Conditional Empirical Likelihood Approach to Statistical Analysis with Missing Data

by

Peisong Han

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2013

Doctoral Committee:

Professor Peter X.-K. Song, Co-Chair
Assistant Professor Lu Wang, Co-Chair
Professor Xuming He
Emeritus Professor Jack D. Kalbfleisch

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

Introduction

1.1 Overview

Data collected from statistical studies often have missing values. This is especially true for biostatistical research where human study subjects are involved. Some subjects may simply refuse to respond to certain questions, despite the researchers' intention of collecting complete data. Very often, budget or technique limitations also restrict researchers to design studies that collect complete data only from a sub-group of subjects. Two-stage study design (e.g. Pepe 1992; Pepe et al. 1994) is a commonly seen example. In such design, some surrogates of the variables of interest are measured at the first stage, based on which a sub-group of subjects is selected to enter the second stage, where the measurements of the variables of interest are recorded. For longitudinal studies where subjects are followed over a certain time period, an important reason for missing data is that subjects do not comply with the protocol, due to, for example, schedule conflicts. This reality leads to incomplete data in irregular patterns. Dropout is another major reason for missing data in longitudinal studies, which occurs when subjects move out of area, or are lost of contact, or experience severe side effect of the treatment that prevents them from staying in the study. Dropout results in monotonic missing data pattern. Missing data usually

bring great challenges to statistical analysis. In general, a direct application of the existing methods for fully observed data to analyzing data with missing values leads to biased estimation and misleading conclusions.

During the past three decades or so, after Rubin (1976) defined the missing data problem from a formal statistical point of view, numerous methods for statistical analysis with missing data have been proposed. Roughly, the majority of the existing methods can be categorized into two classes: the likelihood-based methods and the semiparametric methods. Likelihood-based methods essentially specify a joint distribution of the data, and corresponding estimation and inference are conducted based on the maximum likelihood theory. The problem of missing data can be dealt with in many different ways, including the celebrated EM algorithm (Dempster et al. 1977; Wu 1983; Meng and Rubin 1993; McLachlan and Krishnan 2008) and the seminal multiple imputation method (Rubin 1978, 1987, 1996). Both frequentist and Bayesian approaches can be taken by the likelihood-based methods to conduct statistical analysis. Recently, the calibrated Bayesian approach that combines both frequentist and Bayesian approaches has attracted some attention (Box 1980; Rubin 1984; Little 2006, 2011). A comprehensive coverage of the likelihood-based methods for missing data problems can be found in Little and Rubin (2002), Little (2008) and references therein.

Semiparametric methods, on the contrary, consider models specified by a set of estimating equations (or moment conditions) that involve both the data and the unknown parameters of interest. The relevant theory of estimating functions can be found, for example, in Godambe (1991), Newey and McFadden (1994) and Heyde (1997). Compared to the likelihood-based methods, semiparametric methods are more flexible, as they only require to model certain characteristics of the joint distri-

bution rather than the distribution itself. As a result, such methods are more robust against model misspecifications. For many data types, including the longitudinal data with discrete outcome variables, it may be very difficult to explicitly specify the joint distribution. In such cases, semiparametric methods have become useful alternatives to the likelihood-based methods. The methods developed in this dissertation are semiparametric. Semiparametric methods mainly rely on the large sample theory to establish the properties of estimation and inference procedures. Some references on large sample theory include Newey and McFadden (1994), Lehmann (1998), van der Vaart (1998) and Shao (2003).

A prominent semiparametric method dealing with missing data is the augmented inverse probability weighting (AIPW) method proposed by Robins, Rotnitzky and colleagues in a series of seminal papers, including Robins et al. (1994, 1995), Robins and Rotnitzky (1995), Rotnitzky et al. (1998), Scharfstein et al. (1999), Bang and Robins (2005) and Rotnitzky et al. (2012). The AIPW estimator has attracted much research interest because of its "double robustness" property, which provides double protection on estimation consistency against model misspecification. Tsiatis (2006) gives a detailed coverage of the AIPW method and its application to longitudinal data analysis. An important issue associated with the AIPW method is the estimation efficiency. In the context of regression analysis, to achieve full efficiency, the AIPW method requires to correctly model certain second order moments of the data. This becomes especially demanding for longitudinal data analysis. Even in the simplest case where the longitudinal outcomes are missing completely at random (Little and Rubin 2002), the variance-covariance of the longitudinal outcomes needs to be reasonably modeled in order to achieve satisfactory efficiency for the estimation of regression coefficients (Liang and Zeger 1986). The central focus of this disserta-

tion is to develop new methodologies that are free from modeling the second order moments, yet achieve full estimation efficiency.

The proposed methodology has its root in the empirical likelihood (EL) methodology. First introduced by Owen (1988, 1990), EL is a nonparametric likelihood that does not require the specification of a parametric distribution, yet possesses desirable properties of parametric likelihood. For example, Wilks' theorem still holds for the EL ratio statistic, and Bartlett correction can be conducted as in the case of parametric likelihood to achieve higher order precision (DiCiccio et al. 1991; Chen and Cui, 2007). One groundbreaking work in the literature of EL methodology is Qin and Lawless (1994), who studied how to use EL in estimation and inference for models defined by estimating equations (unconditional moment restrictions). Their EL estimator attains the semiparametric efficiency bound for the corresponding model, in the sense of Bickel et al. (1993). Kolaczyk (1994) and Chen and Cui (2003) investigated the application of EL to generalized linear models (McCullagh and Nelder 1990). Discussions on many other desirable properties of EL, such as invariance under transformation of moment conditions and automatic determination of the shape of confidence regions, can be found in Hall (1990), Hall and LaScala (1990) and Owen (2001). It is worth pointing out that, the EL methodology can be embedded into a much larger family of estimation procedures that corresponds to the Cressie-Read power divergence family of discrepancies (Cressie and Read 1984). Related details regarding this aspect can be found in Newey and Smith (2004), which established the optimality of the EL estimator in terms of higher order efficiency compared to estimators derived from other members of the Cressie-Read family, as well as to the generalized method of moments (GMM) estimator (Hansen 1982). Bayesian approach can also be adopted within the EL framework. For related discussions, refer

to Lazar (2003), Schennach (2005) and Yang and He (2012).

The proposed methodology in this dissertation is most closely connected to the work of Zhang and Gijbels (2003) and Kitamura et al. (2004). These authors independently made the interesting extension of the EL methodology to the conditional EL (CEL) methodology, which provides an estimation procedure under models defined by conditional moment restrictions. Such models include the parametric regression model, which is the model of main interest in this dissertation, as a special case. The CEL methodology was named as "sieve empirical likelihood" by Zhang and Gijbels (2003) and "smoothed empirical likelihood" by Kitamura et al. (2004). We adopt the name "conditional empirical likelihood" following Kitamura (2007). It has been shown that the CEL estimator attains the semiparametric efficiency bound for the model defined by the conditional moment restrictions in the sense of Bickel et al. (1993). See also Chamberlain (1987). Based on the CEL methodology, Tripathi and Kitamura (2003) proposed a test for the validity of conditional moment restrictions. Smith (2007) generalized the CEL idea to information theoretic criteria based on the Cressie-Read power divergence family of discrepancies. Otsu (2007, 2011) studied the CEL inference when unknown functions are present in the conditional moment restrictions.

EL methodology has already been applied to a variety of statistical fields, including missing data problems, longitudinal data analysis and survival data analysis. For a comprehensive coverage on survival data analysis, see Kalbfleisch and Prentice (2002). Despite the great popularity and success, there remain important issues to be addressed. This dissertation focuses on some of the issues arising in the fields of missing data and longitudinal data analysis. These issues, together with related literature review, will be given and discussed separately in the first section of each

chapter. By doing so, the presentation of this dissertation can be tied cohesively under a focused research aim. Some applications of the EL methodology in survival data analysis can be found in Li et al. (2005). To learn more about the EL methodology and its applications in other statistical fields, refer to Owen (2001), Kitamura (2007), Chen and Van Keilegom (2009) and references therein.

1.2 Organization of the Dissertation

The rest of the dissertation consists of three chapters. Chapter II considers regression analysis with cross-sectional data. We propose a CEL method for estimation and inference within the framework of parametric regression when the outcome is subject to missingness while some surrogate variables are available. Unlike the existing estimating functions based estimators, whose efficiency depends on the specific forms of the estimating functions, our proposed estimator can achieve the semiparametric efficiency bound with no explicit specification of any estimating functions. Assuming the missing at random (MAR) mechanism (Little and Rubin 2002), we study the CEL-based inverse probability weighted (CEL-IPW) and CEL-based augmented inverse probability weighted (CEL-AIPW) estimators in detail. Under some regularity conditions, the CEL-IPW estimator is consistent if the missingness mechanism is correctly modeled, whereas the CEL-AIPW estimator is doubly robust, in the sense that it is consistent if either the missingness mechanism or the conditional mean of the outcome given surrogate variables and covariates is correctly modeled. When both are correctly modeled, the CEL-AIPW estimator attains the semiparametric efficiency bound. Numerical implementation through nested optimization routines using Newton-Raphson algorithm is discussed. Asymptotic distributions are derived. Finite sample performance with comparisons to some existing estimators is demon-

strated through simulation experiments. As an application example, data collected from an intervention study for adolescents of parents with HIV are analyzed.

Chapter III provides an extension of Chapter II by considering regression analysis with longitudinal data in the presence of dropout. Under the MAR mechanism, we investigate the CEL-AIPW estimator analytically and numerically. This estimator is doubly robust, in the sense that it is consistent if either the missingness probabilities or the conditional expectations of the longitudinal outcomes given the observed data at each level of missingness are correctly modeled. When both quantities are correctly modeled, the CEL-AIPW estimator achieves the semiparametric efficiency bound. Therefore, the proposed CEL-AIPW estimator does not require to model any second moments of the data in order to achieve full efficiency. This is not the case for the existing AIPW estimator (Robins and Rotnitzky 1995; Tsiatis 2006), whose efficiency depends on modeling certain second moments. We derive the asymptotic distributions. We also discuss issues related to the numerical implementation, and run simulation studies to assess the finite sample performance of the proposed estimator.

