z-estimation for finite-dimensional Z-estimators was generalized to infinite-dimensional Z-estimators [van der Vaart, 1995], also seen in [van der Vaart and Wellner, 1996, theorem 3.3.1]. The Z-estimators they considered were

very general. In the case of i.i.d. observations, the stochastic process in the stochastic condition in their theorems became an empirical process indexed with a class of functions. As a result, this condition was simplified to lemma 3.3.5 in van der Vaart and Wellner [1996], leading to theorem 19.26 in van der Vaart [1998], which is the Z-theorem we choose to rely on. In summary, Z-estimation was created in 1960s by Huber, recognized in 1980s, generalized in 1990s, and then simplified and ready for our use.

In this dissertation, I generalize a consistency theorem for finite-dimensional Z-estimators [van der Vaart, 1998, theorem 5.9] to infinite-dimensional Z-estimators and create a new consistency theorem. Then I simplify the conditions required in the new theorem for the special case of i.i.d. observations. Next I extend my consistency theorem and van der Vaart [1998]'s Z-theorem 19.26 with random sampling to the complex two-phase sampling scenario and the even more complex setting with use of auxiliary variables. As a result, I create a collection of new theoretical tools for semiparametric inference with general two-phase epidemiologic studies.

Systematic

My theoretical results combined with van der Vaart [1998] are referred to Z-estimation system throughout this dissertation. I use this single system to develop various Z-estimators systematically.

This system covers various estimators that were considered separately. In the past, semiparametric inference with random sampling and two-phase sampling was studied separately due to the failure of martingale theory in the later scenario. Semiparametric inference in the presence or absence of model-misspecification were also considered independently. For example, different approaches were used to make inference for Cox's proportional hazards model when this model was assumed to be the true model of observations [Andersen and Gill, 1982] and when this assumption was relaxed [Lin and Wei, 1989]. Now we can use Z-estimation alone to develop methods for all of these scenarios. Since Z-estimation for infinite-dimensional parameters also covers the classical Z-estimators on a finite-dimensional space, it can be used for both parametric and semiparametric models developments. Be-

fore, a two-phase sampling estimator and an improved two-phase sampling estimator by incorporating Phase I cohort information were developed separately. For example, Breslow and Wellner [2007, 2008] considered improving weighted likelihood estimators for two-phase sampling designs by estimation of weights through a parametric model. For this proposed method, Breslow and Wellner [2008] first estimated the regression parameters in a prediction model for weights, and then used a new theoretical tool to study the asymptotics of improved estimators. For my methods development, I instead use calibration to incorporate additional information from Phase I to improve two-phase sampling estimators as [Breslow et al., 2009a]. Since the calibration technique incorporates the cohort information by equating estimated totals of auxiliary variables from phase II subsamples to the known totals from Phase I cohort through weights adjustment, incorporating calibration into the estimating procedure only requires adding another set of weighted estimating equations to the original equations for estimators. Therefore, estimators obtained after weights calibration are still Z-estimators and Z-estimation applies. Because the Z-estimation system allows infinite-dimensional parameters, I am able to jointly estimate both Euclidean and non-Eulidean parameters in a semiparametric model as pointed out in Bickel et al. [1998]. I will show this capacity remains for the two-phase designs. The same joint estimation approach is also seen in a recent paper by Breslow and Lumley [2013]. The benefit of joint estimation is that I can conduct association analysis and risk prediction based on a semiparametric model at the same time by Z-estimation since the inference on regression coefficients is a byproduct of joint estimation of all parameters.

Not only can the Z-estimation system study asymptotic behaviors of various estimators, but also study these estimators systematically and quickly. We do not need to derive the limiting distributions of these estimators separately if they are motivated from the same model. I discovered this systematic approach during my development of the AH model for two-phase designs. For this model, I was interested in estimating both regression parameters and the prediction function of cumulative hazards from two-phase sampling data, with or without using auxiliary information embedded in the Phase I sample. I first considered

estimation of regression parameters and cumulative hazards separately, due to my concern that the second problem was a much harder problem than the first. I also considered different techniques for estimation with and without calibration. However, in the process of developing these estimators, I found that a lot of steps were shared among the derivation of these estimators, and results for one estimator were easily extended to others. I also realized the whole estimation procedure could be restructured and results could be integrated with one and another, so that a collection of estimators would be obtained systematically and almost at one time.

This systematic approach of studying estimators' asymptotic behaviors was facilitated by the systematic nature of the Z-estimation theorem [van der Vaart, 1998, 19.26], on which my theoretical tools rely, and the preservation theorems of Glivenko-Cantelli and Donsker classes [van der Vaart and Wellner, 1996, 2000]. The Z-estimation theorem decomposes a relatively large study on asymptotics to different small components. The preservation theorems make the results in the components of stochastic conditions easy to extend from a simple scenario to more complicated settings. As a result, the procedure to develop a collection of various Z-estimators becomes systematic and easy to follow, which will be revealed in Chapter 2 $\sim$ 4.

Modern empirical process theory has started to be commonly used for semiparametric inference. However, the convenience of this theory for a systematic development of a collection of estimators has not been explored. As current information age demands a large amount of statistical results in a short time, a systematic approach to develop statistics may become one trend of the future. My results and examples in this dissertation confirm the belief in Wellner [1992] that modern empirical processes may shorten the lag time between creation of statistics and the establishment of their asymptotics from a new perspective of systematic development of statistics.

Ease of access

I developed my theoretical results as simply as possible through assuming two-phase i.i.d. variable probability sampling [Lawless et al., 1999], time-invariant weights, and use of the simplest Z-theorem for i.i.d. data. I also find applying modern empirical process theory can be difficult for researchers who are interested in but have just started to use it. It is my intention to explain my application of this theory in great detail and at the same time simplify the application for those researchers. With this purpose, I deliberately chose simple methods and theoretical tools to create my Z-estimation system for two-phase studies. Ease of access is prioritized over seeking the optimal solutions. In return for accessibility I sacrifice some generalizability.

I assume i.i.d. variable probability sampling (VPS), also called Bernoulli sampling, to construct the phase II subsample. In VPS, the sample is drawn by independent Bernoulli trials for all subjects within each phase I stratum, thus members in selected sample are independent to each other and the sample size is random. Due to the independence of observations, VPS makes statistical methods development straightforward and simple. In contrast, sampling without replacement, which is often used in practice, draws a fixed size of phase II subsamples within each stratum, so observations are correlated and thus more difficult to analyze. Theoretical tools for two-phase sampling under sampling without replacement were developed in Breslow and Wellner [2007] and Saegusa and Wellner [2013]. In VPS, the inclusion probability of each Bernoulli trial is a function of variables, thus VPS can also be used in the situation when these variables are continuous; sampling without replacement, however, is limited to the variables that are discrete as in stratified sampling. Nevertheless, variation in the sizes makes the estimators obtained by two-phase VPS slightly less efficient. The amount of this efficiency loss is small as shown in the simulations studies by Kulich [1997]. In addition, this small loss of efficiency can be partially recovered by calibration if I use the stratum membership indicator as my auxiliary variable. Then the calibrated weights are the same as the weights used for analyzing data obtained by sampling without replacement. Iwill demonstrate that calibration techniques are able to recover some efficiency loss in Chapter 5 by simulation studies. Lastly, it is also possible to generalize my results with

VPS to sampling without replacement by using the results on weighted bootstrap empirical process theory from Præstgaard and Wellner [1993, theorem2.2] as shown in Breslow and Wellner [2007].

For simplicity, I only consider time-invariant weights. Further efficiency improvement is possible if time-dependent weights are used. For example, Borgan et al. [2000] showed a slight improvement in efficiency using time-dependent weights for exposure stratified case-cohort studies. Kulich and Lin [2004] also proposed more complex time-dependent weights than Borgan et al. [2000] for case-cohort studies.

In addition I adopt the simplest Z-theorem that assumes the independence of observations, so that important stochastic conditions for showing consistency and asymptotic normality are transformed to Glivenko-Cantelli and Donsker conditions, for which many existing results can be used. My system also requires estimators to have $\sqrt{n}$ rate of convergence. When parameters are not estimable at $\sqrt{n}$-rate such as making inference based on the interval censoring data [van der Vaart, 1998, chap. 25] or for mixture models [Bickel et al., 1998, chap. 4.5], my estimation system does not apply. Alternatively, a more advanced Z-estimation method for the case-cohort design is seen in Nan and Wellner [2013]. Their theory allows correlation of the data and convergence of nuisance parameters at a different rate. They considered bundled parameters, which included a Euclidean parameter of interest θ and an infinite-dimensional nuisance parameter η that was allowed to be a function of θ and needed to be estimated prior to estimating θ.

When using my Z-estimation system, two important steps are to establish Donsker and Glivenko-Cantelli properties for classes of functions. For a broad audience, I find that many Donsker or Glivenko-Cantelli classes that have been discovered can be integrated together by the Donsker and Glivenko-Cantelli preservation theorems for new use and this recycling approach may be one solution to ease difficulties in verifying the stochastic conditions in the Z-estimation system. I do not show Donsker or Glivenko-Cantelli properties of a class of functions by directly showing that the size of the class is finite through the use of bracketing or covering number [van der Vaart and Wellner, 1996, chap. 2.5]. However, this funda-

mental approach should be more straightforward for statisticians with a strong mathematics background.

I consider efficiency gain from oversampling the most informative subjects and from utilizing the correlates of covariates by stratified sampling or calibration techniques rather than sophisticated statistical techniques. In fact, methods of weighted estimating equation and calibration are probably the simplest tools we can find to develop statistics and incorporate additional information for two-phase designs. These simple tools are easy to understand and follow. In return for the ease of access and implementation, some efficiency is lost. This drawback and different methods of improvement were discussed in Breslow and Wellner [2007]. In certain scenarios, semiparmetric efficient estimators do exist. For example, maximum likelihood or profile likelihood estimators are efficient and feasible if a semiparametric model can be partitioned into parametric and non-parametric parts provided phase I data are discrete [Lawless et al., 1999, Breslow et al., Robins et al., 1995] also proposed optimal estimators for the conditional mean model when two-phase sampling is considered as a missing data problem with missing by design.

I summary, in order to make the Z-estimation system accessible, I chose straightforward estimation methods and simplify my applications of modern empirical process theory in several places with some limitations. Readers who are interested in developing new statistical methods for analyzing two-phase sampling data can begin with this Z-estimation system to study the problem, because this system connects methods development for two-phase sampling to results in random sampling and this system provides a thorough consideration of potential problems that may occur in practice such as model misspecification and the presence of auxiliary variables. If the system does not apply or more efficiency gain is desired, readers can advance to other tools that are available and part of them are listed above.

2.2 Sampling, Data, Notation

After reviewing previous works and explaining my motivation, I will proceed to create the Z-estimation system for tow-phase sampling. For clarity, the estimation theory is divided into two parts for two major problems that arise in two-phase designs. The first problem is fitting a (possibly) semiparametric model to the small subsample of a two-phase study, for which we obtain complete information. The second problem is incorporating additional information from the initial large cohort to further improve estimation precision. For both problems, solutions are arranged in the order of constructing estimators, defining parameters, establishing consistency of the estimators, and showing their asymptotic distributions. In the end, strengths and limitations of our theory will be discussed. This chapter can be read side by side with Chapters 3 & 4, which illustrates this chapter's abstract theoretical construction and results by an example.

2.2.1 Sampling

Two-phase sampling design, originated from Neyman [1938], reduces cost substantially by arranging sampling in two phases, so that the more informative subjects in the Phase I samples will be selected to constitute the Phase II sub-samples for costly measurements.

In this dissertation, I consider at phase I, we randomly sample N subjects from an infinite superpopulation. The random vector $X \in \mathcal{X}$ is potentially available for all N subjects. However, we only measure a portion of X and some auxiliary variables U for all N subjects. We denote this portion of X by X^I and $X = (X^I, X^{II})$.

Next according to $V = (X^I, U) \in \mathcal{V}$, we independently generate a phase II selection indicator R for each individual from a Bernoulli distribution with a probability of $P(R = 1|V) = \pi_0(V)$, in which π_0 is an a priori function of V. We then measure X^{II} of X for only phase II samples.

This type of sampling method for selecting phase II sub-samples is also seen in Manski and Lerman [1977] for econometric problems. When V indicates strata and $\pi_0(V)$ is a

constant function within each stratum, this sampling method is also similar to the familiar stratified sampling and the same as variable probability sampling 1 described in Kalbfleisch and Lawless [1988]. However, V can be continuous and $\pi_0(V)$ is not restricted to be discrete. Therefore we consider more general two-phase sampling designs than two-phase stratified sampling.

2.2.2 Data and Assumptions on the Data

Two-phase VPS generates our data in the form of $(R_i, V_i, X_i^{II} R_i), i = 1, \ldots, N$. This is the dataset we intent to analyze. For this data, we make the following assumptions:

A. 2.2.1. *$X_1, \ldots, X_N$ are i.i.d. random samples from a probability distribution P on a measurable space $(\mathcal{X}, \mathcal{A})$. $P \in \mathcal{P}$, which is the set of all possible distributions P for X such that $E(X^T X) < \infty$.*

A. 2.2.2. *Sampling probability bounded from 0: $0 < \sigma \leq \pi_0(v) \leq 1$ for all $v \in \mathcal{V}$.*

A. 2.2.3. *Whether or not a subject is selected into phase II depends only on the observed data V at phase I, in other words, missing at random(MAR) [Rubin, 1976]:*
$P(R_i = 1|X_i, U_i) = P(R_i = 1|V_i) = \pi_0(V_i)$.

A. 2.2.4. *We assume $(R_i, V_i, X_i^{II} R_i)$ are i.i.d. following a probability distribution Q on a measurable space $(\mathcal{R} \times \mathcal{V} \times \mathcal{X}, \mathcal{B})$. $Q \in \mathcal{Q}$, which is the set of all possible distributions for (R, V, X) such that $E[(R, V, X)^T (R, V, X)] < \infty$.*

Note by two-phase VPS, R_i is independently generated. If we assume $V_i, i = 1, 2, \ldots, N$ are i.i.d., then under the assumption of A. 2.2.1 and A.2.2.3, $(R_i, V_i, X_i^{II} R_i)$ is ensured to be i.i.d.. Assumptions A. 2.2.2 and A.2.2.3 are also assumed in Robins et al. [1995] for their missing data problems. The assumptions of $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$ in A. 2.2.1 and A.2.2.4 are similar to [Newey, 1994] and allow us to estimate under general misspecification.

2.2.3 Notation

Throughout the dissertation, we use linear operators for the expectations. We define

$$Pf = \int f(X)dP, \qquad Qf = \int f(X, R, V)dQ.$$

To indicate P and Q are functionals, I adopt square brackets when needed. For example, the regularity conditions in assumptions A.2.2.1 & A. 2.2.4 will be written as $P[X^T X] < \infty$ and $Q[(R, V, X)^T (R, V, X)] < \infty$.

According to van der Vaart [1998, p.269] and van der Vaart and Wellner [1996, p.80], the empirical measure $\mathbb{P}_N$ is defined as the discrete uniform measure that puts mass $1/N$ on each observation. Similar to the operator notation P, given a measurable function $f : \mathcal{X} \mapsto \mathbb{R}$, we write $\mathbb{P}_N$ for the expectation of f under the empirical measure $\mathbb{P}_N$:

$$\mathbb{P}_N f = \frac{1}{N} \sum_{i=1}^{N} f(X_i).$$

With two-phase sampling, for a measurable function f: $\mathcal{R} \times \mathcal{V} \times \mathcal{X} \mapsto \mathbb{R}$

$$\mathbb{P}_N f = \frac{1}{N} \sum_{i=1}^{N} f(R_i, V_i, X_i).$$

We define the empirical process evaluated at f as

$$\mathbb{G}_N f = \sqrt{N}(\mathbb{P}_N f - Pf).$$

Suppose f belongs to a class of functions $\mathcal{F}$. We consider $\mathbb{P}_N f$ as a random map in $l^\infty(\mathcal{F})$. By definition, the space $l^\infty(\mathcal{F})$ is defined as the set of all uniformly bounded real functions on $\mathcal{F}$, i.e., the space $l^\infty(\mathcal{F})$ consists of all functions $z : \mathcal{F} \mapsto \mathbb{R}$ such that

$$\|z\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |z(f)| < \infty$$

[van der Vaart and Wellner, 1996, p.34]. Then the empirical process $\{\mathbb{G}_N f, f \in \mathcal{F}\}$ is the centered and scaled version of this random map. Thus, each empirical process is associated with a class of functions $\mathcal{F}$. To specify which empirical process under consideration, we

index each particular empirical process by an index set $\mathcal{F}$. For example, the empirical process theory in the preceding display is called "$\mathcal{F}$-indexed empirical process". The modern empirical process since Dudley [1978]'s key paper differs from the classic empirical process by how abstract this index set $\mathcal{F}$ is. Therefore, modern empirical process theory studies a much larger class of empirical processes than before and provides a wealth of important results for our use. In this dissertation, since we are concerned with semiparametric inference, the index set $\mathcal{F}$ for the empirical process we are interested in is abstract. Thus, modern empirical process theory is desired.

Throughout this dissertation, we use $\|\cdot\|_{\mathcal{H}}$ to denote the sup norm: $\|x\|_{\mathcal{H}} = sup_{h\in\mathcal{H}}|x(h)|$ and $\|\cdot\|_{\mathbb{E}}$ to denote the Euclidean norm. We define the norm of a product space as a sum: $\|(x,y)\| = \|x\| + \|y\|$.

2.3 The First Problem

Epidemiologists usually collect data to study association between two factors based on an assumed regression model. This model can be parametric or semiparametric. The study of the association is then translated to a problem of estimating the regression parameters of this model. When data are collected by two-phase VPS design, we ask how do we estimate regression parameters in a possibly semiparametric model?

Under this general problem, there are two important issues we will address. First of all, clinical researchers are sometimes interested in using the model to do individual based outcome prediction. Then all the parameters in the model become our interest in order to create a risk prediction function. These parameters include the regression parameters as well as the nuisance parameter, which is often infinite-dimensional in a semiparametric model and not estimated. In this sense, the nuisance parameter is not nuisance anymore.

Furthermore, in real applications, the assumed model is not always correct in describing the true distribution of observations. When a model is misspecified, how do we describe targets of our estimators and these estimators' precision?

The first problem we are concerned with is therefore *estimating all parameters in a*

possibly semiparametric model using two-phase VPS data allowing for model misspecification.

2.4 Solution: A Z-estimation System

Broadly speaking, our solution to the first problem is to estimate all parameters simultaneously using inverse probability weighted estimating equations (IPW-EE) on and off an assumed model.

More specifically, to solve the problem that parameters can be infinite-dimensional, we let our IPW-EE be infinite-dimensional as well. We obtain our estimators by solving this set of zero-valued estimation equations. As a result, our estimators are Z-estimators. Since modern empirical process theory studies empirical processes on abstract infinite-dimensional spaces and the Z-estimation theorem [van der Vaart, 1998, 19.26] allows Z-estimators to range over an infinite-dimensional space, we use them together to study the asymptotic behaviors of our estimators. To address the issue that the model may be misspecified, we consider the targets (parameters) of our estimators as functionals that are estimated nonparametrically.

In the following sections we organize our solution "A Z-estimation system" into four parts: estimators, parameters, consistency of estimators and limiting distributions of estimators. We call this solution a system because each part of it connects to the other parts, working together as a complex whole. Since our estimating procedure for two-phase sampling estimators is closely related to the estimating procedure for random sampling estimators, our Z-estimation system contains theoretical results for developing random sampling estimators as well.

2.4.1 Estimators

Before we formally define our parameter α in section 2.6.2, in this section parameter is used to refer to a numerical characteristic of a superpopulation we are interested to estimate from a sample. We first consider estimating a parameter from a sample that is collected by random sampling. We use the same set of notations introduced in section 2.2 for two-phase VPS to describe this sample. In contrast to two-phase VPS, with random sampling, $X_1, X_2, \ldots, X_N$

are fully observed and i.i.d. following a probability distribution P where $P \in \mathcal{P}$.

If α is a finite-dimensional parameter, i.e., a Euclidean parameter, then one common estimating approach is to solve a zero-valued estimating equation (EE) for an estimator. The challenge is how to construct and describe an infinite number of estimating equations desired for estimating infinite-dimensional parameters α. The solution was provided in van der Vaart and Wellner [1996, theorem 3.3.1]. The authors extended Z-estimation for Z-estimators on the Euclidean space [Huber, 1967, Pollard, 1985] to the infinite-dimensional space. They considered the estimating equations as a zero-valued random map in a Banach space. In the following, I present and further explain their idea on treating estimating equations as a map since it requires a lot of conceptual thinking. For simplicity, I restrict the Banach space to be $l^\infty(\mathcal{H})$ where $\mathcal{H}$ is an arbitrary index set and our estimating equations are the average of i.i.d. functions. As a result, a simplified Z-estimation theorem [van der Vaart, 1998, theorem 19.26] can be used. I then extend this theorem to two-phase VPS for developing the estimators of our interest on a possibly infinite-dimensional space.

Let $\mathbb{L}$ be the Banach space $l^\infty(\mathcal{H})$ and $\mathbb{L}_0$ be a subset of the Banach space $\mathbb{L}$. For random sampling (RS), we consider a random map $\Psi_N : \mathbb{L}_0 \mapsto \mathbb{L}$ of the form

$$\Psi_N(\alpha) = \mathbb{P}_N \psi_\alpha$$

in which ψ_α is a known map. We assume for each fixed x and $\alpha \in \mathbb{L}_0$, the map $h \mapsto \psi_{\alpha,h}$, denoted as ψ_α, is uniformly bounded, and so is the map $h \mapsto P\psi_{\alpha,h}$, denoted as $P\psi_\alpha$. We construct our estimating equation as

$$\Psi_N(\alpha) = 0. \tag{2.2}$$

Then RS estimator $\hat{\alpha}$ is obtained by solving (2.2). $\hat{\alpha}$ is therefore a Z-estimator [van der Vaart and Wellner, 1996, p. 284].

$\Psi_N(\alpha) \in l^\infty(\mathcal{H})$ implies $\mathbb{P}_N \psi_{\alpha,h} \in \mathbb{R}$ and

$$\|\mathbb{P}_N \psi_\alpha\|_{\mathcal{H}} = \sup_{h \in \mathcal{H}} |\mathbb{P}_N \psi_{\alpha,h}| < \infty.$$

One way to motivate $\psi_\alpha(X)$ is from an assumed model $\{P_\alpha, \alpha \in \mathbb{L}_0\}$. $\psi_\alpha(X)$ is proposed such that $P_\alpha \psi_\alpha(X) = 0$. For example, if P_α is a parametric model, then $\psi_\alpha(X)$ can be the score equation under model P_α.

Because $\Psi_N(\alpha) \in l^\infty(\mathcal{H})$, the estimating equation (2.2) is equivalent to a collection of equations

$$\Psi_N(\alpha)h = \mathbb{P}_N \psi_\alpha(X)h = \mathbb{P}_N \psi_{\alpha,h}(X) = 0 \text{ for every } h \in \mathcal{H}.$$

We see from this display that each h indexes a particular real valued estimating equation, for which reason $\mathcal{H}$ is called an index set. The size of this collection can be infinite depending on the size of $\mathcal{H}$. The classic Z-estimation considers estimating equations on the Euclidean space $\mathbb{R}^p$; now it is extended to a much larger space. By considering space $\mathbb{L} = l^\infty(\mathcal{H})$, on one hand we concisely describe an infinite number of estimating equations with one symbol ψ_α, on the other hand the index h facilitates us to focus on a single dimensional function $\psi_{\alpha,h}$ at one time in this infinite-dimensional problem. In later sections and chapters, we will see an abundance of convenience brought by this construction.

In two-phase VPS sampling, X, however, is not fully observed. We observe all of X only when R=1, i.e., for subjects selected to phase II subsamples. We consider a new random map $\Psi_N^* : \mathbb{L}_0 \mapsto \mathbb{L}$ given by

$$\Psi_N^*(\alpha) = \mathbb{P}_N \psi_\alpha^*(X, V, R) = \mathbb{P}_N \frac{R}{\pi_0(V)} \psi_\alpha(X). \tag{2.3}$$

We construct our estimating equation as

$$\Psi_N^*(\alpha) = 0 \tag{2.4}$$

Our two-phase VPS estimator $\hat{\alpha}^*$ is then obtained by solving (2.4), and it is still a Z-estimator. We call this equation inverse probability weighted estimating equations (IPW-EE) because the contribution from each subject to the overall estimating equation is weighted by the inverse of this subject's selection probability.

2.4.2 Parameters

The RS and two-phase VPS share the same parameter defined as follows. Corresponding to the random map Ψ_N in (2.2), we consider a deterministic map $\Psi : \mathbb{L}_0 \mapsto \mathbb{L}$ given by

$$\Psi(a) = P\psi_\alpha. \tag{2.5}$$

We define our parameter α_0 as the solution to the equation, assumed unique,

$$\Psi(a) = 0. \tag{2.6}$$

This definition of α_0 can be put in a more standard way by considering it as a functional. Let $\alpha : \mathcal{P} \mapsto \mathbb{L}_0$ denote a map that is implicitly identified by (2.6). Then the true parameter is the value of this map evaluated at the true distribution: $\alpha_0 = \alpha(P)$.

Since model $\mathcal{P}$ for P is unrestricted except for the regularity condition that $PXX^T < \infty$ as assumed in A.2.2.1, $\mathcal{P}$ is nonparametric. This definition of α_0 provides us a method to describe the target of our estimator under general model misspecification, since we allow $P \notin \{P_\alpha, \alpha \in \mathbb{L}_0\}$

When the assumed model is correct, $P \in \{P_\alpha, \alpha \in \mathbb{L}_0\}$, i.e., P is contained in the model space. (2.6) becomes $P_\alpha\psi_\alpha(X) = 0$. If ψ_α is motivated from the assumed model such that $P_\alpha\psi_\alpha(X) = 0$, then α defined via (2.6) automatically equals the true value of the parameter defined in model $\{P_\alpha, \alpha \in \mathbb{L}_0\}$.

In two-phase VPS sampling, corresponding to the random map Ψ_N^* in (2.4), we consider a deterministic map $\Psi^* : \mathbb{L}_0 \mapsto \mathbb{L}$:

$$\Psi^*(a) = Q\psi_\alpha^*(R, V, X). \tag{2.7}$$

By assumption A.2.2.3,

$$\Psi^*(\alpha) = Q\psi_\alpha^*(R, V, X) = Q\frac{R}{\pi_0(V)}\psi_\alpha(X) = E_Q\left[E_Q\left[\frac{R}{\pi_0(V)}|X, U\right]\psi_\alpha(X)\right] = E_Q\psi_\alpha(X) = P\psi_\alpha(X). \tag{2.8}$$

Since α_0 is the unique solution to $P\psi_\alpha(X) = 0$, equality in (2.8) implies α_0 is also the unique solution to $\Psi^*(a) = 0$. This result guarantees, to be revealed in the next section, that our two-phase sampling estimator $\hat{\alpha}^*$ will estimate the same quantity as $\hat{\alpha}$ would if data were complete, regardless of the considered model for X.

At the end of this section on parameters, I will explain the type of parameters we are concerned with when we set $\mathbb{L}_0 \subset l^\infty(\mathcal{H})$. I will show most parameters considered in statistical models for epidemiologic studies can be identified as elements in $l^\infty(\mathcal{H})$ for some appropriate $\mathcal{H}$.

Any θ that belongs to a subset of an p-dimensional Euclidean space can be identified as an element in $l^\infty(\mathcal{H})$ where $\mathcal{H}$ is the unit ball in $\mathbb{R}^p$. This is because θ can be identified uniquely by the values of its inner product with an element of the unit ball. For any $\theta \in \mathbb{R}^p$, we can define a map $\alpha : \mathcal{H} \mapsto \mathbb{R}$ by $\alpha h = h^T\theta$. Then each θ can be identified uniquely by an α within $l^\infty(\mathcal{H})$.

One common infinite-dimensional parameter in a survival model is the baseline cumulative hazard Λ. Any Λ can also be identified as an element in $l^\infty(\mathcal{H})$ where $\mathcal{H}$ is the set of bounded functions of bounded variation over a time interval $[0, \tau]$. This is because Λ as an finite measure on $[0, \tau]$ can be identified uniquely by the map $\alpha : \mathcal{H} \mapsto \mathbb{R}$ defined by $\alpha h = \int_0^\tau h d\Lambda$. Then each Λ can be identified uniquely by an α within $l^\infty(\mathcal{H})$.

Sometimes a parameter includes both finite and infinite-dimensional components such as when both θ and Λ in a semiparametric survival model become our interest. We usually consider (θ, Λ) belongs to some product space $\Theta \times \mathbb{A}$ where Θ is a bounded set of $\mathbb{R}^p$ and $\mathbb{A}$ is a collection of finite measures. Alternatively any (θ, Λ) can be identified as an element in $l^\infty(\mathcal{H})$, in which $h = (h_1, h_2)$. h_1 belongs to the unit ball in $\mathbb{R}^p$ and h_2 is a bounded function over $[0, \tau]$. This is because (θ, Λ) can be identified uniquely by the map $\alpha : \mathcal{H} \mapsto \mathbb{R}$ defined by $\alpha h = h_1^T\theta + \int_0^\tau h_2 d\Lambda$. Then each $(\theta, \Lambda) \in \Theta \times \mathbb{A}$ can be identified uniquely by an α within $l^\infty(\mathcal{H})$.

These three examples demonstrate we are concerned with a large range of parameters. These include both finite and infinite-dimensional parameters or even these two types of

parameters together. The fact that we are able to consider two types of parameters together as a single parameter allows us to jointly estimate all parameters at one time.

2.4.3 Consistency

Chapter 5 from van der Vaart [1998] provided an approach to show the consistency of a finite-dimensional Z-estimator. In this section, we will extend this approach to an infinite-dimensional Z-estimator and summarize this extension as a new theorem below. This theorem considers maps between general Banach spaces $\mathbb{L}_0$ and $\mathbb{L}$. Thus we do not restrict $\mathbb{L}$ to be $l^\infty(\mathcal{H})$ in the theorem. The consistency of a parameter estimate is usually established case by case, but in the case of i.i.d. observation, there may be general approach to establish consistency. Following the theorem we present two consistency corollaries for i.i.d. observations, one with RS and the other with two-phase VPS.

Theorem 2.4.1. *Let $\Psi_N : \mathbb{L}_0 \mapsto \mathbb{L}$ be a random map and $\Psi : \mathbb{L}_0 \mapsto \mathbb{L}$ be a deterministic map from parameter space $\mathbb{L}_0$, a subset of Banach space, to another Banach space $\mathbb{L}$. Let $||\cdot||_{\mathbb{L}_0}$ be the norm for parameter space $\mathbb{L}_0$ and $||\cdot||_{\mathbb{L}}$ be the norm for Banach space $\mathbb{L}$. Consider*

Condition 2.4.1.

$$\sup_{\alpha\in\mathbb{L}_0} \|\Psi_N(\alpha) - \Psi(\alpha)\|_{\mathbb{L}} \xrightarrow{p} 0;$$

Condition 2.4.2. *$\Psi(\alpha_0) = 0$ and for every $\epsilon > 0$.*

$$\inf_{\alpha:\|\alpha-\alpha_0\|\geq\epsilon} \|\Psi(\alpha) - \Psi(\alpha_0)\|_{\mathbb{L}} > 0$$

If both conditions are satisfied, then $\hat{\alpha}$ that satisfies $\Psi_N(\hat{\alpha}) = o_p(1)$ converges to α_0 in probability.

Proof. Based on condition 2.4.1,

$$\|\Psi(\hat{\alpha}) - \Psi_N(\hat{\alpha})\|_{\mathbb{L}} \leq \sup_{\alpha\in\mathbb{L}_0} \|\Psi(\alpha) - \Psi_N(\alpha)\|_{\mathbb{L}} = o_p(1). \tag{2.9}$$

By assumption, we have $\Psi(\alpha_0) = 0$ and $\Psi_N(\hat{\alpha}) = o_p(1)$, so

$$\|\Psi_N(\hat{\alpha}) - \Psi(\alpha_0)\|_{\mathbb{L}} = o_p(1). \tag{2.10}$$

Combining (2.9) and (2.10) results in

$$\|\Psi(\hat{\alpha}) - \Psi(\alpha_0)\|_{\mathbb{L}_0} \leq \|\Psi(\hat{\alpha}) - \Psi_N(\hat{\alpha})\|_{\mathbb{L}} + \|\Psi_N(\hat{\alpha}) - \Psi(\alpha_0)\|_{\mathbb{L}} = o_p(1),$$

which means for any $\eta > 0$, $P(\|\Psi(\hat{\alpha}) - \Psi(\alpha_0)\|_{\mathbb{L}} \geq \eta) \to 0$ as $N \to \infty$.

According to condition 2.4.2, for any $\epsilon > 0$, there exists $\eta > 0$ such that event $A \equiv \{\|\hat{\alpha} - \alpha_0\| \geq \epsilon > 0\}$ implies event $B \equiv \{\|\Psi(\hat{\alpha}) - \Psi(\alpha_0)\|_{\mathbb{L}} \geq \eta\}$, in other words, $A \subset B$, so

$$P(\|\hat{\alpha} - \alpha_0\|_{\mathbb{L}_0} \geq \epsilon) \leq P(\|\Psi(\hat{\alpha}) - \Psi(\alpha_0)\|_{\mathbb{L}} \geq \eta).$$

Therefore $\forall \epsilon > 0$, $P(\|\hat{\alpha} - \alpha_0\|_{\mathbb{L}_0} \geq \epsilon) \to 0$ as $N \to \infty$. We conclude $\|\hat{\alpha} - \alpha_0\|_{\mathbb{L}_0} \to_p 0$. □

In this theorem, the establishment of consistency is divided into the stochastic condition 2.4.1 on deviations between Ψ_N and Ψ and the condition 2.4.2 on Ψ alone. It is usually a challenge to verify condition 2.4.1. However, for our estimating equation (2.2), $\Psi_N(\alpha) = \mathbb{P}_N\psi_\alpha$, which belongs to $l^\infty(\mathcal{H})$. Hence

$$\sup_{\alpha\in\mathbb{L}_0} \|\Psi_N(\alpha) - \Psi(\alpha)\|_{\mathbb{L}} = \sup_{\alpha\in\mathbb{L}_0} \|\mathbb{P}_N\psi_\alpha - P\psi_\alpha\|_{\mathcal{H}} = \sup_{\alpha\in\mathbb{L}_0, h\in\mathcal{H}} |\mathbb{P}_N\psi_{\alpha,h} - P\psi_{\alpha,h}|. \tag{2.11}$$

As a result, condition 2.4.1 is simplified to

Condition 2.4.3. $\mathcal{G} \equiv \{\psi_{\alpha,h}, \alpha \in \mathbb{L}_0, h \in \mathcal{H}\}$ *is Glivenko-Cantelli.*

Because Glivenko-Cantelli condition can be effortlessly carried to the new two-phase sampling scenario, this Glivenko-Cantelli type of condition will not only simplify our establishment on the consistency of the RS estimator $\hat{\alpha}$ but also the two-phase VPS estimator $\hat{\alpha}^*$, as shown in the following.

Corollary 2.4.2. *(Consistency of RS estimators)*
Let $x \mapsto \psi_{\alpha,h}(x)$ be a measurable function such that conditions 2.4.2 and 2.4.3 are satisfied. Then $\hat{\alpha}$ defined in (2.2) *converges to α_0 in probability.*

Proof. By condition 2.4.3, $\sup_{\alpha\in\mathbb{L}_0, h\in\mathcal{H}} |\mathbb{P}_N\psi_{\alpha,h} - P\psi_{\alpha,h}| \to_p 0$. According to (2.11), $\sup_{\alpha\in\mathbb{L}_0} \|\Psi_N(\alpha) - \Psi(\alpha)\|_{\mathbb{L}} \to_p 0$. Because condition 2.4.2 holds and $\Psi_N(\hat{\alpha}) = 0$ by (2.2), it follows from theorem 2.4.1 that $\hat{\alpha} \to_p \alpha_0$. □

Corollary 2.4.3. *(Consistency of two-phase VPS estimators)*
Suppose A.2.2.2 and A.2.2.3 hold. Let $x \mapsto \psi_{\alpha,h}(x)$ be a measurable function such that conditions 2.4.2 and 2.4.3 are satisfied and the class $\mathcal{G}$ in condition 2.4.3 has integrable envelope function. Then $\hat{\alpha}^*$ defined in (2.4) *converges to α_0 in probability.*

Proof. We show consistency by adapting theorem 2.4.1 to $\hat{\alpha}^*$. We first show condition 2.4.2 in theorem 2.4.1 is satisfied. Given A. 2.2.3, equality in (2.8) holds, so $\Psi^*(\alpha) = \Psi(\alpha)$. Because condition 2.4.2 on Ψ is satisfied, substituting Ψ^* for Ψ in condition 2.4.2 yields $\Psi^*(\alpha_0) = 0$ and for every $\epsilon > 0$

$$\inf_{\alpha:\|\alpha-\alpha_0\|\geq\epsilon} \|\Psi^*(\alpha) - \Psi^*(\alpha_0)\|_{\mathbb{L}} > 0.$$

Next we show condition 2.4.1 in theorem 2.4.1 is satisfied. Because $\Psi^*_N(\alpha) = \mathbb{P}_N\psi^*$ and $\Psi^*(\alpha) = Q\psi^*_\alpha$ as defined in (2.3) and (2.7),

$$\sup_{\alpha\in\mathbb{L}_0} \|\Psi^*_N(\alpha) - \Psi^*(\alpha)\|_{\mathbb{L}} = \sup_{\alpha\in\mathbb{L}_0} \|\mathbb{P}_N\psi^*_\alpha - P\psi^*_\alpha\|_{\mathcal{H}}, = \sup_{\alpha\in\mathbb{L}_0, h\in\mathcal{H}} |\mathbb{P}_N\psi^*_{\alpha,h} - Q\psi^*_{\alpha,h}|.$$

As a result (2.8) becomes a condition to show $\{\psi^*_{\alpha,h} : \alpha \in \mathbb{L}_0, h \in \mathcal{H}\}$ is Glivenko-Cantelli (GC). In view of function $\psi^*_{\alpha,h}$ as a product of $\{\frac{R}{\pi_0(V)}\}$ and $\psi_{\alpha,h}$, the preservation theorem of GC [van der Vaart and Wellner, 2000, theorem 3] can be applied. By the law of large numbers and the definition of GC, the singleton $\{\frac{R}{\pi_0(V)}\}$ is GC. Given A. 2.2.2, the singleton $\{\frac{R}{\pi_0(V)}\}$ also has integrable envelope function. By assumption, $\mathcal{G}$ in condition 2.4.3 is GC with integrable envelope function. By the preservation theorem of GC, $\{\psi^*_{\alpha,h} : \alpha \in \mathbb{L}_0, h \in \mathcal{H}\}$ is GC and condition 2.4.1 is satisfied.

Finally, definition of $\hat{\alpha}^*$ in (2.4) yields $\Psi^*_N(\hat{\alpha}^*) = 0$. It follows from theorem 2.4.1 that $\hat{\alpha}^* \to_p \alpha_0$.

□

2.4.4 Asymptotic Normality

Tools have been created to show the asymptotic normality of Z-estimators [van der Vaart and Wellner, 1996] when they are obtained from a random sample. For example, theorem 5.21

from van der Vaart [1998] applies to Z-estimators on Euclidean spaces and theorem 19.26 from van der Vaart [1998] to Z-estimators on possibly infinite-dimensional spaces. Theorem 19.26 can be immediately applied to show the asymptotic normality of our RS estimators $\hat{\alpha}$ defined in (2.2). We rephrase this theorem below and refer to it as theorem 2.3.4. We then extend this theorem to two-phase VPS scenarios.

Theorem 2.4.4. *(Asymptotic normality for random sampling)*
For each $\alpha \in \mathbb{L}_0$ of a normed space and every h in an arbitrary set $\mathcal{H}$, let $x \mapsto \psi_{\alpha,h}(x)$ be a measurable function, such that the following conditions are satisfied:

Condition 2.4.4. *the class $\mathcal{F} \equiv \{\psi_{\alpha,h} : \|\alpha - \alpha_0\| < \delta, h \in \mathcal{H}\}$, with finite envelope function, is P-Donsker for some $\delta > 0$;*

Condition 2.4.5. *as a map into $l^\infty(\mathcal{H})$, the map $\alpha \mapsto P\psi_\alpha$ is Fréchet-differentiable at a zero α_0, with a derivative $\dot{\Psi}_0 : lin\mathbb{L}_0 \mapsto l^\infty(\mathcal{H})$ that has a continuous inverse on its range;*

Condition 2.4.6. *$\|P(\psi_{\alpha,h} - \psi_{\alpha_0,h})^2\|_{\mathcal{H}} \to 0$ as $\alpha \to \alpha_0$.*

If $\|\mathbb{P}_N \psi_{\hat{\alpha}}\|_{\mathcal{H}} = o_p(N^{-1/2})$ and $\hat{\alpha} \xrightarrow{P} \alpha_0$, then

$$\dot{\Psi}_0 \sqrt{N}(\hat{\alpha} - \alpha_0) = -\mathbb{G}_N \psi_{\alpha_0} + o_p(1).$$

Notation $\mathbb{G}_N$ is given and explained in section 2.2.3. Because $\mathcal{F}$ is a Donsker class, $\mathcal{F}_0 \equiv \{\psi_{\alpha_0,h}, h \in \mathcal{H}\}$ as a subset of a Donsker class is also Donsker [van der Vaart and Wellner, 1996, theorem 2.10.1]. By the definition of the Donsker class [van der Vaart and Wellner, 2000, p.82], the empirical process $\mathbb{G}_N$ indexed by the Donsker class $\mathcal{F}_0$ converges in distribution to the P-Brownian bridge $\mathbb{G}$ in $l^\infty(\mathcal{F}_0)$. In other words, the limit process $\{\mathbb{G}\psi_{\alpha_0,h}, \psi_{\alpha_0,h} \in \mathcal{F}_0\}$ is a zero-mean Gaussian process with covariance function

$$E\mathbb{G}\psi_{\alpha_0,h}\mathbb{G}\psi_{\alpha_0,g} = P\psi_{\alpha_0,h}\psi_{\alpha_0,g} - P\psi_{\alpha_0,h}P\psi_{\alpha_0,g}$$

where $g \in \mathcal{H}$ and $\psi_{\alpha_0,g} \in \mathcal{F}_0$. $P\psi_{\alpha_0,h} = 0$ and $P\psi_{\alpha_0,g} = 0$ due to the definition of α_0. Thus the asymptotic variance

$$Var_A \left\{ \dot{\Psi}_0 \sqrt{N}(\hat{\alpha} - \alpha_0)h \right\} = P\psi^2_{\alpha_0,h}; \tag{2.12}$$

the asymptotic covariance

$$Cov_A(\dot{\Psi}_0\sqrt{N}(\hat{\alpha}-\alpha_0)h, \dot{\Psi}_0\sqrt{N}(\hat{\alpha}-\alpha_0)g) = P\psi_{\alpha_0,h}\psi_{\alpha_0,g}. \tag{2.13}$$

Next we extend these results for the RS estimator $\hat{\alpha}$ to the two-phase VPS estimator $\hat{\alpha}^*$.

Corollary 2.4.5. *(Aysmptotic normality for two-phase VPS)*
For each $\alpha \in \mathbb{L}_0$ and every h in an arbitrary set $\mathcal{H}$, let $x \mapsto \psi_{\alpha,h}(x)$ be a measurable function such that conditions 2.4.4 $\sim$ 2.4.6 are satisfied. Assume that assumptions A.2.2.2 and A.2.2.3 hold. Also assume that $\mathcal{F}$ in condition 2.4.4 has integrable envelope function. If $\hat{\alpha}^*$ defined in (2.4) *converges to α_0 in probability, then*

$$\dot{\Psi}_0\sqrt{N}(\hat{\alpha}^*-\alpha_0) = -\mathbb{G}_N\psi^*_{\alpha_0} + o_p(1).$$

Proof. We prove this corollary by adapting theorem 2.4.4 to $\psi^*_{\alpha,h}$ and $\hat{\alpha}^*$. We will verify that when we substitute $\psi^*_{\alpha,h}$ for $\psi_{\alpha,h}$ in theorem 2.4.4, conditions in this theorem still hold.

First, given condition 2.4.4 and our assumptions, $\mathcal{F}$ is P-Donsker with integrable envelope function. Given A. 2.2.2, $\frac{R}{\pi_0(V)}$ is a bounded function. Applying example 2.10.10 [van der Vaart and Wellner, 1996] yields the new class of functions $\mathcal{F}' \equiv \{\frac{R}{\pi_0(V)}\psi_{\alpha,h}, \alpha \in \mathbb{L}_0, h \in \mathcal{H}\}$ is Q-Donsker with finite envelope function.