Chapter IV concerns the development of a CEL method for unbalanced longitudinal data analysis. The unbalanced follow-up visits are dealt with via stratification according to distinct follow-up patterns. Such a way of dealing with the unbalancedness implicitly assumes the missing completely at random mechanism, the same mechanism assumed by the popular generalized estimating equations (GEE) method (Liang and Zeger 1986). Compared to the GEE method, our proposed CEL method does not require any explicit modeling of the variance-covariances of the longitudinal outcomes, but only requires a marginal mean regression model. Therefore, our method is robust against misspecification of the second moment structures. We in-

vestigate possible connections between the proposed CEL estimator and the GEE estimator. We show that the CEL estimator achieves the same efficiency as that of the GEE estimator obtained employing the true variance-covariance. Asymptotic distribution of the CEL estimator is derived, and various issues regarding the numerical implementation and applications are discussed. Simulation studies are conducted to assess the finite sample performance. To illustrate the CEL method, we analyze data collected from a longitudinal nutrition study.

Chapter V talks about some possible extensions of the current developments in this dissertation, and gives several future research directions.

CHAPTER II

Locally Efficient and Doubly Robust Estimation with Missing Outcome: Cross-Sectional Data

2.1 Background and Literature Review

In this chapter we study a parametric regression problem where the outcome is subject to missingness. The central interest is the estimation and inference of the regression coefficients. In practice there are various reasons that could lead to missing outcome, such as budget or technique restrictions, subjects' failing to comply with the protocol, or simply the study design. Missing data usually bring great challenges for estimation and inference, and need to be handled delicately, as a direct application of statistical methods developed for data without missing values may lead to biased estimation and misleading conclusions.

In addition to the outcome and covariates, we assume some surrogate variables are available for all subjects. Although the surrogate variables are not of direct statistical interest, they may help to explain the missingness mechanism, and thus reduce the impact of missing data on estimation and inference. Data with this structure arise from many observational studies (e.g. Wang et al. 2010), as well as from two-stage design studies (e.g. Pepe 1992; Pepe et al. 1994), where the second-stage outcome is not observed for all subjects, and the probability of observing this outcome depends on the first-stage outcome (surrogate variable) and covariates.

Let Y denote the outcome, $\boldsymbol{X}$ denote the covariates, $\boldsymbol{\beta}$ denote the p-dimensional vector of regression coefficients, and $\boldsymbol{S}$ denote the surrogate variables. Pepe (1992) proposed the maximum likelihood estimation, which assumes the correct specification of two densities, namely $f(Y|\boldsymbol{X})$ and $f(\boldsymbol{S}|Y,\boldsymbol{X})$. To reduce model assumptions, Pepe et al. (1994) proposed the mean score estimation, which assumes the correct specification of density $f(Y|\boldsymbol{X})$. However, this assumption is still more than necessary, and is likely subject to model misspecification. In our development, we only specify the mean regression model in the following form:

$$\mathrm{E}(Y|\boldsymbol{X}) = \mu(\boldsymbol{X}^T\boldsymbol{\beta}) \quad \text{for some } \boldsymbol{\beta} = \boldsymbol{\beta}_0 \in \mathbb{R}^p, \tag{2.1}$$

where $\mu(\cdot)$ is some known link function, and the expectation is taken under the true density $f(Y|\boldsymbol{X})$. Let R denote the indicator of Y being observed. Specifically, $R = 1$ if Y is observed, and $R = 0$ if Y is missing. The sampled data are then $(R_i, R_i * Y_i, \boldsymbol{S}_i, \boldsymbol{X}_i)$, $i = 1, \cdots, N$, which are independent and identically distributed, where $R_i * Y_i = Y_i$ if $R_i = 1$ and $R_i * Y_i =$ missing if $R_i = 0$. We assume the following missing at random (MAR) mechanism (Little and Rubin, 2002):

$$P(R = 1|Y, \boldsymbol{S}, \boldsymbol{X}) = P(R = 1|\boldsymbol{S}, \boldsymbol{X}) \stackrel{\text{def}}{=} \pi(\boldsymbol{S}, \boldsymbol{X}). \tag{2.2}$$

The model defined by (2.1) and (2.2) is embedded in a more general missing data setting where the pattern of missingness is arbitrary. The general setting has been studied extensively by Robins, Rotnitzky and colleagues using the semiparametric theory as in Bickel, Klaassen, Ritov and Wellner (1993). Applying the theory developed by Robins et al. (1994) and Robins and Rotnitzky (1995) to missing outcome data, Yu and Nan (2006) derived the semiparametric efficiency bound for the above model. Under this model, estimators whose asymptotic variance attains such bound are efficient. Chen and Breslow (2004) independently derived the bound using theory

of optimal estimating functions (Godambe 1960, 1991; Heyde 1988, 1997; Newey and McFadden 1994).

Many existing estimation methods for missing outcome data rely on a set of estimating functions $\boldsymbol{U}(\boldsymbol{\beta}; Y, \boldsymbol{X})$ constructed from (2.1), where $\mathrm{E}\{\boldsymbol{U}(\boldsymbol{\beta}_0; Y, \boldsymbol{X})\} = \mathbf{0}$ when there are no missing data. Under MAR assumption (2.2), the augmented inverse probability weighted (AIPW) estimator was proposed as the solution to the following AIPW estimating equation (Robins et al. 1994, 1995; Robins and Rotnitzky 1995; Tsiatis 2006):

$$\sum_{i=1}^{N} \left\{ \frac{R_i}{\hat{\pi}(\boldsymbol{S}_i, \boldsymbol{X}_i)} \boldsymbol{U}(\boldsymbol{\beta}; Y_i, \boldsymbol{X}_i) - \frac{R_i - \hat{\pi}(\boldsymbol{S}_i, \boldsymbol{X}_i)}{\hat{\pi}(\boldsymbol{S}_i, \boldsymbol{X}_i)} \boldsymbol{\sigma}(\boldsymbol{\beta}; \boldsymbol{S}_i, \boldsymbol{X}_i) \right\} = \mathbf{0}, \tag{2.3}$$

where $\hat{\pi}(\boldsymbol{S}, \boldsymbol{X})$ is an estimator of $\pi(\boldsymbol{S}, \boldsymbol{X})$, and $\boldsymbol{\sigma}(\boldsymbol{\beta}; \boldsymbol{S}, \boldsymbol{X})$ is an arbitrary function of $\boldsymbol{\beta}$, $\boldsymbol{S}$ and $\boldsymbol{X}$. When $\boldsymbol{\sigma}(\boldsymbol{\beta}; \boldsymbol{S}, \boldsymbol{X}) \equiv \mathbf{0}$, the AIPW estimator reduces to the inverse probability weighted (IPW) estimator (Horvitz and Thompson 1952). The AIPW estimator possesses the double robustness property, in the sense that it is consistent if either $\pi(\boldsymbol{S}, \boldsymbol{X})$ is correctly modeled, or $\boldsymbol{\sigma}(\boldsymbol{\beta}; \boldsymbol{S}, \boldsymbol{X})$ is a correct model for $\mathrm{E}\{\boldsymbol{U}(\boldsymbol{\beta}; Y, \boldsymbol{X}) | \boldsymbol{S}, \boldsymbol{X}\}$. For a fixed $\boldsymbol{U}(\boldsymbol{\beta}; Y, \boldsymbol{X})$, the smallest asymptotic variance of the AIPW estimator is achieved when both $\pi(\boldsymbol{S}, \boldsymbol{X})$ and $\mathrm{E}\{\boldsymbol{U}(\boldsymbol{\beta}; Y, \boldsymbol{X}) | \boldsymbol{S}, \boldsymbol{X}\}$ are correctly modeled, and $\boldsymbol{\sigma}(\boldsymbol{\beta}; \boldsymbol{S}, \boldsymbol{X})$ is taken to be the correct model for the latter. But this $\boldsymbol{U}(\boldsymbol{\beta}; Y, \boldsymbol{X})$-dependent variance is usually larger than the semiparametric efficiency bound.

In recent literature, many doubly robust estimators that are alternative to the AIPW estimator have been proposed. These proposals include Tan (2006, 2008, 2010), Kang and Schafer (2007), Robins et al. (2007), Rubin and van der Laan (2008), Cao et al. (2009), Tsiatis et al. (2011), Han (2012) and Rotnitzky et al. (2012). While most of these alternatives concerned a relatively simple setting of estimating the population mean of a response variable with incomplete data, Han

(2012) and Rotnitzky et al. (2012) considered the regression setting, and the latter two estimators are refereed to as HAN estimator and RLSR estimator, respectively, in the rest of this chapter. Along the lines of Tan's (2006, 2010) approach, the HAN estimator solves an estimating equation that employs a particular linear combination of the two terms in (2.3). When $\pi(\boldsymbol{S}, \boldsymbol{X})$ is correctly modeled, this linear combination yields the residual of the projection of the first term on the second, which endows the HAN estimator with improved efficiency over both the IPW and the AIPW estimators, with an exception when $\boldsymbol{\sigma}(\boldsymbol{\beta}; \boldsymbol{S}, \boldsymbol{X})$ is a correct model for $\mathrm{E}\{\boldsymbol{U}(\boldsymbol{\beta}; Y, \boldsymbol{X})|\boldsymbol{S}, \boldsymbol{X}\}$, in which case the HAN and the AIPW estimators have the same efficiency. In addition to the efficiency improvement over both the IPW and the AIPW estimators, the RLSR estimator has the property that, for a given finite set of user-specified functions, each function evaluated at the RLSR estimator has the asymptotic variance no larger than that of the function evaluated at any AIPW estimator using the same model structure for $\boldsymbol{\sigma}(\boldsymbol{\beta}; \boldsymbol{S}, \boldsymbol{X})$. The RLSR estimator solves an outcome regression estimating equation, which, unlike equation (2.3), always has a solution if the estimated value of $\mathrm{E}(Y|\boldsymbol{S}, \boldsymbol{X})$ falls in the sample space of Y.

Empirical likelihood (EL) (Owen 1988, 1990, 2001; Qin and Lawless 1994; Kitamura 2007) has become a popular tool in analyzing data with missing outcome. Chen et al. (2003) assumed an extra set of estimating functions in addition to $\boldsymbol{U}(\boldsymbol{\beta}; Y, \boldsymbol{X})$ to handle missing data, and employed the EL approach to combine the two. But their method is only valid under the missing completely at random mechanism (Little and Rubin, 2002). Under MAR assumption (2.2), Chen et al. (2008) proposed an estimator, referred to as CLQ estimator in the following, by solving the estimating equation

$$\sum_{i=1}^{N} \hat{p}_i \frac{R_i}{\hat{\pi}(\boldsymbol{S}_i, \boldsymbol{X}_i)} \boldsymbol{U}(\boldsymbol{\beta}; Y_i, \boldsymbol{X}_i) = \boldsymbol{0},$$

where $\hat{p}_i$ is the EL probability mass assigned to data point $(R_i = 1, Y_i, \boldsymbol{S}_i, \boldsymbol{X}_i)$ after incorporating information carried by subjects with missing values. Qin et al. (2009) proposed an estimator, referred to as QZL estimator in the following, by solving the over-identified estimating equation

$$\sum_{i=1}^{N} \left\{ \frac{R_i}{\hat{\pi}(\boldsymbol{S}_i, \boldsymbol{X}_i)} \boldsymbol{U}(\boldsymbol{\beta}; Y_i, \boldsymbol{X}_i)^T, \ \frac{R_i - \hat{\pi}(\boldsymbol{S}_i, \boldsymbol{X}_i)}{\hat{\pi}(\boldsymbol{S}_i, \boldsymbol{X}_i)} \boldsymbol{\sigma}(\boldsymbol{\beta}; \boldsymbol{S}_i, \boldsymbol{X}_i)^T \right\}^T = \mathbf{0},$$

for which the EL was used to account for the over-identification in estimation. It has been shown that, when $\pi(\boldsymbol{S}, \boldsymbol{X})$ is correctly modeled, both the CLQ and the QZL estimators are more efficient than the IPW estimator. In addition, when $\boldsymbol{\sigma}(\boldsymbol{\beta}; \boldsymbol{S}, \boldsymbol{X})$ is a correct model for $\mathrm{E}\{\boldsymbol{U}(\boldsymbol{\beta}; Y, \boldsymbol{X}) | \boldsymbol{S}, \boldsymbol{X}\}$, both estimators asymptotically coincide with the corresponding AIPW estimator. However, when $\pi(\boldsymbol{S}, \boldsymbol{X})$ is incorrectly modeled, neither the CLQ nor the QZL estimator is consistent. Wang and Chen (2009) proposed a nonparametric multiple imputation method under MAR assumption (2.2) that can be applied to analyze missing outcome data. However, large number of imputations are required to obtain stable estimates, which makes this method computationally undesirable.

All of the AIPW method, its recently proposed alternatives and the EL-based methods depend on the form of $\boldsymbol{U}(\boldsymbol{\beta}; Y, \boldsymbol{X})$, which needs to be explicitly specified priori. Different forms of $\boldsymbol{U}(\boldsymbol{\beta}; Y, \boldsymbol{X})$ produce different estimators. These estimators have different levels of estimation efficiency, and their numerical performances may differ from each other dramatically. For many EL-based estimators (e.g. CLQ and QZL), another drawback is that they are not robust against model misspecification of the missingness mechanism. Such model misspecification commonly occurs in practice when dealing with missing data. Therefore, the double robustness property as possessed by the AIPW estimator is highly desired. In view of these facts, we propose a conditional empirical likelihood (CEL) (Zhang and Gijbels 2003; Kitamura

et al. 2004) method for estimation with missing outcome data, where the estimation is carried out directly based on (2.1) rather than on a chosen $\boldsymbol{U}(\boldsymbol{\beta}; Y, \boldsymbol{X})$. We propose two CEL-based estimators, namely the CEL-IPW estimator and the CEL-AIPW estimator. The CEL-IPW estimator is consistent if $\pi(\boldsymbol{S}, \boldsymbol{X})$ is correctly modeled. The CEL-AIPW estimator enjoys the double robustness property, in the sense that it is consistent if either $\pi(\boldsymbol{S}, \boldsymbol{X})$ or $\mathrm{E}(Y|\boldsymbol{S}, \boldsymbol{X})$ is correctly modeled. When both models are correct, the CEL-AIPW estimator attains the semiparametric efficiency bound, and thus is asymptotically the most efficient estimator.

This chapter is organized as follows. Section 2.2 describes the CEL estimation procedure and its implementation. Section 2.3 concerns the large sample properties. Section 2.4 contains the results of simulation studies. Section 2.5 illustrates the data application of the proposed CEL method. Section 2.6 consists of some concluding remarks. Technical assumptions and proofs are provided in Section 2.7.

2.2 CEL Estimation

2.2.1 CEL-based estimators

We now introduce the CEL-based estimators. Define the IPW residual and the AIPW residual respectively as follows:

$$\begin{aligned} f(\boldsymbol{\beta}) &= R\left\{Y - \mu(\boldsymbol{X}^T\boldsymbol{\beta})\right\}/\pi(\boldsymbol{S}, \boldsymbol{X}), \\ g(\boldsymbol{\beta}) &= \frac{R}{\pi(\boldsymbol{S}, \boldsymbol{X})}\left\{Y - \mu(\boldsymbol{X}^T\boldsymbol{\beta})\right\} - \frac{R - \pi(\boldsymbol{S}, \boldsymbol{X})}{\pi(\boldsymbol{S}, \boldsymbol{X})}\mathrm{E}\left\{Y - \mu(\boldsymbol{X}^T\boldsymbol{\beta})|\boldsymbol{S}, \boldsymbol{X}\right\}. \end{aligned}$$

Clearly we have that $\mathrm{E}\{f(\boldsymbol{\beta}_0)|\boldsymbol{X}\} = 0$ and $\mathrm{E}\{g(\boldsymbol{\beta}_0)|\boldsymbol{X}\} = 0$. This conditional mean zero property of both residuals serves as the foundation of the proposed CEL-based estimation procedure. In this section we focus on describing the procedure based on the AIPW residual $g(\boldsymbol{\beta})$, which yields the CEL-AIPW estimator. Estimation based on the IPW residual $f(\boldsymbol{\beta})$ that leads to the CEL-IPW estimator then follows

a similar procedure, with no need of modeling $\mathrm{E}(Y|\boldsymbol{S},\boldsymbol{X})$.

Define $p_{ij} = P\{g_j(\boldsymbol{\beta})|\boldsymbol{X}_i\}$, $i,j = 1,\cdots,N$, which are the conditional empirical probabilities on the augmented data points $\{(\boldsymbol{X}_i^T, g_j(\boldsymbol{\beta})) : i,j = 1,\cdots,N\}$. To ensure that the conditional mean zero property is satisfied by the sampled data, we impose the following constraints over p_{ij}'s:

$$p_{ij} \geq 0\ , \sum_{j=1}^{N} p_{ij} = 1, \text{ and } \sum_{j=1}^{N} p_{ij} g_j(\boldsymbol{\beta}) = 0. \tag{2.4}$$

The first two constraints make sure that for each $i = 1,\cdots,N$, p_{ij}'s are properly defined probabilities, whereas the third constraint is the sample version of the conditional mean zero property. Following Zhang and Gijbels (2003) and Kitamura et al. (2004), we construct weighted conditional empirical log-likelihood for subject i as $\sum_{j=1}^{N} w_{ij} \log p_{ij}$, where w_{ij} is certain suitable weight assigned to subject j, $j = 1,\cdots,N$, satisfying $\sum_{j=1}^{N} w_{ij} = 1$. Naturally, subjects who resemble subject i should be assigned high weights. A technique quantifying this resemblance is to utilize a local kernel function. Let $\boldsymbol{X}^c$ and $\boldsymbol{X}_i^d$ denote the continuous and categorical components of $\boldsymbol{X}$, respectively. Then w_{ij} is given by

$$w_{ij} = \frac{\mathcal{K}\left\{(\boldsymbol{X}_i^c - \boldsymbol{X}_j^c)/b_N\right\} I(\boldsymbol{X}_i^d = \boldsymbol{X}_j^d)}{\sum_{j=1}^{N} \mathcal{K}\left\{(\boldsymbol{X}_i^c - \boldsymbol{X}_j^c)/b_N\right\} I(\boldsymbol{X}_i^d = \boldsymbol{X}_j^d)},$$

where we use a product kernel $\mathcal{K}\left\{(\boldsymbol{X}_i^c - \boldsymbol{X}_j^c)/b_N\right\} = \prod_{l=1}^{q} K\left\{(X_i^{(l)} - X_j^{(l)})/b_N\right\}$, with $K(\cdot)$ being a second order kernel function that is symmetric around zero (e.g. standard Gaussian kernel), b_N being the bandwidth parameter, q being the dimension of $\boldsymbol{X}^c$, and $X_i^{(l)}$ denoting the lth component of $\boldsymbol{X}_i^c$. Here $I(\cdot)$ is the indicator function.

Taking the summation over all subjects, we obtain the weighted conditional empirical log-likelihood $L = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} \log p_{ij}$. Constraints in (2.4) imply that L is a function of both $\boldsymbol{\beta}$ and p_{ij}'s. Maximizing L simultaneously with respect to $\boldsymbol{\beta}$ and

p_{ij}'s subject to constraints in (2.4) leads to the CEL-AIPW estimator, denoted as $\hat{\boldsymbol{\beta}}_{AIPW}$.