Second, in view of equation (2.8), the deterministic map $\alpha \to Q\psi^*_\alpha$ is the same as map $\alpha \to P\psi_\alpha$. Because $\alpha \to P\psi_\alpha$ is Fréchet differentiable at a zero α_0, with a derivative $\dot{\Psi}_0$ that has a continuous inverse on its range. When we replace map $\alpha \to P\psi_\alpha$ by map $\alpha \to Q\psi^*_\alpha$, condition 2.4.5 still holds

Third, under A.2.2.3

$$\begin{aligned}
Q(\psi^*_{\alpha,h} - \psi^*_{\alpha_0,h})^2 =& Q\left[\frac{R^2}{\pi_0^2(V)}(\psi_{\alpha,h} - \psi_{\alpha_0,h})^2\right] \\
=& E_Q\left[(\psi_{\alpha,h} - \psi_{\alpha_0,h})^2 E_Q\left[\frac{R}{\pi_0^2(V)}|X,U\right]\right] \\
=& Q\left[(\psi_{\alpha,h} - \psi_{\alpha_0,h})^2 \frac{1}{\pi_0^2(V)} P(R=1|X,U)\right] \\
=& Q\left[(\psi_{\alpha,h} - \psi_{\alpha_0,h})^2 \frac{1}{\pi_0(V)}\right].
\end{aligned}$$

Thus, given A.2.2.2 we obtain

$$\|Q(\psi^*_{\alpha,h} - \psi^*_{\alpha_0,h})^2\|_{\mathcal{H}} \leq \frac{1}{\sigma^2}\|P(\psi_{\alpha,h} - \psi_{\alpha_0,h})^2\|_{\mathcal{H}}.$$

As a result of condition 2.4.6, $\|P(\psi_{\alpha,h} - \psi_{\alpha_0,h})^2\|_{\mathcal{H}} \to 0$. Thus $\|Q(\psi^*_{\alpha,h} - \psi^*_{\alpha_0,h})^2\|_{\mathcal{H}} \to 0$.

We have verified three conditions in theorem 2.4.4 still hold when replacing ψ_α by ψ^*_α. Since $\hat{\alpha}^*$ defined in (2.4) satisfies $\|\mathbb{P}_N\psi^*_{\hat{\alpha}^*}\| = 0$ and $\hat{\alpha}^* \xrightarrow{P} \alpha_0$. It follows from theorem 2.4.4 that

$$\dot{\Psi}_0\sqrt{N}(\hat{\alpha}^* - \alpha_0) = -\mathbb{G}_N\psi^*_{\alpha_0} + o_p(1).$$

□

Because $\mathcal{F}'$ is a Donsker class, $\mathcal{F}'_0 \equiv \{\psi^*_{\alpha_0,h}, h \in \mathcal{H}\}$ as a subset of a Donsker class is also Donsker [van der Vaart and Wellner, 1996, theorem 2.10.1]. By the definition of the Donsker class [van der Vaart and Wellner, 2000, p.82], the empirical process $\mathbb{G}_N$ indexed by the Donsker class $\mathcal{F}'_0$ converges in distribution to the P-Brownian bridge $\mathbb{G}$ in $l^\infty(\mathcal{F}'_0)$. Therefore, the limit process $\{\mathbb{G}\psi^*_{\alpha_0,h}, \psi^*_{\alpha_0,h} \in \mathcal{F}'_0\}$ is a zero-mean Gaussian process with covariance function

$$E\mathbb{G}\psi^*_{\alpha_0,h}\mathbb{G}\psi^*_{\alpha_0,g} = Q\left[\frac{R}{\pi_0(V)}\psi_{\alpha_0,h}\frac{R}{\pi_0(V)}\psi_{\alpha_0,g}\right] - Q\left[\frac{R}{\pi_0(V)}\psi_{\alpha_0,h}\right]Q\left[\frac{R}{\pi_0(V)}\psi_{\alpha_0,g}\right]$$

where $g \in \mathcal{H}$ and $\psi^*_{\alpha_0,g} \in \mathcal{F}'_0$. To simplify the covariance function, we apply the variance

decomposition formula:

$$\begin{aligned}
Var\left[\frac{R}{\pi_0(V)}\psi_{\alpha_0,h}\right] =& Var\left[E\left[\frac{R}{\pi_0(V)}\psi_{\alpha_0,h}|X,U\right]\right] + E\left[Var\left[\frac{R}{\pi_0(V)}\psi_{\alpha_0,h}|X,V\right]\right] \\
=& Var\left[E\left[\frac{1}{\pi_0(V)}\psi_{\alpha_0,h}P(R=1|X,U)\right]\right] + E\left[\frac{\psi^2_{\alpha_0,h}}{\pi_0^2(V)}Var[R|X,V]\right] \\
=& Var[\psi_{\alpha_0,h}] + E\left[\frac{\{1-\pi_0(V)\}\pi_0(V)}{\pi_0^2(V)}\psi^2_{\alpha_0,h}\right] \\
=& P\psi^2_{\alpha_0,h} - (P\psi_{\alpha_0,h})^2 + Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\psi^2_{\alpha_0,h}\right] \\
=& P\psi^2_{\alpha_0,h} + Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\psi^2_{\alpha_0,h}\right].
\end{aligned}$$

Therefore, the asymptotic variance

$$Var_A\left[\dot{\Psi}_0\sqrt{N}(\hat{\alpha}^* - \alpha_0)h\right] = P\psi^2_{\alpha_0,h} + Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\psi^2_{\alpha_0,h}\right]; \tag{2.14}$$

the asymptotic covariance

$$Cov_A\left[\dot{\Psi}_0\sqrt{N}(\hat{\alpha}^* - \alpha_0)h, \dot{\Psi}_0\sqrt{N}(\hat{\alpha}^* - \alpha_0)g\right] = P\psi_{\alpha_0,h}\psi_{\alpha_0,g} + Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\psi_{\alpha_0,h}\psi_{\alpha_0,g}\right]. \tag{2.15}$$

The asymptotic variance in (2.14) has two components. The first equals (2.12), the variance we would obtain if we had collected complete information on X for all N subjects.The second component is a penalty term for the fact that we only observe complete information on X among phase II subsamples.

Note the asymptotic variances results in (2.12) with random sampling and (2.14) with two-phase VPS are robust variances since we estimate allowing $P \in \mathcal{P}$. Suppose an assumed model is believed to be the true model for the observed data, i.e., $P \in \{P_\alpha, \alpha \in \mathbb{L}_0\}$. Then the model-based variances can be obtained by replacing P in (2.12) and the first component of (2.14) by P_{α_0}.

2.4.5 Summary

In view of section 2.4.1-2.4.4, the problem of developing estimators was decomposed into small components. These components consist of: ① estimators (section 2.2), ② parameter

(section 2.3), ③ Glivenko-Cantelli class (condition 2.4.3), ④ Donsker class (condition 2.4.4), ⑤ Fréchet derivative and its continuous inverse (condition 2.4.5), ⑥ convergence in quadratic mean. These components are almost independent of each other and can be established separately. This gives us a lot of convenience and transparency in the asymptotic analysis of our estimators.

When the problem becomes more complex, it won't be daunting any more since we can break down the complex problem to small tasks and tackle each task separately. For example, comparing our procedure of developing the two-phase VPS estimator $\hat{\alpha}^*$ to RS estimator $\hat{\alpha}$, we adapted components ①③④ to the new scenario, while other components were found not to need modification during the adaptation. Integrating these small results together, we developed $\hat{\alpha}^*$ and showed its asymptotics very clearly and quickly.

In the next section, we will encounter an even more complex problem and we propose a new estimator for α_0. We will see how each component in the estimating procedure we established in this section is modified, and then brought together to develop the new estimator and to solve the new problem .

2.5 The Second Problem

In the initial random sample at phase I, we observe auxiliary variable U and X^I for all N subjects. Although some of these variables $V = (U, X^I)$ have been used for deciding phase II subsamples, there may still be a lot of information in V that is potentially useful but is not exploited. Therefore, we ask "can we incorporate these information into our estimation procedure to develop a more efficient two-phase sampling estimator than $\hat{\alpha}^*$?"

2.6 Solution: A Z-estimation System Using Auxiliary Variables

To answer this question, the calibration technique proposed by Deville and Särndal [1992] is borrowed. we adjust the weights in the original IPW-EE (2.4) for $\hat{\alpha}^*$ such that the totals of some auxiliary variables that are fully observed at phase I are exactly estimated by their phase II estimates. We call the new estimator $\hat{\alpha}^{**}$ two-phase calibrated estimator. Through

calibration we incorporate phase I information in auxiliary variables into the estimation and this may further improve our estimation efficiency. We use the Z-estimation again due to its convenience for extending existing results. We will still build our theoretical tools for studying $\hat{\alpha}^{**}$ based on the Z-estimation theorem van der Vaart [1998, theorem 19.26], so that $\hat{\alpha}^{**}$ is allowed to be infinite dimensional and its variance to be estimated on or off an assumed model.

In the following sections, I first discuss construction of the new estimating equation for $\hat{\alpha}^{**}$ with calibration on auxiliary variables. Next I identify the parameter α_0 that $\hat{\alpha}^{**}$ is aiming to estimate. Then I show the consistency of $\hat{\alpha}^{**}$ and its asymptotic distribution.

2.6.1 Estimators

The idea of calibration is to modify the sampling weights, in our case the inverse probability weights $\frac{1}{\pi_0(V_i)}$, to a set of new weights w_i that are as close as possible to the original weights and, at the same time, subject to constraints based on some variables that are fully observed at phase I. Let vector $\tilde{V} = \tilde{V}(V)$ of q-dimension be the quantity we choose to calibrate on and be called auxiliary variables. $\tilde{V}$ is a function of V, so $\tilde{V}$ is available for all N subjects. Here we make two assumptions on $\tilde{V}$:

A. 2.6.1. *$\tilde{V}$ is bounded.*

A. 2.6.2. *$Q\tilde{V}\tilde{V}^T$ is positive definite.*

Suppose we use Poisson deviance

$$G\{w, \frac{1}{\pi_0(v)}\} = wlog\{\frac{w}{1/\pi_0(v)}\} - \{w - \frac{1}{\pi_0(v)}\}$$

as our distance measure for comparing the calibrated weight w to the original weights weight $\frac{1}{\pi_0(v)}$. Then according to Deville and Särndal [1992], our goal is to find a new series of weights $w_i, i = 1, 2, \cdots, N$ such that $\sum_{i=1}^{N} R_i G\{w_i, \frac{1}{\pi_0(v_i)}\}$ is minimized subject to the constraints

$$\tilde{v}_{tot} = \sum_{i=1}^{N} R_i w_i \tilde{v}_i \tag{2.16}$$

where $\tilde{v}_{tot} = \sum_{i=1}^{N} \tilde{v}_i$.

We use Lagrange multipliers method to solve this optimization problem. Suppose the size of phase II subsample is n. Let $\gamma \in \Gamma$ be a q-vector of Lagrange multipliers corresponding to the constraints (2.16) and let $f(w_i, \gamma)$ be the Lagrange functions:

$$f(w_i, \gamma) = G\{w_i, \frac{1}{\pi_0(v_i)}\} + \gamma^T (\sum_{i=1}^{N} \tilde{v}_i - \sum_{i=1}^{N} R_i w_i \tilde{v}_i)$$

for $i = 1, 2, \ldots, n$. Solving

$$\nabla_{w_i} f(w_i, \gamma) = 0$$

yields

$$log(\frac{w_i}{1/\pi_0(v)}) - \gamma^T R_i \tilde{v}_i = 0.$$

Then

$$w_i = \frac{exp(-\gamma^T \tilde{v}_i)}{\pi_0(v_i)}. \tag{2.17}$$

Next solving

$$\nabla_{\gamma} f(w_i, \gamma) = 0$$

yields

$$\tilde{v}_{tot} = \sum_{i=1}^{N} \tilde{v}_i = \sum_{i=1}^{N} R_i w_i \tilde{v}_i. \tag{2.18}$$

Equations (2.17) and (2.18) together generate an estimating equation for γ:

$$\sum_{i=1}^{N} \frac{R_i}{\pi_0(v_i)} exp(-\gamma^T \tilde{v}_i) \tilde{v}_i = \tilde{v}_{tot}.$$

Replacing the original weight in (2.4) by the new weight in (2.17) generates another estimating equation for (α, γ):

$$\mathbb{P}_N \frac{R}{\pi_0(V)} exp(-\gamma^T \tilde{V}) \psi_\alpha(X) = 0.$$

Solving these two equations together yields the improved two-phase VPS estimator $\hat{\alpha}^{**}$.

To adapt the Z-estimation results in section 2.4 for studying the asymptotic behaviors of $\hat{\alpha}^{**}$, we introduce following Banach spaces and the maps between these Banach spaces.

We consider γ ranges over the parameter space Γ where $\Gamma \subset \mathbb{R}^q$ and (α, γ) ranges over the parameter space $\mathbb{L}_0 \times \Gamma$. Recall $\mathbb{L}_0 \subset \mathbb{L}$ and $\mathbb{L} = l^\infty(\mathcal{H})$. We define random maps

$$\Psi_N^{**} : \mathbb{L}_0 \times \Gamma \mapsto \mathbb{L} \times \mathbb{R}^q$$

by $\Psi_N^{**} = (\Psi_{N,1}^{**}, \Psi_{N,2}^{**})$ with

$$\Psi_{N,1}^{**}(\alpha, \gamma)h = \mathbb{P}_N \psi_{1,\alpha,\gamma}^{**}(X, V, R)h = \mathbb{P}_N \psi_{1,\alpha,\gamma,h}^{**}(X, V, R) = \mathbb{P}_N \frac{R}{\pi_0(V)} exp(-\gamma^T \tilde{V}) \psi_{\alpha,h}(X) \tag{2.19}$$

$$\Psi_{N,2}^{**}(\gamma) = \mathbb{P}_N \psi_{2,\gamma}^{**}(V, R) = \mathbb{P}_N \frac{R}{\pi_0(V)} exp(-\gamma^T \tilde{V}) \tilde{V} - \tilde{V}. \tag{2.20}$$

Then estimators $(\hat{\alpha}^{**}, \hat{\gamma})$ are obtained by solving the new inverse probability weighted estimating equation (IPW-EE):

$$\Psi_N^{**}(\alpha, \gamma) = 0. \tag{2.21}$$

2.6.2 Parameters

We consider deterministic maps $\Psi^{**} = (\Psi_1^{**}, \Psi_2^{**}) : \mathbb{L}_0 \times \Gamma \mapsto \mathbb{L} \times \mathbb{R}^q$ that are the expectation of $\Psi_N^{**}(\alpha, \gamma)$:

$$\Psi^{**}(\alpha, \gamma) = (Q\psi_{1,\alpha,\gamma}^{**}, Q\psi_{2,\gamma}^{**}).$$

Then our parameters are defined by the unique solution to the equation

$$\Psi^{**}(\alpha, \gamma) = 0. \tag{2.22}$$

This unique solution, as shown in the following lemma, in fact equals $(\alpha_0, 0)$ where α_0 is given in (2.6). This result guarantees that our improved two-phase VPS estimator $\hat{\alpha}^{**}$ will estimate the same quantity as $\hat{\alpha}$ would if complete information on X were available for all N subjects, regardless of the model for X and the auxiliary variable $\tilde{V}$ we choose to calibrate on.

Lemma 2.6.1. *Assume Γ is a compact convex subset of $\mathbb{R}^q$ with 0 as an interior point. Assume A. 2.6.1 and 2.6.2 hold. If α_0 is the unique solution to (2.6), then $(\alpha_0, 0)$ is the unique solution to (2.22).*

Proof. We first show 0 is the unique solution to $Q\psi^{**}_{2,\gamma}(V,R) = 0$. In view of (2.20) and assumption A.2.2.3, we see 0 is one solution to $Q\psi^{**}_{2,\gamma}(V,R) = 0$. Now we show $\gamma_0 = 0$ is the only solution. Suppose there exists another solution $\gamma_1 \in \Gamma$ such that $\gamma_1 \neq 0$ and $Q\psi^{**}_{2,\gamma_1}(V,R) = 0$. Then by the mean value theorem for multiple variables

$$\begin{aligned}\psi^{**}_{2,\gamma_1}(V,R) - \psi^{**}_{2,0}(V,R) =& \frac{R}{\pi_0(V)} \exp(-\gamma_1^T \tilde{V})\tilde{V} - \frac{R}{\pi_0(V)} exp(-\overrightarrow{0}^T \tilde{V})\tilde{V} \\ =& \frac{R}{\pi_0(V)} \exp(-\gamma_v^{*T} \tilde{V})\tilde{V}\tilde{V}^T(\gamma_1 - \overrightarrow{0})\end{aligned}$$

where γ_v^* is on the line segment between 0 and γ_1. We use notation γ_v^* to suggest this value depends on V. As a result,

$$Q \exp(-\gamma_v^{*T} \tilde{V})\tilde{V}\tilde{V}^T(\gamma_1 - \overrightarrow{0}) = Q\left[\psi^{**}_{2,\gamma_1}(V,R) - \psi^{**}_{2,0}(V,R)\right] = 0.$$

Given that $Q\tilde{V}\tilde{V}^T$ is positive definite, $\gamma_1^T Q\tilde{V}\tilde{V}^T\gamma_1 > 0$ unless $\gamma_1 = 0$. Since Γ is a compact convex subset of $\mathbb{R}^q$ with 0 as an interior point, γ_v^* belongs to Γ and is bounded. By assumption A.2.6.1, $\tilde{V}$ is bounded. Thus $\exp(-\gamma_v^{*T}\tilde{V})$ is always positive, bounded and bounded from 0. Therefore, $\gamma_1^T Q \exp(-\gamma_v^{*T}\tilde{V})\tilde{V}\tilde{V}^T\gamma_1 > 0$ unless $\gamma_1 = 0$. By contradiction, we prove $\gamma = 0$ is the unique solution to $Q\psi^{**}_{2,\gamma}(V,R) = 0$.

When $\gamma_0 = 0$,

$$Q\psi^{**}_{1,\alpha,\gamma}(R,V,X) = Q\frac{R}{\pi_0(V)}\psi_\alpha(X) = P\psi_\alpha(X)$$

as shown in (2.8). Since α_0 is the unique solution to (2.6), it is also the unique solution to $Q\psi^{**}_{1,\alpha,0}(R,V,X) = 0$. Thus $(\alpha_0, 0)$ is the unique solution to equation (2.22) □

2.6.3 Consistency

Before proving the consistency of $\hat{\alpha}^{**}$, we first prove a lemma that will be used repeatedly.

Lemma 2.6.2. *Suppose Γ is a compact convex subset of $\mathbb{R}^q$ with 0 as an interior point. Assume A.2.2.2 and A.2.6.1. Let $f_\gamma(V,R) = \frac{R}{\pi_0(V)} exp(-\gamma^T\tilde{V})$. Then $\{f_\gamma, \gamma \in \Gamma\}$ is uniformly Lipschitz and uniformly bounded.*

Proof. Because Γ is convex, for every $\gamma_1, \gamma_2 \in \Gamma$ we can find a $\gamma_v^* \in \Gamma$ such that

$$|f_{\gamma_1}(R, X, V) - f_{\gamma_2}(R, X, V)| = |\frac{R}{\pi_0(V)} \exp(-\gamma_v^{*T} \tilde{V}) \tilde{V}^T (\gamma_1 - \gamma_2)|$$

By Cauchy-Schwarz inequality,

$$|\frac{R}{\pi_0(V)} \exp(-\gamma_v^{*T} \tilde{V}) \tilde{V}^T (\gamma_1 - \gamma_2)| \leq \|\frac{R}{\pi_0(V)} \exp(-\gamma_v^{*T} \tilde{V}) \tilde{V}^T\|_{\mathbb{E}} \|\gamma_1 - \gamma_2\|_{\mathbb{E}}.$$

Since Γ is compact, thus it is bounded. Under the assumption of A.2.2.2 and A.2.6.1, both $1/\pi_0(V)$ and $\tilde{V}$ are bounded. Therefore, $\{f_\gamma, \gamma \in \Gamma\}$ is uniformly bounded and there exists a positive constant C such that

$$|f_{\gamma_1}(R, X, V) - f_{\gamma_2}(R, X, V)| \leq C \|\gamma_1 - \gamma_2\|_{\mathbb{E}}.$$

Hence the class of functions $\{f_\gamma, \gamma \in \Gamma\}$ is also uniformly Lipschitz. □

Theorem 2.6.3. *(Consistency of improved two-phase VPS estimators)*
Suppose Γ *is a compact convex subset of* $\mathbb{R}^q$ *with 0 as an interior point. Assume that A.2.2.2, A.2.2.3, A.2.6.1 and A.2.6.2 hold. If conditions 2.4.2 and 2.4.3 are satisfied, and class* $\mathcal{G}$ *in condition 2.4.3 has integrable envelope function, then* $\hat{\alpha}^{**} \to_p \alpha_0$.

Proof. We will prove this theorem by adapting theorem 2.4.1 to the new maps

$$\Psi_N^{**}, \Psi^{**} : \mathbb{L}_0 \times \Gamma \mapsto \mathbb{L} \times \mathbb{R}^q.$$

$\mathbb{L}$ is a Banach space with the sup norm and $\mathbb{R}^q$ is the Euclidean space. We first verify condition 2.4.1. By assumption, $\mathcal{G}$ is Glivenko-Cantelli(GC) with integrable envelope function. Lemma 2.6.2 shows under the assumption of A.2.2.2 and A. 2.6.1 that $\{f_\gamma(V, R), \gamma \in \Gamma \subset \mathbb{R}^q\}$ is uniformly Lipschitz and is uniformly bounded. Theorem 2.4.1 and theorem 2.7.11 [van der Vaart and Wellner, 1996] together yield that, $\{f_\gamma, \gamma \in \Gamma\}$ is also GC, with integrable envelope function. By definition and assumption A. 2.6.1, each coordinate of $\tilde{V}$ itself forms a bounded GC class with one element. According to van der Vaart [1998, p. 270], the singleton $\{\tilde{V}\}$ is GC and bounded. It follows from preservation theorem 3 for Glivenko-Cantelli from van der

Vaart and Wellner [2000] that both classes $\{\psi^{**}_{1,\alpha,\gamma,h} = f_\gamma \psi_{\alpha,h}, \alpha \times \gamma \in \mathbb{L}_0 \times \Gamma, h \in \mathcal{H}\}$ and $\{\psi^{**}_{2,\gamma} = f_\gamma(V,R)\tilde{V} - \tilde{V}, \gamma \in \Gamma\}$ are GC. As a result,

$$\sup_{(\alpha,\gamma)\in\mathbb{L}_0\times\Gamma, h\in\mathcal{H}} |\mathbb{P}_N \psi^{**}_{1,\alpha,\gamma,h} - Q\psi^{**}_{1,\alpha,\gamma,h}| \to_p 0,$$

$$\sup_{\gamma\in\Gamma} \|\mathbb{P}_N \psi^{**}_{2,\gamma} - Q\psi^{**}_{2,\gamma}\|_{\mathbb{E}} \to_p 0.$$

Then

$$\begin{aligned}
&\sup_{(\alpha,\gamma)\in\mathbb{L}_0\times\Gamma} \|\Psi^{**}_N(\alpha,\gamma) - \Psi^{**}(\alpha,\gamma)\|_{\mathbb{L}\times\mathbb{R}^q} \\
=& \sup_{(\alpha,\gamma)\in\mathbb{L}_0\times\Gamma} \left\{ \|\Psi^{**}_{N,1}(\alpha,\gamma) - \Psi^{**}_1(\alpha,\gamma)\|_{\mathcal{H}} + \|\Psi^{**}_{N,2}(\alpha,\gamma) - \Psi^{**}_2(\alpha,\gamma)\|_{\mathbb{E}} \right\} \\
\leq& \sup_{(\alpha,\gamma)\in\mathbb{L}_0\times\Gamma} \|\Psi^{**}_{N,1}(\alpha,\gamma) - \Psi^{**}_1(\alpha,\gamma)\|_{\mathcal{H}} + \sup_{\gamma\in\Gamma} \|\Psi^{**}_{N,2}(\alpha,\gamma) - \Psi^{**}_2(\alpha,\gamma)\|_{\mathbb{E}} \\
=& \sup_{(\alpha,\gamma)\in\mathbb{L}_0\times\Gamma, h\in\mathcal{H}} |\mathbb{P}_N \psi^{**}_{1,\alpha,\gamma} - Q\psi^{**}_{1,\alpha,\gamma}| + \sup_{\gamma\in\Gamma} \|\mathbb{P}_N \psi^{**}_{2,\gamma} - Q\psi^{**}_{2,\gamma}\|_{\mathbb{E}} \to_p 0.
\end{aligned}$$

Therefore condition 2.4.1 is still satisfied after we adapt theorem 2.4.1 to the new maps Ψ^{**}_N, Ψ^{**}.

Next we check condition 2.4.2 after we adapt it to Ψ^{**}. We divide the set $\{\alpha, \gamma : \|\alpha - \alpha_0\| + \|\gamma - 0\| \geq \epsilon\}$ into two parts

$$A \equiv \{\gamma = 0\} \cap \{\alpha, \gamma : \|\alpha - \alpha_0\| + \|\gamma - 0\| \geq \epsilon\}$$

$$B \equiv \{\gamma \neq 0\} \cap \{\alpha, \gamma : \|\alpha - \alpha_0\| + \|\gamma - 0\| \geq \epsilon\}$$

We examine infima of $\|\Psi^{**}(\alpha,\gamma) - \Psi^{**}(\alpha_0, 0)\|_{\mathbb{L}\times\mathbb{R}^q}$ over sets A and B separately. Under the assumption of A.2.2.3, (2.8) holds. Then

$$\begin{aligned}
&\inf_A \|\Psi^{**}(\alpha,\gamma) - \Psi^{**}(\alpha_0,0)\|_{\mathbb{L}\times\mathbb{R}^q} \\
=& \inf_{\{\gamma=0\}\cap\{\alpha,\gamma:\|\alpha-\alpha_0\|+\|\gamma-0\|\geq\epsilon\}} \|Q\psi^{**}_{1,\alpha,\gamma} - Q\psi^{**}_{1,\alpha_0,0}\|_{\mathcal{H}} + \inf_{\{\gamma=0\}\cap\{\alpha,\gamma:\|\alpha-\alpha_0\|+\|\gamma-0\|\geq\epsilon\}} \|Q\psi^{**}_{2,\gamma} - Q\psi^{**}_{2,0}\|_{\mathbb{E}} \\
=& \inf_{\alpha:\|\alpha-\alpha_0\|\geq\epsilon} \|Q\psi^{*}_{1,\alpha,0} - Q\psi^{*}_{1,\alpha_0,0}\|_{\mathcal{H}} + 0 \\
=& \inf_{\alpha:\|\alpha-\alpha_0\|\geq\epsilon} \|P\psi_\alpha - P\psi_{\alpha_0}\|_{\mathcal{H}}
\end{aligned}$$

Since condition 2.4.2 is satisfied, $\inf_{\alpha:\|\alpha-\alpha_0\|\geq\epsilon}\|P\psi_\alpha - P\psi_{\alpha_0}\|_{\mathcal{H}} > 0$. Thus,

$$\inf_A \|\Psi^{**}(\alpha,\gamma) - \Psi^{**}(\alpha_0,0)\|_{\mathbb{L}\times\mathbb{R}^q} > 0.$$

Taking infimum over set B we obtain

$$\begin{aligned}
&\inf_B \|\Psi^{**}(\alpha,\gamma) - \Psi^{**}(\alpha_0,0)\|_{\mathbb{L}\times\mathbb{R}^q} \\
=&\inf_{\{\gamma\neq 0\}\cap\{\alpha,\gamma:\|\alpha-\alpha_0\|+\|\gamma-0\|\geq\epsilon\}} \|Q\psi^{**}_{1,\alpha,\gamma} - Q\psi^{**}_{1,\alpha_0,0}\|_{\mathcal{H}} + \inf_{\{\gamma\neq 0\}\cap\{\alpha,\gamma:\|\alpha-\alpha_0\|+\|\gamma-0\|\geq\epsilon\}} \|Q\psi^{**}_{2,\gamma} - Q\psi^{**}_{2,0}\|_{\mathbb{E}} \\
\geq&\inf_{\{\gamma\neq 0\}\cap\{\alpha,\gamma:\|\alpha-\alpha_0\|+\|\gamma-0\|\geq\epsilon\}} \|Q\psi^{**}_{2,\gamma} - Q\psi^{**}_{2,0}\|_{\mathbb{E}} \\
\geq&\inf_{\gamma:\|\gamma-0\|\geq\epsilon} \|Q\psi^{**}_{2,\gamma} - Q\psi^{**}_{2,0}\|_{\mathbb{E}}
\end{aligned}$$

In view of lemma 2.6.1, under our assumptions, 0 is the unique solution to $Q\psi^{**}_{2,\gamma}(V,R) = 0$. According to van der Vaart [1998, p.46], for a compact set Γ and a continuous function $Q\psi^{**}_{2,\gamma}(V,R)$, unique solution 0 implies $\inf_{\gamma:\|\gamma-0\|\geq\epsilon} \|Q\psi^{**}_{2,\gamma} - Q\psi^{**}_{2,0}\|_{\mathbb{E}} > 0$. As a result,

$$\inf_B \|\Psi^{**}(\alpha,\gamma) - \Psi^{**}(\alpha_0,0)\|_{\mathbb{L}} > 0.$$

Therefore,

$$\begin{aligned}
&\inf_{(\alpha,\gamma):\|(\alpha,\gamma)-(\alpha_0,0)\|\geq\epsilon} \|\Psi^{**}(\alpha,\gamma) - \Psi^{**}(\alpha_0,0)\|_{\mathbb{L}\times\mathbb{R}^q} \\
=&\inf_{A\cup B} \|\Psi^{**}(\alpha,\gamma) - \Psi^{**}(\alpha_0,0)\|_{\mathbb{L}\times\mathbb{R}^q} \\
=&\min\left\{\inf_A \|\Psi^{**}(\alpha,\gamma) - \Psi^{**}(\alpha_0,0)\|_{\mathbb{L}\times\mathbb{R}^q}, \inf_B \|\Psi^{**}(\alpha,\gamma) - \Psi^{**}(\alpha_0,0)\|_{\mathbb{L}\times\mathbb{R}^q}\right\} > 0.
\end{aligned}$$

Condition 2.4.2 is satisfied when we adapt it to the new map Ψ^{**} and the new parameter (α,γ). Finally, in view of our estimating equation (2.21), $\Psi^{**}_N(\hat{\alpha}^{**},\hat{\gamma}) = 0$. Hence it follows from theorem 2.4.1 that $(\hat{\alpha}^{**},\hat{\gamma}) \to_p (\alpha_0,0)$, i.e., $\hat{\alpha}^{**} \to_p \alpha_0$. □

Remark 1. *By using preservation theorems of GC [van der Vaart and Wellner, 2000], we are able to establish a GC class systematically by decomposing this class into small classes and establish the GC property of each small class separately. This systematic approach will be useful when a complicated class of functions is considered.*

2.6.4 Limiting Distribution

Corollary 2.6.4. *(Asymptotic normality of improved two-phase VPS estimators) Suppose* Γ *is a compact convex subset of* $\mathbb{R}^q$ *with 0 as an interior point. For each* $\alpha \in \mathbb{L}_0$ *and every* h *in an arbitrary set* $\mathcal{H}$*, let* $x \mapsto \psi_{\alpha,h}(x)$ *be a measurable function such that conditions 2.4.4* $\sim$ *2.4.6 are all satisfied and* $\mathcal{F}$ *in condition 2.4.4 has integrable envelope function. If* $(\hat{\alpha}^{**}, \hat{\gamma}) \to_p (\alpha_0, 0)$*, then*

$$\dot{\Psi}_0^{**}\sqrt{N}\begin{pmatrix} \hat{\alpha}^{**} - \alpha_0 \\ \hat{\gamma} - 0 \end{pmatrix} = -\mathbb{G}_N\psi_{\alpha_0,0}^{**} + o_p(1).$$

$\dot{\Psi}_0^{**}$ *is the Fréchet derivative of the map* $\Psi^{**} : \mathbb{L}_0 \times \Gamma \mapsto \mathbb{L} \times \mathbb{R}^q$ *that has an continuous inverse on its range.* $\dot{\Psi}_0^{**} : lin\mathbb{L}_0 \times lin\Gamma \mapsto \mathbb{L} \times \mathbb{R}^q$ *takes the form*

$$(\alpha - \alpha_0, \gamma - \gamma_0) \mapsto \begin{pmatrix} \dot{\Psi}_{11}^{**} & \dot{\Psi}_{12}^{**} \\ \dot{\Psi}_{21}^{**} & \dot{\Psi}_{22}^{**} \end{pmatrix}\begin{pmatrix} \alpha - \alpha_0 \\ \gamma - \gamma_0 \end{pmatrix}$$

and

$$\dot{\Psi}_0^{**} = \begin{pmatrix} \dot{\Psi}_{11}^{**} & \dot{\Psi}_{12}^{**} \\ \dot{\Psi}_{21}^{**} & \dot{\Psi}_{22}^{**} \end{pmatrix} = \begin{pmatrix} \dot{\Psi}_0 & -Q\psi_{\alpha_0}\tilde{V}^T \\ 0 & -Q\tilde{V}\tilde{V}^T \end{pmatrix}. \tag{2.23}$$

Proof. We prove this theorem by theorem 2.4.4 with several adaptations. In the above theorem, the parameter and the range of the considered maps are all elements in a product space $l^\infty(\mathcal{H}) \times \mathbb{R}^q$. If we consider a new general direction $c = (h, b) \in \mathcal{H} \times B$ where B is a bounded subset of $\mathbb{R}^q$, it is not difficult to verify that the parameter (α, γ) and $(\psi_{1,\alpha,\gamma}^{**}, \psi_{2,\gamma}^{**})$ also belong to $l^\infty(\mathcal{H} \times B)$, and so does Ψ^{**}. Thus theorem 2.4.4 still can be adapted to our new parameters and new maps. In the following we adapt condition 2.4.4 $\sim$ 2.4.6 in theorem 2.4.4 to (α, γ) and $(\psi_{1,\alpha,\gamma,h}^{**}, \psi_{2,\gamma}^{**}) \in \mathbb{R}^{q+1}$, and show these conditions after adaptation are still valid. Then by this theorem, we will obtain the desired results about the improved two-phase sampling estimate $\hat{\alpha}^{**}$ of α_0.

- New condition 2.4.4

Let $\mathcal{F}'' = \mathcal{F}_1 \cup \mathcal{F}_2$ where $\mathcal{F}_1 = \{\psi^{**}_{1;\alpha,\gamma,h}, (\alpha,\gamma) \in \mathbb{L}_0 \times \Gamma, h \in \mathcal{H}\}$ and $\mathcal{F}_2 = \{\psi^{**}_{2,\gamma}, \gamma \in \Gamma\}$. We first verify that $\mathcal{F}_2$ is Donsker. By definition and assumption A.2.6.1, a singleton $\{\tilde{V}\}$ is Donsker with integrable envelope function. In lemma 2.6.2, we have shown $\{f_\gamma, \gamma \in \Gamma\}$ is uniformly bounded and

$$|f_{\gamma_1}(R,X,V) - f_{\gamma_2}(R,X,V)| \leq C\|\gamma_1 - \gamma_2\|$$

where C is a positive constant. Applying example 19.7 [van der Vaart, 1998], $\{f_\gamma, \gamma \in \Gamma\}$ is P-Donsker with integrable envelope function. Since condition 2.4.4 is satisfied, $\{\psi_{\alpha,h} : \alpha \in \mathbb{L}_0, h \in \mathcal{H}\}$ is also P-Donsker. By assumption, it also has integrable envelope function. Because $\psi^{**}_{1;\alpha,\gamma,h} = f_\gamma \psi_{\alpha,h}$ and $\psi^{**}_{2;\gamma} = f_\gamma \tilde{V} - \tilde{V}$, according to example 19.20 [van der Vaart, 1998], one of the preservation theorems of Donsker[van der Vaart and Wellner, 1996], $\mathcal{F}_1$ and $\mathcal{F}_2$ are both Q-Donsker. Applying [van der Vaart and Wellner, 1996, example 2.10.7] yields $\mathcal{F}''$ is Q-Donsker with finite envelope function.

- New condition 2.4.5

  First, we verify $\dot{\Psi}^{**}_{11} = \dot{\Psi}_0$ and $\dot{\Psi}^{**}_{21} = 0$. Based on the definition of Fréchet derivative, condition 2.4.5 implies that $\dot{\Psi}_0$ as a continuous and linear map satisfies

  $$\|P\psi_\alpha(X) - P\psi_{\alpha_0}(X) - \dot{\Psi}_0(\alpha - \alpha_0)\| = o(\|\alpha - \alpha_0\|), \qquad \text{as} \qquad \|\alpha - \alpha_0\| \downarrow 0.$$

  According to (2.8), $Q\psi^*_\alpha(R,V,X) = P\psi_\alpha(X)$. Note $\Psi^{**}_1(\alpha,\gamma_0) = Q\psi^{**}_{1,\alpha,0}(R,V,X) = Q\psi^*_\alpha(R,V,X)$. Hence $\Psi^{**}_1(\alpha,\gamma_0) = P\psi_\alpha(X)$. Then

  $$\|\Psi^{**}_1(\alpha,\gamma_0) - \Psi^{**}_1(\alpha_0,\gamma_0) - \dot{\Psi}_0(\alpha-\alpha_0)\| = \|P\psi_\alpha(X) - P\psi_{\alpha_0}(X) - \dot{\Psi}_0(\alpha-\alpha_0)\| = o(\|\alpha-\alpha_0\|).$$

  As a result,

  $$\begin{aligned}
  &\|\Psi^{**}_1(\alpha,\gamma) - \Psi^{**}_1(\alpha_0,\gamma) - \dot{\Psi}_0(\alpha-\alpha_0)\| \\
  =&\|\Psi^{**}_1(\alpha,\gamma) - \Psi^{**}_1(\alpha,\gamma_0) + \Psi^{**}_1(\alpha,\gamma_0) - \Psi^{**}_1(\alpha_0,\gamma_0) + \Psi^{**}_1(\alpha_0,\gamma_0) - \Psi^{**}_1(\alpha_0,\gamma) - \dot{\Psi}_0(\alpha-\alpha_0)\| \\
  \leq&\|\Psi^{**}_1(\alpha,\gamma) - \Psi^{**}_1(\alpha,\gamma_0)\| + \|\Psi^{**}_1(\alpha_0,\gamma_0) - \Psi^{**}_1(\alpha_0,\gamma)\| + \|\Psi^{**}_1(\alpha,\gamma_0) - \Psi^{**}_1(\alpha_0,\gamma_0) - \dot{\Psi}_0(\alpha-\alpha_0)\| \\
  =&0 + 0 + o(\|\alpha-\alpha_0\|) \text{ at } (\alpha_0,\gamma_0).
  \end{aligned}$$

Thus $\|\Psi_1^{**}(\alpha,\gamma) - \Psi_1^{**}(\alpha_0,\gamma) - \dot{\Psi}_0(\alpha-\alpha_0)\| = o(\|\alpha-\alpha_0\|)$ at (α_0,γ_0) as $\|\alpha-\alpha_0\| \downarrow 0$. By definition, $\dot{\Psi}_0$ is the Fréchet derivative of Ψ_1^{**} with respect to α at (α_0,γ_0). Since Ψ_2^{**} does not involve α, the Fréchet derivative of Ψ_2^{**} with respect to α is 0.

Second, we verify $\dot{\Psi}_{12}^{**} = -Q\psi_{\alpha_0}\tilde{V}^T$ and $\dot{\Psi}_{22}^{**} = -Q\tilde{V}\tilde{V}^T$. For each $h \in \mathcal{H}$, we have

$$\Psi_1^{**}(\alpha,\gamma)h = Q\psi_{1,\alpha,\gamma,h}^{**} = Q\left[\frac{R}{\pi_0(V)}\exp(-\gamma^T\tilde{V})\psi_{\alpha,h}(X)\right] \in \mathbb{R}$$
$$\Psi_2^{**}(\alpha,\gamma) = Q\psi_{2,\gamma}^{**} = Q\left[\frac{R}{\pi_0(V)}\exp(-\gamma^T\tilde{V})\tilde{V} - \tilde{V}\right] \in \mathbb{R}^q.$$

Since $\tilde{V}$ and γ are bounded and $\{\psi_{\alpha,h}, \alpha \in \mathbb{L}_0, h \in \mathcal{H}\}$ has integrable envelope function, by dominated convergence theorem and rules of differentiation on the Euclidean space we obtain

$$\begin{aligned}
\nabla\dot{\Psi}_1^{**}(\alpha_0,0)(\gamma-\gamma_0)h =& \nabla Q\left[\frac{R}{\pi_0(V)}\exp(-\gamma_0^T\tilde{V})\psi_{\alpha,h}(X)\right](\gamma-\gamma_0)\\
=& Q\left[\nabla\frac{R}{\pi_0(V)}\exp(-\gamma_0^T\tilde{V})\psi_{\alpha,h}(X)\right](\gamma-\gamma_0)\\
=& -Q\left[\psi_{\alpha_0,h}(X)\tilde{V}^T\right](\gamma-\gamma_0)
\end{aligned}$$

and

$$\begin{aligned}
\nabla\dot{\Psi}_2^{**}(\alpha_0,0)(\gamma-\gamma_0)h =& \nabla Q\left[\frac{R}{\pi_0(V)}\exp(-\gamma_0^T\tilde{V})\tilde{V} - \tilde{V}\right](\gamma-\gamma_0)\\
=& Q\nabla\left[\frac{R}{\pi_0(V)}\exp(-\gamma^T\tilde{V})\tilde{V} - \tilde{V}\right](\gamma-\gamma_0)\\
=& -Q[\tilde{V}\tilde{V}^T](\gamma-\gamma_0).
\end{aligned}$$

Because on a Euclidean space, Fréchet derivative agrees with differential, we have

$$\dot{\Psi}_{22}^{**}(\gamma-\gamma_0) = -Q[\tilde{V}\tilde{V}^T](\gamma-\gamma_0).$$

For each $h \in \mathcal{H}$, by Taylor expansion and dominated convergence theorem

$$\begin{aligned}
\Psi_1^{**}(\alpha,\gamma)h =& \Psi_1^{**}(\alpha,\gamma_0)h - Q\left[\psi_{\alpha,h}\tilde{V}^T\right](\gamma-\gamma_0)\\
&+ 1/2(\gamma-\gamma_0)^T Q[\tilde{V}\tilde{V}^T\psi_{\alpha,h}\exp(-\gamma_v^T\tilde{V})](\gamma-\gamma_0)
\end{aligned}$$

Then

$$\begin{aligned}
&\|\Psi_1^{**}(\alpha,\gamma) - \Psi_1^{**}(\alpha,\gamma_0) + Q\left[\psi_{\alpha_0}\tilde{V}^T\right](\gamma-\gamma_0)\| \\
=&\sup_{h\in\mathcal{H}}|\Psi_1^{**}(\alpha,\gamma)h - \Psi_1^{**}(\alpha,\gamma_0)h + Q\left[\psi_{\alpha_0,h}(X)\tilde{V}^T\right](\gamma-\gamma_0)| \\
=&(\gamma-\gamma_0)^T\sup_{h\in\mathcal{H}}|1/2Q[\tilde{V}\tilde{V}^T\psi_{\alpha_0,h}\exp(-\gamma_v^T\tilde{V})]|(\gamma-\gamma_0) \\
=&o(\|\gamma-\gamma_0\|) \text{ as } \quad \|\gamma-\gamma_0\|\downarrow 0.
\end{aligned}$$

Therefore, $-Q\psi_{\alpha_0}\tilde{V}^T$ is the Fréchet derivative of Ψ_1^{**} with respect to γ at (α_0,γ_0) and $-Q\tilde{V}\tilde{V}^T$ is the Fréchet derivative of Ψ_2^{**} with respect to γ at (α_0,γ_0).