Note that the AIPW residual $g(\boldsymbol{\beta})$ involves two possibly unknown quantities, $\pi(\boldsymbol{S}, \boldsymbol{X})$ and $\mathrm{E}(Y|\boldsymbol{S}, \boldsymbol{X})$, which may need to be estimated in order to calculate $\hat{\boldsymbol{\beta}}_{AIPW}$. When the missing outcome data are collected from studies where the missingness of Y is due to study design (e.g. two-stage design), $\pi(\boldsymbol{S}, \boldsymbol{X})$ is known. Otherwise, we postulate a parametric model $\pi(\boldsymbol{\alpha}; \boldsymbol{S}, \boldsymbol{X})$, with $\boldsymbol{\alpha}$ being an unknown finite dimensional parameter whose true value is denoted as $\boldsymbol{\alpha}_0$. One example is the logistic model, $\mathrm{logit}\left\{\pi(\boldsymbol{\alpha}; \boldsymbol{S}, \boldsymbol{X})\right\} = \boldsymbol{Z}^T\boldsymbol{\alpha}$, where $\boldsymbol{Z}^T = (\boldsymbol{S}^T, \boldsymbol{X}^T)$. An estimator $\hat{\boldsymbol{\alpha}}$ is given by maximizing the following binomial likelihood:

$$\prod_{i=1}^{N} \left\{\pi(\boldsymbol{\alpha}; \boldsymbol{S}_i, \boldsymbol{X}_i)\right\}^{R_i} \left\{1 - \pi(\boldsymbol{\alpha}; \boldsymbol{S}_i, \boldsymbol{X}_i)\right\}^{1-R_i}. \tag{2.5}$$

On the other hand, to estimate $\mathrm{E}(Y|\boldsymbol{S}, \boldsymbol{X})$, we may postulate another parametric model $h(\boldsymbol{\gamma}; \boldsymbol{S}, \boldsymbol{X})$, where $h(\cdot)$ is a known link function and $\boldsymbol{\gamma}$ is an unknown finite dimensional parameter with true value $\boldsymbol{\gamma}_0$. Choices of this parametric model include the generalized linear model (McCullagh and Nelder 1989) and the quasi-likelihood model (Wedderburn 1974). Under the MAR assumption (2.2), i.e. $R \perp Y|(\boldsymbol{S}, \boldsymbol{X})$, we have $\mathrm{E}(Y|\boldsymbol{S}, \boldsymbol{X}) = \mathrm{E}(Y|\boldsymbol{S}, \boldsymbol{X}, R = 1)$. Therefore, one estimator $\hat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}_0$ is obtained based on complete-case analysis by solving the estimating equation

$$\sum_{i=1}^{N} R_i \boldsymbol{Z}_i \dot{h}(\boldsymbol{Z}_i^T\boldsymbol{\gamma}) \mathrm{Var}(Y_i|\boldsymbol{Z}_i)^{-1} \left\{Y_i - h(\boldsymbol{Z}_i^T\boldsymbol{\gamma})\right\} = \boldsymbol{0}, \tag{2.6}$$

where $\dot{h}(\cdot)$ is the first order derivative function of $h(\cdot)$. Note that although the underlying distribution that generates the data must satisfy

$$\mathrm{E}(Y|\boldsymbol{X}) = \int \mathrm{E}(Y|\boldsymbol{S}, \boldsymbol{X}) f(\boldsymbol{S}|\boldsymbol{X}) d\boldsymbol{S},$$

where $f(\boldsymbol{S}|\boldsymbol{X})$ is the density of $\boldsymbol{S}|\boldsymbol{X}$, the two models, $\mu(\boldsymbol{X}^T\boldsymbol{\beta})$ for $\mathrm{E}(Y|\boldsymbol{X})$ and

$h(\boldsymbol{\gamma}; \boldsymbol{S}, \boldsymbol{X})$ for $\mathrm{E}(Y|\boldsymbol{S}, \boldsymbol{X})$, may not have such a relationship. This is because although we assume that $\mu(\boldsymbol{X}^T\boldsymbol{\beta})$ is a correctly specified model, $h(\boldsymbol{\gamma}; \boldsymbol{S}, \boldsymbol{X})$ may not be. When $h(\boldsymbol{\gamma}; \boldsymbol{S}, \boldsymbol{X})$ is misspecified, the above relationship is unlikely to hold for the two models. For this reason, the relationship actually provides us a practical guidance to rule out certain parametric models $h(\boldsymbol{\gamma}; \boldsymbol{S}, \boldsymbol{X})$ that have apparent incompatibility with $\mu(\boldsymbol{X}^T\boldsymbol{\beta})$.

Given estimators $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$, the AIPW residual can be rewritten as following:

(2.7)

$$g(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) = \frac{R}{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{S}, \boldsymbol{X})} \left\{Y - \mu(\boldsymbol{X}^T\boldsymbol{\beta})\right\} - \frac{R - \pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{S}, \boldsymbol{X})}{\pi(\hat{\boldsymbol{\alpha}}; \boldsymbol{S}, \boldsymbol{X})} \left\{h(\hat{\boldsymbol{\gamma}}; \boldsymbol{S}, \boldsymbol{X}) - \mu(\boldsymbol{X}^T\boldsymbol{\beta})\right\}.$$

Then the proposed CEL estimation is carried out by substituting $g_j(\boldsymbol{\beta})$ in the third constraint in (2.4) with $g_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$. For convenience, we denote the resulting CEL-AIPW estimator still as $\hat{\boldsymbol{\beta}}_{AIPW}$.

2.2.2 Numerical implementation

The calculation of $\hat{\boldsymbol{\beta}}_{AIPW}$ pertains to a constrained optimization problem. Using Lagrange multipliers, the Lagrangian is given by

$$\mathcal{L} = \sum_{i=1}^{N} \left(\sum_{j=1}^{N} w_{ij} \log p_{ij} \right) - \sum_{i=1}^{N} \varpi_i \left(\sum_{j=1}^{N} p_{ij} - 1 \right) - \sum_{i=1}^{N} \lambda_i \left\{ \sum_{j=1}^{N} p_{ij} g_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \right\},$$

where scalars ϖ_i and λ_i are the Lagrange multipliers associated with the second and third constraints in (2.4), respectively. With $\partial\mathcal{L}/\partial p_{ij} = 0$ and (2.4), it can be easily shown that, for a fixed $\boldsymbol{\beta}$,

$$p_{ij}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) = \frac{w_{ij}}{1 + \hat{\lambda}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) g_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})}, \quad i,\ j = 1, \cdots, N, \tag{2.8}$$

where $\hat{\lambda}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$ is the solution to equation $\sum_{j=1}^{N} p_{ij}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) g_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) = 0$. It is easy to see that

$$\hat{\lambda}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) = \arg\min_{\lambda_i} \left[-\sum_{j=1}^{N} w_{ij} \log \left\{1 + \lambda_i g_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\right\} \right].$$

Denote $\Lambda_i(\lambda_i, \boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) = -\sum_{j=1}^{N} w_{ij} \log\{1 + \lambda_i g_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\}$. Then the objective function L can be rewritten as a function of $\boldsymbol{\beta}$ only:

$$L(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{N} \Lambda_i \left\{ \hat{\lambda}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}), \boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}} \right\} + \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \log w_{ij}.$$

Therefore, the CEL-AIPW estimator can be equivalently derived from the following optimization:

$$\hat{\boldsymbol{\beta}}_{AIPW} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left\{ \min_{\lambda_i} \Lambda_i(\lambda_i, \boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \right\}.$$

This representation of the CEL-AIPW estimator essentially suggests a way of numerical implementation, which is done via nested optimization routines. The Newton-Raphson algorithm may be employed for the optimization. For convenience, we suppress $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ in the following description of algorithmic implementation. For a fixed $\boldsymbol{\beta}$, given λ_i^{old}, the inner loop updates λ_i by

$$\lambda_i^{new} = \lambda_i^{old} - \Lambda_{i,\lambda\lambda}^{-1}\left(\lambda_i^{old}, \boldsymbol{\beta}\right) \Lambda_{i,\lambda}\left(\lambda_i^{old}, \boldsymbol{\beta}\right), \; i = 1, \cdots, N,$$

where

$$\Lambda_{i,\lambda}(\lambda_i, \boldsymbol{\beta}) = -\sum_{j=1}^{N} w_{ij} \frac{g_j(\boldsymbol{\beta})}{1 + \lambda_i g_j(\boldsymbol{\beta})} \quad \text{and} \quad \Lambda_{i,\lambda\lambda}(\lambda_i, \boldsymbol{\beta}) = \sum_{j=1}^{N} w_{ij} \frac{g_j(\boldsymbol{\beta})^2}{\{1 + \lambda_i g_j(\boldsymbol{\beta})\}^2}.$$

For each i, an initial value can be taken as $\lambda_i = 0$, and the converged value gives an estimate of $\hat{\lambda}_i(\boldsymbol{\beta})$. To guarantee the positivity of the estimated p_{ij}'s, the updates should be restricted on the legitimate region $\{\lambda_i : 1 + \lambda_i g_j(\beta) \geq w_{ij}\}$. Given $\boldsymbol{\beta}^{old}$ and the estimated $\hat{\lambda}_i(\boldsymbol{\beta}^{old})$'s from the inner loop, the outer loop updates $\boldsymbol{\beta}$ by

$$\boldsymbol{\beta}^{new} = \boldsymbol{\beta}^{old} - \left\{ \sum_{i=1}^{N} \boldsymbol{L}_{i,\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}^{old}) \right\}^{-1} \left\{ \sum_{i=1}^{N} \boldsymbol{L}_{i,\boldsymbol{\beta}}(\boldsymbol{\beta}^{old}) \right\},$$

where

$$\begin{aligned}
\boldsymbol{L}_{i,\boldsymbol{\beta}}(\boldsymbol{\beta}) &= -\hat{\lambda}_i(\boldsymbol{\beta}) \sum_{j=1}^{N} w_{ij} \frac{\boldsymbol{G}_j(\boldsymbol{\beta})^T}{1+\hat{\lambda}_i(\boldsymbol{\beta}) g_j(\boldsymbol{\beta})}, \\
\boldsymbol{L}_{i,\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}) &= -\frac{\boldsymbol{\Lambda}_{i,\lambda\boldsymbol{\beta}}^T \left\{\hat{\lambda}_i(\boldsymbol{\beta}), \boldsymbol{\beta}\right\} \boldsymbol{\Lambda}_{i,\boldsymbol{\lambda}\boldsymbol{\beta}} \left\{\hat{\lambda}_i(\boldsymbol{\beta}), \boldsymbol{\beta}\right\}}{\Lambda_{i,\lambda\lambda} \left\{\hat{\lambda}_i(\boldsymbol{\beta}), \boldsymbol{\beta}\right\}}, \\
\boldsymbol{\Lambda}_{i,\lambda\boldsymbol{\beta}} \left\{\hat{\lambda}_i(\boldsymbol{\beta}), \boldsymbol{\beta}\right\} &= \sum_{j=1}^{N} w_{ij} \left[\frac{\hat{\lambda}_i(\boldsymbol{\beta}) g_j(\boldsymbol{\beta}) \boldsymbol{G}_j(\boldsymbol{\beta})}{\left\{1+\hat{\lambda}_i(\boldsymbol{\beta}) g_j(\boldsymbol{\beta})\right\}^2} - \frac{\boldsymbol{G}_j(\boldsymbol{\beta})}{1+\hat{\lambda}_i(\boldsymbol{\beta}) g_j(\boldsymbol{\beta})} \right],
\end{aligned}$$

and $\boldsymbol{G}_j(\boldsymbol{\beta}) = \partial g_j(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$. Iterate the above nested loops until a certain convergence criterion is satisfied. At the convergence, the algorithm produces $\hat{\boldsymbol{\beta}}_{AIPW}$.

For bandwidth selection, we follow Smith (2007) as a rule of thumb to first determine the order of b_N. That is, $b_N \to 0$, $N^{1-2\nu-2/\delta} b_N^{2q} \to \infty$ and $N^{1-2\nu} b_N^{5q/2} \to \infty$ as $N \to \infty$, where $\nu \in (0, 1/2)$ and $\delta \geq 8$. Then the following cross-validation criterion, which is a modified version of the one suggested by Newey (1993), may be used for data-driven bandwidth selection:

$$CV(b_N) = \sum_{i=1}^{N} \frac{\left\{ g_i(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})^2 - \hat{\sigma}_{-i}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})^2 \right\}^2}{\hat{\sigma}_{-i}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})^6}, \tag{2.9}$$

where $\hat{\sigma}_{-i}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})^2 = \sum_{j=1}^{N} \hat{w}_{ij} g_j(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})^2$, $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(b_N)$ is the CEL-AIPW estimator obtained with a given b_N, and

$$\hat{w}_{ii} = 0, \quad \hat{w}_{ij} = \frac{\mathcal{K}\left\{(\boldsymbol{X}_i^c - \boldsymbol{X}_j^c)/b_N\right\} I(\boldsymbol{X}_i^d = \boldsymbol{X}_j^d)}{\sum_{j=1, j\neq i}^{N} \mathcal{K}\left\{(\boldsymbol{X}_i^c - \boldsymbol{X}_j^c)/b_N\right\} I(\boldsymbol{X}_i^d = \boldsymbol{X}_j^d)} \quad \text{for } j \neq i.$$

The optimal bandwidth b_N is chosen as the minimizer of $CV(b_N)$.

2.3 Large Sample Properties

For the large sample properties presented in this section, primary consideration is given to the CEL-AIPW estimator, and the corresponding results are summarized in a series of theorems. Regularity conditions and proofs are provided in Section

2.7. Properties regarding the CEL-IPW estimator are listed as corollaries, since the CEL-IPW estimator may be treated as a special case of the CEL-AIPW estimator. Proofs of the corollaries are omitted because they are trivially modified versions of proofs of the corresponding theorems.

Based on the results of White (1982), we know that $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ converge in probability to some values $\boldsymbol{\alpha}_*$ and $\boldsymbol{\gamma}_*$ respectively, and that $\sqrt{N}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_*)$ and $\sqrt{N}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_*)$ are bounded in probability. Here $\boldsymbol{\alpha}_*$ and $\boldsymbol{\gamma}_*$ are not necessarily $\boldsymbol{\alpha}_0$ and $\boldsymbol{\gamma}_0$. Only when the model for $\pi(\boldsymbol{S},\boldsymbol{X})$ or $\mathrm{E}(Y|\boldsymbol{S},\boldsymbol{X})$ is correctly specified is $\boldsymbol{\alpha}_*$ or $\boldsymbol{\gamma}_*$ equal to $\boldsymbol{\alpha}_0$ or $\boldsymbol{\gamma}_0$, respectively. Since $\hat{\boldsymbol{\alpha}}$ maximizes (2.5), from White (1982), the asymptotic linear expansion for $\hat{\boldsymbol{\alpha}}$ is given as

$$\sqrt{N}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_*) = -\left[\mathrm{E}\left\{\frac{\partial\boldsymbol{\psi}(\boldsymbol{\alpha}_*)}{\partial\boldsymbol{\alpha}}\right\}\right]^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{\psi}_i(\boldsymbol{\alpha}_*)+o_p(1), \tag{2.10}$$

where $\boldsymbol{\psi}(\boldsymbol{\alpha})=\boldsymbol{\psi}(\boldsymbol{\alpha};\boldsymbol{S},\boldsymbol{X},R)$ is the score function corresponding to (2.5). When a logistic regression model is assumed for $\pi(\boldsymbol{S},\boldsymbol{X})$, we have

$$\boldsymbol{\psi}(\boldsymbol{\alpha}) = \left\{R-\frac{\exp(\boldsymbol{Z}^T\boldsymbol{\alpha})}{1+\exp(\boldsymbol{Z}^T\boldsymbol{\alpha})}\right\}\boldsymbol{Z}.$$

Similarly, the asymptotic linear expansion for $\hat{\boldsymbol{\gamma}}$ is given as

$$\sqrt{N}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_*) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{\phi}_i(\boldsymbol{\gamma}_*)+o_p(1), \tag{2.11}$$

where $\boldsymbol{\phi}(\boldsymbol{\gamma})=\boldsymbol{\phi}(\boldsymbol{\gamma};Y,\boldsymbol{S},\boldsymbol{X},R)$ is the influence function. When $\hat{\boldsymbol{\gamma}}$ is the solution to estimating equation (2.6), we have

$$\boldsymbol{\phi}(\boldsymbol{\gamma}) = -\left[\mathrm{E}\left\{\frac{\partial\boldsymbol{\zeta}(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}\right\}\right]^{-1}\boldsymbol{\zeta}(\boldsymbol{\gamma}) \quad \text{with} \quad \boldsymbol{\zeta}(\boldsymbol{\gamma}) = \frac{R\boldsymbol{Z}\dot{h}(\boldsymbol{Z}^T\boldsymbol{\gamma})\left\{Y-h(\boldsymbol{Z}^T\boldsymbol{\gamma})\right\}}{\mathrm{Var}(Y|\boldsymbol{Z})}.$$

Let us denote the CEL-IPW estimator as $\hat{\boldsymbol{\beta}}_{IPW}$. The following Theorem II.1 and Corollary II.1 provide the consistency of $\hat{\boldsymbol{\beta}}_{AIPW}$ and $\hat{\boldsymbol{\beta}}_{IPW}$, respectively.

Theorem II.1. *For the model defined by (2.1) and (2.2), under Assumptions 1 in Section 2.7, if either $\boldsymbol{\alpha}_*=\boldsymbol{\alpha}_0$ or $\boldsymbol{\gamma}_*=\boldsymbol{\gamma}_0$, we have $\hat{\boldsymbol{\beta}}_{AIPW}\xrightarrow{p}\boldsymbol{\beta}_0$ as $N\to\infty$.*

Corollary II.1. *For the model defined by (2.1) and (2.2), under Assumptions 2 in Section 2.7,* if$\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$, *we have* $\hat{\boldsymbol{\beta}}_{IPW} \xrightarrow{p} \boldsymbol{\beta}_0$ *as* $N \to \infty$.

From Theorem II.1, $\hat{\boldsymbol{\beta}}_{AIPW}$ is doubly robust, in the sense that if either $\pi(\boldsymbol{S}, \boldsymbol{X})$ or $\mathrm{E}(Y|\boldsymbol{S}, \boldsymbol{X})$ is correctly modeled, $\hat{\boldsymbol{\beta}}_{AIPW}$ is a consistent estimator of $\boldsymbol{\beta}_0$.

To describe the asymptotic distribution of $\hat{\boldsymbol{\beta}}_{AIPW}$, denote $V_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \mathrm{E}\left\{g(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})^2 | \boldsymbol{X}\right\}$, $\boldsymbol{G}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \mathrm{E}\left\{\partial g(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})/\partial \boldsymbol{\gamma} | \boldsymbol{X}\right\}$,

$$\boldsymbol{Q}_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \left\{\partial \mu(\boldsymbol{X}^T\boldsymbol{\beta})/\partial \boldsymbol{\beta}^T\right\} V_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})^{-1} g(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}),$$

$$\boldsymbol{I}_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \mathrm{E}\left\{\boldsymbol{Q}_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \boldsymbol{Q}_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})^T\right\},$$

$$\begin{aligned}
&\boldsymbol{V}_{\boldsymbol{\alpha}, AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \\
&= \mathrm{Var}\left[\boldsymbol{Q}_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) - \mathrm{E}\left\{\boldsymbol{Q}_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \boldsymbol{\psi}(\boldsymbol{\alpha})^T\right\} \left[\mathrm{E}\left\{\boldsymbol{\psi}(\boldsymbol{\alpha}) \boldsymbol{\psi}(\boldsymbol{\alpha})^T\right\}\right]^{-1} \boldsymbol{\psi}(\boldsymbol{\alpha})\right],
\end{aligned}$$

$$\boldsymbol{V}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \mathrm{Var}\left[\boldsymbol{Q}_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) + \mathrm{E}\left\{\frac{\partial \mu(\boldsymbol{X}^T\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} V_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})^{-1} \boldsymbol{G}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})\right\} \boldsymbol{\phi}(\boldsymbol{\gamma})\right].$$

Theorem II.2. *For the model defined by (2.1) and (2.2), under Assumptions 1 in Section 2.7, we have the following results of asymptotic distribution.*

(i) If $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$, *then*

$$\begin{aligned}
&\sqrt{N}(\hat{\boldsymbol{\beta}}_{AIPW} - \boldsymbol{\beta}_0) \xrightarrow{d} \\
&\mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}_{AIPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)^{-1} \boldsymbol{V}_{\boldsymbol{\alpha}, AIPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*) \boldsymbol{I}_{AIPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)^{-1}\right).
\end{aligned} \tag{2.12}$$

(ii) If $\boldsymbol{\gamma}_* = \boldsymbol{\gamma}_0$, *then*

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{AIPW} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}_{AIPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_0)^{-1} \boldsymbol{V}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_0) \boldsymbol{I}_{AIPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_0)^{-1}\right).$$

(iii) If both $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$ *and* $\boldsymbol{\gamma}_* = \boldsymbol{\gamma}_0$, *then*

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{AIPW} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}_{AIPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0)^{-1}\right).$$