In summary, the Fréchet derivative of the map $(\alpha,\gamma)\mapsto\Psi^{**}(\alpha,\gamma)$ at (α_0,γ_0) is in the form of

$$\dot{\Psi}_0^{**} = \begin{pmatrix} \dot{\Psi}_{11}^{**} & \dot{\Psi}_{12}^{**} \\ \dot{\Psi}_{21}^{**} & \dot{\Psi}_{22}^{**} \end{pmatrix} = \begin{pmatrix} \dot{\Psi}_0 & -Q\psi_{\alpha_0}(X)\tilde{V}^T \\ 0 & -Q\tilde{V}\tilde{V}^T \end{pmatrix}.$$

Finally, we show $\dot{\Psi}_0^{**}$ has a continuous inverse on its range. Given condition 2.4.5, and the result $\dot{\Psi}_{11}^{**} = \dot{\Psi}_0$, $\dot{\Psi}_{11}^{**}$ has a continuous inverse. $Q\tilde{V}\tilde{V}^T$ is continuous. Under assumption A.2.6.2, the null space of $Q\tilde{V}\tilde{V}^T$ is 0. By assumption A.2.6.1, $\tilde{V}$ is bounded. Since the domain of $Q\tilde{V}\tilde{V}^T$ is LinΓ and it is closed, thus the range of $Q\tilde{V}\tilde{V}^T$ is closed. According to Bickel et al. [1998, proposition 7B in Appendix], $Q\tilde{V}\tilde{V}^T$ has an continuous inverse. Since $\dot{\Psi}_{21}^{**} = 0$,

$$\dot{\Psi}_{22}^{**} - \dot{\Psi}_{21}^{**}\dot{\Psi}_{11}^{**^{-1}}\dot{\Psi}_{12}^{**} = \dot{\Psi}_{22}^{**}$$

Thus $\dot{\Psi}_{22}^{**} - \dot{\Psi}_{21}^{**}\dot{\Psi}_{11}^{**^{-1}}\dot{\Psi}_{12}^{**}$ has a continuous inverse on its range. Applying the argument on p.422 from van der Vaart [1998] for ascertaining continuous invertibility of partitioned $\dot{\Psi}_0^{**}$, we obtain $\dot{\Psi}_0^{**}$ has continuous inverse on its range.

- New condition 2.4.6

  Because condition 2.4.6 is satisfied, we have

$$\sup_{h\in\mathcal{H}} P(\psi_{\alpha,h} - \psi_{\alpha_0,h})^2 \to 0, \alpha\to\alpha_0.$$

Because $\exp(-\gamma^T\tilde{v}) \to \exp(-\gamma_0^T\tilde{v})$ as $\gamma \to \gamma_0$ and $R, 1/\pi_0(V), \gamma$ and $\tilde{V}$ are all bounded, by dominated convergence theorem

$$Q[\exp(-\gamma^T\tilde{V}) - \exp(-\gamma_0^T\tilde{V})]^2 \to 0, \text{ as } \gamma \to \gamma_0,$$
$$Q[\psi^{**}_{2;\gamma} - \psi^{**}_{2;\gamma_0}]^2 = Q\left[\frac{R}{\pi_0(V)^2}\tilde{V}\tilde{V}^T\{\exp(-\gamma^T\tilde{V}) - \exp(-\gamma_0^T\tilde{V})\}\right] \to 0, \text{ as } \gamma \to \gamma_0.$$

Since for each fixed x and α, the map $\psi_\alpha : h \mapsto \psi_{\alpha,h}$ is uniformly bounded,

$$\begin{aligned}
&\sup_{h\in\mathcal{H}} Q(\psi^{**}_{1,\alpha,\gamma,h} - \psi^{**}_{1,\alpha_0,\gamma_0,h})^2 \\
=&\sup_{h\in\mathcal{H}} Q(\psi^{**}_{1,\alpha,\gamma,h} - \psi^{**}_{1,\alpha_0,\gamma,h} + \psi^{**}_{1,\alpha_0,\gamma,h} - \psi^{**}_{1,\alpha_0,\gamma_0,h})^2 \\
\le&\sup_{h\in\mathcal{H}} 2Q(\psi^{**}_{1,\alpha,\gamma,h} - \psi^{**}_{1,\alpha_0,\gamma,h})^2 + \sup_{h\in\mathcal{H}} 2Q(\psi^{**}_{1,\alpha_0,\gamma,h} - \psi^{**}_{1,\alpha_0,\gamma_0,h})^2 \\
=&2\sup_{h\in\mathcal{H}} Q\left[\frac{R^2}{\pi_0(V)^2}\exp(-\gamma^T\tilde{V})^2(\psi_{\alpha,h} - \psi_{\alpha_0,h})^2\right] \\
&\qquad + 2\sup_{h\in\mathcal{H}} Q\left[\frac{R^2}{\pi_0(V)^2}\psi^2_{\alpha_0,h}\left[\exp(-\gamma^T\tilde{V}) - \exp(-\gamma_0^T\tilde{V})\right]^2\right] \\
\le&C_1\sup_{h\in\mathcal{H}}|P(\psi_{\alpha,h} - \psi_{\alpha_0,h})^2| + C_2Q\{\exp(-\gamma^T\tilde{V}) - \exp(-\gamma_0^T\tilde{V})\}^2 \\
\to&0 \text{ as } (\alpha,\gamma) \to (\alpha_0,\gamma_0)
\end{aligned}$$

where C_1 and C_2 are positive constants.

Therefore,

$$\sup_{h\in\mathcal{H}} Q(\psi^{**}_{1,\alpha,\gamma,h} - \psi^{**}_{1,\alpha_0,\gamma_0,h})^2 \to 0 \text{ and } Q[\psi^{**}_{2;\gamma} - \psi^{**}_{2;\gamma_0}]^2 \to 0, \text{ as } (\alpha,\gamma) \to (\alpha_0,\gamma_0).$$

In view of (2.21), $\|\mathbb{P}_N(\psi_{1;\hat{\alpha}^{**},\hat{\gamma}}, \psi_{2;\hat{\gamma}})\| = 0$. It follows from theorem 2.4.4 that if $(\hat{\alpha}^{**}, \hat{\gamma}) \to_p (\alpha_0, 0)$, then

$$\sqrt{N}\dot{\Psi}^{**}_0\begin{pmatrix}\hat{\alpha}^{**} - \alpha_0 \\ \hat{\gamma} - 0\end{pmatrix} = -\mathbb{G}_N\begin{pmatrix}\psi^{**}_{1,\alpha_0,0} \\ \psi^{**}_{2,0}\end{pmatrix} + o_p(1).$$

□

Lastly, we derive the asymptotic variance of $\sqrt{N}\dot{\Psi}_0(\hat{\alpha}^{**} - \alpha_0)$. The preceding display implies

$$\sqrt{N}\begin{pmatrix} \dot{\Psi}_0 & -Q\psi_{\alpha_0}(X)\tilde{V}^T \\ 0 & -Q\tilde{V}\tilde{V}^T \end{pmatrix}\begin{pmatrix} \hat{\alpha}^{**} - \alpha_0 \\ \hat{\gamma} - 0 \end{pmatrix} h = -\mathbb{G}_N \begin{pmatrix} \psi^{**}_{1,\alpha_0,0,h} \\ \psi^{**}_{2,0} \end{pmatrix} + o_p(1) \text{ for all } h \in \mathcal{H}$$

i.e.,

$$\sqrt{N}\dot{\Psi}_0(\hat{\alpha} - \alpha_0)h - \sqrt{N}Q\{\psi_{\alpha_0,h}(X)\tilde{V}^T\}(\hat{\gamma} - 0) = -\mathbb{G}_N\psi^{**}_{\alpha_0,h} + o_p(1) \tag{2.24}$$

$$-\sqrt{N}Q\tilde{V}\tilde{V}^T(\hat{\gamma} - 0) = -\mathbb{G}_N\psi^{**}_{2,0} + o_p(1). \tag{2.25}$$

Applying $Q\{\psi_{\alpha_0,h}(X)\tilde{V}^T\}(Q\tilde{V}\tilde{V}^T)^{-1}$ to both sides of (2.25), we obtain

$$-\sqrt{N}Q\{\psi_{\alpha_0,h}(X)\tilde{V}^T\}(\hat{\gamma} - 0) = -Q\{\psi_{\alpha_0,h}(X)\tilde{V}^T\}(Q\tilde{V}\tilde{V}^T)^{-1}\mathbb{G}_N\psi^{**}_{2,0} + o_p(1). \tag{2.26}$$

Subtracting (2.26) from (2.24) yields

$$\begin{aligned}\sqrt{N}\dot{\Psi}_0(\hat{\alpha}^{**} - \alpha_0)h =& -\mathbb{G}_N\psi^{*}_{\alpha_0,h} + Q\{\psi_{\alpha_0,h}(X)\tilde{V}^T\}(Q\tilde{V}\tilde{V}^T)^{-1}\mathbb{G}_N\psi^{**}_{2,0} + o_p(1) \\ =& -\mathbb{G}_N\frac{R}{\pi_0(V)}\psi_{\alpha_0,h} + \mathbb{G}_N\{Q\psi_{\alpha_0,h}(X)\tilde{V}^T\}(Q\tilde{V}\tilde{V}^T)^{-1}\{\frac{R}{\pi_0(V)} - 1\}\tilde{V} + o_p(1) \\ =& -\mathbb{G}_N\left[\psi_{\alpha_0,h} - \{\frac{R}{\pi_0(V)} - 1\}\left\{\psi_{\alpha_0,h} - \Pi(\psi_{\alpha_0,h}|\tilde{V})\right\}\right] + o_p(1) \end{aligned} \tag{2.27}$$

where $\Pi(\cdot|\tilde{V})$ refers to population least squares projection on the space spanned by the calibration variables $\tilde{V}$:

$$\Pi(\cdot|\tilde{V}) = Q\{\cdot\tilde{V}^T\}(Q\tilde{V}\tilde{V}^T)^{-1}\tilde{V}.$$

Using the orthogonality of the two terms in the square brackets of (2.27), $\sqrt{N}\dot{\Psi}_0(\hat{\alpha}^{**} - \alpha_0)$ converges in distribution to the Q-Brownian bridge in $l^\infty(\mathcal{H})$, which is a zero-mean Gaussian with variance

$$Var_A\{\sqrt{N}\dot{\Psi}_0(\hat{\alpha}^{**} - \alpha_0)h\} = P\psi^2_{\alpha_0,h} + Q\left[\frac{1 - \pi_0(V)}{\pi_0(V)}\left\{\psi_{\alpha_0,h} - \Pi(\psi_{\alpha_0,h}|\tilde{V})\right\}^2\right]. \tag{2.28}$$

This asymptotic variance is the robust variance since this result allows $P \in \mathcal{P}$ without assuming a particular model. Suppose an assumed model is believed to be the true model

for the observed data, i.e., $P \in \{P_\alpha, \alpha \in \mathbb{L}_0\}$. Then the model based variances can be obtained by replacing P in (2.28) by P_α.

Comparing the asymptotic variance for the calibrated two-phase VPS estimator in (2.28) to the two-phase VPS estimator in (2.14), we find that calibration will reduce the variance of a two-phase sampling estimator if

$$Q\left\{\frac{1-\pi_0(V)}{\pi_0(V)}\left[\psi_{\alpha_0,h}-\Pi(\psi_{\alpha_0,h}|\tilde{V})\right]^2\right\} \leq Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\psi^2_{\alpha_0,h}\right].$$

2.7 Summary of the Procedure

Two-phase designs have been proposed for epidemiological studies in order to substantially reduce the cost of a study. This chapter solved two general estimation problems for analyzing this type of data: 1) estimating all parameters in a possibly semiparametric model allowing for model misspecification; 2) incorporating auxiliary information into estimation to further improve estimation efficiency.

For the first problem, the estimation procedure is summarized below:

Step 1. Following section 2.2, assume two-phase VPS Lawless et al. [1999] for an obtained two-phase design dataset and identify variables R, V, X, U, X^I, X^{II}. Examine whether assumptions A.2.2.1$\sim$ A.2.2.3 hold for the data.

Step 2. Propose a function ψ_α, usually motivated (but not defined) by an assumed model. In order to have the desired asymptotic properties, this function is required to be uniformly bounded and to satisfy conditions 2.4.3 $\sim$ 2.4.6. Then following section 2.4.1, write estimating equation as (2.4) and obtain two-phase sampling estimators.

Step 3. Following section 2.6.2, define the parameter as a unique solution to (2.6) and show this solution is well-separated.

Step 4. Establish the consistency property of the two-phase VPS estimators proposed in (2.4) using corollary 2.4.3.

Step 5. Establish the asymptotic normality of these estimators using corollary 2.4.5 and then calculate the asymptotic variance according to (2.14) on and off the assumed model.

For the second problem, the estimation procedure is summarized as follows

Step 1. Choose auxiliary variables $\tilde{V}$ based on Phase I information; Examine the assumptions A.2.6.1 $\sim$ A.2.6.2 about $\tilde{V}$.

Step 2. Following section 2.6.1, write estimating equations as (2.21) and solve these equations for improved two-phase VPS estimators.

Step 3. The parameter remains as the solution to (2.6).

Step 4. Establish the consistency property of the improved two-phase VPS estimators using theorem 2.6.3

Step 5. Establish the asymptotic normality of these estimators using theorem 2.6.4 and then calculate the asymptotic variance according to (2.28) on and off the assumed model.

2.8 Discussion

In this chapter, we transformed several inference problems for two-phase epidemiologic studies to a single problem— estimating a parameter using i.i.d. observations. We then applied the simplest infinite-dimensional Z-estimation theorem to develop the asymptotic properties of these Z-estimators. During the developments of these Z-estimators, we took a systematic approach to derive their asymptotic properties.

The several inference problems we addressed are all practical problems. These problems include: 1)inference on regression coefficients in semiparametric models for association analysis, 2) estimation of all parameters in a semiparametric model for prediction purposes, 3) making inference based on the phase II samples with missing values for some cases, 4) use of auxiliary Phase I information to further improve estimation efficiency, 5) estimation

under model misspecification. Nan and Wellner [2013]'s Z-estimation method focused on the first problem. We had a systematic study on a variety of problems from two-phase sampling. In fact, Z-estimation itself is a practical choice of solving estimation problems, because sometimes estimators are motivated from methods of moment or the least square principle. In these circumstances, we do not have parameterized or partially parametrized family of distributions to start with for applying maximum likelihood estimation (MLE). Sometimes although we are able to write a likelihood based on an assumed model, solving the resulting score equations for estimators can be challenging such as the Lin & Ying's additive hazards model, to be studied in the next chapter.

We were able to solve all these inference problems by Z-estimation because Z-estimation, together with modern empirical process theory, is a very universal asymptotic analysis tool. We were able to use the simplest Z-estimation because we assumed variable probability sampling [Lawless et al., 1999]. We were able to have a systematic study of various Z-estimators because we recognized the Z-estimation theorem we invoked is an extension of Huber [1967]'s Z-estimation. Huber in this paper started a systematic asymptotic study of Z-estimators by decomposing the conditions of establishing asymptotic normality into two separate conditions. This approach was rarely noticed and used, but we exploited this systematic approach. We decomposed each big inference problem into small components ① $\sim$ ⑥ described at the end of section 2.5. These small tasks in each component can be first solved separately. Then these results are integrated together by the theorems and corollaries in this chapter, leading immediately to the desired asymptotic properties of the estimators.

The first benefit of this systematic approach is that the asymptotic analysis of estimators becomes transparent, particularly for solving complex problems. This is because small tasks are easy to follow and check. The transparency, as a result, will improve the quality of methodology research. The second benefit is that a systematic Z-estimation approach helps researches communicate and share their theoretical results. Since Z-estimation is universally applicable, the idea of developing various related estimators systematically using a single Z-estimation system facilitates us to identify existing results and use them for solving new

related problems. The asymptotic analysis based on the Taylor expansion technique does not have these two benefits.

By assuming i.i.d. two-phase VPS, several estimating procedures are simplified and standardized. Without assuming i.i.d. data, consistency is usually established case by case. In contrast, with i.i.d. data, we demonstrated the procedure of establishing consistency becomes systematic, which was decomposed into the verification of Glivenko-Cantelli(GC) condition and an analysis of a deterministic map. For i.i.d. data, the proof of asymptotic normality was also simplified. The stochastic equicontinuity condition was replaced by a Donsker condition and a condition to show convergence in quadratic mean [van der Vaart and Wellner, 1996]. The establishment of these two conditions are fairly standardized. More importantly, GC and Donsker conditions can be easily extended for solving more complex problems. For example, we demonstrated these straightforward extensions when developing two-phase sampling estimators $\hat{\alpha}^*$ and calibrated two-phase sampling estimators $\hat{\alpha}^{**}$ after establishing the GC and Donsker properties for developing RS estimator $\hat{\alpha}^*$. When a new complex scenario appears in the future, researchers can continue extending the existing results to the new circumstances using this Z-estimation system for i.i.d. data. Without assuming i.i.d. VPS, the ease of extension may not exist.

In addition, the establishments of our stochastic conditions, Glivenko-Cantelli (GC) and Donsker properties of classes of functions, also take a systematic approach. We decompose a complex class of functions into small classes and verify the Donsker or GC properties of each small class separately. By applying preservation theorems of Donsker [van der Vaart and Wellner, 1996, chap. 2.10], we integrate these small results together to establish the Donsker or GC of the complex class of functions. Sometimes Donsker or GC property of a basic class may have already been established. Then this Z-estimation system for i.i.d. data allows researchers without a relevant background for directly verifying GC or Donsker conditions to have a way to use modern empirical process results. The notions of GC and Donsker classes allow collaboration between the researchers who have a substantial knowledge of empirical process theory and the researchers whose primary interest lies in the application

of this theory. By this means, the powerful and abstract modern empirical process theory may become more accessible and applicable for a wide audience.

This systematic approach for methodology development will be very useful for teamwork. We see that after function ψ_α was proposed in component ①, each small task in each component ② − ⑥ does not rely on results from other components. This means these tasks can be assigned to different individuals. In the past, statistical methods were usually developed by an individual. However, the complexity of today's statistical problems and the demands for a short schedule may require a team effort for methods development soon. Our systematic approach provides this possibility.

Chapter 3

APPLICATION OF THE Z-ESTIMATION SYSTEM TO ADDITIVE HAZARDS MODELS

This chapter and the following chapter apply our Z-estimation system for two-phase sampling to Lin & Ying's additive hazards (AH) model [Lin and Ying, 1994]. This model specifies the hazard function of a censored failure time T as a sum of a baseline hazard function $\lambda(\cdot)$ and a regression function of Z:

$$\lambda(t|Z) = \lambda(t) + Z^T\theta \tag{3.1}$$

Our aim in this chapter is to estimate the cumulative baseline hazard $\Lambda(\cdot) = \int_0^{\cdot} \lambda(t)dt$ and the regression parameter θ simultaneously for both random sampling (RS) and two-phase variable probability sampling (VPS), so that we are able to estimate an individual's cumulative incidence by different times based on individual risk factors. In addition, estimation under model misspecification will also be considered.

This chapter has both pedagogical and practical purposes. For the pedagogical purposes, I will follow exactly the estimation procedure described in sections 2.4 $\sim$ 2.7 to develop our estimators for the AH model, so that readers will gain a better understanding on the theories from Chapter 2. Since the AH model is indexed by both Euclidean parameters θ and non-Euclidean parameter $\Lambda(\cdot) = \int_0^{\cdot} \lambda(t)dt$, this application demonstrates how to use our theory to develop finite-dimensional, infinite-dimensional estimators or both of them simultaneously. Sometimes the AH model may not correctly specify the distribution of an observed dataset. Through this application, readers can also learn how to develop robust variances of estimators under model misspecification by using our theory, as well as the interpretation of these estimators for two-phase designs.

For the practical purpose, the AH model is an important survival analysis tool for epidemiologic studies and public health prevention. Methods to fit this model to general two-phase sampling are desired and methods to fit this model under model misspecification are important. Since we simultaneously estimate θ and Λ, the cumulative hazard and therefore the survival curve can be derived immediately. In this chapter, we provide a method to predict an individual's cumulative incidence based on his/her risk factors in an AH model with both RS and two-phase VPS. This method will provide clinical researchers a tool to predict an individual's cumulative incidence or disease progression over time when an AH model is believed to be the true model. Although the AH model is recognized among several literature [Cox and Oakes, 1984, Thomas, 1986, Breslow and Day, 1987] as an important analytic tools for medical and epidemiological research, due to the lack of software developments and knowledge of this model among applied research community, this model was rarely used even for random sampling data since its inception in 1994. Thus, we not only develop a collection of useful estimators based on this model for both RS and two-phase VPS studies, but also implement these results in a R program for researchers to apply conveniently.

3.1 Background

As an alternative model to Cox's proportional hazards model, Lin & Ying's AH model describes a different aspect of the association between the failure time outcome and covariates. Cox's regression model provides approximate estimates of relative risks, while the additive hazard model provides approximate estimates for the excess risks. For example, Xie et al. [2013] used Lin & Ying's model to estimate the excess risk of human papillomavirus (HPV) infection associated with immune responses. Their analysis result shows on average an additional 14 oncogenic HPV infections per 100 women-years was associated with CDV < 200 relative to HIV-negative women. For the same data, analysis based on Cox's regression model provides an estimate of hazards ratio at 3.82 between the two groups with different immune responses. The two models offer different perspectives on the association and thus have different interpretations. For public health planning and intervention, the AH model

can be more meaningful. The excess risk tells the amount of absolute risk could be reduced if a risk factor were removed from the population. This interpretation is informative for public health practitioners.

In the regression analysis of cause-specific hazards as well as all-cause hazards, the additive hazard model has also been advocated. This is because under Cox's proportional hazards model if we assume a constant hazards ratio for each cause-specific hazard, the hazards ratio for all-cause mortality will no longer be constant [Klein, 2006]. The additive hazards model, on the other hand, solves this inconsistency problem.

Sometimes instead of using each model exclusively, Cox's model and Lin & Ying's model are used together to provide a more comprehensive understanding of the association between factors and outcomes. One example is that Lim and Zhang [2009] applied both models to the investigation of pediatric firearm injuries and the risk factors associated with repeated intentional injury. Their results show the same group of factors are identified by both models. The use of both models for association analysis makes the authors' conclusions very convincing .

When data are collected with random sampling (RS), Lin and Ying [1994] proposed a semiparametric estimator for θ based on martingale theory [Andersen and Gill, 1982]. However if data are collected with two-phase VPS and the phase II subsampling probability depends on outcomes, martingale theory does not apply. Therefore rather than using this classic theory, two different semiparametric inference methods were proposed [Kulich and Lin, 2000, Nan and Wellner, 2013] to fit the AH model to case-cohort studies [Prentice, 1986], which is a special case of general two-phase VPS. These two methods both derived their estimators from the weighted estimating equation proposed in Kulich and Lin [2000]. To derive the asymptotic properties for the estimator, Kulich and Lin [2000] used the Taylor expansion technique and Nan and Wellner [2013] applied the semi-parametric Z-estimation theory for bundled parameters proposed in the same paper. Later Kang et al. [2013] fitted marginal additive hazards regression models for case-cohort studies with multiple disease outcomes and considered a generalized case-cohort design that allows sampling among cases.

They adopted and extended Kulich and Lin [2000]'s approach to conduct the asymptotic analysis on their estimators.

All these methods for fitting Lin & Ying's model to case cohort designs are based on the estimating equation proposed in equation (2.7) of Lin and Ying [1994], which does not involve the baseline cumulative hazard $\Lambda(\cdot) = \int_0^{\cdot} \lambda(t)dt$. Thus Kulich and Lin [2000] and Kang et al. [2013] developed their estimator for Λ separately from θ, and Nan and Wellner [2013] did not include the baseline hazard function estimation. No work have been done for developing a risk prediction function based on Lin and Ying's model and two-phase sampling data. Furthermore, the equation (2.7) in Lin and Ying [1994] involves an estimate of a nuisance parameter $\eta(t) = \frac{P\{Z1(T>t)\}}{P\{1(T>t)\}}$ where T is the censored failure time and Z is the covariates. Nan and Wellner [2013]'s inference procedure based on equation (2.7) from Lin and Ying [1994] must estimate η and establish its convergence rate prior to the estimation and the asymptotic study of θ. We show in this chapter this approach is not necessary. If they studied estimating equations we proposed, their method could be used to develop both regression parameters and the baseline cumulative hazard function. The theoretical tools provided in Nan and Wellner [2013] can be applied to the circumstances when the nuisance parameter does not have $\sqrt{N}$-rate convergence, but, for the AH model, both baseline cumulative hazard function and regression parameters are $\sqrt{N}$-estimable. Our simpler approach applies.

In contrast to the preceding works, we propose a new weighted estimating equation that is an average of i.i.d. functions of $\Lambda(\cdot)$ and θ. As a result, we are able to estimate $\Lambda(\cdot)$ and θ simultaneously without estimating the nuisance parameter η, and we obtain the joint distribution of $\Lambda(\cdot)$ and θ so that we can estimate the subject-specific survival probability immediately. The whole estimating procedure is organized systematically, so that after establishing the joint distribution of estimates of $\Lambda(\cdot)$ and θ for random sampling, the estimators for θ and Λ alone and the prediction function of individual cumulative hazard, the extension of these estimators to two-phase sampling with or without using additional cohort information, and furthermore these nine estimators' robust variances under model-misspecification are developed immediately.

3.2 Notation

Let $\tilde{T}$ denote a failure time, C a censoring time and Z a p-dimensional covariate. Corresponding to the notations in Chapter 2, in Lin & Ying's additive model, the random vector $X = \{\tilde{T} \wedge C, 1(\tilde{T} \leq C), Z\} = (T, \Delta, Z)$ where T and Δ are the censored failure time and the censoring indicator. Let $Y(t) = \mathbf{1}[T \geq t]$ denote the "at risk" process, where $Y(t) = 1$ if an individual is still at risk at time t and 0 otherwise. Let $N(t)$ denote the counting process that records whether a failure has occurred by time t, $N(t) = 1(T \leq t, \Delta = 1)$. Note that both processes are functions of X. We assume $X_1, X_2, \ldots, X_N$ are i.i.d. governed by a probability distribution $P \in \mathcal{P}$. We further assume there exists a finite maximum censoring time τ such that $Pr(C \geq \tau) = Pr(C = \tau) > 0$ and the failure time $\tilde{T}$ is independent of the censoring time C given the covariate Z.

With random sampling (RS), X is observed for every member in the sample of N. With two-phase variable probability (VPS) , the sampling scheme is arranged in two steps. At phase I, we randomly select N subjects from a population. T, Δ, some covariates Z^I in $Z = (Z^I, Z^{II})$ and some auxiliary variables U that are observed for all members in the phase I sample. At phase II, we independently generate the phase II selection indicator R from a Bernoulli distribution $\pi_0(V) = P(R = 1|V)$ where $V = (Z^I, T, \Delta)$. We then measure Z^{II} for all the individuals in phase II subsamples, for which we have a full collection on Z and therefore X.

3.3 Assumptions

We first introduce the parameter space we consider. We let $\mathbb{L}_0$ be the parameter space for (θ, Λ). Conventionally, $\mathbb{L}_0$ is considered as a product space $\Theta \times \mathbb{A}$ where Θ is a bounded subset of $\mathbb{R}^p$ and $\mathbb{A}$ is a collection of finite nondecreasing and nonnegative functions over the time interval $[0, \tau]$. In order to employ our Z-estimation theory from Chapter 2, we consider our parameters as maps in l^∞ spaces. We consider the index set

$$\mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2 = \{h = (h_1, h_2) : h_1 \in \mathbb{R}^p, h_2 \in BV[0,\tau], \|h_1\|_E \leq 1, \|h_2\| \leq 1, \|h_2\|_{TV} \leq 1\}. \tag{3.2}$$

$\|\cdot\|_E$ is an Euclidean norm: $\|h_1\|_E = \sqrt{\sum_{i=1}^p h_{1,j}^2}$. $\|\cdot\|_{TV}$ is a total variation norm, i.e., $\|h_2\|_{TV} = \sup \sum_{i=1}^n |h_2(x_i) - h_2(x_{i-1})|$ where the supreme is taken over all partitions $0 = x_0 < x_1, \ldots, x_{n-1} < x_n = \tau$. $\mathcal{H}_1$ is the unit ball of the Euclidean space $\mathbb{R}^p$. $\mathcal{H}_2$ is the set of all uniformly bounded functions of uniformly bounded variation over $[0, \tau]$, with both bounds equal to 1.

We consider $\theta \in l^\infty(\mathcal{H}_1)$ identified by

$$\theta h_1 = h_1^T \theta,$$

$\Lambda \in l^\infty(\mathcal{H}_2)$ identified by

$$\Lambda h_2 = \int_0^\tau h_2 d\Lambda,$$

and $(\theta, \Lambda) \in l^\infty(\mathcal{H})$ identified by

$$(\theta, \Lambda)h = h_1^T \theta + \int_0^\tau h_2 d\Lambda.$$

Therefore $\Theta \subset l^\infty(\mathcal{H}_1)$ with a sup norm $\|\cdot\|_{\mathcal{H}_1}$, $\mathbb{A} \subset l^\infty(\mathcal{H}_2)$ with a sup norm $\|\cdot\|_{\mathcal{H}_2}$ and $\mathbb{L}_0 \subset l^\infty(\mathcal{H})$ with a sup norm $\|\cdot\|_{\mathcal{H}}$.

The following assumptions are assumed throughout Chapters 3 & 4. Not all of the assumptions are needed for every result, but we assume them all at once for clarity.

A. 3.3.1. *The parameter space* $\mathbb{L}_0$ *for* (θ, Λ) *is compact, closed and bounded.* θ *is uniformly bounded and* Λ *is a uniformly bounded function of uniformly bounded variation over* $[0, \tau]$.

When data are collected by i.i.d. random sampling, we assume

A. 3.3.2. $Z \in \mathbb{R}^p$, *which is an Euclidean space.* $\|Z\|_E \leq M_0, M_0 < \infty$.

A. 3.3.3. $PY(\tau) \geq m, m > 0$.

A. 3.3.4.

$$A \equiv P \int_0^\tau [Z - \frac{PZY(t)}{PY(t)}]^{\otimes 2} Y(t) dt$$

is positive definite.

When data are collected by two-phase VPS, we further assume

A. 3.3.5. $\pi_0(V) = P(R = 1|X, U) = P(R = 1|V)$.

A. 3.3.6. $0 < \sigma \leq \pi_0(v) \leq 1$ *for all* $v \in \mathcal{V}$.

When we are interested in incorporating auxiliary variables into the estimation procedure for improved two-phase sampling estimators, we add two more assumptions. Let vector $\tilde{V} = \tilde{V}(V)$ of q-dimension be the quantity we choose to calibrate on. We assume

A. 3.3.7. $\tilde{V}$ *is bounded.*

A. 3.3.8. $Q\tilde{V}\tilde{V}^T$ *is positive definite.*

3.4 Random Map and Estimators

In this section, we follow the procedure in section 2.4.1 to construct our estimating equations for estimators. We first construct them based on i.i.d. RS data and then two-phase VPS data.

Since we are concerned with both Euclidean and non-Euclidean parameters, our estimating equations will not be a finite-dimensional vector as usual but of infinite dimension. Hence we consider the collection of our estimation equations as a random map from $\mathbb{L}_0$, a subset of a Banach space into a Banach space $\mathbb{L} = l^\infty(\mathcal{H})$, where $\mathcal{H}$ is the index set we introduced in the previous section.

We consider the random map $\Psi_N : \mathbb{L}_0 \mapsto \mathbb{L}$ defined by

$$\Psi_N(\theta, \Lambda)h = \mathbb{P}_N \psi_{\theta,\Lambda,h}$$

where $\psi_{\theta,\Lambda,h} = \psi_{1,\theta,\Lambda,h_1} + \psi_{2,\theta,\Lambda,h_2}$. In $\psi_{\theta,\Lambda,h}$,

$$\psi_{1,\theta,\Lambda,h_1}(X) = h_1^T \int_0^\tau Z\{dN(t) - Y(t)d\Lambda(t) - Y(t)Z^T\theta dt\} \tag{3.3}$$

$$\psi_{2,\theta,\Lambda,h_2}(X) = \int_0^\tau \{h_2(t)dN(t) - h_2(t)Y(t)d\Lambda(t) - h_2(t)Y(t)Z^T\theta dt\}. \tag{3.4}$$

Motivation behind $\psi_{\theta,\Lambda,h}$ will be given in section 3.6. Our RS estimators $(\hat{\theta}, \hat{\Lambda})$ are then obtained by solving the estimating equations (EE):

$$\Psi_N(\theta, \Lambda)h = 0 \text{ for every } h \in \mathcal{H}. \tag{3.5}$$

With two-phase sampling, X is not fully observed. We only have a complete collection of X when $R = 1$. According to (3.6), we consider a new random map $\Psi_N^* : \mathbb{L}_0 \mapsto \mathbb{L}$ identified by

$$\Psi_N^*(\theta, \Lambda) = \mathbb{P}_N \psi_{\theta,\Lambda}^*(X, V, R) = \mathbb{P}_N \frac{R}{\pi_0(V)} \psi_{\theta,\Lambda}(X).$$

Our two-phase sampling estimators $(\hat{\theta}^*, \hat{\Lambda}^*)$ are then obtained by solving the inverse probability weighted estimating equations (IPW-EE):

$$\Psi_N^*(\theta, \Lambda)h = 0 \text{ for every } h \in \mathcal{H}. \tag{3.6}$$

To solve $(\hat{\theta}, \hat{\Lambda})$, we make use of a fact that EE (3.5) holds for every $h \in \mathcal{H}$. Furthermore since (3.5) is an equation, it holds for every $h \in \mathcal{H} \cdot M$ where M can be any finite positive number. Thus we can first choose an $h \in \mathcal{H} \cdot M$ to remove Λ and solve (3.5) for $\hat{\theta}$, and then we choose a different h for obtaining $\hat{\Lambda}$. For clarity, we suppress notation t in $Y(t)$ in the middle of a derivation and put it back in the result. Since for a particular dataset, $\frac{\mathbb{P}_N ZY(t)}{\mathbb{P}_N Y(t)}$ is fixed, we let

$$h = (h_1, h_2) = (h_1, -h_1^T \frac{\mathbb{P}_N ZY(t)}{\mathbb{P}_N Y(t)})$$

With this h, the LHS of EE (3.5) becomes

$$\begin{aligned}
\mathbb{P}_N \psi_{\theta,\Lambda,h} =& \mathbb{P}_N(\psi_{1,\theta,\Lambda,h_1} + \psi_{2,\theta,\Lambda,h_2}) \\
=& \mathbb{P}_N h_1^T \int_0^\tau Z \left\{ dN(t) - Y d\Lambda(t) - Y Z^T \theta dt \right\} \\
&+ \mathbb{P}_N \int_0^\tau \left\{ h_2(t) dN(t) - h_2(t) Y d\Lambda(t) - h_2(t) Z^T Y \theta dt \right\}. \\
=& \int_0^\tau \left\{ h_1^T \mathbb{P}_N Z dN(t) - h_1^T \mathbb{P}_N ZY d\Lambda(t) - h_1^T \mathbb{P}_N Z Z^T Y \theta dt \right\} \\
&- \int_0^\tau \left\{ h_1^T \frac{\mathbb{P}_N ZY}{\mathbb{P}_N Y} \mathbb{P}_N dN(t) - h_1^T \frac{\mathbb{P}_N ZY}{\mathbb{P}_N Y} \mathbb{P}_N Y d\Lambda(t) - h_1^T \frac{\mathbb{P}_N ZY}{\mathbb{P}_N Y} \mathbb{P}_N Z^T Y \theta dt \right\} \\
=& h_1^T \int_0^\tau \mathbb{P}_N \left(Z - \frac{\mathbb{P}_N ZY}{\mathbb{P}_N Y} \right) dN(t) - h_1^T \int_0^\tau \mathbb{P}_N \left(Z - \frac{\mathbb{P}_N ZY}{\mathbb{P}_N Y} \right) Z^T Y \theta dt.
\end{aligned}$$

Because

$$\int_0^\tau \mathbb{P}_N \left(Z - \frac{\mathbb{P}_N ZY}{\mathbb{P}_N Y} \right) \frac{\mathbb{P}_N Z^T Y}{\mathbb{P}_N Y} Y \theta dt = \int_0^\tau \left[\mathbb{P}_N ZY \frac{\mathbb{P}_N Z^T Y}{\mathbb{P}_N Y} - \frac{\mathbb{P}_N ZY}{\mathbb{P}_N Y} \mathbb{P}_N Z^T Y \right] \theta dt = 0,$$

we can add zero valued $\int_0^\tau \mathbb{P}_N \left(Z - \frac{\mathbb{P}_N ZY}{\mathbb{P}_N Y} \right) \frac{\mathbb{P}_N Z^T Y}{\mathbb{P}_N Y} Y \theta dt$ to the RHS of the preceding display without changing the equality. As a result, we obtain

$$\mathbb{P}_N \psi_{\theta,\Lambda,h} = \mathbb{P}_N \int_0^\tau h_1^T \left(Z - \frac{\mathbb{P}_N ZY}{\mathbb{P}_N Y} \right) dN(t) - \mathbb{P}_N \int_0^\tau h_1^T \left(Z - \frac{\mathbb{P}_N ZY}{\mathbb{P}_N Y} \right) \left(Z - \frac{\mathbb{P}_N ZY}{\mathbb{P}_N Y} \right)^T Y \theta dt = 0.$$

Because this equation holds for every $h_1 \in \mathcal{H}_1$, thus

$$\mathbb{P}_N \int_0^\tau \left(Z - \frac{\mathbb{P}_N ZY}{\mathbb{P}_N Y} \right) dN(t) - \mathbb{P}_N \int_0^\tau \left(Z - \frac{\mathbb{P}_N ZY}{\mathbb{P}_N Y} \right) \left(Z - \frac{\mathbb{P}_N ZY}{\mathbb{P}_N Y} \right)^T Y \hat{\theta} dt = 0.$$

Under our assumption A.3.3.4, when N is large enough $\mathbb{P}_N \int_0^\tau \left[Z - \frac{\mathbb{P}_N ZY(t)}{\mathbb{P}_N Y(t)}\right]^{\otimes 2} Y(t)dt$ is positive definite and has an inverse. Therefore

$$\hat{\theta} = \left[\mathbb{P}_N \int_0^\tau \left\{Z - \frac{\mathbb{P}_N ZY(t)}{\mathbb{P}_N Y(t)}\right\}^{\otimes 2} Y(t)dt\right]^{-1} \mathbb{P}_N \int_0^\tau \left\{Z - \frac{\mathbb{P}_N ZY(t)}{\mathbb{P}_N Y(t)}\right\} dN(t). \tag{3.7}$$

Then given each $s \in [0, \tau]$, we set

$$h = (h_1, h_2) = (0, \frac{1(t \le s)}{\mathbb{P}_N Y(t)}).$$

With this h, EE (3.5) becomes

$$\mathbb{P}_N \psi_{\theta,\Lambda,h} = 0 + \int_0^\tau \left[\frac{1(t \le s)}{\mathbb{P}_N Y} \mathbb{P}_N Y dN(t) - \frac{1(t \le s)}{\mathbb{P}_N Y} \mathbb{P}_N Y d\Lambda(t) - \frac{1(t \le s)}{\mathbb{P}_N Y} \mathbb{P}_N Z^T Y \theta dt\right] = 0.$$

Plugging $\hat{\theta}$ from (3.7) into the above equation yields

$$\hat{\Lambda}(s) = \int_0^s \frac{\mathbb{P}_N[Y(t)dN(t)]}{\mathbb{P}_N Y(t)} - \int_0^s \frac{\mathbb{P}_N[Z^T Y(t)]}{\mathbb{P}_N Y(t)} \hat{\theta} dt. \tag{3.8}$$

In conclusion, we have shown estimating equation (3.5) has a unique solution $(\hat{\theta}, \hat{\Lambda})$ by writing this solution explicitly in (3.7) and (3.8). Although we used a different approach to derive $(\hat{\theta}, \hat{\Lambda})$ from Lin and Ying [1994], our results are the same as theirs.

With two-phase VPS, using the same set of h and procedure as in RS, we obtain our two-phase VPS estimators from (3.6). For simplicity, we create a new notation $\mathbb{P}_N^\pi$ defined by $\mathbb{P}_N^\pi f(X) = \frac{1}{N} \sum_{i=1}^N \frac{R_i}{\pi_0(V_i)} f(X_i)$. Then

$$\hat{\theta}^* = \left[\mathbb{P}_N^\pi \int_0^\tau \left\{Z - \frac{\mathbb{P}_N^\pi ZY(t)}{\mathbb{P}_N^\pi Y(t)}\right\}^{\otimes 2} Y(t)dt\right]^{-1} \mathbb{P}_N^\pi \int_0^\tau \left\{Z - \frac{\mathbb{P}_N^\pi ZY(t)}{\mathbb{P}_N^\pi Y(t)}\right\} dN(t),$$

$$\hat{\Lambda}^*(s) = \int_0^s \frac{\mathbb{P}_N^\pi[Y(t)dN(t)]}{\mathbb{P}_N^\pi Y(t)} - \int_0^s \frac{\mathbb{P}_N^\pi[Z^T Y(t)]}{\mathbb{P}_N^\pi Y(t)} \hat{\theta}^* dt.$$

Remark 2. *Our restriction on the index set $\mathcal{H}$ is mostly for technical reasons. Later we will see our choice on $\mathcal{H}$ facilitates us to establish asymptotic normality of our estimators. This restriction on $\mathcal{H}$, for example, is used for verifying Donsker conditions and showing the continuous invertibility of certain maps we are concerned with in section 3.7. Other potentially useful choices of $\mathcal{H}$ are discussed in Murphy and Van der Vaart [2001] and the basic rule to choose this set, over which we have some control, is briefly discussed in van der Vaart [1998, p.422]. However, there is also a nontechnical reason of our choice on $\mathcal{H}$. $\mathcal{H}$ needs to be large enough to ensure the asymptotic normality of Λ in the usual sense of a function over $[0, \tau]$ [van der Vaart, 1998, p. 430]. This is the case for our choice of $\mathcal{H}_2$. Since $\forall s \in [0, \tau]$, $\Lambda(s)$ can be written as $\int 1_{[0,s]}(t) d\Lambda(t)$, if we let $h_2 = 1_{[0,s]}(t)$, then $h_2 \in \mathcal{H}_2$ and $\hat{\Lambda} h_2 = \hat{\Lambda}(s)$. As a result, our asymptotic results in the later section 3.10 on $\hat{\Lambda} h_2$ with $h_2 \in \mathcal{H}_2$ will include the desired asymptotic results for $\hat{\Lambda}(s), s \in [0, \tau]$.*

3.5 Deterministic Map and Parameters

Corresponding to the random map Ψ_N, we consider a deterministic map $\Psi : \mathbb{L}_0 \mapsto \mathbb{L}$ defined by

$$\Psi(\theta, \Lambda)h = P\psi_{\theta,\Lambda,h}.$$

where $P \in \mathcal{P}$. According to section 2.6, we define our parameter (θ_0, Λ_0) as the solution to

$$\Psi(\theta, \Lambda)h = P\psi_{\Lambda,\theta,h}(X) = 0 \text{ for every } h \in \mathcal{H}. \tag{3.9}$$

This solution is unique because (θ_0, Λ_0) can be solved explicitly from (3.9) as shown in the following.

We apply the same technique used for obtaining $(\hat{\theta}, \hat{\Lambda})$ in section 2.4.1. We first set

$$h = (h_1, h_2) = (h_1, -h_1^T \frac{PZY(t)}{PY(t)}).$$

Applying this h to (3.9) and following each step in the previous section, we obtain $\forall h_1 \in \mathcal{H}_1$

$$P\psi_{\theta,\Lambda,h} = P\int_0^\tau h_1^T \left(Z - \frac{PZY}{PY}\right) dN(t) - P\int_0^\tau h_1^T \left(Z - \frac{PZY}{PY}\right)\left(Z - \frac{PZY}{PY}\right)^T Y\theta_0 dt = 0,$$

which is equivalent to

$$P\int_0^\tau \left(Z - \frac{PZY}{PY}\right) dN(t) - P\int_0^\tau \left(Z - \frac{PZY}{PY}\right)^{\otimes 2} Y\theta_0 dt = 0.$$

By assumption A.3.3.4, $P\int_0^\tau \left[Z - \frac{PZY(t)}{PY(t)}\right]^{\otimes 2} Y(t)dt$ is positive definite. Thus

$$\theta_0 = \left\{P\int_0^\tau \left[Z - \frac{PZY(t)}{PY(t)}\right]^{\otimes 2} Y(t)dt\right\}^{-1} P\int_0^\tau \left[Z - \frac{PZY(t)}{PY(t)}\right] dN(t). \tag{3.10}$$

Next given each $s \in [0, \tau]$ we set

$$h = (h_1, h_2) = (0, \frac{1(t \le s)}{PY(t)}),$$

and (3.9) becomes

$$P\psi_{\theta,\Lambda,h} = 0 + \int_0^\tau \left[\frac{1(t \le s)}{PY} PY dN(t) - \frac{1(t \le s)}{PY} PY d\Lambda_0(t) - \frac{1(t \le s)}{PY} PZ^T Y\theta_0 dt\right] = 0.$$

Plugging θ_0 from (3.10) into the above equation yields

$$\Lambda_0(s) = \int_0^s \frac{PdN(t)}{PY(t)} - \int_0^s \frac{PZ^TY(t)}{PY(t)}\theta_0 dt. \tag{3.11}$$

In this definition, (θ, Λ) can be regarded as a map $\alpha(P) : \mathcal{P} \mapsto l^\infty(\mathcal{H})$ that satisfies (3.9) for each $h \in \mathcal{H}$. Since P is the true underlying distribution from which X is sampled. Then the true parameter is the value of the map evaluated at $P : (\theta_0, \Lambda_0) = \alpha(P)$.

In our derivation of θ_0 and $\Lambda_0(s)$, we do not assume $P = P_{\theta_0,\Lambda_0}$, a probability distribution based on model (3.1). When Lin and Ying's AH model is the correct model for the observed data, parameters defined in (3.10) and (3.11) recover the parameters used in the model (3.1), because our estimating equation is unbiased under the true model. When Lin and Ying's AH model does not hold, our parameters are different from the model parameters in (3.1), but are still well-defined through (3.9). According to our theory in section 2.4.2, the definition of parameters with two-phase sampling is the same as the one used for RS. Thus (θ_0, Λ_0) defined in (3.9) are also the parameters for our two-phase sampling estimators.

3.6 Motivation Behind $\psi_{\theta,\Lambda,h}$

In this section we present how we obtain $\Psi_N(\theta,\Lambda)$ in (3.5). The main idea is to derive the "score functions" for both parameters θ and λ from the AH model (3.1). Based on the score functions, we decide our function $\psi_{\theta,\Lambda,h}$. The derivation of our score functions follows the efficient estimation theory for semiparametric model from Bickel et al. [1998]. McKeague and Sasieni [1994] derived the likelihood equation for a partly parametric additive risk model, which is related to Lin & Ying's model. Adapting their derivation steps to Lin & Ying's model will give the same result.

Based on the theory in Bickel et al. [1998], we first write the density for $X = (T, \Delta, Z)$ as

$$\left[\lambda(t|z)e^{-\Lambda(t|z)}\{1 - F_{C|Z}(t - |z)\}\right]^{\delta} \{e^{-\Lambda(t|z)} f_{C|Z}(t|z)\}^{1-\delta} p_Z(z)$$

where p_Z is the density for Z, and $F_{C|Z}$ and $f_{C|Z}$ are the cumulative distribution function and the density function for censoring time given Z. The density can be factored into two pieces and only one piece depends on (θ,Λ). Dropping the piece in the density that does not involve (θ,Λ) yields the likelihood for the parameter (θ,Λ):

$$L(\theta,\lambda;x) = \left[\lambda(t|z)e^{-\Lambda(t|z)}\right]^{\delta} \{e^{-\Lambda(t|z)}\}^{1-\delta}.$$

When (3.1) holds, the likelihood based on the model $P_{\theta,\Lambda}$ is given by

$$L(\theta,\lambda;x) = \left[\{\lambda(t) + z^T\theta\}e^{-\int_0^t \{\lambda(u)+z^T\theta\}du}\right]^{\delta} \{e^{-\int_0^t \{\lambda(u)+z^T\theta\}du}\}^{1-\delta}.$$

Next we construct one dimensional parametric submodels P_{θ_s,λ_s} that belong to $\mathcal{P}$ and pass through $P_{\theta,\lambda}$. These submodels are built by considering the paths $s \mapsto P_{\theta_s,\lambda_s}$ that satisfy

$$\theta_s = \theta + sh_1$$
$$\lambda_s(t) = \lambda(t) + sh_2(t)$$

where $h_1 \in \mathbb{R}^p$ and h_2 is a bounded function that maps from $[0,\tau]$ to the real line $\mathbb{R}$. As s approaches to 0, these paths pass through $P_{\theta,\Lambda}$. As a result, the likelihood for our submodels

P_{θ_s,λ_s} is given by

$$L(\theta_s, \lambda_s; x) = \left[\{\lambda_s(t) + z^T\theta_s\}e^{-\int_0^t\{\lambda_s(u)+z^T\theta_s\}du}\right]^{\delta}\left[e^{-\int_0^t\{\lambda_s(u)+z^T\theta_s\}du}\right]^{1-\delta}.$$

We can approach the calculation of the "score functions" for our parameters heuristically by calculating each score function g for each member of the smooth families P_{θ_s,λ_s}[van der Vaart, 1998, p.362]:

$$g(x) = \frac{\partial}{\partial s}\log L(\theta_s, \lambda_s; x)\Big|_{s=0}.$$

When $s \mapsto P_{\theta_s,\lambda_s}$ range over all submodels, we obtain a collection of score functions. This collection is of score functions was called the tangent set of the model $\mathcal{P}$ at $P_{\theta,\lambda}$ [Bickel et al., 1998, p.50-p.51]. In the following, we present how this collection of score functions motivates $\Psi_N(\theta, \Lambda)$. Since each P_{θ_s,λ_s} in $\{P_{\theta_s,\lambda_s} : s \mapsto (\theta_s, \Lambda_s)\}$ is one-dimensional parametric model with a single parameter s, the calculation of the score g of P_{θ_s,Λ_s} with respect to s is straightforward. Then

$$\begin{aligned}
g(x)h =& \frac{\partial}{\partial s}\left[\{\lambda_s(t) + z^T\theta_s\}^{\delta} - \int_0^t\{\lambda_s(u) + z^T\theta_s\}du\right]\Big|_{s=0} \\
=& \frac{\partial}{\partial s}\left(\delta\log\left[\lambda(t) + z^T\{\theta + sh_1\}\right] - \int_0^t\left[\lambda(u) + z^T\{\theta + sh_1\}\right]du\right)\Big|_{s=0} \\
&+ \frac{\partial}{\partial s}\left(\delta\log\left[\{\lambda(t) + sh_2(t)\} + z^T\theta\right] - \int_0^t\left[\{\lambda(u) + sh_2(u)\} + z^T\theta\right]du\right)\Big|_{s=0} \\
=& g_1(x)h_1 + g_2(x)h_2
\end{aligned}$$

where

$$
\begin{aligned}
g_1(x)h_1 =& \delta \frac{z^T h_1}{\lambda(t) + z^T\theta} - \int_0^t z^T h_1 du \\
=& h_1^T \int_0^\tau z \left\{ \frac{1}{\lambda(u) + z^T\theta} dN(u) - 1(u \le t) \frac{\lambda(u) + z^T\theta}{\lambda(u) + z^T\theta} du \right\} \\
=& h_1^T \int_0^\tau \frac{1}{\lambda(u|z)} z \left\{ dN(u) - 1(u \le t) d\Lambda(u) - 1(u \le t) z^T\theta du \right\}
\end{aligned}
$$

and

$$
\begin{aligned}
g_2(x)h_2 =& \delta \frac{h_2(t)}{\lambda(t) + z^T\theta} - \int_0^t h_2(u) du \\
=& \int_0^\tau \frac{h_2(u)}{\lambda(u) + z^T\theta} dN(u) - \int_0^\tau 1(u \le t) h_2(u) \frac{\lambda(u) + z^T\theta}{\lambda(u) + z^T\theta} du \\
=& \int_0^\tau \frac{1}{\lambda(u|z)} h_2(u) \left\{ dN(u) - 1(u \le t) d\Lambda(u) - 1(u \le t) z^T\theta du \right\}.
\end{aligned}
$$

Based on $g(x)$, we are able to obtain an efficient estimator by solving $\mathbb{P}_N g(X)h = 0$.

Examining $g(x)$ closely, we find $\frac{1}{\lambda(u|z)}$ in the integrand of $g_1(x)h_1$ and $g_2(x)h_2$ plays a role of a weight function. According to McKeague and Sasieni [1994], to obtain this theoretically more efficient estimator, in the first step we replace the weight function $\frac{1}{\lambda(u|z)}$ by 1 and then solve $\mathbb{P}_N g(X)h = 0$ for an initial estimator $\hat{\theta}$ and $\hat{\Lambda}(\cdot)$. In the second step we estimate $\lambda(u|z)$ using a kernel smoother based on the initial estimators $\hat{\theta}$ and $\hat{\Lambda}(\cdot)$. In the last step we solve $\mathbb{P}_N g(X) = 0$ for (θ, Λ) again with the estimated $\lambda(u|z)$ in the equation.

In theory, there is some efficiency gain in this more elaborate approach. However, in practice, we can stop at the first step of this estimating procedure, because after replacing $\frac{1}{\lambda(u|z)}$ by 1, the function $g(x)h$ remains unbiased at the true parameter, different authors have shown the efficiency loss is small and the calculation for estimators becomes much simpler. We can even write out the explicit forms of the estimators as shown in section 3.4. With this replacement we avoid choosing some smoothing parameters to estimate $\lambda(u|z)$ and the problem that arise when $\hat{\lambda}(u|z)$ is negative or close to zero. The latter will give an unreliable estimator. For the AH model, Lin and Ying [1994] computed the relative efficiency comparing unweighted estimators to the optimal estimators. In the special case

of assuming no censorship or truncation, if there are no covariate effects and $\lambda(\cdot)$ is half-logistic, the efficiency is 0.9609; if there are covariate effects and $\lambda(\cdot) = 1$, the efficiency were found to be 0.999, 0.996, and 0.993 for regression coefficients θ of 0.5, 1, and 1.5. In real examples, the $\lambda(\cdot)$ and β are much smaller than their assumed values, but their conclusion should still hold after scaling down the $\lambda(\cdot)$ and θ at the same time. Potential adaptive estimators to achieve the semiparametric efficiency bound was also suggested in the same paper. For Aalen's additive hazard model [Aalen, 1980], which is closely related to Lin and Ying's additive hazard model, Huffer and McKeague [1991] has shown by simulations the estimators obtained without the weight function do as well as with the maximum likelihood estimators unless the dataset is very large. Same conclusion was drawn by Martinussen and Scheike [2006] from their practice.

These arguments and evidence convince us that it is a good choice to use the unweighted score functions as our estimating equations, which lead to $\psi_{\theta,\Lambda,h}$ we used for our EE as shown in (3.5). Furthermore during our calculation of "score functions", each path $s \mapsto (\theta_s, \lambda_s)$ is associated with a particular direction $h = (h_1, h_2)$ and each score function is associated with this direction. Naturally h is used to to label the score function. As a result, we obtain our index set $\mathcal{H}$ at the same time.

3.7 Preliminary Results on $\psi_{\theta,\Lambda}$

Before studying the asymptotic properties of $(\hat{\theta}, \hat{\Lambda})$, we will establish a few preliminary results on $\psi_{\theta,\Lambda}$ in this section. At the end of section 2.4.5, we explained our Z-estimation theory based on empirical process theory dividing the problem of developing estimators into small tasks. These small tasks can be solved almost independently. In the last three sections, we have created our EE estimators and parameters. In this section, we will complete the rest of these small tasks. These include showing that (ψ_0, Λ_0) is a well-separated solution to $P\psi_{\theta,\Lambda}h = 0$, the Glivenko-Cantelli and Donsker Properties of $\psi_{\theta,\Lambda,h}$, the Fréchet derivative of $P\psi_{\theta,\Lambda}$ at (θ_0, Λ_0) and that this derivative has a continuous inverse on its range, as well as $\|P(\psi_{\theta,\Lambda,h} - \psi_{\theta_0,\Lambda_0,h})^2\|_{\mathcal{H}} \to 0$ as $(\theta, \Lambda) \to (\theta_0, \Lambda_0)$. These results will then be brought together repeatedly for establishing the asymptotics of various estimators.

3.7.1 A Well-separated Solution

Lemma 3.7.1. *Under assumptions A.3.3.1 and A.3.3.2, parameter* (θ_0, Λ_0) *is a well-separated solution. In other words,* $\forall \epsilon > 0$,

$$\inf_{(\theta,\Lambda):\|(\theta,\Lambda)^T - (\theta_0,\Lambda_0)^T\|_{\mathbb{L}_0} \geq \epsilon} \|\Psi(\theta, \Lambda)\|_{\mathbb{L}} > 0$$

Proof. By assumption, the set $\mathbb{L}_0$ is compact, closed and bounded. As a closed subset of a compact set, $\mathbb{L}_0' \equiv \{(\theta, \Lambda) : \theta \in \Theta, \Lambda \in \mathbb{A}, \|(\theta, \Lambda)^T - (\theta_0, \Lambda_0)^T\|_{\mathbb{L}_0} \geq \epsilon\}$ is also compact. $\Psi_{\theta,\Lambda}$ is a continuous function in (θ, Λ) and a norm is also a continuous function. Since the image of a continuous function on a compact set is compact [Marsden and Hoffman, 1999, proposition 1.4.19], over the set $\mathbb{L}_0'$, $\|\Psi(\theta, \Lambda)\|_{\mathbb{L}}$ is compact. By the extreme value theorem [Marsden and Hoffman, 1999, theorem 1.4.20], the real valued compact set contains its infimum. Hence there exists a $(\theta', \Lambda') \in \mathbb{L}_0'$ such that $\inf_{\mathbb{L}_0'} \|\Psi(\theta, \Lambda)\|_{\mathbb{L}} = \|\Psi(\theta', \Lambda')\|_{\mathbb{L}}$. Because (θ_0, Λ_0) is the unique solution to $\Psi(\theta, \Lambda) = 0$ for $(\theta, \Lambda) \in \mathbb{L}_0$, we have $\|\Psi(\theta', \Lambda')\|_{\mathbb{L}} > 0$. Therefore

$$\inf_{(\theta,\Lambda):\|(\theta,\Lambda)^T - (\theta_0,\Lambda_0)^T\|_{\mathbb{L}_0} \geq \epsilon} \|\Psi(\theta, \Lambda)\| = \|\Psi(\theta', \Lambda')\|_{\mathbb{L}} > 0.$$

□

3.7.2 Donsker

Lemma 3.7.2. *Under assumptions A.3.3.1 and A.3.3.2, the class* $\mathcal{F} \equiv \{\psi_{\theta,\Lambda,h}(X), \theta \in \Theta, \Lambda \in \mathbb{A}, h \in \mathcal{H}\}$ *is P-Donsker, with finite envelope function.*

Proof. We will first show each component of function $\psi_{\theta,\Lambda,h}$ belongs to some Donsker class and then apply the Donsker preservation theorem to conclude $\mathcal{F}$ as the summation and multiplication of these P-Donsker classes also P-Donsker.

- By assumption A.3.3.2, Z is a vector-valued bounded function. Because a finite class of square-integrable functions is always Donsker [van der Vaart, 1998, p.270], the class $\{Z\}$ is P-Donsker. Likewise the classes $\{\Delta\}$ and $\{T\}$ are both P-Donsker.

- For every $\theta_1, \theta_2 \in \Theta$ and $\Theta \subset \mathbb{R}^p$, we have

$$|Z^T\theta_1 - Z^T\theta_2| = |Z^T(\theta_1 - \theta_2)| \leq \|Z\|_{\mathbb{E}}\|\theta_1 - \theta_2\|_{\mathcal{H}_1}.$$

  A. 3.3.2 assumes Z is bounded and A.3.3.1 assumes Θ is bounded. By example 19.7 from van der Vaart [1998], $\{Z^T\theta, \theta \in \Theta\}$ is P-Donsker.

- Since h_2 is uniformly bounded and of uniformly bounded variation, according to theorem 2.7.5 [van der Vaart and Wellner, 1996] and example 19.11 [van der Vaart, 1998], $\{h_2(T), h_2 \in \mathcal{H}_2\}$ as a class of functions of T is P-Donsker.

- Under the assumption of A.3.3.1, $\Lambda(t)$ is uniformly bounded and of uniformly bounded variation over $[0, \tau]$. By theorem 2.7.5 [van der Vaart and Wellner, 1996] and example 19.11 [van der Vaart, 1998], $\{\Lambda(T), \Lambda \in \mathbb{A}\}$ is P-Donsker.

- Because $\mathcal{H}_2$ is a subset of $BV[0, \tau]$, for any $h_2 \in \mathcal{H}_2$, we can find two nondecreasing functions $h_2^1(t)$ and $h_2^2(t)$ such that $h_2(t) = h_2^1(t) - h_2^2(t)$ [Dudley, 2002, theorem 7.2.4]. As a result,

$$\int_0^\tau h_2(t)Y(t)d\Lambda(t) = \int_0^T \{h_2^1(t) - h_2^2(t)\}d\Lambda(t) = \int_0^T h_2^1(t)d\Lambda(t) - \int_0^T h_2^2(t)d\Lambda(t).$$

In addition, A.3.3.1 assumes Λ is of uniformly bounded variation and uniformly bounded, so $\int_0^\tau h_2(t)Y(t)d\Lambda(t)$ can be written as the difference of two uniformly bounded non-decreasing functions in T. Therefore $\int_0^\tau h_2(t)Y(t)d\Lambda(t)$ is uniformly bounded and of uniformly bounded variation over $[0,\tau]$ as well. Applying example 19.11 [van der Vaart, 1998, p.273] again, we have $\{\int_0^\tau h_2(t)Y(t)d\Lambda(t), \Lambda \in \mathbb{A}, h \in \mathcal{H}\}$ is P-Donsker. Similarly $\{\int_0^\tau h_2(t)Y(t)dt, h_2 \in \mathcal{H}_2\}$ is P-Donsker.

Recall

$$\psi_{\theta,\Lambda,h} = h_1^T\{Z\Delta - \Lambda(T) - ZZ^T\theta T\} + \{h_2(T)\Delta - \int_0^\tau h_2(t)Y(t)d\Lambda(t) - \int_0^\tau h_2(t)Y(t)dtZ^T\theta\}.$$

In the preceding proofs, we have verified each component of $\psi_{\theta,\Lambda,h}$ belongs to some P-Donsker class. Assumptions A.3.3.1 and A.3.3.2 also guarantee all these classes are either bounded or uniformly bounded. Then according to the Donsker preservation theorems (examples 2.10.7 and 2.10.8 from van der Vaart and Wellner [1996]), $\mathcal{F} \equiv \{\psi_{\theta,\Lambda,h}(X), \theta \in \Theta, \Lambda \in \mathbb{A}, h \in \mathcal{H}\}$ is P-Donsker.

□

3.7.3 Glivenko-Cantelli

Lemma 3.7.3. *Under the assumption of A.3.3.1 and A.3.3.2,* $\{\psi_{\Lambda,\theta,h}, \Lambda \in \mathbb{A}, \theta \in \Theta, h \in \mathcal{H}\}$ *is Glivenko-Cantelli(GC), with finite envelope function.*

Proof. This is an immediate result of Slutsky's Theorem. □

Remark 3. *This shortcut proof applies to the AH model but may not apply to other models. This is because in this case we are able to show the Donsker property of* $\psi_{\theta,\Lambda,h}$ *for* (θ,Λ) *over the whole parameter space. Then the proof for GC of* $\{\psi_{\Lambda,\theta,h}, \Lambda \in \mathbb{A}, \theta \in \Theta, h \in \mathcal{H}\}$ *is simplified. Usually we can not show GC directly from a Donsker result because the establishment of an estimator's asymptotic normality requires* $\psi_{\theta,\Lambda,h}(X)$ *to be contained in some Donsker class for* (θ,Λ) *only in the neighborhood of* (θ_0,Λ_0) *as stated in theorem 2.4.4. Thus for other models, we should consider GC and Donsker properties separately.*

3.7.4 Fréchet Derivative

Lemma 3.7.4. *Assume A.3.3.1 ∼ A.3.3.4 hold. The map* $\Psi : \Theta \times \mathbb{A} \mapsto l^\infty(\mathcal{H})$ *is Fréchet-differentiable at* (Λ_0, θ_0)*, with a derivative*

$$\dot{\Psi}_0 = \begin{pmatrix} \dot{\Psi}_{11} & \dot{\Psi}_{12} \\ \dot{\Psi}_{21} & \dot{\Psi}_{22} \end{pmatrix} \quad : lin\Theta \times lin\mathbb{A} \mapsto l^\infty(\mathcal{H})$$

that has a continuous inverse on its range. Components of $\dot{\Psi}_0$ *are maps identified by*

$$\begin{aligned} \dot{\Psi}_{11}(\theta - \theta_0)h_1 =& - h_1^T P \int_0^\tau Y(t) Z Z^T dt(\theta - \theta_0) \\ \dot{\Psi}_{12}(\Lambda - \Lambda_0)h_1 =& - h_1^T P \int_0^\tau Z Y(t) d(\Lambda - \Lambda_0)(t) \\ \dot{\Psi}_{21}(\theta - \theta_0)h_2 =& - P \int_0^\tau h_2(t) Y(t) Z^T dt(\theta - \theta_0) \\ \dot{\Psi}_{22}(\Lambda - \Lambda_0)h_2 =& - P \int_0^\tau h_2(t) Y(t) d(\Lambda - \Lambda_0)(t). \end{aligned} \tag{3.12}$$

Remark 4. *In lemma 3.7.4, we write* $\dot{\Psi}_0$ *in a partitioned form, which means we consider a partitioned parameter* (θ, Λ) *and a partitioned map* $\Psi = (\Psi_1, \Psi_2)$ *where* $\Psi_1(\theta, \Lambda)h_1 = P\psi_{1,\theta,\Lambda,h_1}$ *and* $\Psi_2(\theta, \Lambda)h_2 = P\psi_{2,\theta,\Lambda,h_2}$*. Then*

$$\begin{aligned} \dot{\Psi}_{11} :& lin\Theta \mapsto l^\infty(\mathcal{H}_1) \\ \dot{\Psi}_{12} :& lin\mathbb{A} \mapsto l^\infty(\mathcal{H}_1) \\ \dot{\Psi}_{21} :& lin\Theta \mapsto l^\infty(\mathcal{H}_2) \\ \dot{\Psi}_{22} :& lin\mathbb{A} \mapsto l^\infty(\mathcal{H}_2). \end{aligned}$$

We define operators $\dot{\Psi}_0$ *as a sum:*

$$\begin{aligned} \dot{\Psi}_0 \begin{pmatrix} \theta - \theta_0 \\ \Lambda - \Lambda_0 \end{pmatrix} h =& \begin{pmatrix} \dot{\Psi}_{11} & \dot{\Psi}_{12} \\ \dot{\Psi}_{21} & \dot{\Psi}_{22} \end{pmatrix} \begin{pmatrix} \theta - \theta_0 \\ \Lambda - \Lambda_0 \end{pmatrix} h \\ =& \dot{\Psi}_{11}(\theta - \theta_0)h_1 + \dot{\Psi}_{12}(\Lambda - \Lambda_0)h_1 + \dot{\Psi}_{21}(\theta - \theta_0)h_2 + \dot{\Psi}_{22}(\Lambda - \Lambda_0)h_2. \end{aligned}$$

Thus for given any $h \in \mathcal{H}$*,* $\dot{\Psi}_0(\theta - \theta_0, \Lambda - \Lambda_0)h \in \mathbb{R}$ *and* $\dot{\Psi}_0 : lin\Theta \times lin\mathbb{A} \mapsto l^\infty(\mathcal{H})$*.*

Proof. By definition [van der Vaart, 1998, p.297], the Fréchet derivative of a map $\phi : \mathbb{D}_\phi \mapsto \mathbb{E}$, where $\mathbb{D}_\phi$ is a subset of a normed space $\mathbb{D}$ that contains α and $\mathbb{E}$ is a normed space, is a continuous and linear map $\psi'_\alpha : \mathbb{D} \to \mathbb{E}$ such that

$$\|\phi(\alpha + g) - \phi(\alpha) - \phi'(g)\|_{\mathbb{E}} = o(\|g\|) \text{ as } \|g\| \downarrow 0. \tag{3.13}$$

In our problem, $\alpha = \theta_0$ or Λ_0, $\alpha + g = \theta$ or Λ. ϕ is either Ψ_1 or Ψ_2. Since Ψ_1 and Ψ_2 are both bounded linear operators, the calculation of Fréchet derivative is simple and straightforward. We can obtain the derivative by setting $\phi'(g) = \phi(\alpha + g) - \phi(\alpha)$.

To calculate $\dot{\Psi}_{11}$, we set $\alpha = \theta_0$, $g = \theta - \theta_0$. Then we let

$$\begin{aligned}
\dot{\Psi}_{11}(\theta - \theta_0)h_1 =& \Psi_1(\theta)h_1 - \Psi_1(\theta_0)h_1 \\
=& h_1^T P \int_0^\tau Z\{dN(t) - Y(t)d\Lambda(t) - Y(t)Z^T\theta dt\} \\
& - h_1^T P \int_0^\tau Z\{dN(t) - Y(t)d\Lambda(t) - Y(t)Z^T\theta_0 dt\} \\
=& - h_1^T P \int_0^\tau Y(t)ZZ^T dt(\theta - \theta_0).
\end{aligned}$$

As a result, $\|\Psi_1(\theta, \Lambda) - \Psi_1(\theta_0, \Lambda_0) - \dot{\Psi}_{11}(\theta - \theta_0)\| = 0$, which is $o(\|\theta - \theta_0\|)$ as $\|\theta - \theta_0\| \downarrow 0$. Since Ψ is a linear map, so is $\dot{\Psi}_{11}$. Since Ψ_1 is uniformly bounded, $\dot{\Psi}_{11}$ is a bounded linear operator. Thus it is continuous. By definition, $\Psi_1(\theta, \Lambda) : \Theta \to \mathbb{R}^p$ is Fréchet differentiable at θ_0. Using the same technique, we obtain

$$\begin{aligned}
& \dot{\Psi}_{12}(\Lambda - \Lambda_0)h_1 \\
=& h_1^T P \int_0^\tau Z\{dN(t) - Y(t)d\Lambda(t) - Y(t)Z^T\theta dt\} - h_1^T P \int_0^\tau Z\{dN(t) - Y(t)d\Lambda_0(t) - Y(t)Z^T\theta dt\} \\
=& - h_1^T P \int_0^\tau ZY(t)d(\Lambda - \Lambda_0)(t);
\end{aligned}$$

$$\begin{aligned}
& \dot{\Psi}_{21}(\theta - \theta_0)h_2 \\
=& P \int_0^\tau h_2(t)\{dN(t) - Y(t)d\Lambda(t) - Y(t)Z^T\theta dt\} - \int_0^\tau h_2(t)\{dN(t) - Y(t)d\Lambda(t) - Y(t)Z^T\theta_0 dt\} \\
=& - P \int_0^\tau h_2(t)Y(t)Z^T(\theta - \theta_0)dt;
\end{aligned}$$

$$
\begin{aligned}
&\dot{\Psi}_{22}(\Lambda-\Lambda_0)h_2 \\
=&P\int_0^\tau h_2(t)\{dN(t)-Y(t)d\Lambda(t)-Y(t)Z^T\theta dt\}-\int_0^\tau h_2(t)\{dN(t)-Y(t)d\Lambda_0(t)-Y(t)Z^T\theta dt\} \\
=&-P\int_0^\tau h_2(t)Y(t)d(\Lambda-\Lambda_0)(t).
\end{aligned}
$$

Thus we have established (3.12).

Next we will show $\dot{\Psi}_0$ has a continuous inverse on its range. van der Vaart [1998, p.422] claims the continuous invertibility of a partitioned $\dot{\Psi}_0$ can be verified by ascertaining the continuous invertibility of the two operators $\dot{\Psi}_{22}$ and $\dot{V}=\dot{\Psi}_{11}-\dot{\Psi}_{12}\dot{\Psi}_{22}^{-1}\dot{\Psi}_{21}$. In the following we will demonstrate $\dot{\Psi}_{22}$ and $\dot{V}$ have continuous inverses.

We begin with showing $\dot{\Psi}_{22}$ has a continuous inverse on its range by giving this inverse explicitly. We first identify the range of $\dot{\Psi}_{22}$ by $\mathbf{R}(\dot{\Psi}_{22})=\{\eta\in l^\infty(\mathcal{H}_2)\}$ such that

$$\eta(h_2)=-P\int_0^\tau h_2(t)PY(t)d(\Lambda-\Lambda_0)(t)$$

for some $\Lambda,\Lambda_0\in\mathbb{A}$. Such an η may be identified also as a signed measure on $[0,\tau]$ defined by

$$d\eta(t)=-PY(t)d(\Lambda-\Lambda_0)(t).$$

Next we define a new operator $\dot{\Psi}_{22}^{-1}:\mathbf{R}(\dot{\Psi}_{22})\mapsto lin\mathbb{A}$ by

$$(\dot{\Psi}_{22}^{-1}\eta)h_2=-\int_0^\tau\frac{h_2(t)}{PY(t)}d\eta(t). \tag{3.14}$$

Then $\dot{\Psi}_{22}^{-1}$ is the inverse of $\dot{\Psi}_{22}$ on its range because

$$
\begin{aligned}
\dot{\Psi}_{22}^{-1}\dot{\Psi}_{22}(\Lambda-\Lambda_0)h_2=&-\dot{\Psi}_{22}^{-1}\int_0^\tau h_2(t)PY(t)d(\Lambda-\Lambda_0)(t)\\
=&\int_0^\tau\frac{1}{PY(t)}h_2(t)PY(t)d(\Lambda-\Lambda_0)(t)\\
=&\int_0^\tau h_2(t)d(\Lambda-\Lambda_0)(t)=(\Lambda-\Lambda_0)h_2.
\end{aligned}
$$

Given A.3.3.6, the operator $\dot{\Psi}_{22}^{-1}$ is bounded. By A.3.3.1, the parameter space $\mathbb{A}$ is complete. According to corollary 3 [Bickel et al., 1998, p.419], $\mathbf{R}(\dot{\Psi}_{22})$ is closed. Since $\dot{\Psi}_{22}$ is a bounded

linear operator, it is continuous. Applying proposition 7B [Bickel et al., 1998, p.418], we obtain $\dot{\Psi}_{22}^{-1}$ is also continuous.

Then we compute $\dot{V}$ and its inverse. Since $\dot{\Psi}_{21} : lin\Theta \mapsto l^\infty(\mathcal{H}_2)$, $\dot{\Psi}_{22}^{-1} : l^\infty(\mathcal{H}_2) \mapsto lin\mathbb{A}$ and $\dot{\Psi}_{12} : lin\mathbb{A} \mapsto l^\infty(\mathcal{H}_1)$ are all linear maps, $\dot{\Psi}_{12}\dot{\Psi}_{22}^{-1}\dot{\Psi}_{21}$ is a linear map : $lin\Theta \to l^\infty(\mathcal{H}_1)$. In addition, $\dot{\Psi}_{11}$ also maps $lin\Theta$ to $l^\infty(\mathcal{H}_1)$, therefore $\dot{V} = \dot{\Psi}_{11} - \dot{\Psi}_{12}\dot{\Psi}_{22}^{-1}\dot{\Psi}_{21}$: $lin\Theta \mapsto l^\infty(\mathcal{H}_1)$. To obtain $\dot{V}(\theta - \theta_0)h_1$, we first compute $\dot{\Psi}_{12}\dot{\Psi}_{21}^{-1}\dot{\Psi}_{22}(\theta - \theta_0)h_1$. By (3.12),

$$\dot{\Psi}_{21}(\theta - \theta_0)h_2 = -\int_0^\tau h_2(t)PZ^TY(t)dt(\theta - \theta_0),$$

so that, considered as an element of $\mathbf{R}(\dot{\Psi}_{22})$,

$$d\{\dot{\Psi}_{21}(\theta - \theta_0)\}(t) = -PZ^TY(t)dt(\theta - \theta_0).$$

By (3.12)

$$\dot{\Psi}_{22}^{-1}\{\dot{\Psi}_{21}(\theta - \theta_0)\}h_2(t) = -\int_0^\tau \frac{h_2(t)}{PY(t)} d\{\dot{\Psi}_{21}(\theta - \theta_0)(t)\} = \int_0^\tau \frac{h_2(t)}{PY(t)} PZ^TY(t)dt(\theta - \theta_0).$$

so that, considered as an element of $lin\mathbb{A}$,

$$d\{\dot{\Psi}_{22}^{-1}\dot{\Psi}_{21}(\theta - \theta_0)\} = \frac{1}{PY(t)} PZ^TY(t)dt(\theta - \theta_0).$$

According to (3.12),

$$\dot{\Psi}_{21}\left\{\dot{\Psi}_{11}^{-1}\dot{\Psi}_{12}(\theta - \theta_0)\right\}h_1 = -h_1^T \int_0^\tau \frac{PZY(t)}{PY(t)} PZ^TY(t)dt(\theta - \theta_0).$$

Therefore

$$\begin{aligned}
\dot{V}(\theta - \theta_0)h_1 =& \dot{\Psi}_{22}(\theta - \theta_0)h_1 - \dot{\Psi}_{21}\dot{\Psi}_{11}^{-1}\dot{\Psi}_{12}(\theta - \theta_0)h_1 \\
=& -h_1^T P \int_0^\tau Y(t)ZZ^Tdt(\theta - \theta_0) + h_1^T \int_0^\tau \frac{PZY(t)PZ^TY(t)}{PY(t)} dt(\theta - \theta_0) \\
=& -h_1^T \left[\int_0^\tau P\{Z - \frac{PZY(t)}{PY(t)}\}^{\otimes 2} Y(t)dt\right](\theta - \theta_0).
\end{aligned}$$

In A. 3.3.4 we assume the matrix $A = \left[\int_0^\tau P\{Z - \frac{PZY(t)}{PY(t)}\}^{\otimes 2} Y(t)dt\right]$ is positive definite, so its inverse A^{-1} exists. Then the inverse operator $\dot{V}^{-1}$ exists, which applies A^{-1} to the range

of $\dot{V}$. By assumption, the domain of $\dot{V}$ is closed and bounded and so is the range of $\dot{V}$. $\dot{V}^{-1}$ as a linear map between Euclidean spaces is automatically continuous. Therefore $\dot{V}$ has a continuous inverse on its range. We have verified the continuous invertibility of the two operators $\dot{V}$ and $\dot{\Psi}_{22}$ in the partitioned form of $\dot{\Psi}_0$. According to van der Vaart [1998, p.422], $\dot{\Psi}_0$ has a continuous inverse on its range. □

3.7.5 Convergence in Quadratic Mean

Lemma 3.7.5. *Assume A.3.3.1 ∼ A.3.3.2 hold. Then* $\|P(\psi_{\theta,\Lambda,h} - \psi_{\theta_0,\Lambda_0,h})^2\|_{\mathcal{H}} \to 0$ *as* $(\theta, \Lambda) \to (\theta_0, \Lambda_0)$.

Proof. In view of

$$\psi_{1,\theta,\Lambda,h_1}(X) - \psi_{1,\theta_0,\Lambda_0,h_1}(X) = -h_1^T Z \int_0^\tau Y(t) d(\Lambda - \Lambda_0)(t) - h_1^T Z \int_0^\tau Y(t) Z^T (\theta - \theta_0) dt$$

and the fact that $\mathcal{H}_1$ is a bounded subset of $\mathbb{R}^p$, $\psi_{1,\theta,\Lambda,h_1}(X) \to \psi_{1,\theta_0,\Lambda_0,h_1}(X)$ pointwise and uniformly in h_1.

Next, since

$$\psi_{2,\theta,\Lambda,h_2} - \psi_{2,\theta_0,\Lambda_0,h_2} = -\int_0^\tau h_2(t) Y(t) d(\Lambda - \Lambda_0)(t) - \int_0^\tau h_2(t) Y(t) dt Z^T (\theta - \theta_0)$$

we have $\psi_{2,\theta,\Lambda,h_2}(X) \to \psi_{2,\theta_0,\Lambda_0,h_2}(X)$ pointwise. There exists two sets M, N, such that

1) $M \cup N = [0, \tau]$ and $M \cap N = \varnothing$

2) for $t \in M$, $h_2(t) d(\Lambda - \Lambda_0)(t) >= 0$

3) for $t \in N, h_2(t) d(\Lambda - \Lambda_0)(t) < 0$.

With this result and the fact that $Y(t) \leq 1$ for all $t \in [0, \tau]$, we have

$$\begin{aligned}
\sup_{h_2 \in \mathcal{H}_2} \left| \int_0^\tau \{h_2(t) Y(t) d(\Lambda - \Lambda_0)(t)\} \right| &= \sup_{h_2 \in \mathcal{H}_2} \left\{ \int_M h_2(t) Y(t) d(\Lambda - \Lambda_0)(t) - \int_N h_2(t) Y(t) d(\Lambda - \Lambda_0)(t) \right\} \\
&\leq \sup_{h_2 \in \mathcal{H}_2} \left\{ \int_M h_2(t) d(\Lambda - \Lambda_0)(t) - \int_N h_2(t) d(\Lambda - \Lambda_0)(t) \right\} \\
&= \sup_{h_2 \in \mathcal{H}_2} | \int_0^\tau h_2(t) d(\Lambda - \Lambda_0)(t)| \\
&= \|\Lambda - \Lambda_0\|_{\mathcal{H}_2}.
\end{aligned}$$

Hence

$$\begin{aligned}\sup_{h_2\in\mathcal{H}_2}|\psi_{2,\theta,\Lambda,h_2}-\psi_{2,\theta_0,\Lambda_0,h_2}| &= \sup_{h_2\in\mathcal{H}_2}|\int_0^\tau h_2(t)Y(t)d(\Lambda-\Lambda_0)(t)+\int_0^\tau h_2(t)Y(t)dtZ^T(\theta-\theta_0)| \\ &\le \sup_{h_2\in\mathcal{H}_2}|\int_0^\tau h_2(t)Y(t)d(\Lambda-\Lambda_0)(t)|+\sup_{h_2\in\mathcal{H}_2}|\int_0^\tau h_2(t)Y(t)dtZ^T(\theta-\theta_0)| \\ &\le \|\Lambda-\Lambda_0\|_{\mathcal{H}_2}+\sup_{h_2\in\mathcal{H}_2}|\int_0^\tau h_2(t)Y(t)dt|\|Z\|_{\mathbb{E}}\|\theta-\theta_0\|_{\mathcal{H}_1}.\end{aligned}$$

Because $\mathcal{H}_2$ is uniformly bounded and of uniformly bounded variation, $\sup_{h_2\in\mathcal{H}_2}|\int_0^\tau h_2(t)Y(t)dt|$ is bounded. As $(\theta,\Lambda)\to(\theta_0,\Lambda_0)$, $\sup_{h_2\in\mathcal{H}_2}|\psi_{2,\theta,\Lambda,h_2}-\psi_{2,\theta_0,\Lambda_0,h_2}|\to 0$. Thus $\psi_{\theta,\Lambda,h}\to\psi_{\theta_0,\Lambda_0,h}$ pointwise and uniformly in h. By dominated convergence theorem, we have $\|P(\psi_{\theta,\Lambda,h}-\psi_{\theta_0,\Lambda_0,h})^2\|_{\mathcal{H}}\to 0$ as $(\theta,\Lambda)\to(\theta_0,\Lambda_0)$ [van der Vaart and Wellner, 1996, p.317]. □

3.8 Consistency of RS and Two-phase VPS Estimators

After studying various properties of $\psi_{\theta,\Lambda}$, we now use them, together with the theoretical tools developed in Chapter 2, to prove consistency and asymptotic normality of our estimators.

Theorem 3.8.1. *Suppose assumptions A.3.3.1 - 3.3.2 hold, Then*

$$\begin{pmatrix} \hat{\theta} \\ \hat{\Lambda} \end{pmatrix} \to_p \begin{pmatrix} \theta_0 \\ \Lambda_0 \end{pmatrix}$$

Note that Λ ranges over a non-Euclidean space. $\hat{\Lambda} \to_p \Lambda_0$ means $\|\hat{\Lambda} - \Lambda_0\|_{\mathcal{H}_2} \to_p 0$.

Proof. Lemma 3.7.3 shows under the assumption of A.3.3.1 and A.3.3.2. $\{\psi_{\Lambda,\theta,h}, \theta \in \Theta, \Lambda \in \mathbb{A}, h \in \mathcal{H}\}$ is Glivenko-Cantelli(GC), with finite envelope function. Lemma 3.7.1 shows under the assumption of A.3.3.1, (θ_0, Λ_0) is a well separated solution to $\Psi(\theta, \Lambda) = 0$. According to corollary 2.4.2, $(\hat{\theta}, \hat{\Lambda})$ defined in (3.5) converges to (θ_0, Λ_0) in probability. □

Theorem 3.8.2. *In addition to A.3.3.1 - A.3.3.2, suppose assumptions A.3.3.5 and A. 3.3.6 also hold. Then*

$$\begin{pmatrix} \hat{\theta}^* \\ \hat{\Lambda}^* \end{pmatrix} \to_p \begin{pmatrix} \theta_0 \\ \Lambda_0 \end{pmatrix}$$

Proof. Because assumptions A. 3.3.5 and A.3.3.6 hold, applying corollary 2.4.3 to results from Lemma 3.7.1 and lemma 3.7.3 gives that $(\hat{\theta}^*, \hat{\Lambda}^*)$ defined in (3.6) converges to (θ_0, Λ_0) in probability. □

3.9 Joint Limiting Distributions of RS and Two-phase VPS Estimators

Theorem 3.9.1. *Under assumptions A.3.3.1 ∼ A.3.3.4,*

$$\dot{\Psi}_0\sqrt{N}\begin{pmatrix}\hat{\theta}-\theta_0\\ \hat{\Lambda}-\Lambda_0\end{pmatrix} = -\mathbb{G}_N\psi_{\theta_0,\Lambda_0} + o_P(1). \tag{3.15}$$

Proof. Theorem 3.9.1 is an immediate result from theorem 2.4.4 when we replace α in Theorem 2.4.4 by (θ,Λ). In our preliminary results on $\psi_{\theta,\Lambda}$, Lemma 3.7.2, lemma 3.7.4, and lemma 3.7.5 have verified, under the assumption of A.3.3.1∼A.3.3.4, conditions 2.4.4∼2.4.6 in theorem 2.4.4 are satisfied. Theorem 3.8.1 also proved $(\hat{\theta},\hat{\Lambda}) \to_p (\theta_0,\Lambda_0)$. By construction, estimators in (3.5) satisfy $\|\mathbb{P}_N\psi_{\hat{\theta},\hat{\Lambda}}\|_{\mathcal{H}} = 0$. Therefore it follows from theorem 2.4.4 that (3.15) holds. □

Let $\mathcal{F}_0 \equiv \{\psi_{\Lambda_0,\theta_0,h}(X), h\in\mathcal{H}\}$. Since $\mathcal{F}_0$ is a subset of the P-Donsker class $\mathcal{F}$ defined in lemma 3.7.2, $\mathcal{F}_0$ is also P-Donsker [van der Vaart and Wellner, 1996, theorem 2.10.1]. On the RHS of (3.15), the empirical process $\mathbb{G}_N$ indexed by $\mathcal{F}_0$ converges in distribution to a P-Brownian bridge $\mathbb{G}$ in the space $l^\infty(\mathcal{F}_0)$. According to (2.12), the P-Brownian bridge is a zero-mean Gaussian with covariance

$$Cov_A\left\{\dot{\Psi}_0\sqrt{N}\begin{pmatrix}\hat{\theta}-\theta_0\\ \hat{\Lambda}-\Lambda_0\end{pmatrix}h, \dot{\Psi}_0\sqrt{N}\begin{pmatrix}\hat{\theta}-\theta_0\\ \hat{\Lambda}-\Lambda_0\end{pmatrix}g\right\} = P(\psi_{\theta_0,\Lambda_0,h}\psi_{\theta_0,\Lambda_0,g}), \forall g,h\in\mathcal{H} \tag{3.16}$$

and variance

$$Var_A\left\{\dot{\Psi}_0\sqrt{N}\begin{pmatrix}\hat{\theta}-\theta_0\\ \hat{\Lambda}-\Lambda_0\end{pmatrix}h\right\} = P\psi^2_{\theta_0,\Lambda_0,h}, \forall h\in\mathcal{H} \tag{3.17}$$

Theorem 3.9.2. *Suppose assumptions A.3.3.1 ∼ A. 3.3.6 hold. Then*

$$\dot{\Psi}\sqrt{N}\begin{pmatrix}\hat{\Lambda}^*-\Lambda_0\\ \hat{\theta}^*-\theta_0\end{pmatrix} = -\mathbb{G}_N\psi^*_{\theta_0,\Lambda_0} + o_P(1). \tag{3.18}$$

Proof. Under the assumption of A.3.3.1 ∼3.3.4, Lemma 3.7.2, lemma 3.7.4, and lemma 3.7.5 hold. Thus conditions 2.4.4∼2.4.6 are satisfied. Theorem 3.8.2 shows $(\hat{\theta}, \hat{\Lambda}^*)$ defined in (3.6) satisfy $(\hat{\theta}, \hat{\Lambda}^*) \xrightarrow{P} (\theta_0, \Lambda_0)$. Then applying corollary 2.4.5 we obtain (3.18) holds. □

Let $\mathcal{F}_0' \equiv \{\frac{R}{\pi_0(V)}\psi_{\theta_0,\Lambda_0,h}(X), h \in \mathcal{H}\}$. According to corollary 2.4.5, on the RHS of (3.18), the empirical process $\mathbb{G}_N$ indexed by $\mathcal{F}_0'$ converges in distribution to a Q-Brownian bridge process $\mathbb{G}$ in the space $l^\infty(\mathcal{F}_0')$ with mean 0 and covariance

$$Cov_A\left\{\dot{\Psi}_0\sqrt{N}\begin{pmatrix}\hat{\theta}^*-\theta_0\\ \hat{\Lambda}^*-\Lambda_0\end{pmatrix}h, \dot{\Psi}_0\sqrt{N}\begin{pmatrix}\hat{\theta}^*-\theta_0\\ \hat{\Lambda}^*-\Lambda_0\end{pmatrix}g\right\}$$
$$= P(\psi_{\theta_0,\Lambda_0,h}\psi_{\theta_0,\Lambda_0,g}) + Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\psi_{\theta_0,\Lambda_0,h}\psi_{\theta_0,\Lambda_0,g}\right], \forall g, h \in \mathcal{H} \quad (3.19)$$

and variance

$$Var_A\left\{\dot{\Psi}_0\sqrt{N}\begin{pmatrix}\hat{\theta}^*-\theta_0\\ \hat{\Lambda}^*-\Lambda_0\end{pmatrix}h\right\} = P\psi^2_{\theta_0,\Lambda_0,h} + Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\psi^2_{\theta_0,\Lambda_0,h}\right], \forall h \in \mathcal{H}. \quad (3.20)$$

3.10 Limiting Distributions of Some Interesting Statistics

Based on the asymptotic joint distribution of $\dot{\Psi}_0\sqrt{N}\left(\hat{\theta}-\theta_0, \hat{\Lambda}-\Lambda_0\right)$ from last section, we are able to find limiting distributions for various interesting statistics. These include the regression parameter estimators $\sqrt{N}(\hat{\theta}-\theta_0)$ and $\sqrt{N}(\hat{\theta}^*-\theta_0)$, the cumulative baseline hazards estimators $\sqrt{N}(\hat{\Lambda}-\Lambda_0)$ and $\sqrt{N}(\hat{\Lambda}^*-\Lambda_0)$, and the individual specific cumulative hazards $\sqrt{N}\{\hat{\Lambda}(t|Z=z)-\Lambda_0(t|Z=z)\}$ and $\sqrt{N}\{\hat{\Lambda}^*(t|Z=z)-\Lambda_0(t|Z=z)\}$ from both RS and two-phase VPS data. Because we are also concerned with model misspecification, two types of variances, the model-based and the robust variances will be provided for each statistic.