In the case that the missing outcome data are collected based on a two-stage design, $\boldsymbol{\alpha}_0$ is known. When the known $\boldsymbol{\alpha}_0$ is used instead of the estimator $\hat{\boldsymbol{\alpha}}$ for CEL estimation, following the same arguments as in the proof of Theorem II.2, the asymptotic variance of $\hat{\boldsymbol{\beta}}_{AIPW}$ has the same structure as that in (2.12), but with $\boldsymbol{V}_{\boldsymbol{\alpha},AIPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)$ in the middle replaced by $\mathrm{Var}\left\{\boldsymbol{Q}_{AIPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)\right\} = \boldsymbol{I}_{AIPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)$. The new asymptotic variance is no smaller than that given by (2.12), in the sense that the corresponding difference of the two asymptotic variance matrices is nonnegative-definite. This is because $\boldsymbol{V}_{\boldsymbol{\alpha},AIPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)$ is the variance of a population regression residual in the form of $\mathrm{Var}(\boldsymbol{Q} - \boldsymbol{C}\boldsymbol{\psi})$, where $\boldsymbol{C} = \mathrm{E}(\boldsymbol{Q}\boldsymbol{\psi}^T)\left\{\mathrm{E}(\boldsymbol{\psi}\boldsymbol{\psi}^T)\right\}^{-1}$ is the regression coefficient matrix. Obviously $\mathrm{Var}(\boldsymbol{Q}) \geq \mathrm{Var}(\boldsymbol{Q} - \boldsymbol{C}\boldsymbol{\psi})$ in the nonnegative-definite sense. So in practice even if $\boldsymbol{\alpha}_0$ is known, using an estimator $\hat{\boldsymbol{\alpha}}$ has the advantage of potential efficiency gain for the CEL-AIPW estimator. This counterintuitive phenomenon is well known in the literature of parametric regression with missing data (e.g. Robins et al. 1995). For nonparametric regression with missing data this does not hold any more (Wang et al. 2010). When $\mathrm{E}(Y|\boldsymbol{S}, \boldsymbol{X})$ is correctly modeled, however, using $\hat{\boldsymbol{\alpha}}$ or $\boldsymbol{\alpha}_0$ will make no difference asymptotically, as in both cases the asymptotic variance of $\hat{\boldsymbol{\beta}}_{AIPW}$ is $\boldsymbol{I}_{AIPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0)^{-1}$. From Chen and Breslow (2004) and Yu and Nan (2006), $\boldsymbol{I}_{AIPW}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0)^{-1}$ is the semiparametric efficiency bound for the model defined by (2.1) and (2.2). Therefore, Theorem II.2 implies that the CEL-AIPW estimator is efficient when both $\pi(\boldsymbol{S}, \boldsymbol{X})$ and $\mathrm{E}(Y|\boldsymbol{S}, \boldsymbol{X})$ are correctly modeled.

To describe the asymptotic distribution of $\hat{\boldsymbol{\beta}}_{IPW}$, denote $V_{IPW}(\boldsymbol{\beta},\boldsymbol{\alpha}) = \mathrm{E}\{f(\boldsymbol{\beta},\boldsymbol{\alpha})^2|\boldsymbol{X}\}$,

$$\begin{aligned}
\boldsymbol{Q}_{IPW}(\boldsymbol{\beta},\boldsymbol{\alpha}) &= \mathrm{E}\left\{\frac{R}{\pi(\boldsymbol{\alpha};\boldsymbol{S},\boldsymbol{X})}|\boldsymbol{X}\right\}\frac{\partial\mu(\boldsymbol{X}^T\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T}V_{IPW}(\boldsymbol{\beta},\boldsymbol{\alpha})^{-1}f(\boldsymbol{\beta},\boldsymbol{\alpha}),\\
\boldsymbol{I}_{IPW}(\boldsymbol{\beta},\boldsymbol{\alpha}) &= \mathrm{E}\left\{\boldsymbol{Q}_{IPW}(\boldsymbol{\beta},\boldsymbol{\alpha})\boldsymbol{Q}_{IPW}(\boldsymbol{\beta},\boldsymbol{\alpha})^T\right\},\\
\boldsymbol{V}_{\boldsymbol{\alpha},IPW}(\boldsymbol{\beta},\boldsymbol{\alpha}) &= \mathrm{Var}\left[\boldsymbol{Q}_{IPW}(\boldsymbol{\beta},\boldsymbol{\alpha}) - \mathrm{E}\left\{\boldsymbol{Q}_{IPW}(\boldsymbol{\beta},\boldsymbol{\alpha})\boldsymbol{\psi}(\boldsymbol{\alpha})^T\right\}\left[\mathrm{E}\left\{\boldsymbol{\psi}(\boldsymbol{\alpha})\boldsymbol{\psi}(\boldsymbol{\alpha})^T\right\}\right]^{-1}\boldsymbol{\psi}(\boldsymbol{\alpha})\right].
\end{aligned}$$

Corollary II.2. *For the model defined by (2.1) and (2.2), under Assumptions 2 in Section 2.7, if $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$, we have*

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{IPW} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}_{IPW}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)^{-1}\boldsymbol{V}_{\boldsymbol{\alpha},IPW}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)\boldsymbol{I}_{IPW}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)^{-1}\right). \tag{2.13}$$

If $\boldsymbol{\alpha}_0$ is known by design and is used in the CEL estimation instead of $\hat{\boldsymbol{\alpha}}$, the asymptotic variance of $\hat{\boldsymbol{\beta}}_{IPW}$ has the same structure as that in (2.13), but with $\boldsymbol{V}_{\boldsymbol{\alpha},IPW}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)$ in the middle replaced by $\mathrm{Var}\{\boldsymbol{Q}_{IPW}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)\} = \boldsymbol{I}_{IPW}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)$. The new asymptotic variance is no smaller than that given by (2.13) due to the same reason given before. So using an estimator $\hat{\boldsymbol{\alpha}}$ is still preferred for the CEL-IPW estimator even if $\boldsymbol{\alpha}_0$ is known.

The following Theorem II.3 provides consistent estimators for the asymptotic variance of the CEL-AIPW estimator given in Theorem II.2. Denote $\hat{g}_i(\boldsymbol{\beta}) = g_i(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})$, $\hat{V}_{i,AIPW}(\boldsymbol{\beta}) = \sum_{j=1}^N p_{ij}(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})\hat{g}_j(\boldsymbol{\beta})^2$, $\hat{\boldsymbol{G}}_{i,\boldsymbol{\gamma}}(\boldsymbol{\beta}) = \sum_{j=1}^N p_{ij}(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})\{\partial g_j(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})/\partial\boldsymbol{\gamma}\}$,
$\hat{\boldsymbol{Q}}_{i,AIPW}(\boldsymbol{\beta}) = \left\{\partial\mu(\boldsymbol{X}_i^T\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T\right\}\hat{V}_{i,AIPW}(\boldsymbol{\beta})^{-1}\hat{g}_i(\boldsymbol{\beta})$,
$\hat{\boldsymbol{I}}_{AIPW}(\boldsymbol{\beta}) = (1/N)\sum_{i=1}^N \hat{\boldsymbol{Q}}_{i,AIPW}(\boldsymbol{\beta})\hat{\boldsymbol{Q}}_{i,AIPW}(\boldsymbol{\beta})^T$, and

$$\hat{\boldsymbol{r}}_i(\boldsymbol{\beta}) = \hat{\boldsymbol{Q}}_{i,AIPW}(\boldsymbol{\beta}) - \left\{\frac{1}{N}\sum_{i=1}^N \hat{\boldsymbol{Q}}_{i,AIPW}(\boldsymbol{\beta})\boldsymbol{\psi}_i(\hat{\boldsymbol{\alpha}})^T\right\}\left\{\frac{1}{N}\sum_{i=1}^N \boldsymbol{\psi}_i(\hat{\boldsymbol{\alpha}})\boldsymbol{\psi}_i(\hat{\boldsymbol{\alpha}})^T\right\}^{-1}\boldsymbol{\psi}_i(\hat{\boldsymbol{\alpha}}),$$

$$\hat{\boldsymbol{v}}_i(\boldsymbol{\beta}) = \hat{\boldsymbol{Q}}_{i,AIPW}(\boldsymbol{\beta}) + \left\{\frac{1}{N}\sum_{i=1}^N \frac{\partial\mu(\boldsymbol{X}_i^T\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T}\hat{V}_{i,AIPW}(\boldsymbol{\beta})^{-1}\hat{\boldsymbol{G}}_{i,\boldsymbol{\gamma}}(\boldsymbol{\beta})\right\}\boldsymbol{\phi}_i(\hat{\boldsymbol{\gamma}}).$$

Theorem II.3. *Under Assumptions 1 in Section 2.7, we have the following results.*

(i) If $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$*, then*

$$\left[\hat{\boldsymbol{I}}_{AIPW}(\boldsymbol{\beta})^{-1}\left\{\frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{r}}_i(\boldsymbol{\beta})\hat{\boldsymbol{r}}_i(\boldsymbol{\beta})^T\right\}\hat{\boldsymbol{I}}_{AIPW}(\boldsymbol{\beta})^{-1}\right]\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{AIPW}}$$
$$\xrightarrow{p} \boldsymbol{I}_{AIPW}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*)^{-1}\boldsymbol{V}_{\boldsymbol{\alpha},AIPW}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*)\boldsymbol{I}_{AIPW}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*)^{-1}.$$

(ii) If $\boldsymbol{\gamma}_* = \boldsymbol{\gamma}_0$*, then*

$$\left[\hat{\boldsymbol{I}}_{AIPW}(\boldsymbol{\beta})^{-1}\left\{\frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{v}}_i(\boldsymbol{\beta})\hat{\boldsymbol{v}}_i(\boldsymbol{\beta})^T\right\}\hat{\boldsymbol{I}}_{AIPW}(\boldsymbol{\beta})^{-1}\right]\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{AIPW}}$$
$$\xrightarrow{p} \boldsymbol{I}_{AIPW}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_0)^{-1}\boldsymbol{V}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_0)\boldsymbol{I}_{AIPW}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_0)^{-1}.$$

(iii) If both $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$ *and* $\boldsymbol{\gamma}_* = \boldsymbol{\gamma}_0$*, then* $\hat{\boldsymbol{I}}_{AIPW}(\hat{\boldsymbol{\beta}}_{AIPW}) \xrightarrow{p} \boldsymbol{I}_{AIPW}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0)$.

The asymptotic variance estimators given in Theorem II.3 can be employed for inference if we know which one of $\pi(\boldsymbol{S},\boldsymbol{X})$ and $\mathrm{E}(Y|\boldsymbol{S},\boldsymbol{X})$ is correctly modeled. In many practical studies, however, such knowledge is unavailable. Therefore, it is desirable to have a consistent estimator of the asymptotic variance without knowing which one of $\pi(\boldsymbol{S},\boldsymbol{X})$ and $\mathrm{E}(Y|\boldsymbol{S},\boldsymbol{X})$ is correctly modeled. The following Theorem II.4 provides such a consistent estimator. Denote $\hat{\boldsymbol{G}}_{i,\boldsymbol{\alpha}}(\boldsymbol{\beta}) = \sum_{j=1}^{N} p_{ij}(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})\left\{\partial g_j(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})/\partial\boldsymbol{\alpha}\right\}$ and

$$\begin{aligned}
&\hat{\boldsymbol{m}}_i(\boldsymbol{\beta})\\
=\;&\hat{\boldsymbol{Q}}_{i,AIPW}(\boldsymbol{\beta}) - \left\{\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\mu(\boldsymbol{X}_i^T\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T}\hat{V}_{i,AIPW}(\boldsymbol{\beta})^{-1}\hat{\boldsymbol{G}}_{i,\boldsymbol{\alpha}}(\boldsymbol{\beta})\right\}\left\{\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\boldsymbol{\psi}_i(\hat{\boldsymbol{\alpha}})}{\partial\boldsymbol{\alpha}}\right\}^{-1}\boldsymbol{\psi}_i(\hat{\boldsymbol{\alpha}})\\
&+\left\{\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\mu(\boldsymbol{X}_i^T\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T}\hat{V}_{i,AIPW}(\boldsymbol{\beta})^{-1}\hat{\boldsymbol{G}}_{i,\boldsymbol{\gamma}}(\boldsymbol{\beta})\right\}\boldsymbol{\phi}_i(\hat{\boldsymbol{\gamma}}).
\end{aligned}$$

Theorem II.4. *Under Assumptions 1 in Section 2.7, we have that*

$$\left[\hat{\boldsymbol{I}}_{AIPW}(\boldsymbol{\beta})^{-1}\left\{\frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{m}}_i(\boldsymbol{\beta})\hat{\boldsymbol{m}}_i(\boldsymbol{\beta})^T\right\}\hat{\boldsymbol{I}}_{AIPW}(\boldsymbol{\beta})^{-1}\right]\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{AIPW}}$$

$$\xrightarrow{p}\begin{cases}\boldsymbol{I}_{AIPW}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*)^{-1}\boldsymbol{V}_{\boldsymbol{\alpha},AIPW}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*)\boldsymbol{I}_{AIPW}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*)^{-1} & \text{if } \boldsymbol{\alpha}_*=\boldsymbol{\alpha}_0\\ \boldsymbol{I}_{AIPW}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_0)^{-1}\boldsymbol{V}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_0)\boldsymbol{I}_{AIPW}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_0)^{-1} & \text{if } \boldsymbol{\gamma}_*=\boldsymbol{\gamma}_0\\ \boldsymbol{I}_{AIPW}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0)^{-1} & \text{if } \boldsymbol{\alpha}_*=\boldsymbol{\alpha}_0 \text{ and } \boldsymbol{\gamma}_*=\boldsymbol{\gamma}_0.\end{cases}$$

To consistently estimate the asymptotic variance of the CEL-IPW estimator given in Corollary II.2, denote $\hat{f}_i(\boldsymbol{\beta})=f_i(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}})$, $\hat{V}_{i,IPW}(\boldsymbol{\beta})=\sum_{j=1}^{N}p_{ij}(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}})\hat{f}_j(\boldsymbol{\beta})^2$, $\hat{M}_i(\boldsymbol{\beta})=\sum_{j=1}^{N}p_{ij}(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}})R_j/\pi(\hat{\boldsymbol{\alpha}};\boldsymbol{S}_j,\boldsymbol{X}_j)$,
$\hat{\boldsymbol{Q}}_{i,IPW}(\boldsymbol{\beta})=\hat{M}_i(\boldsymbol{\beta})\left\{\partial\mu(\boldsymbol{X}_i^T\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T\right\}\hat{V}_{i,IPW}(\boldsymbol{\beta})^{-1}\hat{f}_i(\boldsymbol{\beta})$,
$\hat{\boldsymbol{I}}_{IPW}(\boldsymbol{\beta})=(1/N)\sum_{i=1}^{N}\hat{\boldsymbol{Q}}_{i,IPW}(\boldsymbol{\beta})\hat{\boldsymbol{Q}}_{i,IPW}(\boldsymbol{\beta})^T$, and

$$\hat{\boldsymbol{t}}_i(\boldsymbol{\beta})=\hat{\boldsymbol{Q}}_{i,IPW}(\boldsymbol{\beta})-\left\{\frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{Q}}_{i,IPW}(\boldsymbol{\beta})\boldsymbol{\psi}_i(\hat{\boldsymbol{\alpha}})^T\right\}\left\{\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\psi}_i(\hat{\boldsymbol{\alpha}})\boldsymbol{\psi}_i(\hat{\boldsymbol{\alpha}})^T\right\}^{-1}\boldsymbol{\psi}_i(\hat{\boldsymbol{\alpha}}).$$

Here $p_{ij}(\boldsymbol{\beta},\boldsymbol{\alpha})$ is similarly defined to that in (2.8), but is based on the IPW residual instead. The following Corollary II.3 provides a consistent estimator for the asymptotic variance of $\hat{\boldsymbol{\beta}}_{IPW}$.

Corollary II.3. *Under Assumptions 2 in Section 2.7, if $\boldsymbol{\alpha}_*=\boldsymbol{\alpha}_0$, we have*

$$\left[\hat{\boldsymbol{I}}_{IPW}(\boldsymbol{\beta})^{-1}\left\{\frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{t}}_i(\boldsymbol{\beta})\hat{\boldsymbol{t}}_i(\boldsymbol{\beta})^T\right\}\hat{\boldsymbol{I}}_{IPW}(\boldsymbol{\beta})^{-1}\right]\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{IPW}}$$
$$\xrightarrow{p}\quad\boldsymbol{I}_{IPW}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)^{-1}\boldsymbol{V}_{\boldsymbol{\alpha},IPW}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)\boldsymbol{I}_{IPW}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)^{-1}.$$

2.4 Simulation Experiments

We evaluate the finite sample performance of the proposed CEL estimators using simulation experiments in this section. The simulation model contains two covariates, $X_1\sim\mathcal{N}(0,2^2)$ and $X_2\sim$ Bernoulli(0.5), as well as a surrogate variable generated by

$S = 1 + X_1 + X_2 + \epsilon_S$ with $\epsilon_S \sim \mathcal{N}(0, 2^2)$. The outcome of interest Y is generated by a linear model $Y = 1 + S + 0.6X_1 + 2X_2 + \epsilon_Y$, where $\epsilon_Y \sim \mathcal{N}(0, \sigma_Y^2)$ with covariate-dependent variance $\sigma_Y^2 = \exp(0.2 + 0.4S + 0.4X_1)$. Here X_1, X_2 and ϵ_S are independently sampled. A straightforward calculation shows that the conditional distribution of $Y|\boldsymbol{X}$ is a normal distribution, with mean $\mathrm{E}(Y|\boldsymbol{X}) = 2 + 1.6X_1 + 3X_2$ and variance $\mathrm{Var}(Y|\boldsymbol{X}) = 4 + \exp(0.92 + 0.8X_1 + 0.4X_2)$. The missing data mechanism is set to be $\mathrm{logit}\{\pi(S, \boldsymbol{X})\} = 0.5 - 0.2S + 0.6X_1 - 0.2X_2$, under which approximately 50% of subjects have missing Y in our generated data. Therefore, the true parameter values used in our simulation are $\boldsymbol{\beta}_0 = (\beta_1, \beta_2, \beta_3)^T = (2, 1.6, 3)^T$, $\boldsymbol{\alpha}_0 = (0.5, -0.2, 0.6, -0.2)^T$, and $\boldsymbol{\gamma}_0 = (1, 1, 0.6, 2)^T$.

We compare the proposed CEL estimators with the IPW, AIPW, HAN, RLSR, CLQ and QZL estimators under three different scenarios: (i) both $\pi(S, \boldsymbol{X})$ and $\mathrm{E}(Y|S, \boldsymbol{X})$ are correctly modeled; (ii) only $\pi(S, \boldsymbol{X})$ is correctly modeled; and (iii) only $\mathrm{E}(Y|S, \boldsymbol{X})$ is correctly modeled. For the second scenario, $\mathrm{E}(Y|S, \boldsymbol{X})$ is incorrectly modeled as $\mathrm{E}(Y|S, \boldsymbol{X}) = \gamma_1 + \gamma_2 X_1$, and for the third scenario, $\pi(S, \boldsymbol{X})$ is incorrectly modeled as $\mathrm{logit}\{\pi(S, \boldsymbol{X})\} = \alpha_1 + \alpha_2 S + \alpha_3 X_2$.

In each scenario, the six competitors are derived based on the estimating function $\boldsymbol{U}(\boldsymbol{\beta}; Y, \boldsymbol{X}) = \boldsymbol{X}\mathrm{Var}(Y|\boldsymbol{X})^{-1}(Y - \boldsymbol{X}^T\boldsymbol{\beta})$, where $\mathrm{Var}(Y|\boldsymbol{X})$ is specified in three different ways; namely, $V_1 = 1$, $V_2 = \theta_1 + \exp(\theta_2 + \theta_3 X_1 + \theta_4 X_2)$ and $V_3 = 4 + \exp(0.92 + 0.8X_1 + 0.4X_2)$, with $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)^T$ in V_2 being unknown. To estimate $\boldsymbol{\theta}$, we first calculate the residual $\tilde{\epsilon} = Y - \boldsymbol{X}^T\tilde{\boldsymbol{\beta}}$ for subjects whose outcome is observed, where $\tilde{\boldsymbol{\beta}}$ is the IPW estimator based on $\boldsymbol{U}(\boldsymbol{\beta}; Y, \boldsymbol{X}) = \boldsymbol{X}(Y - \boldsymbol{X}^T\boldsymbol{\beta})$ with weight $R/\pi(\boldsymbol{\alpha}_0; S, \boldsymbol{X})$. Note that the true value $\pi(S, \boldsymbol{X})$ is employed here to ensure that $\tilde{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}_0$. In practical studies where $\pi(S, \boldsymbol{X})$ is unknown, the six competitors can not take this advantage any more. We then minimize the least

square objective function $\left[\log\tilde{\epsilon}^2 - \log\left\{\theta_1 + \exp(\theta_2 + \theta_3 X_1 + \theta_4 X_2)\right\}\right]^2$ with respect to $\boldsymbol{\theta}$ over all subjects whose residual has been calculated. Here the log transformation is used to ensure that the estimated value of $\mathrm{Var}(Y|\boldsymbol{X})$ is always positive. In practice, $\mathrm{Var}(Y|\boldsymbol{X})$ may also be jointly estimated with the mean model through the generalized method of moments (Hansen, 1982).