We used the same approach to develop the six statistics above. Taking the developments of RS estimators as an example, the developments are arranged in the following steps. $h'=(h_1', h_2')$ such that $\dot{\Psi}_0\sqrt{N}\begin{pmatrix}\hat{\theta}-\theta_0\\ \hat{\Lambda}-\Lambda_0\end{pmatrix}h'$ equals a desired statistic, which could be $\sqrt{N}(\hat{\theta}-\theta_0)$, $\sqrt{N}(\hat{\Lambda}-\Lambda_0)$, or $\sqrt{N}\{\hat{\Lambda}(t|Z=z)-\Lambda_0(t|Z=z)\}$. Since θ, Λ and $\Lambda(t|Z=z)$ can all be written as a linear combination of (θ, Λ), it is likely h' is identifiable. In step 2, we verify $h' \in \mathcal{H}\cdot M$ where $M<\infty$. As a result (3.9.1) holds for $h=h'$. According to (3.9.1), we conclude the statistic of interest converges to a zero-mean Gaussian distribution or process. In step 3, we compute $\psi_{\theta,\Lambda,h'}$ and the asymptotic variances of the statistic based on(3.17).

For developing two-phase VPS estimators and calibrated two-phase VPS estimators in, we can use the same procedure and the same h' to derive the limiting distribution of these new estimators.The only changes are the theorem we will use for establishing asymptotic normality and the formula for calculating asymptotic variances. Since the procedure is very standardized, a collection of estimators can be developed rapidly. Since the procedure in the following six sections will be similar to each other, understanding section 3.10.1, the simplest scenario among these six derivations, will facilitate the comprehension of others.

3.10.1 Limiting Distribution of $\sqrt{N}(\hat{\theta}-\theta_0)$

In section 2.4.1, we considered θ as an element in $l^\infty(\mathcal{H}_1)$ identified by $\theta h_1 = h_1^T\theta$. To find the limiting distribution of $\sqrt{N}(\hat{\theta}-\theta_0)$ is equivalent to finding the asymptotic distribution of $h_1^T\sqrt{N}(\hat{\theta}-\theta_0)$ for any $h_1 \in \mathcal{H}_1$.

Step 1:

Let

$$h' = (h_1', h_2') = \left(-A^{-1}h_1, \frac{(A^{-1}h_1)^T PY(t)Z}{PY(t)}\right), t \in [0,\tau] \tag{3.21}$$

where A is the $p \times p$ matrix defined in A. 3.3.4. We will verify with this h', the LHS of theorem 3.9.1 equals $h_1'^T(\hat{\theta}-\theta_0)$. Based on (3.12), we have

$$\begin{aligned}
\dot{\Psi}_{11}(\hat{\theta}-\theta_0)h_1' =& (A^{-1}h_1)^T P\{Y(t)ZZ^T\}dt(\hat{\theta}-\theta_0)\\
\dot{\Psi}_{21}(\hat{\theta}-\theta_0)h_2' =& -(A^{-1}h_1)^T\int_0^\tau \frac{P\{Y(t)Z\}}{PY(t)}P\{Y(t)Z^T\}dt(\hat{\theta}-\theta_0)\\
\dot{\Psi}_{12}(\hat{\Lambda}-\Lambda_0)h_1' =& (A^{-1}h_1)^T\int_0^\tau P\{Y(t)Z\}d(\hat{\Lambda}-\Lambda_0)\\
\dot{\Psi}_{22}(\hat{\Lambda}-\Lambda_0)h_2' =& -(A^{-1}h_1)^T\int_0^\tau \frac{P\{Y(t)Z\}}{PY(t)}PY(t)d(\hat{\Lambda}-\Lambda_0).
\end{aligned}$$

Hence

$$\sqrt{N}\dot{\Psi}_{12}(\hat{\Lambda}-\Lambda_0)h_1' + \sqrt{N}\dot{\Psi}_{22}(\hat{\Lambda}-\Lambda_0)h_2' = 0$$

and the LHS of theorem 3.9.1 becomes

$$\begin{aligned}
&\sqrt{N}\dot{\Psi}_{11}(\hat{\theta}-\theta_0)h_1' + \sqrt{N}\dot{\Psi}_{21}(\hat{\theta}-\theta_0)h_2' + \sqrt{N}\dot{\Psi}_{12}(\hat{\Lambda}-\Lambda_0)h_1' + \sqrt{N}\dot{\Psi}_{22}(\hat{\Lambda}-\Lambda_0)h_2'\\
=&\sqrt{N}h_1^T A^{-1}\int_0^\tau \left[P\{Y(t)ZZ^T\} - \sqrt{N}\frac{P\{Y(t)Z\}}{PY(t)}P\{Y(t)Z^T\}\right]dt(\hat{\theta}-\theta_0) + 0\\
=&\sqrt{N}h_1^T A^{-1}\int_0^\tau P\{Z - \frac{PZY(t)}{PY(t)}\}^{\otimes 2}Y(t)dt(\hat{\theta}-\theta_0)\\
=&\sqrt{N}h_1^T A^{-1}A(\hat{\theta}-\theta_0)\\
=&h_1^T\sqrt{N}(\hat{\theta}-\theta_0).
\end{aligned} \tag{3.22}$$

Step 2:

In this step we will show $h' \in \mathcal{H} \cdot M$. First, A^{-1} is a $p \times p$ bounded matrix, so $h_1 = -A^{-1}h_1$ is a bounded p-dimensional vector and thus $h_1' \in \mathcal{H}_1 \cdot M$. Next, we prove $PY(t)Z$ is $BV[0,\tau]$. We show it for each coordinate of Z. We consider $k = 1, 2, \ldots, p$. Then

$$\begin{aligned} PY(t)Z_k &= \int_{\mathbb{R}^p} E[1(T \geq t)|z] z_k dG(z) \\ &= \int_{\mathbb{R}^p} S(t|z) z_k dG(z) \\ &= \int_{\mathbb{R}^p, z_k > 0} S(t|z) z_k dG(z) + \int_{\mathbb{R}^p, z_k < 0} S(t|z) z_k dG(z). \end{aligned}$$

Since the first term of the above display is monotonic decreasing in t and the second term monotonic increasing in t, $PY(t)Z_k$ as the summation of monotonic functions is of bounded variation. By assumption, $PY(t)Z_k$ is also uniformly bounded. Because $1/PY(t)$ is monotonic bounded, $h_2'(t) = \frac{(A^{-1}h_1)^T PY(t)Z}{PY(t)}$ is uniformly bounded with uniformly bounded variation and both bounds are less than M. Hence $h_2' \in \mathcal{H}_2 \cdot M$ and $h' \in \mathcal{H} \cdot M$. We obtain (3.15) holds for $h = h'$. Therefore, according to (3.15) and (3.17), $h_1^T \sqrt{N}(\hat{\theta} - \theta_0)$ converges to a normal distribution with mean 0 and variance $P\psi^2_{\theta_0,\Lambda_0,h'}$.

Step 3:

In the last step, we evaluate $\psi^2_{\theta_0,\Lambda_0,h'}$ and calculate the asymptotic variance $P\psi^2_{\theta_0,\Lambda_0,h'}$. Let $M(t) = N(t) - \Lambda_0(t) - Z^T\theta_0 t$. Note $M(t)$ is not a martingale unless $P = P_{\theta_0,\Lambda_0}$: the AH model (3.1) holds. Then

$$\begin{aligned} \psi_{\theta_0,\Lambda_0,h'} =& h_1'^T \psi_{1,\theta_0,\Lambda_0} + \psi_{2,\theta_0,\Lambda_0,h_2'} \\ =& h_1'^T \int_0^\tau Z\{dN(t) - d\Lambda_0(t) - Z^T\theta_0 dt\} + \int_0^\tau h_2'(t)\{dN(t) - d\Lambda_0(t) - Z^T\theta_0 dt\} \\ =& h_1'^T \int_0^\tau Z dM(t) + \int_0^\tau h_2'(t) dM(t) \end{aligned} \tag{3.23}$$

and

$$\begin{aligned} P\psi^2_{\theta_0,\Lambda_0,h'} =& P\{h_1'^T \int_0^\tau Z dM(t) + \int_0^\tau h_2'(t) dM(t)\}^2 \\ =& P\left\{-h_1^T A^{-1} \int_0^\tau Z dM(t) + \int_0^\tau \frac{h_1^T A^{-1} PY(t) Z}{PY(t)} dM(t)\right\}^2 \\ =& h_1^T A^{-1} P\left[\int_0^\tau \left\{Z - \frac{PY(t)Z}{PY(t)}\right\} dM(t)\right]^{\otimes 2} A^{-1} h_1. \end{aligned}$$

Since this result holds for all h_1 that belong to $\mathcal{H}_1$, we obtain $\sqrt{N}(\hat{\theta} - \theta_0)$ converges in distribution to a p-variate Gaussian distribution with mean zero and variance

$$rVar_A\left\{\sqrt{N}(\hat{\theta} - \theta_0)\right\} = A^{-1} P\left[\int_0^\tau \left\{Z - \frac{PY(t)Z}{PY(t)}\right\} dM(t)\right]^{\otimes 2} A^{-1}. \tag{3.24}$$

This variance is a robust variance because it does not use any result derived from assuming (3.1). We use notation r in front of Var_A to denote this feature.

When the AH model (3.1) holds, $Y(t)d\Lambda_0(t) + Y(t)z^T\theta_0 dt = Y(t)d\Lambda(t|Z = z)$, which is an integrated intensity process. $M(t)$ as the difference of a counting process $N(t)$ and an integrated intensity process $Y(t)d\Lambda(t|Z = z)$ is a martingale under $P = P_{\theta_0,\Lambda_0}$. In consequence, $\int_0^\tau \left\{Z - \frac{PY(t)Z}{PY(t)}\right\} Y(t)dM(t)$ is a martingale integral with a predictable and locally bounded integrand. It follows from the principle of martingale integral variance [Fleming and Harrington, 1991, theorem 2.4.2] that

$$\begin{aligned} P\left[\int_0^\tau \left\{Z - \frac{PY(t)Z}{PY(t)}\right\} Y(t)dM(t)\right]^{\otimes 2} =& Var\left[\int_0^\tau \left\{Z - \frac{PY(t)Z}{PY(t)}\right\} dM(t)\right] \\ =& P\int_0^\tau \left\{Z - \frac{PY(t)Z}{PY(t)}\right\}^{\otimes 2} Y(t)\left\{d\Lambda_0(t) + Z^T\theta_0 dt\right\} \end{aligned} \tag{3.25}$$

$$\equiv B. \tag{3.26}$$

As a result, (3.24) becomes

$$Var_A\left\{\sqrt{N}(\hat{\theta} - \theta_0)\right\} = A^{-1}BA^{-1}. \tag{3.27}$$

Note using martingale theory result, we implicitly assume under the AH model the compensator for the counting process is absolutely continuous, in other words, the existence of

intensity process for the counting process [Martinussen and Scheike, 2006, p.24]. In conclusion, when the AH model holds, $\sqrt{N}(\hat{\theta}-\theta_0)$ converges in distribution to a p-variate Gaussian distribution with mean zero and covariance matrix $A^{-1}BA^{-1}$. This result agrees with Lin and Ying [1994].

Reviewing the approach in this section, we have demonstrated the convenience resulted from the index set $\mathcal{H}$. Compared to other approaches we have attempted to derive the limiting distribution of $\sqrt{N}(\hat{\theta}-\theta_0)$ (methods not shown), we avoid an explicit computation of $\dot{\Psi}_0^{-1}$, which can be complicated, and we avoid using delta methods. Instead we chose a particular h' in step 1 such that the LHS of (3.15) becomes the desired statistic, for which we aim to derive the limiting distribution. We used the fact that we had shown the Donsker property for a class of functions that contains $\psi_{\theta_0,\Lambda_0,h}$ for "all" $h \in \mathcal{H}$ in section 3.7. By demonstrating the new h' is an element of $\mathcal{H} \cdot M$, it follows that $\{\psi_{\theta_0,\Lambda_0,h'}\}$ is a P-Donsker class, from which the limiting distribution results. More importantly, the same approach and choice of h' will be carried on for developing two-phase VPS and calibrated two-phase VPS estimators. These benefits will be revealed more in the coming examples.

3.10.2 Limiting Distribution of $\sqrt{N}(\hat{\theta}^*-\theta_0)$

For two-phase VPS, we use the same h' in (3.21) to develop the limiting distribution of $\sqrt{N}(\hat{\theta}^*-\theta_0)$. Since the Fréchet derivative in (3.18) is the same as (3.15), according to result (3.22),

$$\sqrt{N}\dot{\Psi}_0 \begin{pmatrix} \hat{\Lambda}^*-\Lambda_0 \\ \hat{\theta}^*-\theta_0 \end{pmatrix} h' = h_1^T\sqrt{N}(\hat{\theta}^*-\theta_0).$$

Thus based on (3.18) and (3.20), $h_1^T\sqrt{N}(\hat{\theta}^*-\theta_0)$ converges to a normal distribution with mean 0 and variance $P(\psi_{\theta_0,\Lambda_0,h'})^2 + Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\psi^2_{\theta_0,\Lambda_0,h'}\right]$.

Based on the derivations for $\psi_{\theta_0,\Lambda_0,h'}$ and $P\psi^2_{\theta_0,\Lambda_0,h'}$ in step 3 from section 3.10.1, we

obtain

$$P(\psi_{\theta_0,\Lambda_0}h')^2+Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\psi^2_{\theta_0,\Lambda_0,h'}\right]=h_1^TA^{-1}P\left[\int_0^\tau\left\{Z-\frac{PY(t)Z}{PY(t)}\right\}dM(t)\right]^{\otimes 2}A^{-1}h_1$$
$$+Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}h_1^TA^{-1}\left[\int_0^\tau\left\{Z-\frac{PY(t)Z}{PY(t)}\right\}dM(t)\right]^{\otimes 2}A^{-1}h_1\right]$$

Since this result holds for all $h_1 \in \mathcal{H}_1$, we conclude $\sqrt{N}(\hat{\theta}^*-\theta_0)$ converges in distribution to a p-variate Gaussian distribution with mean zero and variance matrix

$$rVar_A\left\{\sqrt{N}(\hat{\theta}^*-\theta_0)\right\}=A^{-1}P\left\{\int_0^\tau\left[Z-\frac{PY(t)Z}{PY(t)}\right]dM(t)\right\}^{\otimes 2}A^{-1}$$
$$+Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}A^{-1}\left[\int_0^\tau\left\{Z-\frac{PY(t)Z}{PY(t)}\right\}dM(t)\right]^{\otimes 2}A^{-1}\right] \quad (3.28)$$

Equation (3.28) is our robust variance because we do not require the assumed AH model to hold. When model (3.1) holds, the first term in (3.28) is replaced by the model-based variance for the RS estimator $Var_A\left\{\sqrt{N}(\hat{\theta}-\theta_0)\right\}$ in (3.27). Thus

$$Var_A\left\{\sqrt{N}(\hat{\theta}^*-\theta_0)\right\}=A^{-1}BA^{-1}+Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}A^{-1}\left[\int_0^\tau\left\{Z-\frac{PY(t)Z}{PY(t)}\right\}dM(t)\right]^{\otimes 2}A^{-1}\right]. \quad (3.29)$$

3.10.3 Limiting distribution of $\sqrt{N}(\hat{\Lambda}-\Lambda_0)$

Step 1:

We redefine h' as

$$h'=(h_1',h_2')=\left(A^{-1}Dh_2,-\frac{h_2(t)}{PY(t)}-\frac{(A^{-1}Dh_2)^TPY(t)Z}{PY(t)}\right),t\in[0,\tau] \quad (3.30)$$

where $D \in l^\infty(\mathcal{H}_2):\mathcal{H}_2\to\mathbb{R}^p$, defined by

$$Dh_2=\int_0^\tau h_2(t)\frac{PY(t)Z}{PY(t)}dt. \quad (3.31)$$

Let h in (3.15) equal h'. Then the LHS of (3.15) becomes

$$\begin{aligned}
&\sqrt{N}\dot{\Psi}_{11}(\hat{\theta}-\theta_0)h_1' + \sqrt{N}\dot{\Psi}_{21}(\hat{\theta}-\theta_0)h_2' + \sqrt{N}\dot{\Psi}_{12}(\hat{\Lambda}-\Lambda_0)h_1' + \sqrt{N}\dot{\Psi}_{22}(\hat{\Lambda}-\Lambda_0)h_2' \\
=&\sqrt{N}\dot{\Psi}_{11}(\hat{\theta}-\theta_0)A^{-1}Dh_2 + \sqrt{N}\dot{\Psi}_{21}(\hat{\theta}-\theta_0)\left[-\frac{h_2(t)}{PY(t)} - \frac{(A^{-1}Dh_2)^T PY(t)Z}{PY(t)}\right] \\
&+ \sqrt{N}\dot{\Psi}_{12}(\hat{\Lambda}-\Lambda_0)A^{-1}Dh_2 + \sqrt{N}\dot{\Psi}_{22}(\hat{\Lambda}-\Lambda_0)\left[-\frac{h_2(t)}{PY(t)} - \frac{(A^{-1}Dh_2)^T PY(t)Z}{PY(t)}\right].
\end{aligned}$$

Based on (3.12), the sum of the first two terms in the preceding display equals 0 as follows.

$$\begin{aligned}
&\dot{\Psi}_{11}(\hat{\theta}-\theta_0)A^{-1}Dh_2 + \dot{\Psi}_{21}(\hat{\theta}-\theta_0)\left\{-\frac{h_2(t)}{PY(t)} + \frac{(A^{-1}Dh_2)^T PY(t)Z}{PY(t)}\right\} \\
=&-(A^{-1}Dh_2)^T\int_0^\tau P\{Y(t)ZZ^T\}(\hat{\theta}-\theta_0) \\
&\quad + \int_0^\tau \left\{\frac{h_2(t)}{PY(t)} + \frac{(A^{-1}Dh_2)^T PY(t)Z}{PY(t)}\right\} P\{Y(t)Z^T\}dt(\hat{\theta}-\theta_0) \\
=&-(Dh_2)^T A^{-1}\int_0^\tau P\{Y(t)ZZ^T\}(\hat{\theta}-\theta_0) + (Dh_2)^T(\hat{\theta}-\theta_0) \\
&\quad + (Dh_2)^T A^{-1}\int_0^\tau \frac{PY(t)ZPY(t)Z^T}{PY(t)}dt(\hat{\theta}-\theta_0) \\
=&-(Dh_2)^T A^{-1}\int_0^\tau \left\{PY(t)ZZ^T - \frac{PY(t)ZPY(t)Z^T}{PY(t)}\right\}(\hat{\theta}-\theta_0) + (Dh_2)^T(\hat{\theta}-\theta_0) \\
=&-(Dh_2)^T A^{-1}A(\hat{\theta}-\theta_0) + (Dh_2)^T(\hat{\theta}-\theta_0) \\
=&0.
\end{aligned}$$

The sum of the third and fourth terms is $(\hat{\Lambda}-\Lambda_0)h_2$ because

$$\begin{aligned}
&\dot{\Psi}_{12}(\hat{\Lambda}-\Lambda_0)A^{-1}Dh_2 + \dot{\Psi}_{22}(\hat{\Lambda}-\Lambda_0)\left[-\frac{h_2(t)}{PY(t)} - \frac{(A^{-1}Dh_2)^T PY(t)Z}{PY(t)}\right] \\
=&-(Dh_2)^T A^{-1}\int_0^\tau P\{Y(t)Z\}d(\hat{\Lambda}-\Lambda_0) + \int_0^\tau h_2(t)d(\hat{\Lambda}-\Lambda_0) \\
&\quad + (Dh_2)^T A^{-1}\int_0^\tau P\{Y(t)Z\}d(\hat{\Lambda}-\Lambda_0) \\
=&\int_0^\tau h_2(t)d(\hat{\Lambda}-\Lambda_0) \\
=&(\hat{\Lambda}-\Lambda_0)h_2.
\end{aligned}$$

Hence

$$\sqrt{N}\dot{\Psi}_{11}(\hat{\theta}-\theta_0)h_1'+\sqrt{N}\dot{\Psi}_{21}(-\hat{\theta}-\theta_0)h_2'+\sqrt{N}\dot{\Psi}_{12}(\hat{\Lambda}-\Lambda_0)h_1'+\sqrt{N}\dot{\Psi}_{22}(\hat{\Lambda}-\Lambda_0)h_2' = \sqrt{N}(\hat{\Lambda}-\Lambda_0)h_2. \tag{3.32}$$

Step 2:

Now we demonstrate $h' \in \mathcal{H} \cdot M$. Because $Dh_2 = \int_0^\tau \frac{h_2(t)PY(t)Z}{PY(t)}dt$ is a bounded vector in $\mathbb{R}^p$ and A^{-1} is a $p \times p$ matrix, Dh_2, $A^{-1}Dh_2 \in \mathcal{H}_1 \cdot M$. In section 3.10.1, we have shown for any $h_1 \in \mathcal{H}_1$, $\frac{(A^{-1}h_1)^T PY(t)Z}{PY(t)} \in \mathcal{H}_2 \cdot M$. Thus $\frac{(A^{-1}Dh_2)^T PY(t)Z}{PY(t)} \in \mathcal{H}_2 \cdot M$. Since both $h_2(t)$ and $PY(t)$ are of bounded variation and are bounded functions, $\frac{h_2(t)}{PY(t)} \in \mathcal{H}_2 \cdot M$. Thus $h_2' \equiv -\frac{h_2(t)}{PY(t)} - \frac{(A^{-1}Dh_2)^T PY(t)Z}{PY(t)} \in \mathcal{H}_2 \cdot M$. Therefore $h' \in \mathcal{H} \cdot M$ and (3.15) holds for h'. Let

$$g' = (g_1', g_2') = \left(A^{-1}Dg_2, -\frac{g_2(t)}{PY(t)} - \frac{(A^{-1}Dg_2)^T PY(t)Z}{PY(t)}\right), t \in [0,\tau] \tag{3.33}$$

where $g_2 \in \mathcal{H}_2$. According to (3.32), (3.15) and (3.17), $\sqrt{N}(\hat{\Lambda}-\Lambda_0)$ converges to a Brownian bridge process in $l^\infty(\mathcal{H}_2)$ with mean 0 and covariance process $P[\psi_{1,\theta_0,\Lambda_0,h'}\psi_{1\theta_0,\Lambda_0,g'}]$ which depends only on $h_2, g_2 \in \mathcal{H}_2$.

Step 3:

In the final step, we compute $\psi_{\theta_0,\Lambda_0,h'}$ for the h' we set in (3.30). By (3.23), $\psi_{\theta_0,\Lambda_0,h'} = h_1'^T \int_0^\tau Z dM + \int_0^\tau h_2'(t)dM(t)$. Then

$$\begin{aligned}\psi_{\theta_0,\Lambda_0,h'} =& (Dh_2)^T A^{-1}\int_0^\tau Z dM + \int_0^\tau \left\{-\frac{h_2(t)}{PY(t)} + \frac{(A^{-1}Dh_2)^T PY(t)Z}{PY(t)}\right\} dM(t) \\ =& -\int_0^\tau \frac{h_2(t)}{PY(t)} + (Dh_2)^T A^{-1}\int_0^\tau \left\{Z - \frac{PY(t)Z}{PY(t)}\right\} dM(t).\end{aligned} \tag{3.34}$$

Therefore the covariance process

$$
\begin{aligned}
&rCov_A[\sqrt{N}(\hat{\Lambda}-\Lambda_0)h_2, \sqrt{N}(\hat{\Lambda}-\Lambda_0)g_2] \\
&= P\left[\int_0^\tau \frac{h_2(t)}{PY(t)}dM(t)\int_0^\tau \frac{g_2(t)}{PY(t)}dM(t)\right] \\
&- P\left\{\int_0^\tau \frac{h_2}{PY(t)}dM(t)(Dg_2)^T A^{-1}\int_0^\tau \left[Z-\frac{PY(t)Z}{PY(t)}\right]dM(t)\right\} \\
&- P\left\{\int_0^\tau (Dh_2)^T A^{-1}\int_0^\tau \left[Z-\frac{PY(t)Z}{PY(t)}\right]dM(t)\int_0^\tau \frac{g_2(t)}{PY(t)}dM(t)\right\} \\
&+ (Dh_2)^T A^{-1} P\left\{\int_0^\tau \left[Z-\frac{PY(t)Z}{PY(t)}\right]dM(t)\right\}^{\otimes 2} A^{-1}Dg_2.
\end{aligned}
\tag{3.35}
$$

(3.35) is the robust covariance since we derive it without assuming the AH model (3.1). When (3.1) holds, $M(t)$ becomes a martingale. It follows from the principle of martingale integral variance [Fleming and Harrington, 1991, theorem 2.4.2] that the model-based covariance process

$$
\begin{aligned}
&Cov_A\{\sqrt{N}(\hat{\Lambda}-\Lambda_0)h_2, \sqrt{N}(\hat{\Lambda}-\Lambda_0)g_2\} \\
=&P\left[\int_0^\tau \frac{h_2(t)g_2(t)}{P^2Y(t)}Y(t)d\left\{\Lambda_0(t)+Z^T\theta_0 dt\right\}\right] \\
&- P\left[\int_0^\tau \frac{h_2}{PY(t)}(Dg_2)^T A^{-1}\left\{Z-\frac{PY(t)Z}{PY(t)}\right\}d[\Lambda_0(t)+Z^T\theta_0 dt]\right] \\
&- P\left[\int_0^\tau (Dh_2)^T A^{-1}\int_0^\tau \left\{Z-\frac{PY(t)Z}{PY(t)}\right\}\frac{g_2(t)}{PY(t)}d\left\{\Lambda_0(t)+Z^T\theta_0 dt\right\}\right] \\
&\quad + (Dh_2)^T A^{-1}BA^{-1}Dg_2.
\end{aligned}
\tag{3.36}
$$

Given our results in (3.35) and (3.36), we are able to calculate the limiting distribution of $\sqrt{N}(\hat{\Lambda}-\Lambda_0)(s)$ for every $s\in[0,\tau]$. We let $h_2(t)=g_2(t)=1(t\le s)$ Then

$$
D(s)\equiv Dh_2 = Dg_2 = \int_0^s \frac{P\{Y(t)Z\}}{PY(t)}dt.
$$

Under model misspecification, by (3.35) $\sqrt{N}(\hat{\Lambda}-\Lambda_0)(s)$ converges to a normal distribution

with mean 0 and variance

$$
\begin{aligned}
&rVar_A\{\sqrt{N}(\hat{\Lambda}-\Lambda_0)(s)\}\\
=&P\left[\int_0^s \frac{1}{PY(t)}dM(t)\right]^2 - 2P\int_0^s \frac{1}{PY(t)}dM(t)D(s)^TA^{-1}\int_0^\tau \left[Z-\frac{PY(t)Z}{PY(t)}\right]dM(t)\\
&+D(s)^TA^{-1}P\left\{\int_0^\tau \left[Z-\frac{PY(t)Z}{PY(t)}\right]dM(t)\right\}^{\otimes 2}A^{-1}D(s).
\end{aligned}
$$

When the AH model holds, by (3.36) $\sqrt{N}(\hat{\Lambda}-\Lambda_0)(s)$ converges to a normal distribution with mean 0 and variance

$$
\begin{aligned}
&Var_A\{\sqrt{N}(\hat{\Lambda}-\Lambda_0)(s)\}\\
=&P\left[\int_0^s \frac{1}{\{PY(t)\}^2}Y(t)d\{\Lambda_0(t)+Z^T\theta_0 dt\}\right]\\
&-2P\int_0^s \frac{1}{PY(t)}D(s)^TA^{-1}\left\{Z-\frac{PY(t)Z}{PY(t)}\right\}Y(t)\{d\Lambda_0(t)+Z^T\theta_0 dt\}\\
&+D(s)^TA^{-1}BA^{-1}D(s).
\end{aligned}
\tag{3.37}
$$

Given two time points $s_1, s_2 \in [0,\tau]$, if we let $h(t)=1(t\le s_1)$ and $g(t)=1(t\le s_1)$, then (3.35) and (3.36) also provide us both model-based and robust covariances between cumulative baseline hazards at two different time points.

3.10.4 Limiting distribution of $\sqrt{N}(\hat{\Lambda}^*-\Lambda_0)$

To derive limiting distribution of $\sqrt{N}(\hat{\Lambda}^*-\Lambda_0)$ with two-phase VPS, we use the same sets of h' and g' in (3.30) and (3.33) as section 3.10.3. According to our result (3.32),

$$
\sqrt{N}\dot{\Psi}_0\begin{pmatrix}\hat{\Lambda}^*-\Lambda_0\\ \hat{\theta}^*-\theta_0\end{pmatrix}h' = h_1^T\sqrt{N}(\hat{\Lambda}^*-\Lambda_0).
$$

Thus based on (3.18) and (3.20), $\sqrt{N}(\hat{\Lambda}^*-\Lambda_0)$ converges to a Brownian bridge process in $l^\infty(\mathcal{H}_2)$ with mean 0 and covariance process $P\psi_{\theta_0,\Lambda_0,h'}\psi_{\theta_0,\Lambda_0,g'}+Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\psi_{\theta_0,\Lambda_0,h'}\psi_{\theta_0,\Lambda_0,g'}\right]$, which depends only on $h_2, g_2 \in \mathcal{H}_2$. Based on the result of $\psi_{\theta_0,\Lambda_0,h'}$ derived in step 3 from section 3.10.1, we obtain the asymptotic robust covariance of two cumulative baseline hazards.

We omit the presentation of this result in the general form. Instead we focus on deriving the variance for $\sqrt{N}(\hat{\Lambda}^* - \Lambda_0)(s)$ since it is more familiar and useful in real applications. For every $s \in [0, \tau]$, we let $h_2(t) = g_2(t) = 1(t \leq s)$. Then when the AH model is misspecified, by (3.38) $\sqrt{N}(\hat{\Lambda}^* - \Lambda_0)(s)$ converges to a normal distribution with mean 0 and variance

$$\begin{aligned}
&rVar_A\{\sqrt{N}(\hat{\Lambda}^* - \Lambda_0)(s)\} \\
=&P\psi^2_{\theta_0,\Lambda_0,h'} + Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\psi^2_{\theta_0,\Lambda_0,h'}\right] \\
=&rVar_A\{\sqrt{N}(\hat{\Lambda} - \Lambda_0)(s)\} \\
&+ Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\left\{\int_0^s \frac{1}{PY(t)}dM(t)\right\}^2\right] \\
&- 2Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\int_0^s \frac{1}{PY(t)}dM(t)\int_0^\tau \left\{Z - \frac{PY(t)Z}{PY(t)}\right\}^T dM(t)A^{-1}D(s)\right] \\
&+ Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}D(s)^T A^{-1}\left\{\int_0^\tau \left[Z - \frac{PY(t)Z}{PY(t)}\right]dM(t)\right\}^{\otimes 2} A^{-1}D(s)\right].
\end{aligned} \tag{3.38}$$

When the AH model holds, $rVar_A\{\sqrt{N}(\hat{\Lambda} - \Lambda_0)(s)\}$ in this equation is replaced by the model-based variance $Var_A\{\sqrt{N}(\hat{\Lambda} - \Lambda_0)(s)\}$ provided in (3.37) .

3.10.5 Limiting distribution of $\sqrt{N}\{\hat{\Lambda}(s|Z=z) - \Lambda_0(s|Z=z)\}$

Covariate specific cumulative function is an important statistic to describe a patient's cumulative risk of a disease. Since survival function $S(s|Z=z) = \exp\{-\Lambda(s|Z=z)\}$, if we have obtained $\sqrt{N}\{\hat{\Lambda}(s|Z=z) - \Lambda_0(s|Z=z)\}$, then we can derive $\sqrt{N}\{\hat{S}(s|Z=z) - S_0(s|Z=z)\}$ immediately by applying the delta method. Thus we are interested to find the limiting distribution of $\sqrt{N}\{\hat{\Lambda}(s|Z=z) - \Lambda_0(s|Z=z)\}$. For this section and the following section, we assume the AH model hold. Then $\Lambda_0(s|Z=z) = \Lambda_0(s) + \int_0^s z^T\theta_0 dt$, we let

$$\hat{\Lambda}(s|Z=z) = \hat{\Lambda}(s) + \int_0^s z^T\hat{\theta}dt. \tag{3.39}$$

Because $\hat{\Lambda}(s|Z=z)$ is the sum of $\hat{\Lambda}(s)$ and $\int_0^s z^T\hat{\theta}dt$, we set h' to be the sum of h' in (3.21) used for estimating $\sqrt{N}(\hat{\Lambda} - \Lambda_0)$ and h' in (3.30) used for estimating $\sqrt{N}(\hat{\Lambda} - \Lambda_0)$, and we

also set $h_1 = zs$ in (3.21) and $h_2 = 1(t \le s)$ in (3.30). Then the new index function

$$\begin{aligned} h' = (h_1', h_2') &= \left(-A^{-1}h_1 + A^{-1}Dh_2, \frac{(A^{-1}h_1)^T PY(t)Z}{PY(t)} - \frac{h_2(t)}{PY(t)} - \frac{(A^{-1}Dh_2)^T PY(t)Z}{PY(t)} \right) \\ &= \left(-A^{-1}\{zs - D(s)\}, \frac{\{zs - D(s)\}^T A^{-1} PY(t)Z}{PY(t)} - \frac{h_2(t)}{PY(t)} \right), t \in [0, \tau]. \end{aligned} \tag{3.40}$$

Note z and s are specified a priori. z are the values of an individual's risk factors and s is the time by which we are interested to learn about the cumulative risk. Based on results in (3.22) and (3.32), we obtain

$$\begin{aligned} \sqrt{N}\dot{\Psi}_0 \begin{pmatrix} \hat{\Lambda} - \Lambda_0 \\ \hat{\theta} - \theta_0 \end{pmatrix} h' =& \sqrt{N}(\hat{\Lambda} - \Lambda_0)(s) + \sqrt{N} z^T (\hat{\theta} - \theta_0) s \\ =& \sqrt{N} \left\{ \hat{\Lambda}(s|Z = z) - \Lambda_0(s|Z = z) \right\}. \end{aligned}$$

Then (3.15) holds for h'. According to (3.15) and (3.17), $\sqrt{N}\left\{\hat{\Lambda}(s|Z = z) - \Lambda_0(s|Z = z)\right\}$ converges to a normal distribution with mean 0 and variance $P\psi^2_{\theta_0,\Lambda_0,h'}$. Recall $\psi_{\theta_0,\Lambda_0,h'} = h_1'^T \int_0^\tau Z dM(t) + \int_0^\tau h_2'(t) dM(t)$ as shown in (3.23). We have

$$\begin{aligned} \psi_{\theta_0,\Lambda_0,h'} =& - \{zs - D(s)\}^T A^{-1} \int_0^\tau Z dM(t) + \int_0^\tau \left\{ \frac{\{zs - D(s)\}^T A^{-1} PY(t)Z}{PY(t)} - \frac{h_2(t)}{PY(t)} \right\} dM(t) \\ =& - \{zs - D(s)\}^T A^{-1} \int_0^\tau \left\{ Z - \frac{PY(t)Z}{PY(t)} \right\} dM(t) - \int_0^\tau \frac{h_2(t)}{PY(t)} dM(t). \end{aligned} \tag{3.41}$$

When model (3.1) does not hold,

$$\begin{aligned} & rVar_A \left\{ \sqrt{N}\{\hat{\Lambda}(s|Z = z) - \Lambda_0(s|Z = z)\} \right\} \\ =& P \left[\{zs - D(s)\}^T A^{-1} \int_0^\tau \left\{ Z - \frac{PY(t)Z}{PY(t)} \right\} dM(t) - \int_0^\tau \frac{h_2(t)}{PY(t)} dM(t) \right]^2 \\ =& \{zs - D(s)\}^T A^{-1} P \left\{ \int_0^\tau \left[Z - \frac{PY(t)Z}{PY(t)} \right] dM(t) \right\}^{\otimes 2} A^{-1} \{zs - D(s)\} \\ & + 2\{zs - D(s)\}^T A^{-1} P \left[\int_0^\tau \left\{ Z - \frac{PY(t)Z}{PY(t)} \right\} dM(t) \int_0^s \frac{1}{PY(t)} dM(t) \right] \\ & + P \left\{ \int_0^s \frac{1}{PY(t)} dM(t) \right\}^2. \end{aligned}$$

When (3.1) holds, with the same arguments as in section 3.10.1 and 3.10.3, the model-based variance is

$$\begin{aligned}
&Var_A\left\{\sqrt{N}\{\hat{\Lambda}(s|Z=z)-\Lambda_0(s|Z=z)\}\right\}\\
=&\{zs-D(s)\}^T A^{-1}BA^{-1}\{zs-D(s)\}\\
&+2\{zs-D(s)\}^T A^{-1}P\int_0^\tau\left\{Z-\frac{PY(t)Z}{PY(t)}\right\}\frac{Y(t)}{PY(t)}\left\{d\Lambda_0(t)+Z^T\theta_0 dt\right\}\\
&+P\int_0^s\frac{1}{P^2Y(t)}Y(t)\left\{d\Lambda_0(t)+Z^T\theta_0 dt\right\}.
\end{aligned}\tag{3.42}$$

3.10.6 Limiting distribution of $\sqrt{N}\{\hat{\Lambda}^*(s|Z=z)-\Lambda_0(s|Z=z)\}$

In two-phase VPS, we use the same h' given in (3.40). Then the LHS of (3.18) is reduced to $\sqrt{N}\left\{\hat{\Lambda}^*(s|Z=z)-\Lambda_0(s|Z=z)\right\}$. According to (3.18)and (3.20), $\sqrt{N}\left\{\hat{\Lambda}^*(s|Z=z)-\Lambda_0(s|Z=z)\right\}$ converges to a normal distribution with mean 0 and variance

$$\begin{aligned}
&rVar_A\left\{\sqrt{N}\{\hat{\Lambda}(s|Z=z)-\Lambda_0(s|Z=z)\}\right\}\\
=&P\psi^2_{\theta_0,\Lambda_0,h'}+Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\psi^2_{\theta_0,\Lambda_0,h'}\right]\\
=&rVar_A\sqrt{N}\{\hat{\Lambda}(s|Z=z)-\Lambda_0(s|Z=z)\}\\
&+Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\left[\{zs-D(s)\}^T A^{-1}\int_0^\tau\left\{Z-\frac{PY(t)Z}{PY(t)}\right\}dM(t)-\int_0^\tau\frac{h_2(t)}{PY(t)}dM(t)\right]^2\right].
\end{aligned}$$

When the AH model (3.1) holds, the model-based asymptotic variance

$$\begin{aligned}
&Var_A\left\{\sqrt{N}\{\hat{\Lambda}^*(s|Z=z)-\Lambda_0(s|Z=z)\}\right\}\\
=&Var_A\sqrt{N}\{\hat{\Lambda}(s|Z=z)-\Lambda_0(s|Z=z)\}\\
&+Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\left[\{zs-D(s)\}^T A^{-1}\int_0^\tau\left\{Z-\frac{PY(t)Z}{PY(t)}\right\}dM(t)-\int_0^\tau\frac{h_2(t)}{PY(t)}dM(t)\right]^2\right].
\end{aligned}$$

3.11 Weights for the Generalized Case-Cohort Study

The classic case-cohort study assembles covariate measurements for a random subcohort and all cases. It is widely used when the clinical outcomes are event times and the event rate is low. However, sometimes due to the lack of samples or the consent for use of the samples, not all of the cases can be obtained for covariate measurements. When the disease is not rare, it is also not necessary to sample all the cases. In these circumstances, it is desirable to generalize the case-cohort design and consider sampling only a subset of cases in addition to the subcohort. This design was referred to as "generalized case-cohort" by Cai and Zeng [2004].

The common estimating approach in such a sampling design is to use the method of inverse probability weighted estimating equation due to its robustness and convenience for implementation. For this method, different weighting schemes can be adopted and they will provide different estimators of different variances. A careful choice on the weights themselves may enhance efficiency in estimation. In this section, we will compare our weights assignment to Kang et al. [2013]'s method .

3.11.1 Weighting Scheme I

Kang et al. [2013] considered the generalized case-cohort sampling sequentially: data were collected by first selecting a subcohort from the initial cohort of size N followed by subsequent sampling of cases outside the subcohort. The authors proposed following weights for use: the censored (controls) subcohort members were weighted by the inverse of their sampling probability p to the subcohort; the uncensored (cases) subcohort members were weighted by 1; cases outside subcohort were weighted by the inverse of their sampling probability q. Let ξ denote the indicator for being selected into the subcohort, η denote the indicator for the cases outside the subcohort being selected into the sample, and R be the indicator that a member in the initial cohort was selected in the final generalized case-cohort dataset. Let

$V = (\Delta, \xi)$. According to Kang et al. [2013],

$$P(\xi = 1) = p \text{ and } P(\eta = 1|\Delta = 1, \xi = 0) = q.$$

The sampling rate function was

$$\pi_0(V) = P(R = 1|V) = \pi_1(\Delta, \xi) = \begin{cases} p, & \Delta = 0 \\ 1, & \Delta = 1, \xi = 1 \\ q, & \Delta = 1, \xi = 0 \end{cases} . \tag{3.43}$$

Different from the single disease problem we are interested, Kang et al. [2013] were concerned with multivariate failure times. The authors also selected their samples by finite population sampling where a fixed size of subcohort and cases outside the subcohort were selected, rather than two-phase VPS. In addition, the authors considered estimators using both time-variant and time-invariant weights. Since this section's aim is to investigate the consequence of two different weighting schemes resulted from different considerations of strata in the initial cohort, we will only study their time-invariant weighting method , apply it to the simple scenario of one disease, and compare the results generated from their weighting scheme to ours.

3.11.2 Weighting Scheme II

Our two-phase VPS contains the generalized case-cohort design if we consider this sampling design as follows. At phase I, we select a cohort of size N and observe one outcome variable Δ for all cohort members. Based on Δ, we stratify the cohort to two groups of cases ($\Delta = 1$) versus controls ($\Delta = 0$); at phase II we draw a random sample within each stratum using variable sampling probability $\pi_0(V) = \pi_2(\Delta)$. We use notation π_2 in contrast with the rate function π_1 in Kang et al. [2013]'s weighting method. Let R indicate that subjects being selected to phase II, i.e., $R = 1$ when $\xi = 1$ or $\eta = 1$. Given the fact that the subcohort indicator ξ is independent of the outcome variable Δ, we obtain our phase II subsamples

selection probabilities as:

$$\begin{aligned}
&P(R=1|\Delta=0)\\
&=P(R=1,\xi=1|\Delta=0)+P(R=1,\xi=0|\Delta=0)\\
&=P(R=1|\xi=1,\Delta=0)P(\xi=1|\Delta=0)+0\\
&=1\times P(\xi=1)=p.
\end{aligned}$$

and

$$\begin{aligned}
&P(R=1|\Delta=1)\\
&=P(R=1,\xi=1|\Delta=1)+P(R=1,\xi=0|\Delta=1)\\
&=P(R=1|\xi=1,\Delta=1)P(\xi=1|\Delta=1)+P(R=1|\xi=0,\Delta=1)P(\xi=0|\Delta=1)\\
&=1\times P(\xi=1)+qP(\xi=0)=p+q(1-p).
\end{aligned}$$

Thus, based on two-phase VPS,

$$\pi_0(V)=P(R=1|V)=\pi_2(\Delta)=\begin{cases} p, & \Delta=0\\ p+q(1-p), & \Delta=1\end{cases}. \tag{3.44}$$

3.11.3 Theoretical Efficiency Comparison

Since Kang et al. [2013] also considered weighted estimating equations approach for developing their estimators, we find that their estimators agree with ours presented in section 3.4 except for the differences in the sampling probability $\pi_0(V)$ assigned to each member in the initial cohort. $\pi_0(V)=\pi_1(\Delta,\xi)$ for Kang et al. [2013]'s weighting method (approach I) and $\pi_0(V)=\pi_2(\Delta)$ for our method (approach II). Recall $\mathbb{P}^\pi f=\frac{1}{N}\sum_{i=1}^{N}\frac{R}{\pi_0(V)}f(X)$. Both our and Kang et al. [2013] can be written as

$$\begin{aligned}
\hat{\theta}^*&=\left[\mathbb{P}_N^\pi\int_0^\tau\left\{Z-\frac{\mathbb{P}_N^\pi ZY(t)}{\mathbb{P}_N^\pi Y(t)}\right\}^{\otimes 2}Y(t)dt\right]^{-1}\mathbb{P}_N^\pi\int_0^\tau\left\{Z-\frac{\mathbb{P}_N^\pi ZY(t)}{\mathbb{P}_N^\pi Y(t)}\right\}dN(t),\\
\hat{\Lambda}^*(s)&=\int_0^s\frac{\mathbb{P}_N^\pi\{Y(t)dN(t)\}}{\mathbb{P}_N^\pi Y(t)}-\int_0^s\frac{\mathbb{P}_N^\pi\{Z^TY(t)\}}{\mathbb{P}_N^\pi Y(t)}\hat{\theta}^*dt.
\end{aligned}$$

Let $\hat{\theta}^{*I}$ denote the estimates of regression parameters obtained from approach I and $\hat{\theta}^{*II}$ denote those from approach II. We can consider both $\hat{\theta}^{*I}$ and $\hat{\theta}^{*II}$ were solution to our EE (3.5) and they estimate parameter θ_0 defined in (3.9). Then, we can use our systematic estimation results on the AH model to study the properties of these two estimators together. According to our results in section 3.5, both $\hat{\theta}^{*I}$ and $\hat{\theta}^{*II}$ are unbiased estimators for θ in the AH model (3.1) if this assumed model correctly specifying underlying distribution for the observed data. Suppose all the assumptions in section 3.3 hold. Then, based on our results in section 3.10.2, under the model (3.1) $\sqrt{N}(\hat{\theta}^{*I} - \theta_0)$ converges in distribution to a p-variate Gaussian distribution with mean zero and variance matrix

$$Var_A\left\{\sqrt{N}(\hat{\theta}^{*I} - \theta_0)\right\} = A^{-1}BA^{-1} + Q\left[\frac{1-\pi_1(\Delta,\xi)}{\pi_1(\Delta,\xi)}A^{-1}\left[\int_0^\tau \left\{Z - \frac{PY(t)Z}{PY(t)}\right\} dM(t)\right]^{\otimes 2} A^{-1}\right]$$

and $\sqrt{N}(\hat{\theta}^{*II} - \theta_0)$ converges in distribution to a p-variate Gaussian distribution with mean zero and variance matrix

$$Var_A\left\{\sqrt{N}(\hat{\theta}^{*II} - \theta_0)\right\} = A^{-1}BA^{-1} + Q\left[\frac{1-\pi_2(\Delta)}{\pi_2(\Delta)}A^{-1}\left[\int_0^\tau \left\{Z - \frac{PY(t)Z}{PY(t)}\right\} dM(t)\right]^{\otimes 2} A^{-1}\right].$$

We let

$$\Sigma \equiv A^{-1}\left[\int_0^\tau \left\{Z - \frac{PY(t)Z}{PY(t)}\right\} dM(t)\right]^{\otimes 2} A^{-1}.$$

Then,

$$
\begin{aligned}
&Var_A\left\{\sqrt{N}(\hat{\theta}^{*I}-\theta_0)\right\}-Var_A\left\{\sqrt{N}(\hat{\theta}^{*II}-\theta_0)\right\}\\
=&A^{-1}BA^{-1}+Q\left[\frac{1-\pi_1(\Delta,\xi)}{\pi_1(\Delta,\xi)}\Sigma\right]-A^{-1}BA^{-1}-Q\left[\frac{1-\pi_2(\Delta)}{\pi_2(\Delta)}\Sigma\right]\\
=&E_Q\left[\left\{\frac{1-\pi_1(\Delta,\xi)}{\pi_1(\Delta,\xi)}-\frac{1-\pi_2(\Delta)}{\pi_2(\Delta)}\right\}E_Q[\Sigma|\Delta,\xi]\right]\\
=&\left\{\frac{1-q}{q}-\frac{(1-p)(1-q)}{p+q(1-p)}\right\}E(\Sigma|\Delta=1,\xi=0)P(\Delta=1,\xi=0)\\
&+\left\{0-\frac{(1-p)(1-q)}{p+q(1-p)}\right\}E(\Sigma|\Delta=1,\xi=1)P(\Delta=1,\xi=1)\\
&+(\frac{1-p}{p}-\frac{1-p}{p})E(\Sigma|\Delta=0).
\end{aligned}
$$

By the definition of the case-cohort design, $X=(T,\Delta,Z)$ and ξ are independent. Thus

$$E(\Sigma|\Delta=1,\xi=0)=E(\Sigma|\Delta=1,\xi=1)=E(\Sigma|\Delta=1)$$

and

$$P(\Delta=1,\xi=0)=P(\Delta=1)P(\xi=0).$$

Let $\alpha=P(\Delta=1)$. We obtain

$$
\begin{aligned}
&Var_A\left\{\sqrt{N}(\hat{\theta}^{*I}-\theta_0)\right\}-Var_A\left\{\sqrt{N}(\hat{\theta}^{*II}-\theta_0)\right\}\\
=&\left\{\frac{1-q}{q}-\frac{(1-p)(1-q)}{p+q(1-p)}\right\}E(\Sigma|\Delta=1)\alpha(1-p)+\left\{0-\frac{(1-p)(1-q)}{p+q(1-p)}\right\}E(\Sigma|\Delta=1)\alpha p\\
=&\alpha\left\{\frac{(1-p)(1-q)}{q}-\frac{(1-p)(1-q)}{p(1-q)+q}\right\}E(\Sigma|\Delta=1)
\end{aligned}
$$

Because p is the subcohort selection probability, $p>0$. When $q=1$, i.e., a case-cohort design is implemented, the above equation equals 0 and $Var_A\left\{\sqrt{N}(\hat{\theta}^{*I}-\theta_0)\right\}=Var_A\left\{\sqrt{N}(\hat{\theta}^{*II}-\theta_0)\right\}$. When $q<1$, i.e., a generalized case-cohort design is implemented, $\frac{(1-p)(1-q)}{q}>\frac{(1-p)(1-q)}{p(1-q)+q}$ and, as a consequence, $Var_A\left\{\sqrt{N}(\hat{\theta}^{*I}-\theta_0)\right\}>Var_A\left\{\sqrt{N}(\hat{\theta}^{*II}-\theta_0)\right\}$.

In summary, estimation using the weights in scheme I is not as efficient as scheme II using the weights considered in our general two-phase VPS design. When $q=1$, the generalized

case-cohort design becomes a case-cohort design. The asymptotic variances resulted from the two weighting methods are then the same.

A careful examination of our comparison study shows our comparison results hold regardless of Σ, which is related to the influence function of an estimator. Thus our conclusions do not restrict to the estimation of regression parameters in the AH model and can be applied to other statistics of interest. Our conclusions can be generalized to finite finite sampling, multiple outcomes and time-variant weights as well. Furthermore our conclusions hold in spite of the model under study.

In conclusion, for efficiency purposes, our two-phase sampling methods should be considered for the generalized case-cohort study. It is more convenient and transparent if we consider the generalized case-cohort design as a special case of two-phase case-control sampling.

Chapter 4

APPLICATION OF THE Z-ESTIMATION SYSTEM TO IMPROVED ESTIMATORS FOR ADDITIVE HAZARDS MODELS WITH TWO-PHASE SAMPLING

Chapter 4 continues Chapter 3 to further improve the inference precision from two-phase VPS data. In Chapter 3, we solved the problem of fitting the entire semiparametric additive hazards (AH) model to two-phase VPS data. In Chapter 4, we incorporate auxiliary information, a large amount of which is usually available in phase I data, to our inference procedure for the AH model. In addition to the practical purpose of reducing costs in observational studies, this chapter is another illustration of applying the Z-estimation theory introduced in Chapter 2 to create a method for analyzing two-phase design data. Our estimation procedures will follow closely the procedure in Chapters 2.5∼2.6.

4.1 Estimators and Parameters

Let $\tilde{V} = \tilde{V}(V)$ of q-dimension be the auxiliary variables we choose to calibrate on. V is available for all phase I samples. We use $\tilde{V}$ to denote both a function of V and the new variable created after applying this function to V. Suppose assumptions A.3.3.7 and A. 3.3.8 hold for $\tilde{V}$.

Let $\gamma \in \Gamma$ and Γ be a compact convex subset of $\mathbb{R}^q$ with 0 as an interior point. We consider a new parameter space $\mathbb{L}_0 \times \Gamma$ and a new random map $\Psi_N^{**} : \mathbb{L}_0 \times \Gamma \mapsto \mathbb{L} \times \mathbb{R}^q$ given by $\Psi_N^{**}(\theta, \Lambda, \gamma) = \mathbb{P}_N \psi_{\theta,\Lambda,\gamma}^{**}(X, V, R) = \mathbb{P}_N \left(\psi_{1,\theta,\Lambda,\gamma}^{**}(X, V, R), \psi_{2,\gamma}^{**}(V, R)\right)$ where

$$\psi_{1,\theta,\Lambda,\gamma}^{**}(X, V, R)h = \frac{R}{\pi_0(V)} exp(-\gamma^T \tilde{V}) \psi_{\theta,\Lambda,h}(X), \qquad \psi_{1,\theta,\Lambda,\gamma}^{**}(X, V, R) \in l^\infty(\mathcal{H}) \tag{4.1}$$

$$\psi_{2,\gamma}^{**}(V, R) = \frac{R}{\pi_0(V)} exp(-\gamma^T \tilde{V})\tilde{V} - \tilde{V}, \qquad \psi_{2,\gamma}^{**}(V, R) \in \mathbb{R}^q. \tag{4.2}$$

$\mathbb{L}_0$, $\mathbb{L}$, $\psi_{\theta,\Lambda,h}$ and $\mathcal{H}$ were defined in section 3.4.

Based on our random maps, we construct a new modified inverse probability weighted estimating equation (IPW-EE):

$$\Psi_N^{**}(\theta,\Lambda,\gamma)h = 0, \forall h \in \mathcal{H}. \tag{4.3}$$

Estimator $\hat{\gamma}$ is obtained first by solving $\mathbb{P}_N\psi_{2,\gamma}^{**}(V,R) = 0$ numerically. We then set

$$h = (h_1, h_2) = \left(h_1, -h_1^T \frac{\mathbb{P}_N \frac{R}{\pi_0(V)} exp(-\hat{\gamma}^T\tilde{V})ZY(t)}{\mathbb{P}_N \frac{R}{\pi_0(V)} exp(-\hat{\gamma}^T\tilde{V})Y(t)}\right)$$

and follow exactly the computation procedure for $(\hat{\theta}, \hat{\Lambda})$ in section 3.4. This yields the new calibrated two-phase VPS estimators

$$\hat{\theta}^{**} = \left[\mathbb{P}_N \frac{R}{\pi_0(V)} exp(-\hat{\gamma}^T\tilde{V}) \int_0^\tau \left\{Z - \frac{\mathbb{P}_N \frac{R}{\pi_0(V)} exp(-\hat{\gamma}^T\tilde{V})ZY(t)}{\mathbb{P}_N \frac{R}{\pi_0(V)} exp(-\hat{\gamma}^T\tilde{V})Y(t)}\right\}^{\otimes 2} Y(t)dt\right]^{-1}$$
$$\left[\mathbb{P}_N \frac{R}{\pi_0(V)} exp(-\hat{\gamma}^T\tilde{V}) \int_0^\tau \left\{Z - \frac{\mathbb{P}_N \frac{R}{\pi_0(V)} exp(-\hat{\gamma}^T\tilde{V})ZY(t)}{\mathbb{P}_N \frac{R}{\pi_0(V)} exp(-\hat{\gamma}^T\tilde{V})Y(t)}\right\} dN(t)\right]$$

$$\hat{\Lambda}^{**}(s) = \int_0^s \frac{\mathbb{P}_N \frac{R}{\pi_0(V)} exp(-\hat{\gamma}^T\tilde{V})\{dN(t) - Y(t)Z^T\hat{\theta}^{**}dt\}}{\mathbb{P}_N \frac{R}{\pi_0(V)} exp(-\hat{\gamma}^T\tilde{V})Y(t)}, s \in [0,\tau].$$

According to section 2.6.2 and lemma 2.6.1, the unique solution to $\Psi^{**}(\theta,\Lambda,\gamma) = Q\psi_{\theta,\Lambda,\gamma}^{**}(X,V,R) = 0$ is $(\theta_0, \Lambda_0, 0)$ where (θ_0, Λ_0) is defined in (3.9).

4.2 Consistency and Asymptotic Normality

We will use the preliminary results we established on $\psi_{\theta,\Lambda}$ in section 3.7 to show consistency and asymptotic normality of our improved two-phase sampling estimators.

Theorem 4.2.1. *Suppose assumptions A.3.3.1 - 3.3.2 and A.3.3.7 hold, then*

$$\begin{pmatrix} \hat{\theta}^{**} \\ \hat{\Lambda}^{**} \\ \hat{\gamma} \end{pmatrix} \to_p \begin{pmatrix} \theta_0 \\ \Lambda_0 \\ 0 \end{pmatrix}$$

Proof. With the same arguments used for proving consistency for random sampling estimators, it follows from theorem 2.6.3 that $(\hat{\theta}^{**}, \hat{\Lambda}^{**}, \hat{\gamma})$ converges to $(\theta_0, \Lambda_0, 0)$ in probability. □

Theorem 4.2.2. *Suppose assumptions A.3.3.1 - 3.3.4 and the additional assumptions A. 3.3.7 and A.3.3.8 on auxiliary variable $\tilde{V}$ hold, then*

$$\dot{\Psi}_0^{**}\sqrt{N}\begin{pmatrix} \hat{\theta}^{**} - \theta_0 \\ \hat{\Lambda}^{**} - \Lambda_0 \\ \hat{\gamma} - 0 \end{pmatrix} = -\mathbb{G}_N \psi^{**}_{\theta_0,\Lambda_0,0} + o_P(1) \tag{4.4}$$

where $\dot{\Psi}_0^{**}$ is the Fréchet derivative of $\Psi^{**} = (Q\psi^{**}_{1,\theta,\Lambda,\gamma}(X,V,R), Q\psi^{**}_{2,\gamma}(V,R))$ at $(\theta_0, \Lambda_0, 0)$ and

$$\dot{\Psi}_0^{**} = \begin{pmatrix} \dot{\Psi}_0 & -Q\psi_{\theta_0,\Lambda_0}(X)\tilde{V}^T \\ 0 & -Q\tilde{V}\tilde{V}^T \end{pmatrix}. \tag{4.5}$$

Proof. According to corollary 2.6.4, this theorem is an immediate result from lemma 3.7.2∼ 3.7.5 and theorem 4.2.1. □

According to (2.28), $\sqrt{N}\dot{\Psi}_0 \begin{pmatrix} \hat{\theta}^{**} - \theta_0 \\ \hat{\Lambda}^{**} - \Lambda_0 \end{pmatrix}$ converges in distribution to the Q-Brownian bridge in $l^\infty(\mathcal{H})$, which is a zero-mean Gaussian with variance

$$Var_A\left\{\dot{\Psi}_0\sqrt{N}\begin{pmatrix} \hat{\theta}^{**} - \theta_0 \\ \hat{\Lambda}^{**} - \Lambda_0 \end{pmatrix} h\right\} = P\psi^2_{\theta_0,\Lambda_0,h} + Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\{\psi_{\theta_0,\Lambda_0,h} - \Pi(\psi_{\theta_0,\Lambda_0,h}|\tilde{V})\}^2\right] \tag{4.6}$$

where $\dot{\Psi}_0$ is given in lemma 3.7.4 and $\Pi(\cdot|\tilde{V})$ is population least squares projection on the space spanned by the calibration variables $\tilde{V}$: $\Pi(\cdot|\tilde{V}) = Q\{\cdot\tilde{V}^T\}(Q\tilde{V}\tilde{V}^T)^{-1}\tilde{V}$.

4.3 Limiting Distributions of Some Interesting Statistics

Based on the asymptotic results in the previous section, we will develop the limiting distribution of the improved two-phase sampling estimators $\sqrt{N}(\hat{\theta}^{**} - \theta_0)$, $\sqrt{N}(\hat{\Lambda}^{**} - \Lambda_0)$ and $\sqrt{N}\{\hat{\Lambda}^{**}(t|Z=z) - \Lambda_0(t|Z=z)\}$. Both model-based and robust variances will be provided. Since the derivation of these limiting distribution are very similar to section 3.10, the same

detailed steps will not be repeated. We only present the differences in the derivation from section 3.10.

Using h' in (3.21), we obtain

$$\sqrt{N}\dot{\Psi}_0 \begin{pmatrix} \hat{\theta}^{**} - \theta_0 \\ \hat{\Lambda}^{**} - \Lambda_0 \end{pmatrix} h' = h_1^T \sqrt{N}(\hat{\theta}^{**} - \theta_0)$$

and $h' \in \mathcal{H} \cdot M$. According to theorem 4.2.2 and (4.6), $h_1^T\sqrt{N}(\hat{\theta}^{**} - \theta_0)$ converges to a normal distribution with mean 0 and variance $P(\psi_{\theta_0,\Lambda_0,h'})^2 + Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\{\psi_{\theta_0,\Lambda_0,h'} - \Pi(\psi_{\theta_0,\Lambda_0,h'}|\tilde{V})\}^2\right]$.

Step 3 from section 3.10.1 shows

$$\psi_{\theta_0,\Lambda_0,h'} = h_1^T A^{-1} \int_0^\tau \left\{ Z - \frac{PY(t)Z}{PY(t)} \right\} dM. \tag{4.7}$$

Note this result holds for all $h_1 \in \mathcal{H}_1$ and $\mathcal{H}_1 \subset \mathbb{R}^p$. Thus when we do not assume the AH model (3.1), $\sqrt{N}(\hat{\theta}^{**} - \theta_0)$ converges in distribution to a p-variate Gaussian distribution with mean zero and variance matrix

$$\begin{aligned} rVar_A\left\{\sqrt{N}(\hat{\theta}^* - \theta_0)\right\} &= A^{-1}P\left\{\int_0^\tau \left[Z - \frac{PY(t)Z}{PY(t)}\right] dM(t)\right\}^{\otimes 2} A^{-1} \\ &\quad + Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\left(A^{-1}\int_0^\tau \left\{Z - \frac{PY(t)Z}{PY(t)}\right\} dM \right.\right. \\ &\quad \left.\left. -Q\left[A^{-1}\int_0^\tau \left\{Z - \frac{PY(t)Z}{PY(t)}\right\} dM\tilde{V}^T\right](Q\tilde{V}\tilde{V}^T)^{-1}\tilde{V}\right)^{\otimes 2}\right] \end{aligned} \tag{4.8}$$

and when (3.1) holds, this variance matrix becomes

$$\begin{aligned} Var_A\left\{\sqrt{N}(\hat{\theta}^* - \theta_0)\right\} &= A^{-1}BA^{-1} \\ &\quad + Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\left(A^{-1}\int_0^\tau \left\{Z - \frac{PY(t)Z}{PY(t)}\right\} dM \right.\right. \\ &\quad \left.\left. -Q\left[A^{-1}\int_0^\tau \left\{Z - \frac{PY(t)Z}{PY(t)}\right\} dM\tilde{V}^T\right](Q\tilde{V}\tilde{V}^T)^{-1}\tilde{V}\right)^{\otimes 2}\right] \end{aligned} \tag{4.9}$$

where A and B are given in section 3.10.1.

Using h' in (3.30), we obtain

$$\sqrt{N}\dot{\Psi}_0 \begin{pmatrix} \hat{\theta}^{**} - \theta_0 \\ \hat{\Lambda}^{**} - \Lambda_0 \end{pmatrix} h' = \sqrt{N}(\hat{\Lambda}^* - \Lambda_0)h_2$$

where $h' \in \mathcal{H} \cdot M$. According to theorem 4.2.2 and (4.6), $\sqrt{N}(\hat{\Lambda}^{**} - \Lambda_0)$ converges to a Q-Brownian bridge process in $l^\infty(\mathcal{H}_2)$ with mean 0 and variance process

$$P\psi^2_{\theta_0,\Lambda_0,h'} + Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\{\psi_{\theta_0,\Lambda_0,h'} - \Pi(\psi_{\theta_0,\Lambda_0,h'}|\tilde{V})\}^2\right].$$

Based on this result, to derive the limiting distribution of $\sqrt{N}(\hat{\Lambda}^{**} - \Lambda_0)(s)$, we set $h_2(t)$ in $h'(t)$ in (3.30) by $h_2(t) = 1(t \leq s)$. Then according to (3.34).

$$\psi_{\theta_0,\Lambda_0,h'}(X) = -\int_0^s \frac{1}{PY(t)} dM(t) + D(s)^T A^{-1} \int_0^\tau \left\{Z - \frac{PY(t)Z}{PY(t)}\right\} dM(t).$$

When we do not assume the AH model (3.1), $\sqrt{N}(\hat{\Lambda}^* - \Lambda_0)(s)$ converges to a normal distribution with mean 0 and variance

$$rVar_A\{\sqrt{N}(\hat{\Lambda}^{**} - \Lambda_0)(s)\} = P\psi^2_{\theta_0,\Lambda_0,h'} + Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\left\{\psi_{\theta_0,\Lambda_0,h'} - Q(\psi_{\theta_0,\Lambda_0,h'}\tilde{V}^T)(Q\tilde{V}\tilde{V}^T)^{-1}\tilde{V}\right\}^2\right]$$

where $\psi_{\theta_0,\Lambda_0,h'}$ is given in the preceding display. When (3.1) holds, $Var_A\{\sqrt{N}(\hat{\Lambda}^{**} - \Lambda_0)(s)$ is obtained through replacing the term $P\psi^2_{\theta_0,\Lambda_0,h'}$ in the above equation by $Var_A\{\sqrt{N}(\hat{\Lambda} - \Lambda_0)(s)\}$ in (3.37).

We define

$$\hat{\Lambda}^{**}(s|Z = z) = \hat{\Lambda}^{**}(s) + \int_0^s z^T \hat{\theta}^{**} dt.$$

Next we derive the limiting distribution of $\sqrt{N}\left\{\hat{\Lambda}^{**}(s|Z = z) - \Lambda_0(s|Z = z)\right\}$. When the AH model holds, we obtain the limiting distribution of the individual cumulative hazard function estimates by different time points. With h' in (3.40), the LHS of (4.4) is reduced to $\sqrt{N}\left\{\hat{\Lambda}^{**}(s|Z = z) - \Lambda_0(s|Z = z)\right\}$. According to theorem 4.2.2 and (4.6), $\sqrt{N}\left\{\hat{\Lambda}^{**}(s|Z = z)\right.$

$-\Lambda_0(s|Z=z)\}$ converges to a normal distribution with mean 0 and variance

$$\begin{aligned} & rVar_A\left\{\sqrt{N}\{\hat{\Lambda}^{**}(s|Z=z)-\Lambda_0(s|Z=z)\}\right\} \\ =& P\psi^2_{\theta_0,\Lambda_0,h'}+Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\left\{\psi_{\theta_0,\Lambda_0,h'}-Q(\psi_{\theta_0,\Lambda_0,h'}\tilde{V}^T)(Q\tilde{V}\tilde{V}^T)^{-1}\tilde{V}\right\}^2\right] \end{aligned}$$

where $\psi_{\theta_0,\Lambda_0,h'}$ is given in (3.41). The above result is a robust variance since we do not assume the AH model (3.1) holds. When this model does not hold, $Var_A\left[\sqrt{N}\{\hat{\Lambda}^{**}(s|Z=z)-\Lambda_0(s|Z=z)\}\right]$ is obtained by substituting $Var_A\left[\sqrt{N}\{\hat{\Lambda}(s|Z=z)-\Lambda_0(s|Z=z)\}\right]$ given in (3.42) for the term $P\psi^2_{\theta_0,\Lambda_0,h'}$ in the robust variance.

4.4 Variance Estimators

In sections 3.10 and 4.3, we obtained the limiting distributions for three RS estimators, three two-phase VPS estimators and three calibrated two-phase VPS estimators. Each of these estimators is associated with two types of asymptotic variances: model-based and robust variances. We programed these nine statistics and the eighteen variances in an R package called "ah", which is the abbreviation for the additive hazards model. Our package was built upon the R package ahaz [Gorst-Rasmussen and Scheike, 2012], which provided computationally efficient procedures for regularized estimation with Lin & Ying's additive hazards model and used for analyzing high dimensional data. Since the program ahaz only provides regression parameter estimates from a random sample, in our package we also included estimation of cumulative baseline hazard function, prediction of individual cumulative hazards, various two-phase sampling estimators with or without calibration, and the robust variances for all of our estimators.

For random sampling, estimators $\hat{\theta}$ and $\hat{\Lambda}(s)$ was computed based on our results in section

3.4. $\hat{\Lambda}(s)$ were calculated for all of the observed failure times and censoring times. We let

$$\begin{aligned}
\hat{A} =& \mathbb{P}_N \int_0^\tau \{Z - \frac{\mathbb{P}_N ZY(t)}{\mathbb{P}_N Y(t)}\}^{\otimes 2} Y(t) dt \\
\hat{B} =& \mathbb{P}_N \int_0^\tau \{Z - \frac{\mathbb{P}_N ZY(t)}{\mathbb{P}_N Y(t)}\}^{\otimes 2} dN(t) \\
\hat{M}(t) =& N(t) - \hat{\Lambda}(t) - Z^T \hat{\theta} t \\
\hat{D}(s) =& \int_0^s \frac{\mathbb{P}_N \{Y(t) Z^T\}}{\mathbb{P}_N Y(t)} dt.
\end{aligned}$$

The variance estimators were then obtained by substituting $\mathbb{P}_N$, $\hat{M}(t)$, $\hat{A}$, $\hat{B}$, $\hat{D}(s)$ and $dN(t)$ for P, $M(t), A, B, D$ and $\{d\Lambda_0(t) + Z^T \theta_0 dt\}$ in the mode-based and robust variance formula for random sampling estimators in section 3.10.

For two-phase sampling, estimators $\hat{\theta}^*$ and $\hat{\Lambda}(t)^*$ were computed according to the results in section 3.4. We let

$$\begin{aligned}
\hat{A}^\pi =& \mathbb{P}_N^\pi \int_0^\tau \{Z - \frac{\mathbb{P}_N^\pi ZY(t)}{\mathbb{P}_N^\pi Y(t)}\}^{\otimes 2} Y(t) dt \\
\hat{B}^\pi =& \mathbb{P}_N^\pi \int_0^\tau \{Z - \frac{\mathbb{P}_N^\pi ZY(t)}{\mathbb{P}_N^\pi Y(t)}\}^{\otimes 2} dN(t) \\
\hat{M}^\pi(t) =& N(t) - \hat{\Lambda}^*(t) - Z^T \hat{\theta}^* t \\
\hat{D}^\pi(s) =& \int_0^s \frac{\mathbb{P}_N^\pi \{Y(t) Z^T\}}{\mathbb{P}_N^\pi Y(t)} dt
\end{aligned}$$

where $\mathbb{P}_N^\pi f(R, V, X) = \frac{1}{N} \sum_{i=1}^N \frac{R}{\pi_0(V)} f(X, V)$. We then estimated the model-based and robust variances by substituting $\mathbb{P}_N^\pi$, $\hat{M}^\pi(t)$, $\hat{A}^\pi$, $\hat{B}^\pi$, and $\hat{D}^\pi(s)$ for P, Q, $M(t), A, B$, and $D(s)$ in the variances formula for two-phase VPS estimators in section 3.10. $\{d\Lambda_0(t) + Z^T \theta_0 dt\}$ was still replaced by $dN(t)$. When calibration techniques were used, estimators and their variances were computed based on the results in section 4.1. The standard weights $\frac{R}{\pi_0(V)}$ in $\mathbb{P}_N^\pi$, $\hat{M}^\pi(t)$, $\hat{A}^\pi$, $\hat{B}^\pi$, and $\hat{D}^\pi(s)$ were replaced by the new weights $\frac{R}{\pi_0(V)} \exp(-\hat{\gamma}^T \tilde{V})$ for estimating P, Q, $M(t), A, B$, and $D(s)$ in the model-based and robust variances for calibrated two-phase sampling estimators.

Chapter 5
NUMERICAL STUDIES

This chapter aims to verify a collection of theoretical results presented in Chapters $2 \sim 4$ by simulation studies, and to demonstrate numerically the substantial efficiency gain we can obtain by implementing two-phase sampling and calibration techniques. In this chapter, I first conduct simulations to verify my theoretical results on Lin and Ying's additive hazards (AH) models for association studies and individual risk prediction. Then I use simulations to show two-phase sampling and calibration's ability to substantially improve estimation efficiency compared to random sampling. Lastly, I demonstrate the performance of my estimating methods under model misspecification.

5.1 Theoretical Results Evaluation

In Chapters $4 \sim 5$, we developed a series of estimators for the AH model with random sampling (RS) and two-phase variable probability sampling (VPS). We programmed these estimators in our R package "ah". In this section, we will use simulations to evaluate a few important ones among these estimators. They include $\hat{\theta}^*$, $\hat{\theta}^{**}$, the regression coefficient estimates without and with calibration, and $\hat{\Lambda}^*(t|z)$, $\hat{\Lambda}^{**}(t|z)$, the individual cumulative hazard estimates by a given time point without and with calibration. We will evaluate the bias of these estimators and the accuracy of their model-based standard errors. These new estimators have not been implemented in any computer software before.

5.1.1 Simulation Procedures

Our major simulation setting mimics a cohort in National Wilms Tumor Study (NWTS) [D'angio et al., 1989, Green et al., 1998]. This cohort consists of 3378 patients who are over

1 year old and followed until relapse before 2 years. Removing patients less than 1 year old is because the AH model may not fit the original cohort. The baseline variables include age, tumor weight, tumor diameter, stage of disease, central histology and institutional histology. The two types of histology are both classified as "favorable" vs. "unfavorable". Based on this dataset, we designed various simulation settings.

Let N denote the phase I sample size. In each simulation, we began with generating N sets of covariates Z_1, Z_2, Z_3 and an auxiliary variable W. The distributions of Z_1, Z_2, Z_3 and W were determined by the distributions of central histology, tumor stage, age and institutional histology in the NWTS cohort. Z_1 and Z_2 were binary variables simulated from Bernoulli distributions with $p_{z_1} = P(Z_1 = 1) = 0.1$ and $p_{z_2} = P(Z_2 = 1) = 0.4$. Z_3 was a continuous variable. $log(Z_3)$ followed a normal distribution $N(1.2, 0.6)$.

In the NWTS cohort, histology type was initially evaluated at a local institution and then re-evaluated by experienced pathologists in a central laboratory. The central laboratory histology evaluation is considered more accurate, but more difficult to obtain and expensive. Thus, institutional histology can be considered as a surrogate of central laboratory histology. We generated our auxiliary variable W based on the distribution of institutional histology in the NWTS dataset. Because both institutional and central laboratory were measured for participants in the NWTS cohort, the sensitivity and specificity of institutional histology with respect to central laboratory histology can be estimated. They were approximately 74% and 96% respectively. Hence we generated W using Bernoulli distributions according to each subject's Z_1 level. When $Z_1 = 1$, the Bernoulli success rate was set at $P(W = 1|Z_1 = 1) = 0.74$. When $Z_1 = 0$, this rate became $P(W = 1|Z_1 = 0) = 1 - 0.96 = 0.04$.

Based on the covariates we simulated, next we generated the outcome variable (T, Δ). We first simulated uncensored failure time $\tilde{T}$ according to the AH model

$$\lambda(t|Z = z) = \lambda_0(t) + 0.16Z_1 + 0.04Z_2 + 0.009Z_3 \tag{5.1}$$

using different baseline hazard functions in different simulation settings. These coefficient values were obtained by fitting an AH model to the NWTS cohort using central laboratory

histology, tumor stage and age as the covariates and time to relapse before 2 years as the outcome. We then generated the censoring time C. We simulated $C_1 \sim N(10, 5)$. We let $C_2 = max(C_1, 0)$ and $C = min(C_2, 2)$. With this simulation, the distribution of C was the same as the censored time in the NWTS cohort. Finally, we obtained each subject's outcome (T, Δ) by letting $T = min(\tilde{T}, C)$ and $\Delta = 1(\tilde{T} \leq C)$.

Phase II subsamples were generated using case-control (CC) sampling. We stratified our phase I samples to case ($\Delta = 1$) and control ($\Delta = 0$) groups, and then we generated the phase II membership indicator R using Bernoulli distributions with $P(R = 1|\Delta = 1) = 0.9$ and $P(R = 1|\Delta = 0) = 0.15$. As a result, for each simulation run, we generated a dataset $(T, \Delta, Z_1, Z_2, Z_3, W, R)$ of size N based on the AH model (5.1). When we treated the status of Z_1 for subjects with $R = 0$ as unknown, then we obtained a desired two-phase sampling dataset.

5.1.2 Designs of the Simulation Studies

In simulation setting A, we let $\lambda_0(t) = 0.018$ and $N = 5000$. With this constant baseline hazard function, the event rate was approximately 14%, which was similar to the relapse rate of the NWTS cohort. We then let the phase I sample size N vary from 1250 to 20,000. These settings were denoted as A-1250, A-2500, A-10000 and A-20000.

In simulation setting B, we let $\lambda_0(t) = c\frac{10.5(3t)^{2.5}}{1+(3t)^{3.5}}$, c=0.005, N=5000, so that the baseline hazard varied across time. $\frac{10.5(3t)^{2.5}}{1+(3t)^{3.5}}$ was the hazard function of the log-logistic distribution. This unimodal hazard function mimicked the scenario when an event rate increases initially and decreases later, for instance, the mortality rate of cancer patients after treatment or the mortality rate of heart failure patients after heart transplantation. We set $c = 0.005$ so that the shape of the cumulative baseline hazard function and the overall failure rate in the phase I sample of the simulation setting B were the same as the NWTS cohort.

5.1.3 Results of the Simulation Studies

Evaluation of coefficient estimates

2000 datasets were generated for each simulation setting. Table 5.1 presents results from simulation setting A, Table 5.2 from setting B and Table 5.3 from settings A, A-1250, A-2500, A-10000 and A-20000.

The part (a) of these tables evaluate our two-phase sampling estimating method with standard weights. For each dataset under each of these settings, we first fit the AH model

$$\lambda(t|Z=z) = \lambda_0(t) + \theta_1 Z_1 + \theta_2 Z_2 + \theta_3 Z_3$$

to the two-phase dataset $(T, \Delta, Z_1 R, Z_2, Z_3)$. We then obtained the two-phase sampling estimators $\hat{\theta}^* = (\hat{\theta}_1^*, \hat{\theta}_2^*, \hat{\theta}_3^*)$ for the regression coefficients and their model-based standard errors denoted as SE $\hat{\theta}^*$. We calculated the following statistics based on the 2000 pairs of $\hat{\theta}^*$ and SE $\hat{\theta}^*$ computed using our R package "ah": the mean of $\hat{\theta}^*$, the bias of this mean, the empirical standard error of $\hat{\theta}^*$ (ESE), the average of SE $\hat{\theta}^*$ (ASE), and the empirical coverage probability of 95% confidence intervals (95% CI Cov) obtained based on $\hat{\theta}^*$ and SE $\hat{\theta}^*$. We reported these results together with the true θ in part (a) of Tables 5.1 $\sim$ 5.3.

Part (b) of these tables evaluate our two-phase sampling estimating method with calibrated weights. Similar to part (a), for each dataset under each simulation setting, we obtained the calibrated two-phase sampling estimators $\hat{\theta}^{**} = (\hat{\theta}_1^{**}, \hat{\theta}_2^{**}, \hat{\theta}_3^{**})$ and their model-based standard errors SE $\hat{\theta}^{**}$. The calibration was based on the auxiliary variables $W \times \Delta$ and $W \times (1-\Delta)$. Based on 2000 datasets, we reported the same set of statistics in part (b) as in part (a). Note in practice when a surrogate W of Z_1 is available, we can conduct stratified sampling and calibration using the surrogate W of Z_1 in order to improve the coefficient estimate for Z_1 instead of calibration techniques.

In Table 5.1 (setting A), on average 14.7 % of cohort members were observed experiencing events, and on average 662 cases and 640 controls were selected to phase II. In Table 5.2 (setting B), on average 14.3 % of cohort members were observed experiencing events, and

on average 644 cases and 643 controls were selected to phase II.

Results in Tables 5.1 and 5.2 show both $\hat{\theta}^*$ and $\hat{\theta}^{**}$ have small biases, the ASE of them are similar to the corresponding ESE, and 95% CI Cov are close to 95%. Comparing Table 5.2 to Table 5.1, we find our estimating methods perform equally well with or without a constant baseline hazard.

In Table 5.3, we report Mean, ASE/ESE and 95% CI Cov for $\hat{\theta}_1^*$ and $\hat{\theta}_1^{**}$ under settings A, A-1250, A-2500, A-10000 and A-20000. The ratios of ASE and ESE are always close to 1 and the 95% CI Cov are close to 95% across different sample sizes. As N increases, the biases of $\hat{\theta}_1^*$ and $\hat{\theta}_1^{**}$ become smaller, and the bias of $\hat{\theta}_1^{**}$ is consistently smaller than $\hat{\theta}_1^*$. Tables 5.1$\sim$5.3 together confirm that our two-phase sampling estimating methods for regression coefficients perform well with or without calibration.

Table 5.1: Simulation results of $\hat{\theta}^*$ and $\hat{\theta}^{**}$ under simulation setting A

(a) Before calibration

	True θ	Mean $\hat{\theta}^*$	Bias $\hat{\theta}^*$	ESE $\hat{\theta}^*$	ASE $\hat{\theta}^*$	95% CI Cov
θ_1	0.16000	0.16217	0.00217	0.03187	0.03110	95.0%
θ_2	0.04000	0.04027	0.00027	0.00976	0.00998	95.9%
θ_3	0.00900	0.00929	0.00029	0.00260	0.00249	93.8%

(b) After calibrtion

	True θ	Mean $\hat{\theta}^{**}$	Bias $\hat{\theta}^{**}$	ESE $\hat{\theta}^{**}$	ASE $\hat{\theta}^{**}$	95% CI Cov
θ_1	0.16000	0.16133	0.00133	0.02645	0.02638	95.5%
θ_2	0.04000	0.04028	0.00028	0.00975	0.00997	95.7%
θ_3	0.00900	0.00929	0.00029	0.00260	0.00249	93.7%

Simulation setting A:$\lambda_0(t) = 0.018, N = 5000$; ESE: empirical standard error; ASE: average standard error; 95% CI Cov:empirical coverage probability of 95% confidence intervals; $\hat{\theta}^*$: regular two-phase sampling coefficient estimates; $\hat{\theta}^{**}$: calibrated two-phase sampling coefficient estimates.

Table 5.2: Simulation results of $\hat{\theta}^*$ and $\hat{\theta}^{**}$ under simulation setting B

(a) Before calibration

	True θ	Mean $\hat{\theta}^*$	Bias $\hat{\theta}^*$	ESE $\hat{\theta}^*$	ASE $\hat{\theta}^*$	95% CI Cov
θ_1	0.16000	0.16456	0.00456	0.03172	0.03121	95.5%
θ_2	0.04000	0.04033	0.00033	0.00994	0.00985	95.0%
θ_3	0.00900	0.00938	0.00038	0.00249	0.00247	94.7%

(b) After calibrtion

	True θ	Mean $\hat{\theta}^{**}$	Bias $\hat{\theta}^{**}$	ESE $\hat{\theta}^{**}$	ASE $\hat{\theta}^{**}$	95% CI Cov
θ_1	0.16000	0.16316	0.00316	0.02662	0.02640	95.1%
θ_2	0.04000	0.04032	0.00032	0.00993	0.00983	95.3%
θ_3	0.00900	0.00937	0.00037	0.00249	0.00247	95.0%

Simulation setting B:$\lambda_0(t) = 0.0525\frac{(3t)^{2.5}}{1+(3t)^{3.5}}$, $N = 5000$. See notation definition in Table 5.1.

Table 5.3: Simulation results of $\hat{\theta}_1^*$ and $\hat{\theta}_1^{**}$ using standard and calibrated weights under simulation settings A-1250, A, A-10000 and A-20000

Phase I sample size (N)	1250	2500	5000	10000	20000
Bias $\hat{\theta}_1^*$	0.01252	0.00739	0.00217	0.00156	0.00134
Bias $\hat{\theta}_1^{**}$	0.00853	0.00495	0.00133	0.00104	0.00092
ASE/ESE $\hat{\theta}_1^*$	0.903	0.994	0.976	0.984	0.988
ASE/ESE $\hat{\theta}_1^{**}$	0.899	0.993	0.997	0.995	0.991
95% CI Cov using standard weights	94.1%	95.7%	95.0%	94.6 %	94.6%
95% CI Cov using calibrated weights	93.8%	95.4%	95.5%	94.6 %	94.4%

Simulation setting A: $\lambda_0(t) = 0.018$. See notation definition in Table 5.1.

Evaluation of individual cumulative hazard function prediction

We used simulation settings A and B to evaluate our cumulative hazard function prediction for an individual based on the AH model. This will be useful when clinical practitioners obtain each individual's risk factors and are interested in using them for risk prediction. Suppose an individual's covariates $z = (z_1, z_2, z_3) = (1, 1, 3.3)$. We are interested in predicting this individual's cumulative hazards $\Lambda(t|Z = z)$ by time $t = 0.5, 1, 1.5$. For the NWTS cohort, this individual corresponds to a cohort member of unfavorable central histology, tumor stage III/IV and an age of 3.3 years old, and the obtained prediction values correspond to this individual's cumulative risk estimates by the time 0.5, 1 and 1.5 years.

We generated 2000 datasets under simulation settings A and B respectively and reported the simulation results in Tables 5.4 and 5.5. Based on (5.2), we calculated the true $\Lambda(t|z)$ under setting A by $\Lambda(t|z) = 0.018t + (0.16z_1 + 0.04z_2 + 0.009z_3)t$ and under setting B by $\Lambda(t|z) = 0.005log\{1 + (3t)^{3.5}\} + (0.16z_1 + 0.04z_2 + 0.009z_3)t$. In part (a) of Tables 5.4 and 5.5, we observe the Bias $\hat{\Lambda}^*(t|z)$ by different time points are all small and positive, the ASE $\hat{\Lambda}^{**}(t|z)$ are close to and consistently smaller than ESE $\hat{\Lambda}^*(t|z)$, and as time increases, the 95% CI Cov get closer to 95%. In part (b) of the two tables, the Bias $\hat{\Lambda}^{**}(t|z)$ are also small and positive, the ASE $\hat{\Lambda}^*(t|z)$ are close to and consistently larger than ESE $\hat{\Lambda}^{**}(t|z)$, and the 95% CI Cov are close to 96%. These results suggest our prediction methods with or without calibration both perform well. However, the small inflation of SE $\hat{\Lambda}^{**}(t|z)$ requires further investigation.

Table 5.6 reports simulation results under settings A, A-1250, A-2500, A-10000 and A-20000. We are interested in evaluating our estimators $\hat{\Lambda}^*(t = 1|z)$ and $\hat{\Lambda}^{**}(t = 1|z)$ when the sample size N varies. We observe the biases of both estimators get smaller as N increases. The calibrated two-phase sampling estimators consistently have lower biases than estimators using standard weights. We also observe the ratio of ASE/ESE remains around 1 as N varies. The empirical 95% CI Cov are all close to 95%. The 95% CI Cov based on our estimation using calibration fluctuate between 95% and 96%. The results in this table further convince

us our prediction on individual specific cumulative hazard function based on the AH model is valid with or without calibration.

Table 5.4: Simulation results of $\hat{\Lambda}^*(t|z)$ and $\hat{\Lambda}^{**}(t|z)$ under simulation setting A

(a) Before calibration

	True $\Lambda(t\|z)$	Mean $\hat{\Lambda}(t\|z)^*$	Bias $\hat{\Lambda}(t\|z)^*$	ESE $\hat{\Lambda}(t\|z)^*$	ASE $\hat{\Lambda}(t\|z)^*$	95% CI Cov
t=0.5	0.11440	0.11549	0.00109	0.01668	0.01557	94.1%
t=1	0.22880	0.23095	0.00215	0.03319	0.03123	94.2%
t=1.5	0.34320	0.34631	0.00311	0.04969	0.04703	94.5%

(b) After calibration

	True $\Lambda(t\|z)$	Mean $\hat{\Lambda}(t\|z)^{**}$	Bias $\hat{\Lambda}(t\|z)^{**}$	ESE $\hat{\Lambda}(t\|z)^{**}$	ASE $\hat{\Lambda}(t\|z)^{**}$	95% CI Cov
t= 0.5	0.11440	0.11507	0.00067	0.01399	0.01401	95.4%
t= 1	0.22880	0.23012	0.00132	0.02773	0.02810	96.0%
t= 1.5	0.34320	0.34506	0.00186	0.04140	0.04236	96.2%

Simulation setting A: $\lambda_0(t) = 0.018, N = 5000$. See notation definition in Table 5.1.

Table 5.5: Simulation results of $\hat{\Lambda}^*(t|z)$ and $\hat{\Lambda}^{**}(t|z)$ under simulation setting B

(a) Before calibration

	True $\Lambda(t\|z)$	Mean $\hat{\Lambda}^*(t\|z)$	Bias $\hat{\Lambda}^*(t\|z)$	ESE $\hat{\Lambda}^*(t\|z)$	ASE $\hat{\Lambda}^*(t\|z)$	95% CI Cov
t=0.5	0.11358	0.11575	0.00217	0.01647	0.01561	93.8%
t= 1	0.23013	0.23437	0.00424	0.03283	0.03133	94.6%
t=1.5	0.34255	0.34902	0.00647	0.04918	0.04717	94.8%

(b) After calibration

	True $\Lambda(t\|z)$	Mean $\hat{\Lambda}^{**}(t\|z)$	Bias $\hat{\Lambda}^{**}(t\|z)$	ESE $\hat{\Lambda}^{**}(t\|z)$	ASE $\hat{\Lambda}^{**}(t\|z)$	95% CI Cov
t=0.5	0.11358	0.11506	0.00148	0.01393	0.01400	95.9%
t= 1	0.23013	0.23299	0.00285	0.02763	0.02816	95.9%
t=1.5	0.34255	0.34693	0.00438	0.04139	0.04239	96.2%

Simulation setting B:$\lambda_0(t) = 0.0525\frac{(3t)^{2.5}}{1+(3t)^{3.5}}$, $N = 5000$. See notation definition in Table 5.1.

Table 5.6: Simulation results of $\hat{\Lambda}^*(t = 1|z)$ and $\hat{\Lambda}^{**}(t = 1|z)$ using standard and calibrated weights under simulation settings A-1250, A, A-10000 and A-20000

Phase I sample size (N)	1250	2500	5000	10000	20000
Bias $\hat{\Lambda}^*(t = 1\|z)$	0.01271	0.00683	0.00215	0.00158	0.00140
Bias $\hat{\Lambda}^{**}(t\|z)$	0.00877	0.00440	0.00132	0.00106	0.00098
ASE/ESE $\hat{\Lambda}^*(t = 1\|z)$	0.890	0.967	0.941	0.959	0.972
ASE/ESE $\hat{\Lambda}^{**}(t = 1\|z)$	0.939	1.023	1.013	1.015	1.021
95% CI Cov using standard weights	93.9%	95.0%	94.1%	93.8 %	94.9%
95% CI Cov using calibrated weights	94.2%	95.6%	96.0%	95.1 %	96.1%

Simulation setting A: $\lambda_0(t) = 0.018$, $N = 5000$. See notation definition in Table 5.1.

Evaluation of variance decomposition

We conducted simulations to verify our conclusion in (2.14) that the variance of a two-phase sampling estimator is composed of two components: $Var(\hat{\theta}^*) = VarI(\hat{\theta}^*) + VarII(\hat{\theta}^*)$. $VarI(\hat{\theta}^*)$ is the variance we would obtain if we had complete information for all phase I samples and $VarII(\hat{\theta}^*)$ is the additional variance resulted from the fact that we only have complete information for phase II subsamples. The calculation of both variances was programmed in our package.

In the simulation study, 2000 datasets were generated under setting B. For each dataset, we fit the AH model to the complete phase I sample and obtained RS estimator $\hat{\theta}_N$, and then we fit the AH model to the two-phase sampling data and obtained $\hat{\theta}^*$, $\sqrt{Var(\hat{\theta}^*)}$, $\sqrt{VarI(\hat{\theta}^*)}$, and $\sqrt{VarII(\hat{\theta}^*)}$. Based on these 2000 replicates, we calculated the average of $\sqrt{VarI(\hat{\theta}^*)}$ (ASE I), the average of $\sqrt{VarII(\hat{\theta}^*)}$ (ASE II), the empirical standard error (ESE) of $\hat{\theta}_N$ and the ESE of $(\hat{\theta}^* - \hat{\theta}_N)$. If the first component of $Var(\hat{\theta}^*)$ is the variance we would obtain from a complete phase I cohort, we expect ASE I $(\hat{\theta}^*)$ = ESE $(\hat{\theta}_N)$ and different choices of phase II sampling probabilities will not affect the value of ASE I $(\hat{\theta}^*)$. Thus we added an additional simulation setting B', under which the phase I sample was still generated according to setting B, but the phase II sampling probabilities change to 0.6 for cases and 0.074 for controls. We present our simulation results in Table 5.7.

In Table 5.7, for both settings B and B', we observe ASE I $(\hat{\theta}^*) \approx$ ESE $(\hat{\theta}_N)$. In addition, comparing Table 5.7 (b) to (a) we find although the selection rules to Phase II differ in the two settings, their ASE I $(\hat{\theta}^*)$ are similar. These results show that $VarI(\hat{\theta}^*)$ is the variance we would obtain from a complete phase I sample and it does not depend on how phase II subsamples are selected. In Table 5.7, we also observe ASE II $(\hat{\theta}^*) \approx$ ESE $(\hat{\theta}^* - \hat{\theta}_N)$, which implies that $VarII(\hat{\theta}^*)$ is resulted from the extra variability introduced during phase II subsampling. Finally, we compare the squared ESE $(\hat{\theta}^*)$ to the sum of squared ESE $(\hat{\theta}_N)$ and squared ESE $(\hat{\theta}^* - \hat{\theta}_N)$. The last column of Table 5.6 shows $\text{ESE}^2(\hat{\theta}^*) \approx$ ESE $\hat{\theta}_N^2$ + ESE $(\hat{\theta}^* - \hat{\theta}_N)^2$. This simulation result verifies our theoretical conclusion in (2.14).

Table 5.7: Simulation results for studying decomposition of variance

(a) Setting B: $\bar{n}_{case} \approx \bar{n}_{control} \approx 645$

	ESE $\hat{\theta}_N$	ASE I $\hat{\theta}^*$	ESE $(\hat{\theta}^* - \hat{\theta}_N)$	ASE II $\hat{\theta}^*$	ESE $\hat{\theta}^*$	$\sqrt{ESE^2\hat{\theta}_N + ESE^2(\hat{\theta}^* - \hat{\theta}_N)}$
θ_1	0.01717	0.01771	0.02606	0.02559	0.03158	0.03121
θ_2	0.00641	0.00656	0.00759	0.00733	0.00992	0.00993
θ_3	0.00152	0.00154	0.00195	0.00191	0.00247	0.00247

(b) Setting B': $\bar{n}_{case} \approx \bar{n}_{control} \approx 430$

	ESE $\hat{\theta}_N$	ASE I $\hat{\theta}^*$	ESE $(\hat{\theta}^* - \hat{\theta}_N)$	ASE II $\hat{\theta}^*$	ESE $\hat{\theta}^*$	$\sqrt{ESE^2\hat{\theta}_N + ESE^2(\hat{\theta}^* - \hat{\theta}_N)}$
θ_1	0.01717	0.01784	0.03497	0.03402	0.03925	0.3896
θ_2	0.00641	0.00658	0.01043	0.01011	0.01226	0.01224
θ_3	0.00152	0.00155	0.00269	0.00257	0.00307	0.00309

$N = 5000$; $\hat{\theta}_N$: estimators from a random sample of N subjects; $\hat{\theta}^*$: two-phase sampling estimators using standard weights; ESE: empirical standard errors; ASE I: average of $\sqrt{VarI(\hat{\theta}^*)}$; ASE II: average of $\sqrt{VarII(\hat{\theta}^*)}$.

5.2 Efficiency is Improved

In this section we will use simulations to show how two-phase sampling designs and calibration improve the efficiency of estimating association parameters. We first study effects of different two-phase sampling designs on estimation efficiency. This investigation includes how the correlation between the surrogate W and the covariate Z_1 will influence estimation efficiency when W is used for stratified sampling. Then we will examine how efficiency gains depend on the choice of calibration variables. Finally, we investigate whether efficiency gains due to calibration becomes larger when the covariates in a model are correlated with each other.

5.2.1 Definition of Efficiency

A reasonable definition of the efficiency gain could be the relative efficiency of a two-phase finite sampling estimator to a random sampling estimator of the same sample size. Since the expensive variable is only measured for phase II subsamples in a two-phase design, the same sample size refers to an equivalence between the size of a random sample and the phase II sample size in a two-phase design so that the two sampling designs have the same costs on the expensive variables. However, in variable probability sampling (VPS), or so called Bernoulli sampling, we do not fix the size of phase II samples as finite sampling does. Instead, we fix the sampling fraction $\pi_0(V) = P(R = 1|V)$ for each individual depending on the values of their variable V. Thus to have a fair comparison, we first set up a new two-phase design called design I that mimics the random sampling by selecting phase II subsample using a constant sampling fraction π_c for each individual. Then we define the efficiency gain of a two-phase VPS design by the relative efficiency of estimators obtained from the two-phase VPS design compared to design I, subject to the constraint that the average sampling fraction used for the two-phase VPS equals $\pi_c : E[\pi_0(V)] = \pi_c$.

In view of (3.29), for a two-phase VPS estimator of regression model a coefficient, the efficiency gain (eff) is defined as the corresponding diagonal term of

$A^{-1}BA^{-1} + Q\left[\frac{1-\pi_c}{\pi_c}A^{-1}\left[\int_0^\tau \left\{Z - \frac{PY(t)Z}{PY(t)}\right\} dM(t)\right]^{\otimes 2} A^{-1}\right]$ divided by the corresponding diagonal term of $A^{-1}BA^{-1} + Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}A^{-1}\left[\int_0^\tau \left\{Z - \frac{PY(t)Z}{PY(t)}\right\} dM(t)\right]^{\otimes 2} A^{-1}\right]$.

In view of (4.9), for a calibrated two-phase VPS estimator, eff is defined as the corresponding diagonal term of $A^{-1}BA^{-1}+Q\left[\frac{1-\pi_c}{\pi_c}\left[f_{\theta_0,\Lambda_0}(T,\Delta,Z) - Q\{f_{\theta_0,\Lambda_0}(T,\Delta,Z)\tilde{V}^T\}(Q\tilde{V}\tilde{V}^T)^{-1}\tilde{V}\right]^{\otimes 2}\right]$ divided by the corresponding diagonal term of $A^{-1}BA^{-1} + Q\left[\frac{1-\pi_0(V)}{\pi_0(V)}\left[f_{\theta_0,\Lambda_0}(T,\Delta,Z) - Q\{f_{\theta_0,\Lambda_0}(T,\Delta,Z)\tilde{V}^T\}(Q\tilde{V}\tilde{V}^T)^{-1}\tilde{V}\right]^{\otimes 2}\right]$ where

$$f_{\theta_0,\Lambda_0}(T,\Delta,Z) = A^{-1}\int_0^\tau \left\{Z - \frac{PY(t)Z}{PY(t)}\right\} dM(t).$$

5.2.2 The Simulation Procedure for Phase I Samples

In section 5.1.2, we described our simulation procedure for setting B. The simulation procedure for phase I samples in this section followed the same procedure as setting B with a few modifications. First we reduced coefficient values including $\lambda_0(t)$ in the AH model from setting B by 80%:

$$\lambda(t|Z=z) = \lambda_0(t) + 0.032Z_1 + 0.08Z_2 + 0.018Z_3, \tag{5.2}$$

$\lambda_0(t) = 0.001c\frac{10.5(3t)^{2.5}}{1+(3t)^{3.5}}$, so that the failure rate was lowered from 14% to approximately 3%. We did not use the event rate of 14% in the NWTS because in practice it is more likely that a two-phase design is adopted for studying rare diseases. Second we increased the phase I sample size N to 10,000 since the event rates were lowered. Third we modified our approach to generate Z_1 and Z_3 in order to introduce different degrees of correlation between them. We simulated (Z_1', Z_3') from a bivariate normal distribution of size $N = 10,000$ with mean (1.2,1.2) and covariance matrix $\begin{bmatrix} 0.36 & 0.36\rho \\ 0.36\rho & 0.36 \end{bmatrix}$. $Z_1 = 1$ if Z_1' was greater than the 90% quantile of Z_1' and 0 otherwise; $Z_3 = exp(Z_3')$. By this means, although Z_1 and Z_3 were correlated, their marginal distributions remained the same as setting B. Fourth we modified the sensitivity and specificity of W with respective to Z_1 to generate new surrogates of Z with different degrees of correlation.

5.2.3 The Simulation Procedure for Phase II Subsamples

In order to study the efficiency gain by different study designs, we selected phase II subsamples with different sampling schemes. These included one random sampling (RS), one case-control (CC) sampling and three stratified case-control (SCC) sampling. RS used a constant phase II sampling probability π_c for every phase I member. With CC sampling, the probability depended on outcomes only: $\pi_0(V) = \pi_0(\Delta)$. In SCC sampling, the probability depended on both outcomes Δ and the surrogate W: $\pi_0(V) = \pi_0(\Delta, W)$.

The phase II selection probabilities were designed to satisfy several constraints so that different study designs were comparable to each other. The first constraint for the five sampling designs was $E[\pi_0(V)] = \pi_c$. Secondly, to ensure the efficiency gain of SCC versus CC sampling design was due to the stratification on auxiliary variables as opposed to other elements, we required the binary outcome Δ to be balanced among the phase II subsamples. Hence the potential efficiency gain would not be due to the difference in the distribution of cases and controls among phase II. Likewise we required W to be balanced among the controls selected to phase II for all three SCC designs.

Let $p_\Delta = P(\Delta = 1)$, $p_{00} = P(\Delta = 0, W = 0)$ and $p_{01} = P(\Delta = 0, W = 1)$. To meet our constraints, for RS we set

$$\pi_c = 2p_\Delta, \tag{5.3}$$

as a result,

$$P(\Delta = 1|R = 1) = 0.5.$$

For CC sampling, we set

$$\pi_0(\Delta) = \begin{cases} 1, & \Delta = 1 \\ p_\Delta/(1 - p_\Delta), & \Delta = 0 \end{cases} \tag{5.4}$$

so that

$$\begin{aligned} P(\Delta = 1|R = 1) =& \frac{P(R = 1|\Delta = 1)P(\Delta = 1)}{P(R = 1|\Delta = 1)P(\Delta = 1) + P(R = 1|\Delta = 0)P(\Delta = 0)} \\ =& \frac{p_\Delta * 1}{p_\Delta * 1 + (1 - p_\Delta) * p_\Delta/(1 - p_\Delta)} \\ =& 0.5 \end{aligned}$$

and

$$\begin{aligned} E[\pi_0(V)] =& P(R = 1|\Delta = 1)P(\Delta = 1) + P(R = 1|\Delta = 0)P(\Delta = 0) \\ =& p_\Delta + p_\Delta/(1 - p_\Delta) * (1 - p_\Delta) \\ =& 2p_\Delta = \pi_c. \end{aligned}$$

For SCC sampling, we set

$$\pi_0(\Delta, W) = \begin{cases} 1, & \Delta = 1 \\ \frac{0.5p_\Delta}{p_{00}} & \Delta = 0, W = 1 \\ \frac{0.5p_\Delta}{p_{01}} & \Delta = 0, W = 0 \end{cases}, \tag{5.5}$$

so that

$$\begin{aligned} & P(\Delta = 1|R = 1) \\ =& \frac{P(R = 1|\Delta = 1)P(\Delta = 1)}{P(R = 1, \Delta = 1) + P(R = 1, \Delta = 0, W = 0) + P(R = 1, \Delta = 0, W = 1)} \\ =& \frac{p_\Delta * 1}{p_\Delta * 1 + p_{00} * \frac{0.5p_\Delta}{p_{00}} + p_{01} * \frac{0.5p_\Delta}{p_{01}}} \\ =& 0.5, \end{aligned}$$

$$\begin{aligned} & P(W = 1|R = 1, \Delta = 0) \\ =& \frac{P(R = 1|\Delta = 0, W = 1)P(\Delta = 0, W = 1)}{P(R = 1|\Delta = 0, W = 1)P(\Delta = 0, W = 1) + P(R = 1|\Delta = 0, W = 0)P(\Delta = 0, W = 0)} \\ =& \frac{\frac{0.5p_\Delta}{p_{00}} * p_{00}}{\frac{0.5p_\Delta}{p_{00}} * p_{00} + \frac{0.5p_\Delta}{p_{01}} * p_{01}} \\ =& 0.5, \end{aligned}$$

and

$$\begin{aligned}
E[\pi_0(V)] =& P(R=1|\Delta=1)P(\Delta=1) + P(R=1|\Delta=0,W=1)P(\Delta=0,W=1) \\
&+ P(R=1|\Delta=0,W=0)P(\Delta=0,W=0) \\
=& p_\Delta + \frac{0.5p_\Delta}{p_{00}} * p_{00} + \frac{0.5p_\Delta}{p_{01}} * p_{01} \\
=& 2p_\Delta = \pi_c.
\end{aligned}$$

Because the given sampling probabilities $\pi_0(V)$ in (5.3), (5.4) and (5.5) depend on the joint distribution of (Δ, W), p_Δ, p_{00} and p_{01} need to be estimated prior to the calculation of $\pi_0(V)$. In our simulation procedure, we first simulated a single phase I dataset with $N = 100,000$ to estimate p_Δ, p_{00} and p_{01}. With these estimates, we computed the phase II selection probabilities according to (5.3), (5.4) or (5.5). Lastly, we used the obtained sampling probabilities to select the phase II subsamples from the simulated phase I samples.

5.2.4 Design of the Simulation Study

According to our study aims, three factors were varied in our simulation studies. The first one is the sampling design at phase II where setting I used RS, setting II used CC sampling and setting III-V used SCC sampling. The second factor is the accuracy of our surrogate W. The sensitivity and specificity of W in setting III, IV and V were (0.74,0.74), (0.74,0.96) and (0.96,0.96) respectively. The third factor is the correlation between Z_1 and Z_3. The correlation ρ between the latent variable (Z_1', Z_3') of (Z_1, Z_3) equaled 0 in setting C and 0.3 in setting D. Overall 10 simulation scenarios were designed (5.8). Based on (5.3), (5.4) and (5.5), the phase II sampling probabilities for various settings were computed in Table 5.9.

2000 dataset were generated for each of the 10 simulation settings. For settings C-I and D-I (Table 5.8), we used standard random sampling estimating method. For the other eight settings (Settings II-V) in Table 5.8, we used four different two-phase analyzing methods. Method 1 applied our two-phase sampling method using the standard weights; Methods 2-4 used calibrated weights with different choices of auxiliary variables.

Table 5.8: Simulation scenarios

Settings	Correlation between Covariates	Surrogate W	Sampling Designs
C-I	None	not available	RS
C-II	None	not available	CC
C-III	None	poor	SCC
C-IV	None	average	SCC
C-V	None	good	SCC
D-I	low	not available	RS
D-II	low	not available	CC
D-III	low	poor	SCC
D-IV	low	average	SCC
D-V	low	good	SCC

Table 5.9: Phase II sampling probabilities

	Formula	Values
I	π_c	0.062
II	$\pi_0(\Delta)$	$\pi_0(1) = 1, \pi_0(0) = 0.032$
III	$\pi_0(\Delta, W)$	$\pi_0(1,1) = \pi_0(1,0) = 1, \pi_0(0,1) = 0.052, \pi_0(0,0) = 0.023$
IV	$\pi_0(\Delta, W)$	$\pi_0(1,1) = \pi(1,0) = 1, \pi_0(0,1) = 0.15, \pi_0(0,0) = 0.018$
V	$\pi_0(\Delta, W)$	$\pi_0(1,1) = \pi_0(1,0) = 1, \pi_0(0,1) = 0.13, \pi_0(0,0) = 0.018$

In Method 2, we calibrated to the stratum frequencies. The auxiliary variables were $(1 - \Delta)$ for the CC design and $\{(1 - \Delta)W, (1 - \Delta)(1 - W)\}$ for the SCC designs, so that the estimated stratum sizes based on phase II subsamples agreed with the stratum sizes at phase I. Since in our simulation studies, all the cases were selected to phase II according to

(5.3),(5.4) and (5.5), calibration to the total of case frequencies was unnecessary. Although we used Bernoulli sampling to select phase II samples, through this calibration we would acquire estimators that were as if obtained from finite sampling. Since finite sampling is slightly more efficient than Bernoulli sampling (see the discussion in 2.1), we expect efficiency improvement from Method 2 compared to Method 1.

In Method 3, in addition to calibrating to stratum frequencies, we included the interaction terms $\{Z_2\Delta,\ Z_2(1-\Delta),\ Z_3\Delta,\ Z_3(1-\Delta)\}$. Z_2, Z_3 and Δ were fully observed at phase I. According to Chapter 8.5 [Lumley, 2011], calibration to the total of the outcomes, the covariates or their correlates will not be very helpful for improving the precision of the slope estimates in a regression model. The more effective auxiliary variables are the correlates of the scores for the regression model. This conclusion is also revealed in our formula (4.6). Thus we included the interaction terms between auxiliary variables and outcomes since they were simple and straightforward to obtain and correlated with the scores.

In Method 4, we first fit the AH model to the phase I data using (W, Z_2, Z_3) as our covariates and obtained the integrated martingale residuals for Z_2, Z_3:

$$\psi_{Z_2} = \int_0^\tau \left[Z_2 - \frac{\mathbb{P}_N Y(t) Z_2}{\mathbb{P}_N Y(t)} \right] d\hat{M}(t), \psi_{Z_3} = \int_0^\tau \left[Z_3 - \frac{\mathbb{P}_N Y(t) Z_3}{\mathbb{P}_N Y(t)} \right] d\hat{M}(t)$$

where $d\hat{M}(t) = dN(t) - \hat{\Lambda}(t) - Z^T\hat{\theta}$. Note here $Z = (W, Z_2, Z_3)$. $\hat{\theta}$ and $\hat{\Lambda}$ are the parameter estimates in the fitted AH model. ψ_{Z_2} and ψ_{Z_3} can be obtained in the R ahaz program [Gorst-Rasmussen and Scheike, 2012]. Next we used ψ_{Z_2} and ψ_{Z_3} as our auxiliary variables and fit the AH model to the two-phase sampling data with calibration. Note if $W = Z_1$, then ψ_{Z_2} and ψ_{Z_3} are the estimating functions in Lin and Ying [1994] and called pseudoscore in Kulich [1997] since they are the modified score functions for regression parameters in the AH model. Thus we expect efficiency improvement from the use of these auxiliary variables, especially when W is highly correlated with Z_1. Breslow et al. [2009b] proposed another calibration method, which can be adopted for improving the AH model with two-phase sampling as well. They first used the phase II data to develop a prediction function for the partially missing variables from the variables available for every cohort member,

Table 5.10: Two-phase sampling analysis methods

Methods	Calibration variables	Settings applied to
1	None	II, III, IV, V
2	$(1-\Delta)$	II
2	$(1-\Delta)W$, $(1-\Delta)(1-W)$	III, IV, V
3	$(1-\Delta)$,$Z_2\Delta$, $Z_2(1-\Delta)$, $Z_3\Delta$, $Z_3(1-\Delta)$	II
3	$(1-\Delta)W$, $(1-\Delta)(1-W)$, $Z_2\Delta$, $Z_2(1-\Delta)$, $Z_3\Delta$, $Z_3(1-\Delta)$	III, IV, V
4	$(1-\Delta)W$, $(1-\Delta)(1-W)$, ψ_{Z_2}, ψ_{Z_3}	III, IV, V

and then imputed the partially missing variables in the phase I based on the prediction function and the phase I information. In the last step, they fit the model to the whole cohort using the imputed variables as the replacement of partially missing variables and obtained the estimated influence function for each cohort subject as the auxiliary variables. We did not take this approach because in our simulation studies, we did not have much additional information other than W at phase I to impute Z_1. Fitting a prediction function is unnecessary in our case and may introduce extra variation.

We summarize our two-phase sampling analysis methods in Table 5.10. To evaluate these methods, we calculated the Mean, Bias, ESE, ASE, 95% C.I. Cov and the efficiency gain (eff). Table 5.11 $\sim$ 5.15 summarize our simulation results for setting C and Table 5.16 for setting D.

5.2.5 Results of the Simulation Study

The average phase II subsample size of the simulated 2000 datasets for each of the 10 simulation settings in Table 5.8 ranged from 620 to 624. Thus results of these simulation studies are comparable to each other since the costs of the measurement of the expensive variable would be approximately the same.

Bias

We used setting C to study the bias of our estimation methods. Table 5.11 presents RS estimation results while Table 5.12 $\sim$ 5.15 provides two-phase sampling estimation results. Comparing Table 5.12 $\sim$ 5.15 to Table 5.11, we find the biases of estimators increase but the ESE and ASE of estimators dramatically decrease. As a result, the efficiency gain from two-phase designs ranges from 3.55 to 16.16 depending on the covariates, the sampling designs, the quality of the surrogates, and the calibration methods. There is a bias-variance trade-off when implementing the two-phase sampling.

Although biases were introduced for the efficiency gain by two-phase sampling, these biases are not large. Furthermore, the stratified sampling and calibration methods, both using additional information in auxiliary variables for further improvement of estimators, reduced these biases. In Table 5.12, the bias of the θ_1 estimate decreases from 0.00184 to 0.00126 when stratified sampling was conducted based on W, a surrogate of Z_1. The bias is further reduced from 0.00126 to 0.00071, and to 0.00010 when the sensitivity and specificity of W increase from (0.74,0.74) to (0.74, 0.96), and to (0.96, 0.96). From this observation we conclude stratification on the correlate of a covariate will reduce the bias of the coefficient estimate for this covariate and the reduction of this bias is substantial when their correlation is high. When comparing the biases of θ_2 and θ_3 estimates across Table 5.12 $\sim$ 5.15 for each simulating setting respectively, we find calibration is able to reduce the bias as well. In general, the biases for θ_2 and θ_3 analyzed by Method 4 $<$ Method 3 $<$ Method 2 $<$ Method 1.

Variance Estimation

We first examine the results under setting C, in which covariates are independent of each other. With random sampling (Table 5.11), 95% CI Cov for θ_1 and θ_2 are 91.5% and 94.4% respectively. The ASE of $\hat{\theta}_2$ is closer to the ESE of $\hat{\theta}_2$ than the ASE of $\hat{\theta}_1$ to the ESE of $\hat{\theta}_1$. This difference may be due to the difference of the distributions of binary variables Z_1 and Z_3 since $P(Z_1 = 1) = 0.1$ and $P(Z_2 = 1) = 0.4$ in the phase I cohort. With two-phase samplings (Table 5.12 $\sim$ 5.15), the 95% CI Cov for θ_1 are in general close to 95% even though Z_1 is not as balanced as Z_2 in the phase I cohort. This improvement of the coverage probabilities is possibly due to the oversampling of more informative subjects by two-phase designs and the stratified sampling based on the correlate of Z_1. With random sampling, the 95% CI Cov for θ_3 is only 89.7% and the ASE of $\hat{\theta}_3$ is about 10% smaller than the ESE of $\hat{\theta}_3$. Note Z_3 is a continuous variable. With two-phase sampling, the coverage probabilities for θ_3 are all in between 93% and 94% except the results in Table 5.15, in which the coverage probabilities are all around 95%. These results show that when the sample size is approximately 620, the estimation of variances may have not been stabilized for the coefficients of continuous variable. However, two-phase sampling performs better than random sampling at this sample size in term of variance estimation, especially when Method 4 is used for calibration.

Results in setting D (Table 5.16) follow the same pattern as setting C except that both Methods 3 and 4 improved the variance estimation for θ_3 estimate. It worth noting in the presence of correlation between covariates under setting D, Method 3 performs better than under setting C when comparing Table 5.16 to Table 5.14. The coverage probabilities for parameters in all of the five sampling designs are close to 95% in the column of Method 3 in Table 5.16.

Efficiency

In the following, we investigate how sampling designs, the quality of surrogates, the choice of auxiliary variables and the correlation between covariates will influence the efficiency gain of

our estimation. The efficiency gains (eff) in Table 5.11 $\sim$ 5.15 are calculated based on ESE. The efficiency gains from various two-phase designs range from 3.55 in setting C-II of Table 5.12 to 16.16 in setting C-V of Table 5.15. These large values of eff are a strong evidence of the ability of two-phase sampling to improve estimation efficiency compared to random sampling.

In Table 5.12, the last column shows eff in C-II (3.55) $<$ C-III(5.06) $<$ C-IV (6.05) $<$ C-V(9.19) for θ_1. This comparison result suggests when a surrogate of a covariate is used for stratified sampling, SCC sampling is more efficient than CC sampling for estimating the coefficient of this covariate. As the correlation of this surrogate with the covariate increases, the efficiency gain becomes larger. In the same table, we also observe that SCC sampling based on W will not improve the estimation efficiency for θ_2 and θ_3, which can be explained by the independence between W and covariates Z_2 and Z_3. The above three patterns appear in Table 5.13 $\sim$ 5.15 as well.

We then study how the choice of calibration variables affect our estimation efficiency. We begin with setting C. As expected, calibration to stratum frequencies, which recovers the efficiency loss in Bernoulli sampling compared to finite sampling, slightly improves the estimation efficiency for every parameter regardless of sampling designs (Table 5.13 vs. Table 5.12). Especially when a very good surrogate is available such as in setting C-V, calibration to stratum frequencies alone is able to increase eff from 9.19 to 12.29 for θ_1. Calibration on the interaction terms between Z_2, Z_3 and outcomes (Table 5.14) increases the eff for θ_2 and θ_3 but not θ_1. Taking setting C-IV as an example, the eff for θ_2 increases from 5.83 in Table 5.13 to 11.85 in Table 5.14 and the eff for θ_3 increases from 3.79 to 9.4, while the eff for θ_1 does not increase. The efficiency gain by calibrating on the integrated martingale residuals is even more prominent. Eff reaches 12.21 for θ_2 and 14.56 for θ_3 in setting C-IV as shown in Table 5.15.

Finally, we investigate the effect of correlation between covariates on efficiency gains. According to the design of our simulation studies, ρ, the correlation between the latent variables (Z_1', Z_3') of (Z_2, Z_3) was set at 0.3 in setting D, which lead to a correlation of about

0.17 between Z_1 and Z_3. This value mimics real scenarios. For example, in the NWTS cohort, the correlation between the cancer stage and age is about 0.16. From Table 5.16 we find our conclusions about the effects of various factors on efficiency gains for setting C remain the same for setting D.

Comparing results of Method 3 and Method 4 to Method 2 in Table 5.16, we do not observe an increase in eff for θ_1 in general when we calibrate on factors involving Z_2, Z_3. The only exception is setting D-V when the sensitivity and specificity of a surrogate are high. The eff for θ_1 increases from 12.43 to 12.59 comparing Method 3 to Method 2 in this table. On the other hand, for setting C-V, the eff for θ_1 is 12.29 by Method 2 in Table 5.13 and 12.25 by Method 3 in Table 5.14. When the correlation between Z_1 and Z_3 increases to 0.47 in simulations, this increase in eff by Method 3 is more evident. The eff for θ_1 is 10. 29 in Method 2 and 11.04 in Method 3 (results not shown). We do not observe the efficiency gain by Method 4 for θ_1 and by Method 3 for other simulation settings other than settings C-V and -V.

Table 5.11: Simulation results from a random sampling design

Setting	Coef.	True	Mean	Bias	ESE	ASE	95% CI Cov	eff
C-I	θ_1	0.03200	0.03222	0.00022	0.02058	0.02012	91.5%	1
	θ_2	0.00800	0.00822	0.00022	0.00816	0.00807	94.4%	1
	θ_3	0.00180	0.00185	0.00005	0.00187	0.00176	89.7%	1

Simulation setting C: $\lambda_0(t) = 0.001\frac{10.5(3t)^{2.5}}{1+(3t)^{3.5}}$, $N = 10,000$. See descriptions of settings I-V in Table 5.8.
ESE: empirical standard errors. ASE: average standard errors. eff: efficiency gain, which is defined in section 5.2.1.

Table 5.12: Simulation results from two-phase sampling designs analyzed without calibration (Method 1)

Settings	Coef.	True	Mean	Bias	ESE	ASE	95% CI Cov	eff
C-II	θ_1	0.03200	0.03384	0.00184	0.01093	0.01019	94.8%	3.55
	θ_2	0.00800	0.00812	0.00012	0.00324	0.00321	95.1%	6.34
	θ_3	0.00180	0.00193	0.00013	0.00089	0.00081	93.3%	4.41
C-III	θ_1	0.03200	0.03326	0.00126	0.00915	0.00925	95.4%	5.06
	θ_2	0.00800	0.00815	0.00015	0.00321	0.00322	95.8%	6.46
	θ_3	0.00180	0.00194	0.00014	0.00088	0.00082	94.2%	4.52
C-IV	θ_1	0.03200	0.03271	0.00071	0.00837	0.00815	94.6%	6.05
	θ_2	0.00800	0.00826	0.00026	0.00347	0.00342	95.3%	5.53
	θ_3	0.00180	0.00202	0.00022	0.00098	0.00090	93.5%	3.64
C-V	θ_1	0.03200	0.03210	0.00010	0.00679	0.00683	94.9%	9.19
	θ_2	0.00800	0.00824	0.00024	0.00326	0.00328	95.8%	6.27
	θ_3	0.00180	0.00200	0.00020	0.00096	0.00087	92.8%	3.79

See the explanations of notations and settings in Table 5.11.

Table 5.13: Simulation results from two-phase sampling designs analyzed with calibration on stratum frequencies (Method 2)

Settings	Coef.	True	Mean	Bias	ESE	ASE	95% CI Cov	eff
C-II	θ_1	0.03200	0.03374	0.00174	0.01065	0.00998	95.0%	3.73
	θ_2	0.00800	0.00810	0.00010	0.00318	0.00316	94.7%	6.58
	θ_3	0.00180	0.00192	0.00012	0.00088	0.00080	93.1%	4.52
C-III	θ_1	0.03200	0.03315	0.00115	0.00874	0.00887	95.5%	5.54
	θ_2	0.00800	0.00812	0.00012	0.00314	0.00316	95.6%	6.75
	θ_3	0.00180	0.00194	0.00014	0.00087	0.00081	94.1%	4.62
C-IV	θ_1	0.03200	0.03275	0.00075	0.00803	0.00772	93.8%	6.57
	θ_2	0.00800	0.00820	0.00020	0.00338	0.00334	95.5%	5.83
	θ_3	0.00180	0.00201	0.00021	0.00096	0.00088	92.9%	3.79
C-V	θ_1	0.03200	0.03206	0.00006	0.00587	0.00595	95.3%	12.29
	θ_2	0.00800	0.00819	0.00019	0.00318	0.00320	95.6%	6.58
	θ_3	0.00180	0.00199	0.00019	0.00094	0.00086	92.8%	3.96

See the explanations of notations and settings in Table 5.11.

Table 5.14: Simulation results from two-phase sampling designs analyzed with calibration on stratum frequencies and interactions between outcomes and covariates (Method 3)

Settings	Coef.	True	Mean	Bias	ESE	ASE	95% CI Cov	eff
C-II	θ_1	0.03200	0.03376	0.00176	0.01067	0.00997	95.0%	3.72
	θ_2	0.00800	0.00813	0.00013	0.00244	0.00241	94.7%	11.18
	θ_3	0.00180	0.00187	0.00007	0.00063	0.00058	93.5%	8.81
C-III	θ_1	0.03200	0.03320	0.00120	0.00878	0.00887	95.5%	5.49
	θ_2	0.00800	0.00811	0.00011	0.00236	0.00237	95.2%	11.96
	θ_3	0.00180	0.00187	0.00007	0.00061	0.00059	93.8%	8.54
C-IV	θ_1	0.03200	0.03287	0.00087	0.00806	0.00772	93.5%	6.51
	θ_2	0.00800	0.00813	0.00013	0.00237	0.00236	95.1%	11.85
	θ_3	0.00180	0.00190	0.00010	0.00064	0.00061	93.9%	9.40
C-V	θ_1	0.03200	0.03221	0.00021	0.00588	0.00597	95.5%	12.25
	θ_2	0.00800	0.00813	0.00013	0.00226	0.00229	95.1%	13.04
	θ_3	0.00180	0.00189	0.00009	0.00063	0.00060	93.6%	8.81

See the explanations of notations and settings in Table 5.11.

Table 5.15: Simulation results from two-phase sampling designs analyzed with calibration on stratum frequencies and integrated martingale residuals (Method 4)

Settings	Coef.	True	Mean	Bias	ESE	ASE	95% CI Cov	eff
C-III	θ_1	0.03200	0.03327	0.00127	0.00875	0.00890	95.7%	5.53
	θ_2	0.00800	0.00809	0.00009	0.00223	0.00226	95.4%	13.39
	θ_3	0.00180	0.00180	0.00000	0.00051	0.00053	95.0%	13.44
C-IV	θ_1	0.03200	0.03304	0.00104	0.00807	0.00778	93.9%	6.50
	θ_2	0.00800	0.00806	0.00006	0.00216	0.00219	95.5%	12.21
	θ_3	0.00180	0.00180	0.00000	0.00049	0.00054	95.1%	14.56
C-V	θ_1	0.03200	0.03234	0.00034	0.00589	0.00599	95.5%	12.21
	θ_2	0.00800	0.00807	0.00007	0.00203	0.00209	96.0%	16.16
	θ_3	0.00180	0.00180	-0.00000	0.00048	0.00052	95.0%	15.18

See the explanations of notations and settings in Table 5.11.

Table 5.16: Simulation results of different sampling designs under simulation setting D

Setting	Coef.	ESE	ASE	Cov	eff
D-I	θ_1	0.02200	0.02083	91.4%	1
	θ_2	0.00813	0.00801	94.2%	1
	θ_3	0.00199	0.00184	89.4%	1

Analysis Methods		1				2				3				4			
Settings	Coef.	ESE	ASE	Cov	eff	ESE	ASE	Cov	eff	ESE	ASE	Cov	eff	ESE	ASE	Cov	eff
D-II	θ_1	0.01044	0.01050	94.9%	4.44	0.01024	0.01031	95.1%	4.62	0.01025	0.01025	95.2%	4.61	NA	NA	NA	NA
	θ_2	0.00333	0.00324	94.9%	5.96	0.00328	0.00319	94.7%	6.14	0.00245	0.00244	95.0%	11.01	NA	NA	NA	NA
	θ_3	0.00094	0.00089	93.9%	4.48	0.00093	0.00088	93.0%	4.58	0.00067	0.00065	95.0%	8.82	NA	NA	NA	NA
D-III	θ_1	0.00991	0.00972	95.3%	4.93	0.00950	0.00932	94.8%	5.36	0.00954	0.00929	94.7%	5.32	0.00961	0.00937	94.8%	5.24
	θ_2	0.00325	0.00324	95.2%	6.26	0.00320	0.00318	95.1%	6.45	0.00242	0.00240	95.5%	11.29	0.00224	0.00225	95.6%	13.17
	θ_3	0.00095	0.00089	93.9%	4.39	0.00093	0.00087	93.0 %	4.58	0.00067	0.00064	94.3%	8.82	0.00054	0.00058	95.8%	13.58
D-IV	θ_1	0.00893	0.00858	93.5%	6.07	0.00835	0.00809	94.2%	6.94	0.00838	0.00808	94.3%	6.89	0.00853	0.00829	94.6%	6.65
	θ_2	0.00341	0.00340	95.6%	5.68	0.00335	0.00332	95.4%	5.89	0.00240	0.00237	94.8%	11.48	0.00218	0.00218	95.2%	13.91
	θ_3	0.00099	0.00092	93.0%	4.04	0.00097	0.00090	93.3%	4.21	0.00067	0.00065	94.8%	8.82	0.00053	0.00057	95.8%	14.10
D-V	θ_1	0.00730	0.00721	94.9%	9.08	0.00624	0.00627	95.3%	12.43	0.00620	0.00623	95.5%	12.59	0.00641	0.00648	95.5%	11.78
	θ_2	0.00326	0.00325	95.6%	6.22	0.00319	0.00317	95.5%	6.50	0.00232	0.00229	95.1%	12.28	0.00207	0.00209	95.8%	15.43
	θ_3	0.00093	0.00088	93.4%	4.58	0.00091	0.00086	95.5%	4.78	0.00064	0.00063	95.1%	9.67	0.00050	0.00054	95.6%	15.84

See the explanations of notations and settings in Table 5.11.Cov: 95% CI Cov.

5.3 Estimation Under Model Misspecification

In our method development for the AH model, we also considered estimation under model misspecification. Hence in the last section of our numeric studies, we will briefly examine the performance of our robust standard errors under model misspecification.

5.3.1 Design of simulation studies

In the following simulation studies, we evaluate the performance of our estimating method using robust standard errors by fitting the AH model to a data simulated based on the Cox proportional hazards model.

In simulation setting E, we modified the AH model (5.2) in setting B to the following Cox model while keeping other factors unchanged:

$$\lambda(t|z) = \lambda_0(t)\exp(1.3z_1 + 0.5z_2 + 0.3z_3).$$

Thus, the baseline hazard is still the same as setting B: $\lambda_0(t) = 0.005\frac{10.5(3t)^{2.5}}{1+(3t)^{3.5}}$ and $N = 5000$. The coefficients in this Cox model were determined by fitting a Cox model to the original NWTS cohort. In simulation setting F, we also used the same Cox model but reduced the baseline hazard function in setting E by 80% in order to simulate a rare disease scenario. For both settings E and F, 2000 datasets were generated following the simulation procedure described in section 5.1.1.

For both simulation settings, we first fit the AH model to the phase I random sample using Z_1, Z_2 and Z_3 as our covariates. We computed the mean of random sampling estimators across 2000 datasets. Since the mean of these random sampling estimators was computed based on N subjects, we denote this estimator by $\hat{\theta}_N$. Next, we fit the same AH model to the two-phase sampling dataset and calculated the mean, ESE, ASE of $\hat{\theta}^*$ as well as the average robust standard error of $\hat{\theta}^*$ (ARSE) across 2000 datasets. We presented these results in Table 5.17 for simulation setting E and in Table 5.18 for simulation setting F.

5.3.2 Results of the simulation study

When the model is misspecified, both Table 5.17 and 5.18 exhibit biases of our two-phase sampling estimators $\hat{\theta}^*$. However, $\hat{\theta}^*$ are close to the phase I random sampling estimator $\hat{\theta}_N$, which supports our theoretical results in theorems 3.8.1 and 3.8.2.

In both settings E and F in Table 5.17 and 5.18, comparing ASE $\hat{\theta}^*$ and ARSE $\hat{\theta}^*$ to ESE $\hat{\theta}^*$ we find the model-based standard errors of $\hat{\theta}_3^*$ are inflated but the robust standard errors correct this inflation. We do not find similar deviations of ASE from ESE for $\hat{\theta}_1^*$ or $\hat{\theta}_2^*$, probably because Z_1 and Z_2 are binary variables. The variance of estimating the difference between the two levels of a binary variable should be close when assuming an additive effect or a proportional effect of this variable on the outcome. Z_3, on the other hand, is a continuous variable. Assuming an additive effect of Z_3 when this effect is actually proportional will bias our estimation of the standard error.

Table 5.17: Simulation results under simulation setting E

	True	Mean $\hat{\theta}_N$	Mean $\hat{\theta}^*$	ESE $\hat{\theta}^*$	ASE $\hat{\theta}^*$	ARSE $\hat{\theta}^*$
θ_1	1.30000	0.18669	0.19008	0.03604	0.03520	0.03505
θ_2	0.50000	0.04795	0.04855	0.01277	0.01278	0.01283
θ_3	0.30000	0.05187	0.05227	0.00508	0.00523	0.00499

Table 5.18: Simulation results under simulation setting F

	True	Mean $\hat{\theta}$	Mean $\hat{\theta}^*$	ESE $\hat{\theta}^*$	ASE $\hat{\theta}^*$	ARSE $\hat{\theta}^*$
θ_1	1.30000	0.04530	0.04612	0.01223	0.01201	0.01194
θ_2	0.50000	0.01161	0.01175	0.00493	0.00486	0.00488
θ_3	0.30000	0.01608	0.01621	0.00230	0.00240	0.00231

Chapter 6

ANALYSIS OF TWO-PHASE ARIC DATA USING ADDITIVE HAZARDS MODELS

Atherosclerosis Risk in Communities ARIC is one of the few major cohort studies to make systematic use of stratified sampling to select participants within the main cohort for biomarker studies of the risk of Coronary Heart Disease (CHD) and stroke. Current biostatistical research is concerned with efficient ways of combining the marker data available for disease cases and the selected sample with information on standard risk factors that is available for a much larger number of subjects in the main cohort. The new statistical theory and methods we develop in Chapter 2∼4 can be used to address this concern. In this chapter, using ARIC data to illustrate our new statistical methods serves the dual purpose of demonstrating their utility, or lack thereof, when applied to important studies, and of providing new insights about disease risk.

6.1 Background and Aims

In this chapter, we will use the new methods developed in Chapter 3∼4 to perform risk prediction of CHD based on an additive hazards model fitted to a two-phase ARIC study introduced in Ballantyne et al. [2004]. This will involve joint estimation of the coefficients of excess risk associated with each risk factor and of the baseline CHD risk curve, as a function of time on study. A method that predicts individual CHD risk at 5 and 8 years in the future, for example, could provide physicians a practical tool to explain clinical results to patients and to plan for prevention. The intent is to use not only data on hs-CRP available for the limited number of CHD cases and members of the Cohort Random Sample (CRS), but also to incorporate the information on standard risk factors available for a much large number

of subjects remaining in the main cohort. Breslow and Lumley [2013] demonstrated that such risk prediction based on the Cox proportional hazards model was feasible, using 10,000 simulated stratified samples from a cohort of participants in the National Wilms Tumor Study. They showed empirically calibration of the weights from information available for all subjects in the main cohort reduced the mean squared error (MSE) for risk estimates. No such risk prediction methodology has yet been developed for the additive hazards model with stratified sampling, either using standard or calibrated weights. No such risk prediction has been applied to an actual two-phase data.

The inflammation marker hs-CRP has been shown in multiple epidemiological studies to predict cardiovascular disease. A statement from the Centers for Disease Control and Prevention and the American Heart Association concluded that hs-CRP was an independent predictor of increased coronary risk and recommended the optional use of hs-CRP to identify asymptomatic people, already known to be at intermediate risk on the basis of standard risk factors, who might be at even higher risk [Pearson et al., 2003]. Development of new statistical tools to improve the accuracy of risk prediction in this context would be an important contribution.

Two statistical studies using data from Ballantyne et al. (2004) have been published already. Breslow et al. [2009b] studied methods of fitting Cox proportional hazards models to evaluate the effect of lipoprotein-associated phospholipase A2 (Lp-PLA2) levels in combination with standard risk factors on CHD outcomes. Calibration of the design weights using information available for the entire cohort increased the efficiency of most regression coefficients, though not those for Lp-PLA2 itself. Kang et al. (2013) demonstrated that increased levels of high-sensitivity C-reactive protein (hs-CRP) acted additively (added to), rather than proportionally, on the baseline risks of both CHD and stroke. They mentioned in the conclusion of their paper that efficiency gains might be possible using techniques similar to those developed by Breslow et al. We are interested to compare our inference results on hs-CRP to Kang et al. [2013] and calibration results to Breslow et al. [2009b].

After reviewing earlier works we consider the aims of our data analysis are

1. Combination of the estimated coefficients and baseline risk function to yield a risk prediction equation based on risk factors measured for the stratified sample and the time since their measurement. This will include calculation of standard errors of regression coefficient estimates and risk estimates.

2. Calibration of the design (inverse probability of sampling) weights used in the above methods using information on the association between disease outcomes and standard risk factors known for most subjects in the cohort. The goal is to include more cohort data and thereby improve the accuracy of the inference on major risk factors and the overall risk prediction.

3. Based on the prediction equations we built, two sets of CHD predictors will be constructed based on hs-CRP and standard risk factors, one with and one without calibration of the weights. We are interested to examine whether standard errors for major risk factors, hs-CRP and interaction terms involving main risk factors (e.g., LDL-C) will be reduced by calibration, and whether the standard errors of the predicted risks will also be reduced by calibration.

6.2 Study Population

ARIC is a prospective epidemiologic study that began in 1987 and followed healthy middle-age men and women in four U.S. communities. The detailed design, objectives and data collection of ARIC have been described previously Investigators et al. [1989]. The main cohort was followed up for the subsequent (after visit 2) development of a CHD event, including CHD-related death. Subjects alive and event-free at the end of 1998, or who had been lost to follow-up, had their records censored at that time. A cohort random sample (CRS) was selected using a stratified random sampling design based on 8 strata defined by sex, race (black versus white) and age at baseline (≥ 55 versus < 55). Hs-Crp and Lp-PLA2 were assessed only for the CRS members and the subsequently identified CHD cases using plasma stored from visit 2.

After removing subjects with missing information in baseline variables, N=12,345 sub-

jects remained in the full cohort. We call this full cohort our phase I sample. Our phase II sample was constructed by combining CRS and all identified CHD cases. After removing the subjects with missing information in hs-CRP and Lp-PLA2, n=1336 subjects remained, including 604 CHD cases and 732 controls. Of the 1336 subjects, 777 subjects were from CRS. Our phase II dataset matches the study population analyzed by Breslow et al. [2009b] and Kang et al. [2013]. It is slightly different from the one analyzed by Ballantyne et al. [2004]. Their dataset included 608 CHD cases and 740 controls. The slight differences arise because some participants had not given proper consent or had missing information in Body Mass Index (BMI), which is needed in our later data analysis.

6.3 Analysis Methods

6.3.1 Weighting schemes

Case-cohort data can be considered within the two-phase sampling framework. Following our introduction on two-phase variable probability sampling in Section 2.1, we consider the main cohort was divided to 9 strata in this ARIC study. 8 strata were obtained by stratifying on sex, race and age in the control group and the additional, ninth stratum was the case group. Because CRS membership was independent of outcomes, sampling proportions for the 8 strata in controls were the same as 8 sampling proportions used for selecting CRS. In two-phase sampling analysis, if all the cases are selected to phase II such as a case-cohort study, weights of 1 should be assigned to the cases. However in ARIC, although a case-chort design was implemented, a fraction of cases did not have hs-CRP/Lp-PLA_2 measurements available and these cases were removed from the sample for analysis. Hence our phase II dataset, as well as the study population in Ballantyne et al. [2004]Breslow et al. [2009b] and Kang et al. [2013], did not include all the cases. Using CRS, we estimated the proportion of missingness on biomarkers among cases was 10/57= 17.5%. Thus the weight of 1/(1-17.5%)=1.21 was assigned to cases. Table 1 illustrates the sampling weights we used for analyzing ARIC. These sampling weights were provided by ARIC investigators in the dataset

prepared for our use. They are close to those used by Breslow et al. [2009b] listed in their Table 1. The difference is that Breslow et al. [2009b] used the observed sampling fraction to construct their sampling weights while we used the sampling proportions determined prior to the phase II subsampling. Both Breslow (2009) and our methods are able to handle sampling among cases, in contrast to Ballantyne et al. [2004]. It is unclear whether Kang et al. [2013] considered missing values in hs-CRP among cases.

Table 6.1: Variable probability sampling weights for ARIC

	CHD Controls								CHD Cases
	Black				White				
	Female		Male		Female		Male		
	Age <55	Age ≥ 55	Age <55	Age ≥ 55	Age <55	Age ≥ 55	Age <55	Age ≥ 55	
Stratum	1	2	3	4	5	6	7	8	9
Weights	19.5	14.3	16.1	6.1	32.0	15.1	17.7	12. 3	1.21

6.3.2 Association study

To study the association between hs-CRP and CHD incidents, the additive hazards model with two-phase sampling developed in Chapter 3 was used to investigate both traditional and new risk factors. Explanatory variables in the model included the stratification variables sex, race, and age; the traditional risk factors smoking status, diabetes, systolic blood pressure (SBP), low density lipoprotein cholesterol (LDL-C), and high density lipoprotein cholesterol (HDL-C); the new risk factor hs-CRP; and the interaction term between hs-CRP and LDL-C.

According to the cutpoints defined by American Heart Association /Center for Disease Control and Prevention (CDC) guidelines, hs-CRP was categorized to three groups (<1.0 mg/L, 1.0-3.0mg/L, > 3.0mg/L)[Ballantyne et al., 2004]. For details on the treatments of other explanatory variables in this dataset, see Ballantyne et al. [2004].

The excess hazard associated with increased hs-CRP will be reported with two standard errors (SE), one assuming that the additive hazards model is correct, with respect both to

the selected covariates and the additivity of their effect on the baseline hazard, and one that is still valid when the model is wrong. We call the second standard error robust standard error since it is robust against model misspecification. Previous studies of additive hazards model only reported the first, model based standard error (Lin and Ying [1994]; Kulich and Lin [2000]; Kang et al. [2013]), while we offer robust standard errors as well for practitioners to use.

6.3.3 Risk prediction

Because our new methods from Chapter 3 estimate the entire additive hazards (AH) model, including both cumulative baseline hazard and regression coefficients, we are able to predict CHD risk curves over a period for individuals having specific covariate values based on the fitted AH model. We will plot this risk curve (the estimated cumulative CHD hazard curves) for subjects having fixed levels of traditional risk factors but with different hs-CRP levels and LDL-C levels over a period of 3000 days (around 8 years). Since our prediction assumes the AH model holds, only regular standard errors will be used to construct 95% pointwise confidence bands. The risk curves and their 95% confidence bands together may provide additional information for physicians when the traditional risk factors cannot identify the patient at a high risk of CHD.

6.3.4 Improving precision

A second set of analyses will incorporate additional information available for the full cohort by calibration of the weights. The main idea is to adjust the weights used in the IPW-EE method by multipliers such that the estimates based on the sampled data agree with known totals for auxiliary variables in the full cohort. We expect the confidence intervals obtained for regression coefficients and risk estimates using this method will be narrower than those obtained using only the sampled data, because we now utilize more information than before.

Based on the results of simulation studies from Chapter 5, we will adopt the best performed calibration technique (method 5 from section 5.2) for analysis of ARIC. When the

study aim is to improve the estimation precision for hs-CRP coefficients, the method has four steps:

1. Find correlates of hs-CRP that are available for every main cohort member. These correlates were determined based on both literature reviews and the computation of weighted correlation coefficients using CRS.

2. Fit the AH regression model to the main cohort using hs-CRP correlates as explanatory variables using our R package "ah" or "ahaz".

3. Obtain integrated martingale residuals for these explanatory variables from an output of the software and use this output as the auxiliary variables for calibration.

4. Refit the AH-model with calibration to the sampled data by the methods developed in Chapter 4.

When the study aim is to improve risk estimates, traditional risk factors, available for every main cohort member, will be included in the AH model fitted in step 2, so that the estimation for other regression coefficients and thereby the entire model will be improved.

6.4 Results

The baseline characteristics of subjects in the case-cohort sample and the full cohort were summarized in Table 1 from Ballantyne et al. [2004] and Table 3 from Kang et al. [2013].

Reviewing previous literatures we find obesity, diabetes and BMI are associated with increased levels of hs-CRP [Ford, 1999, Visser et al., 1999, Rawson et al., 2003]. Smokers tend to have elevated levels of hs-CRP, but long term smoking cessation may reduce the elevated levels of hs-CRP [Ohsawa et al., 2005, Ridker, 2001]. Table 6.2 computed the weighted correlation coefficients between hs-CRP and other risk factors from CRS. Results in Table 6.2 confirmed the positive correlations of BMI and diabetes with hs-CRP ($r = 0.408, 0.186$). However, the correlation between smoking status and hs-CRP was small (r= -0.023). The

negative correlation resulted from comparing current and former smokers to nonsmokers. Table 6.2 also suggested hypertension history (r=0.215), triglycerides levels (r=0.212) and systolic blood pressure (SBP) (r=0.197) might be associated with hs-CRP as well. Thus we considered all these main cohort variables as potential hs-CRP correlates. By trying various combination of these correlates, we found calibration based on BMI, triglycerides levels, hypertension history and smoking status together reduced our estimation variance for hs-CRP most. The removal of SBP and diabetes was possibly due to their association with hypertension history and triglycerides levels. According to the calibration procedures we explained in the method section, auxiliary variables were obtained from fitting an AH regression model to the main cohort using hs-CRP correlates as covariates. When SBP and diabetes were correlated with other variables in this regression model, including all of them into the model would inflate our estimation variances at this step.

Table 6.2: Weighted correlation between hs-CRP and main cohort variables

Variables	Correlation Coefficients r
Hypertension history	0.215
Triglycerides	0.212
Total cholesterol	0.079
LDL-C	0.043
HDL-C	-0.064
Systolic blood pressure (SBP)	0.197
Diastolic blood pressure (DBP)	0.093
Diabetes	0.186
Body Mass Index (BMI)	0.408
Smoking (current, former, never)	-0.023

HDL-C, high density lipoprotein cholesterol (mg/L); LDL-C, low density lipoprotein cholesterol (mg/L).

Table 6.3 and 6.4 show results on hazards difference (HD) associated with hs-CRP levels. Results in subtable (a) were obtained from the sampled 1336 subjects alone using standard weights for our IPW-EE procedure, while results in subtable (b) incorporated the main cohort information using calibrated weights. Table 6.3 was comparable to Table 3 in Ballantyne et al. [2004]. Our results based on the AH model for the new risk factor hs-CRP gave the same conclusion as theirs based on the Cox's proportional hazards model. Both tables find average and high hs-CRP were associated with a significant increase in CHD risks when adjusting for age, sex and race. Both tables show these association were attenuated after further adjustment for standard risk factors, but risk remained significantly elevated for high hs-CRP group. The difference of the two tables lies in the interpretation. We explain the change of CHD risk by hazards difference compared to theirs using hazards ratio. We see the Cox's proportional hazards model and the additive hazards offer different perspectives on an association. They can be used together for a more complete understanding on risk factors.

Table 6.4 (a) shows with elevated LDL-C, the CHD hazards in high hs-CRP group ($>$ 3.0 mg/L) was 1.756×10^{-5} (95 % CI: $0.649 \times 10^{-5}, 2.859 \times 10^{-5}$) per person-day higher than low hs-CRP group (< 1.0 mg/L). Without elevated LDL-C, this difference decreased to 0.798×10^{-5} (95 % CI: $-0.007 \times 10^{-5}, 1.603 \times 10^{-5}$) per person-day and this difference was marginal (p-value=0.053). Our results agreed with the findings reported in Kang et al. [2013]. Their hazard differences associated with high hs-CRP group with and without elevated LDL-C were 1.770×10^{-5} and 0.566×10^{-5} respectively. Without elevated LDL-C, their estimation on the difference between high and low hs-CRP group was also concluded marginal (p-value =0.053). Note Kang et al. [2013] fitted a multivariate failure time additive hazard model to study both stroke and CHD events, and it was unclear whether they considered about 17% missing values in hs-CRPs among cases. Thus our results and theirs would not match exactly.

Table 6.4 (a) also shows with current sample size of 1336, we were not able to detect the difference between the effects of middle (1.0-3.0mg/L) and low (< 1.0 mg/L) hs-CRP on

CHD incidents regardless of an individual's LDL-C levels. When LDL-C level < 130 mg/dL, the hazard difference was close to 0 (p-value=0.992).

Comparing Table 6.3 (b) to Table 6.3 (a) and Table 6.4 (b) to Table 6.4 (a), we observe SE and confidence intervals for coefficient estimates were all reduced, suggesting the ability of calibration to improve estimation. However these shrinkages in SE and CIs were small. This may be caused by the weak association between hs-CRP correlates we discovered in the main cohort and hs-CRP.It is worth noting the p-value for examining the association between high hs-CRP reduced from 0.052 to 0.046 when LDL-C < 130 mg/dL (Table 6.4). Our conclusion on the effect of hs-CRP were changed without increasing the sample size due to calibration. This example shows the potential of using additional information from the main cohort for a two-phase design study. The benefit would be more evident if moderately or highly correlated auxiliary variables were present for hs-CRP.

Comparing Table 6.4 (b) to Table 6.4 (a) we also find all the coefficients estimates decreased after calibration. Our simulation results in Chapter 5 suggest two-phase sampling tends to introduce small positive biases to coefficients estimates, and calibration can reduce these biases. Hence it is likely our calibrated estimates in Table 6.4 (b) are closer to the true hazards differences between hs-CRP groups than Table 6.4. More simulation studies and real data investigation will be needed to study the small biases introduced by two-phase sampling and the role of calibration in correcting them.

Although we did not have highly correlated surrogates for hs-CRP in this dataset, traditional risk factors and demographic variables were all measured for every main cohort member. These additional information in the main cohort were used to improve our inference precision for the whole model. Table 6.5 shows the estimates and their standard errors for all the regression coefficients when fitting an AH model to the ARIC data. Three groups of SE were reported. SE I were estimated using standard weights. SE II and SE III were obtained using calibrated weights. For SE II, we used traditional risk factors and demographic variables when constructing auxiliary variables at our calibration step 2. For SE III, BMI, triglycerides levels, and hypertension history that might improve hs-CRP inference

Table 6.3: Excess hazards associated with hs-CRP adjusted for demographic variables and standard risk factors

(a) Before Calibration, Hazards Difference (95% CI) ($\times 10^{-5}$)

hs-CRP Categories	Average risk (1.0-3.0 mg/L)	High risk (>3.0 mg/L)
Model 1†	0.84 (0.32,1.36)	1.92 (1.27,2.57)
Model 2‡	0.38 (-0.20,0.96)	1.26 (0.56,1.94)

(b) After Calibration, Hazards Difference (95% CI) ($\times 10^{-5}$)

hs-CRP Categories	Average risk (1.0-3.0 mg/L)	High risk (>3.0 mg/L)
Model 1	0.79 (0.28, 1.30)	1.86 (1.22, 2.49)
Model 2	0.36 (-0.20, 0.93)	1.23 (0.57, 1.90)

†Adjsuted for age, sex, and race.

‡Adjsuted for age, sex, race,smoking status, systolic blood pressure, LDL-C, HDL-C and diabetes.

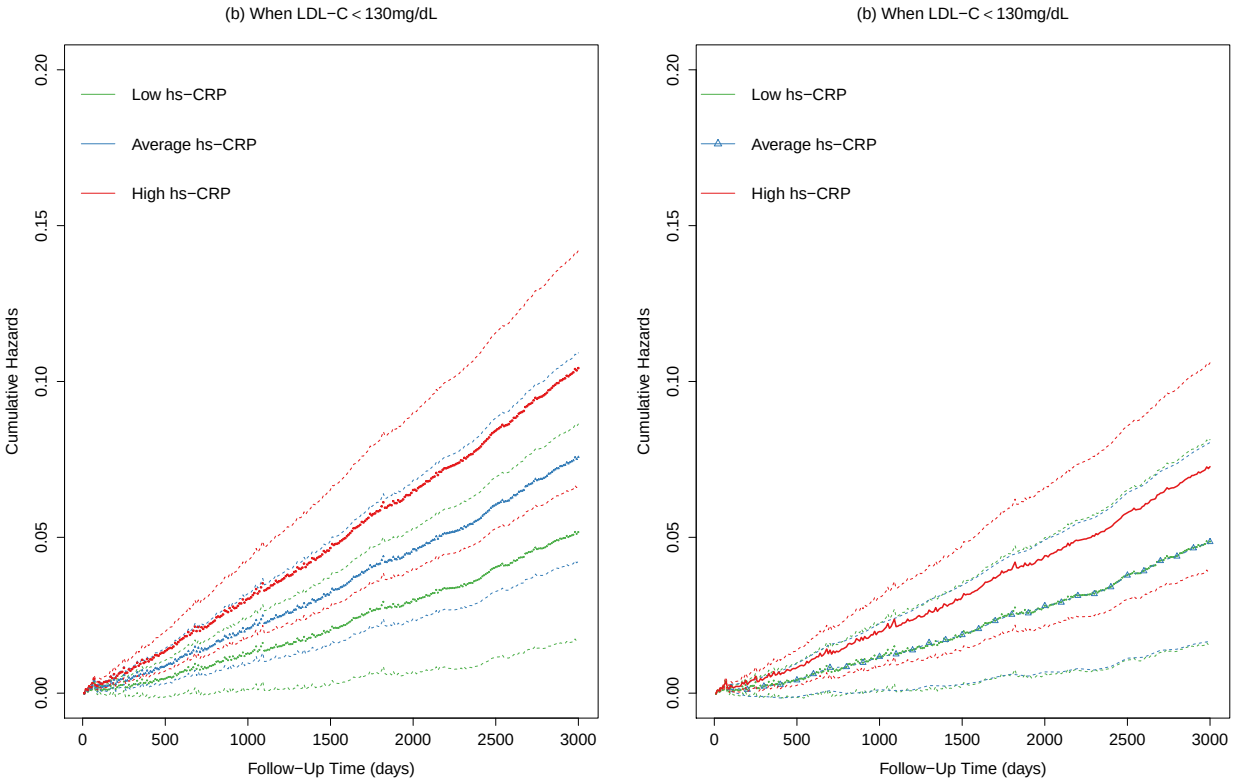
were also included in the calibration procedure in addition to the factors considered for SE II. Comparing SE II to SE I from Table 6.5 we find the SE of coefficient estimates for most variables were reduced by more than 1/3, though the SE for hs-CRP increased. This result suggests model fitting in our calibration step 2 introduced extra variability. Without additional information on a risk factor from the main cohort, variances for coefficient estimation will be inflated by calibration. Comparing SE III to SE II in Table 6.5 confirmed this finding. Adding correlates of hs-CRP to the calibration procedure brought down the inflated SE but these hs-CRP correlates were not informative enough to counteract the inflation induced. Similar observations were found in Table 2 from Breslow et al. [2009b] when fitting the Cox's regression models to the same dataset using different weight adjustment techniques.

Figure 6.1 and 6.2 present results on individual- specific risk prediction based on the risk function from estimating the entire AH model using the two-phase ARIC data. Figure 6.1 plotted a series of cumulative hazard prediction curve for an individual when his/her

hs-CRP and LDL-C were at different levels. Predictions were performed for every 10 days over a period of 8 years, since the follow-up time in the ARIC dataset was about a little less than 9 years. The individual we chose to predict the CHD risk was a 60 year old white male smoker without diabetes. His SBP=120 mm/Hg and HDL-C=45mg/dL. When this individual's LDL-C =145 mg/DL as shown in Figure 6.1 (a), three hs-CRP risk curves were well separated. This separation became larger as time increased. The differences of the risk curves between adjacent hs-CRP groups were almost identical across the time. This result suggests three levels of hs-CRP can be treated continuously instead of categorically. When LDL =120 mg/DL as shown in Figure 6.1 (b), the risk curve for the low hs-CRP group was unchanged compared to elevated LDL-C levels in Figure 6.1 (a), but both high and average hs-CRP risk curves were lowered. As a result, we could not distinguish the risk curves for low and average hs-CRP groups without elevated LDL-C.

Figure 6.2 shows two sets of prediction results based on standard and calibrated weights for six individuals of different ages, HDL-C levels and smoking status. The risk curves and their 95% pointwise confidence band based on the AH model provides a new visual tool to examine traditional risk factors. Comparing the right panel to the left panel in Figure 6.2, it is evident that age and smoking increased the CHD risk while HDL-C was negatively associated with CHD risks regardless of using standard or calibrated weights. Comparing results between calibrated and standard weights we observe calibration narrowed our 95% pointwise confidence band for five out of six individuals' cumulative hazards prediction as shown in Figure 6.2 (a),(b),(c), (d) and (f). However, no precision improvement was seen for individual (e). We conclude calibration in general improves risk estimates' precision, but the power of this improvement depends on individuals' covariate values. Future research on the influence of covariates values and auxiliary variable choices on improving risk prediction will be important and helpful.

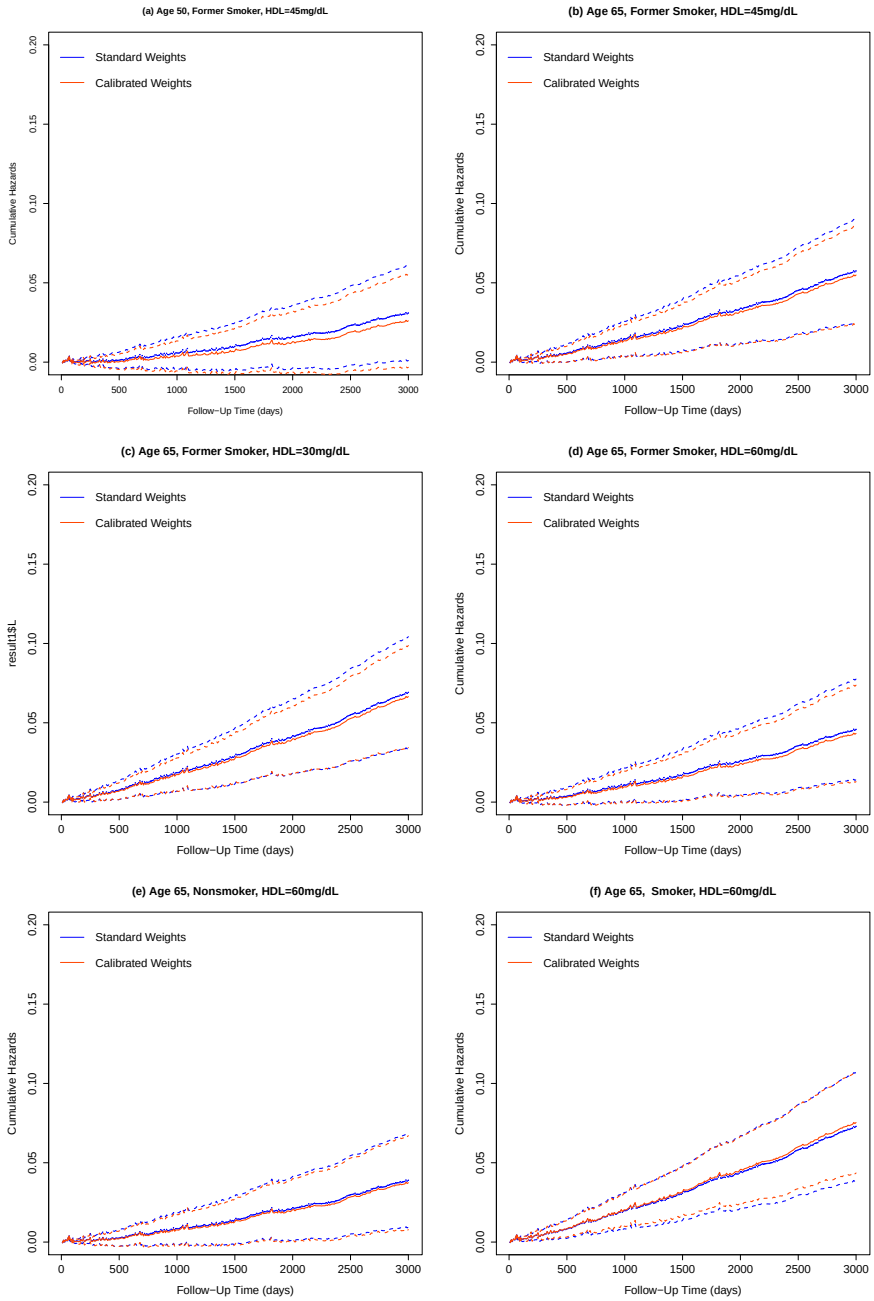
Figure 6.1: Individual cumulative hazards prediction by hs-CRP levels



Individual profile: age=60, male, white, current smoker, no diabetes, SBP=120 mm/Hg, HDL-C=45 mg/DL. 95% pointwise CI band were estimated based using standard weights.

Low hs-CRP ($<$ 1 mg/L); Average hs-CRP (1-3 mg/L); High hs-CRP ($>$ 3 mg/L).

Figure 6.2: Individual cumulative hazards prediction with standard and calibrated weights



Shared covariates among individual (a) to (f): male, white, no diabetes, SBP=120 mm/Hg, LDL-C=145 mg/L, hs-CRP (1-3 mg/L).

Variables used for calibration: age, sex, race, diabetes, smoking, SBP, LDL-C, HDL-C, BMI, triglycerides, and hypertension history.

Table 6.4: CHD hazards differences (95% CI) by hs-CRP with and without elevated LDL-C

(a) Before Calibration

Categories	Estimates ($\times 10^{-5}$)	SE ($\times 10^{-5}$)	Robust SE ($\times 10^{-5}$)	95% CI † ($\times 10^{-5}$)	p-value†
When LDL-C ≥130 mg/dL					
hs-CRP (1.0-3.0 mg/L) *	0.803	0.468	0.467	(-0.114, 1.721)	0.086
hs-CRP (>3.0 mg/L)	1.754	0.564	0.563	(0.649, 2.859)	0.0018
When LDL-C <130 mg/dL					
hs-CRP (1.0-3.0 mg/L)	-0.004	0.344	0.344	(-0.678, 0.671)	0.992
hs-CRP (>3.0 mg/L)	0.798	0.411	0.411	(-0.007, 1.603)	0.052

(b) After Calibration‡

Categories	Estimates ($\times 10^{-5}$)	SE ($\times 10^{-5}$)	Robust SE ($\times 10^{-5}$)	95% CI † ($\times 10^{-5}$)	p-value†
When LDL-C ≥130 mg/dL					
hs-CRP (1.0-3.0 mg/L)	0.759	0.456	0.456	(-0.135, 1.653)	0.096
hs-CRP (>3.0 mg/L)	1.705	0.544	0.544	(0.638, 2.773)	0.0017
When LDL-C <130 mg/dL					
hs-CRP (1.0-3.0 mg/L)	-0.047	0.334	0.335	(-0.701, 0.607)	0.888
hs-CRP (>3.0 mg/L)	0.795	0.399	0.399	(0.013, 1.577)	0.046

The model is adjusted for sex, race, age, smoking status, diabetes, systolic blood pressure, LDL-C, and HDL-C.

To assess the effects of hs-CRP for individuals with different LDL-C levels, interaction term between hs-CRP and a dichotomized LDL-C was added to the model.

*hs-CRP <1.0 mg/L is the reference group

†95% CI and p-value are calculated based on model-based variances

‡ Calibration is based on BMI, triglycerides levels, hypertension history and smoking

Table 6.5: Standard errors of CHD hazards differences from different calibration choices

Model Terms	Estimates ($\times 10^{-5}$)	SE I * ($\times 10^{-5}$)	SE II † ($\times 10^{-5}$)	SE III ‡ ($\times 10^{-5}$)
hs-CRP 1.0-	0.803	0.468	0.499	0.498
hs-CRP 3.0-	1.754	0.564	0.591	0.581
LDL-Low♯	0.250	0.552	0.577	0.575
Age in years	0.059	0.026	0.018	0.018
Male sex	1.804	0.321	0.217	0.218
White Race	0.670	0.323	0.226	0.226
Former smokers	-0.904	0.435	0.280	0.281
Nonsmokers	-1.130	0.406	0.268	0.269
SBP	0.033	0.009	0.006	0.006
LDL-C	0.014	0.007	0.006	0.006
HDL-C	-0.026	0.008	0.005	0.005
Diabetes	1.630	0.534	0.366	0.367
hs-CRP 1.0- × LDL-Low	-0.808	0.570	0.604	0.603
hs-CRP 3.0- × LDL-Low	-0.960	0.685	0.724	0.720

*No calibration.

†Calibration based on sex, race, age, smoking status, diabetes, systolic blood pressure, LDL-C, and HDL-C.

‡Adding BMI, triglycerides levels, hypertension history to the calibration procedures for SE†.

♯LDL-Low: LDL-C < 130 mg/L.

6.5 Conclusion

In conclusion, this dissertation makes four contributions: 1) I provided new solutions to several practical problems that arise in two-phase sampling. 2) I developed a computer software for fitting Lin and Ying's additive hazards model to two-phase sampling data that can be immediately implemented for association studies or identifying risk susceptible patients for preventative health care. 3) I connected abstract but powerful modern empirical process theory to a real-life problem and showed the importance of this theory in applications on "function"-value data [Wellner, 1992]. I demonstrated how to simplify the application of this theory with some restrictions. This approach will be helpful for readers who are interested in but have just started to use this theory. 4) I built a transparent Z-estimation system for developing a series of methods for various two-phase sampling problems.

At a practice level, my dissertation will perhaps popularize the use of the two-phase sampling design in epidemiology. Most statistical tools improve efficiency after obtaining data through sophisticated techniques, but the study design we made tools for considers efficiency at the data-collecting step. If costs can be dramatically reduced through a carefully designed sampling scheme, then investigators may afford and even welcome a simple and transparent statistical tool with a minor loss of efficiency in analyzing the data. In consequence, with two-phase sampling and their new tools, the quality of medical research can be improved and the costs will be reduced.

BIBLIOGRAPHY

O. Aalen. A model for nonparametric regression analysis of counting processes. In *Mathematical statistics and probability theory*, pages 1–25. Springer, 1980.

P.K. Andersen and R.D. Gill. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, 10(4):1100–1120, 1982.

C.M. Ballantyne, R.C. Hoogeveen, H. Bang, J. Coresh, A.R. Folsom, G. Heiss, and A.R. Sharrett. Lipoprotein-associated phospholipase A2, high-sensitivity C-reactive protein, and risk for incident coronary heart disease in middle-aged men and women in the atherosclerosis risk in communities (ARIC) study. *Circulation*, 109(7):837–842, 2004.

W. E. Barlow, L. Ichikawa, D. Rosner, and S. Izumi. Analysis of case-cohort designs. *Journal of clinical epidemiology*, 52(12):1165–1172, 1999.

W.E. Barlow. Robust variance estimation for the case-cohort design. *Biometrics*, 50(4): 1064–1072, 1994.

P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and adaptive estimation for semiparametric models*. Springer-Verlag, New York, 1998. Reprint of the 1993 original.

D.A. Binder. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review/Revue Internationale de Statistique*, 51(3):279–292, 1983.

D.A. Binder. Fitting Cox's proportional hazards models from survey data. *Biometrika*, 79 (1):139–147, 1992.

O. Borgan, B. Langholz, S.O. Samuelsen, L. Goldstein, and J. Pogoda. Exposure stratified case-cohort designs. *Lifetime data analysis*, 6(1):39–58, 2000.

N. E. Breslow and N. E. Day. The design and analysis of cohort studies. *IARC Scientific Publications No. 82*, page 446, 1987. International Agency for Research on Cancer, Lyon.

N. E. Breslow and J. A. Wellner. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics*, 34(1):86–102, 2007.

N.E. Breslow and T. Lumley. Semiparametric models and two-phase samples: applications to Cox regression. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pages 65–77. Institute of Mathematical Statistics, 2013.

N.E. Breslow and J.A. Wellner. A Z-theorem with estimated nuisance parameters and correction note for: "Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression". *Scandinavian Journal of Statistics*, 35(1): 186–192, 2008.

N.E. Breslow, B. McNeney, and J.A. Wellner. Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *Annals of Statistics*, 31.

N.E. Breslow, T. Lumley, C.M. Ballantyne, L.E. Chambless, and M. Kulich. Improved horvitz–thompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology. *Statistics in Biosciences*, 1(1):32–49, 2009a.

N.E. Breslow, T. Lumley, C.M. Ballantyne, L.E. Chambless, and M. Kulich. Using the whole cohort in the analysis of case-cohort data. *American journal of epidemiology*, 169 (11):1398–1405, 2009b.

J. Cai and D. Zeng. Sample size/power calculation for case–cohort studies. *Biometrics*, 60 (4):1015–1024, 2004.

D.R. Cox and D. Oakes. *Analysis of survival data*. CRC Press, 1984.

G.J. D'angio, N.E. Breslow, J.B. Beckwith, A. Evans, E. Baum, A. Delorimier, D. Fernbach, E. Hrabovsky, B. Jones, P. Kelalis, et al. Treatment of Wilms' tumor. Results of the third National Wilms' Tumor Study. *Cancer*, 64(2):349–360, 1989.

J.C. Deville and C.E. Särndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382, 1992.

R. M. Dudley. Central limit theorems for empirical measures. *The Annals of Probability*, pages 899–929, 1978.

R. M. Dudley. *Real analysis and probability*, volume 74. Cambridge University Press, 2002.

T. Fleming and D. Harrington. *Counting Processes and Survival Analysis*. Wiley, 1991.

E.S. Ford. Body mass index, diabetes, and C-reactive protein among us adults. *Diabetes care*, 22(12):1971–1977, 1999.

V.P. Godambe and M. E. Thompson. Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review/Revue Internationale de Statistique*, 54(2):127–138, 1986.

A. Gorst-Rasmussen and T.H. Scheike. Coordinate descent methods for the penalized semi-parametric additive hazards model. *Journal of Statistical Software*, 47(9):1–17, 2012.

R.J. Gray. Weighted analyses for cohort sampling designs. *Lifetime data analysis*, 15(1): 24–40, 2009.

D.M. Green, N.E. Breslow, J.B. Beckwith, J.Z. Finklestein, P.E. Grundy, P.R. Thomas, T. Kim, S.J. Shochat, G.M. Haase, M.L. Ritchey, et al. Comparison between single-dose and divided-dose administration of dactinomycin and doxorubicin for patients with Wilms' tumor: a report from the National Wilms' Tumor Study group. *Journal of clinical oncology*, 16(1):237–245, 1998.

D.G. Horvitz and D.J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

P.J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

P.J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233, 1967.

P.J. Huber. Robust statistics. *Wiley Series in Probability and Mathematical Statistics, New York: Wiley*, 1981.

F.W. Huffer and I.W. McKeague. Weighted least squares estimation for Aalen's additive risk model. *Journal of the American Statistical Association*, 86(413):114–129, 1991.

Aric Investigators et al. The Atherosclerosis Risk In Communities (ARIC) study: design and objectives. *American journal of epidemiology*, 129(4):687–702, 1989.

J.D. Kalbfleisch and J.F. Lawless. Likelihood analysis of multi-state models for disease incidence and mortality. *Statistics in Medicine*, 7(1-2):149–160, 1988.

S. Kang and J. Cai. Marginal hazards model for case-cohort studies with multiple disease outcomes. *Biometrika*, 96(4):887–901, 2009.

S. Kang, J. Cai, and L. Chambless. Marginal additive hazards model for case-cohort studies with multiple disease outcomes: an application to the Atherosclerosis Risk In Communities (ARIC) study. *Biostatistics*, 14(1):28–41, 2013.

J. P. Klein. Modelling competing risks in cancer studies. *Statistics in Medicine*, 25(6): 1015–1034, 2006.

M. Kulich. *Additive hazards regression with incomplete covariate data*. PhD thesis, University of Washignton, 1997.

M. Kulich and D.Y. Lin. Additive hazards regression for case-cohort studies. *Biometrika*, 87(1):73–87, 2000.

M. Kulich and D.Y. Lin. Improving the efficiency of relative-risk estimation in case-cohort studies. *Journal of the American Statistical Association*, 99(467):832–844, 2004.

J.F. Lawless, J.D. Kalbfleisch, and C.J. Wild. Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):413–438, 1999.

H.J. Lim and X. Zhang. Semiparametric additive risk models: application to injury duration study. *Accident Analysis & Prevention*, 41(2):211–216, 2009.

D.Y. Lin. On fitting Cox's proportional hazards models to survey data. *Biometrika*, 87(1): 37–47, 2000.

D.Y. Lin and L.J. Wei. The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, 84(408):1074–1078, 1989.

D.Y. Lin and Z. Ying. Semiparametric analysis of the additive risk model. *Biometrika*, 81 (1):61–71, 1994.

T. Lumley. *Complex surveys: A guide to analysis using R*, volume 565. John Wiley & Sons, 2011.

C. F. Manski and S. R. Lerman. The estimation of choice probabilities from choice based samples. *Econometrica*, 45(8):1977–1988, 1977.

J. E. Marsden and M. J. Hoffman. *Basic complex analysis*. Macmillan, 1999.

T. Martinussen and T.H. Scheike. *Dynamic regression models for survival data*, volume 1. Springer, New York, 2006.

I.W. McKeague and P.D. Sasieni. A partly parametric additive risk model. *Biometrika*, 81 (3):501–514, 1994.

S.A. Murphy and A.W. Van der Vaart. Semiparametric mixtures in case-control studies. *Journal of Multivariate Analysis*, 79(1):1–32, 2001.

B. Nan and J.A. Wellner. A general semiparametric Z-estimation approach for case-cohort studies. *Statistica Sinica*, 23(3):1155–1180, 2013.

W.K. Newey. The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, 62(6):1349–1382, 1994.

J. Neyman. Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33(201):101–116, 1938.

M. Ohsawa, A. Okayama, M. Nakamura, T. Onoda, K. Kato, K. Itai, Y. Yoshida, A. Ogawa, K. Kawamura, and K. Hiramori. CRP levels are elevated in smokers but unrelated to the number of cigarettes and are decreased by long-term smoking cessation in male smokers. *Preventive Medicine*, 41(2):651–656, 2005.

T.A. Pearson, G.A. Mensah, R.W. Alexander, J.L. Anderson, R.O. Cannon, M. Criqui, Y.Y. Fadl, S.P. Fortmann, Y. Hong, G.L. Myers, et al. Markers of inflammation and cardiovascular disease application to clinical and public health practice: a statement for healthcare professionals from the centers for disease control and prevention and the American Heart Association. *Circulation*, 107(3):499–511, 2003.

D. Pollard. *Convergence of stochastic processes*. Springer Series in Statistics. Springer-Verlag, New York, 1984.

D. Pollard. New ways to prove central limit theorems. *Econometric Theory*, 1(3):295–313, 1985.

D. Pollard. Asymptotics via empirical processes. *Statistical Science*, 4(4):341–354, 1989.

J. Præstgaard and J.A. Wellner. Exchangeably weighted bootstraps of the general empirical process. *The Annals of Probability*, 21(4):2053–2086, 1993.

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

E.S. Rawson, P.S. Freedson, S.K. Osganian, C.E. Matthews, G. Reed, and I.S. Ockene. Body mass index, but not physical activity, is associated with C-reactive protein. *Medicine and Science in Sports and Exercise*, 35(7):1160–1166, 2003.

P. M. Ridker. High-sensitivity C-reactive protein potential adjunct for global risk assessment in the primary prevention of cardiovascular disease. *Circulation*, 103(13):1813–1818, 2001.

J.M. Robins, A. Rotnitzky, and L.P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

R.M. Royall. Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review/Revue Internationale de Statistique*, 54(2):221–226, 1986.

D.B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

T. Saegusa and J.A. Wellner. Weighted likelihood estimation under two-phase sampling. *Annals of statistics*, 41(1):269, 2013.

C.E. Särndal. The calibration approach in survey theory and practice. *Survey Methodology*, 33(2):99–119, 2007.

S.G. Self and R.L. Prentice. Asymptotic distribution theory and efficiency results for case-cohort studies. *The Annals of Statistics*, 16(1):64–81, 1988.

D.C. Thomas. Use of auxiliary information in fitting nonproportional hazards models. *Modern Statistical Methods in Chronic Disease Epidemiology*, pages 197–210, 1986.

A.W. van der Vaart. Efficiency of infinite dimensional M-estimators. *Statistica Neerlandica*, 49(1):9–30, 1995.

A.W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

A.W. van der Vaart and J.A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

A.W. van der Vaart and J.A. Wellner. Preservation theorems for Glivenko-Cantelli and uniform Glivenko-Cantelli classes. In *High dimensional probability, II (Seattle, WA, 1999)*, volume 47 of *Progr. Probab.*, pages 115–133. Birkhäuser, Boston, MA, 2000.

M. Visser, L.M. Bouter, G.M. McQuillan, M.H. Wener, and T.B. Harris. Elevated C-reactive protein levels in overweight and obese adults. *The Journal of the American Medical Association*, 282(22):2131–2135, 1999.

A.M. Walker. Anamorphic analysis: sampling and estimation for covariate effects when both exposure and disease are known. *Biometrics*, 38(4):1025–1032, 1982.

J.A. Wellner. Empirical processes in action: a review. *International Statistical Review/Revue Internationale de Statistique*, 60(3):247–269, 1992.

J.E. White. A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, 115(1):119–128, 1982.

X. Xie, H.D. Strickler, and X. Xue. Additive hazard regression models: an application to the natural history of human papillomavirus. *Computational and Mathematical Methods in Medicine*, 2013.

VITA

Jie Hu was born to Qian Zhou and Wei Hu on July 4th, 1984, in Beijing, China. She enrolled in Peking University in 2002 and was recruited the following year with a full scholarship by the University of Hong Kong, where she earned a Bachelor of Science in Biochemistry in 2006. During her senior year, she attended the University of California at Los Angeles as an exchange student and decided to study Biostatistics in the United States. In 2008, she received a Master of Science in Biostatistics from Harvard University in Cambridge, Massachusetts. Since then, she has been a graduate student at the University of Washington in Seattle, working toward a PhD degree in Biostatistics and enjoying the great outdoors.