To establish the benchmark for the comparison, we also include an estimator based on fully observed data. This estimator, denoted as IDEAL, is derived based on $\boldsymbol{U}(\boldsymbol{\beta}; Y, \boldsymbol{X}) = \boldsymbol{X}V_2^{-1}(Y - \boldsymbol{X}^T\boldsymbol{\beta})$, where V_2 is estimated by a procedure similar to what has been described above. We consider two cases of sample size, $N = 200$ and $N = 800$, and the corresponding results are summarized in Table 2.1 and Table 2.2, respectively, based on 500 simulations. We employ the Gaussian kernel to calculate the weights for the CEL method, where X_1 is standardized to have mean 0 and variance 1 to facilitate the bandwidth selection via cross-validation criterion (2.9).

Table 2.1: Comparison of different estimators (N=200). The numbers have been multiplied by 100. For each estimator, three statistics are reported: bias, empirical standard error (the number in ( )), and mean square error (the number in [ ]). For CEL-AIPW and CEL-IPW estimators, the number in { } is the mean of estimated standard error based on either Theorem II.3 or Corollary II.3. For CEL-AIPW estimator, the number in ⟨ ⟩ is the mean of estimated standard error based on Theorem II.4.

		both models correct			correct $\pi(\boldsymbol{S}, \boldsymbol{X})$			correct $\mathrm{E}(Y|\boldsymbol{S}, \boldsymbol{X})$		
method	$\mathrm{Var}(Y|\boldsymbol{X})$	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
IDEAL		-1	-2	-4	-1	-2	-4	-1	-2	-4
		(29)	(13)	(41)	(29)	(13)	(41)	(29)	(13)	(41)
		[9]	[2]	[17]	[9]	[2]	[17]	[9]	[2]	[17]
CEL-IPW		-1	0	-4	-1	0	-4	-49	7	-12
		(40)	(22)	(68)	(40)	(22)	(68)	(39)	(21)	(63)
		[16]	[5]	[47]	[16]	[5]	[47]	[39]	[5]	[41]
		{40}	{19}	{63}	{40}	{19}	{63}	{38}	{17}	{56}
IPW	V_1	-1	0	-5	-1	0	-5	-49	6	-12
		(44)	(28)	(76)	(44)	(28)	(76)	(44)	(32)	(76)
		[19]	[8]	[57]	[19]	[8]	[57]	[43]	[11]	[60]
	V_2	-1	0	-5	-1	0	-5	-49	5	-13
		(42)	(23)	(70)	(42)	(23)	(70)	(39)	(21)	(60)
		[17]	[5]	[50]	[17]	[5]	[50]	[39]	[5]	[38]
	V_3	0	2	-5	0	2	-5	-49	8	-12
		(41)	(21)	(70)	(41)	(21)	(70)	(37)	(18)	(59)
		[16]	[5]	[49]	[16]	[5]	[49]	[38]	[4]	[36]
CEL-AIPW		0	-2	-3	1	1	-7	-1	-1	-2
		(36)	(18)	(57)	(39)	(23)	(71)	(37)	(18)	(53)
		[13]	[3]	[33]	[15]	[5]	[51]	[14]	[3]	[29]
		{39}	{15}	{57}	{40}	{21}	{68}	{36}	{15}	{49}

		⟨36⟩	⟨16⟩	⟨53⟩	⟨41⟩	⟨22⟩	⟨71⟩	⟨36⟩	⟨15⟩	⟨49⟩
		0	-2	-4	0	0	-7	-1	-1	-3
	V_1	(41)	(26)	(67)	(44)	(30)	(81)	(42)	(31)	(72)
		[17]	[7]	[45]	[19]	[9]	[66]	[18]	[10]	[52]
		-5	-2	3	-2	-2	-1	-4	-2	3
AIPW	V_2	(59)	(33)	(117)	(110)	(82)	(225)	(54)	(27)	(96)
		[35]	[11]	[138]	[120]	[67]	[507]	[29]	[7]	[92]
		1	-1	-4	3	3	-8	-1	-1	-3
	V_3	(33)	(16)	(52)	(46)	(31)	(88)	(35)	(17)	(49)
		[11]	[3]	[27]	[21]	[10]	[78]	[12]	[3]	[24]
		0	-2	-3	-1	0	-5	-1	-2	-4
	V_1	(41)	(27)	(68)	(44)	(28)	(74)	(44)	(34)	(82)
		[16]	[7]	[46]	[19]	[8]	[55]	[19]	[12]	[67]
		-4	-1	3	1	2	-9	-5	-2	2
HAN	V_2	(56)	(33)	(113)	(44)	(30)	(77)	(54)	(24)	(97)
		[32]	[11]	[127]	[20]	[9]	[60]	[29]	[6]	[94]
		0	0	-3	1	3	-6	-1	-1	-3
	V_3	(34)	(16)	(52)	(40)	(39)	(68)	(35)	(17)	(55)
		[11]	[3]	[27]	[16]	[15]	[46]	[12]	[3]	[30]
		0	-2	-2	-	-	-	0	-2	-3
	V_1	(42)	(25)	(63)	-	-	-	(45)	(31)	(75)
		[18]	[6]	[40]	-	-	-	[20]	[10]	[56]
		-5	-2	5	-	-	-	-5	-2	5
RLSR	V_2	(57)	(24)	(111)	-	-	-	(58)	(25)	(112)
		[33]	[6]	[123]	-	-	-	[34]	[6]	[126]
		0	-1	-2	-	-	-	0	-1	-3
	V_3	(32)	(14)	(45)	-	-	-	(32)	(13)	(46)
		[10]	[2]	[20]	-	-	-	[10]	[2]	[21]
		1	-1	-4	-8	4	-7	-3	-2	1
	V_1	(43)	(28)	(72)	(104)	(52)	(147)	(128)	(44)	(182)
		[19]	[8]	[53]	[109]	[27]	[215]	[163]	[19]	[331]
		-5	-1	2	-2	2	-11	-11	-4	3
CLQ	V_2	(73)	(84)	(163)	(64)	(35)	(122)	(149)	(38)	(211)
		[54]	[70]	[265]	[41]	[12]	[149]	[222]	[15]	[444]
		2	2	-5	0	3	-4	-30	-4	0
	V_3	(40)	(16)	(67)	(40)	(21)	(70)	(39)	(17)	(64)
		[16]	[3]	[45]	[16]	[5]	[49]	[24]	[3]	[41]
		-3	-3	-4	1	4	-8	-31	-45	-23
	V_1	(42)	(30)	(72)	(45)	(32)	(77)	(46)	(31)	(83)
		[18]	[9]	[51]	[20]	[11]	[60]	[31]	[30]	[74]
		-3	-2	-4	-1	4	-6	-20	-21	-5
QZL	V_2	(36)	(20)	(58)	(43)	(26)	(70)	(59)	(32)	(89)
		[13]	[4]	[33]	[18]	[7]	[50]	[38]	[15]	[79]
		-2	0	-2	-1	6	-4	-8	-10	-3
	V_3	(34)	(18)	(54)	(40)	(24)	(69)	(42)	(20)	(61)
		[11]	[3]	[29]	[16]	[6]	[47]	[18]	[5]	[38]

Table 2.2: Comparison of different estimators (N=800). The numbers have been multiplied by 100. For each estimator, three statistics are reported: bias, empirical standard error (the number in ( )), and mean square error (the number in [ ]). For CEL-AIPW and CEL-IPW estimators, the number in { } is the mean of estimated standard error based on either Theorem II.3 or Corollary II.3. For CEL-AIPW estimator, the number in ⟨ ⟩ is the mean of estimated standard error based on Theorem II.4.

		both models correct			correct $\pi(\boldsymbol{S},\boldsymbol{X})$			correct $\mathrm{E}(Y\|\boldsymbol{S},\boldsymbol{X})$		
method	$\mathrm{Var}(Y\|\boldsymbol{X})$	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
		1	0	-1	1	0	-1	1	0	-1
IDEAL										

		(15)	(6)	(19)	(15)	(6)	(19)	(15)	(6)	(19)
		[2]	[0]	[3]	[2]	[0]	[3]	[2]	[0]	[3]
CEL-IPW		3	0	-3	3	0	-3	-46	7	-12
		(21)	(11)	(32)	(21)	(11)	(32)	(19)	(10)	(29)
		[4]	[1]	[10]	[4]	[1]	[1]	[24]	[2]	[10]
		{20}	{10}	{33}	{20}	{10}	{33}	{19}	{9}	{28}
IPW	V_1	3	1	-4	3	1	-4	-46	8	-12
		(23)	(14)	(37)	(23)	(14)	(37)	(22)	(16)	(37)
		[5]	[2]	[14]	[5]	[2]	[14]	[26]	[3]	[15]
	V_2	3	0	-3	3	0	-3	-46	7	-12
		(22)	(12)	(35)	(22)	(12)	(35)	(19)	(10)	(29)
		[5]	[1]	[12]	[5]	[1]	[12]	[25]	[1]	[10]
	V_3	3	1	-2	3	1	-2	-46	8	-11
		(21)	(11)	(34)	(21)	(11)	(34)	(19)	(9)	(27)
		[5]	[1]	[12]	[5]	[1]	[12]	[25]	[1]	[9]
CEL-AIPW		3	0	-3	3	1	-4	2	0	-2
		(18)	(9)	(26)	(20)	(11)	(33)	(18)	(9)	(24)
		[3]	[1]	[7]	[4]	[1]	[11]	[3]	[1]	[6]
		{19}	{7}	{28}	{20}	{11}	{34}	{18}	{8}	{23}
		⟨18⟩	⟨8⟩	⟨25⟩	⟨20⟩	⟨11⟩	⟨35⟩	⟨18⟩	⟨7⟩	⟨23⟩
AIPW	V_1	3	1	-3	3	1	-4	3	1	-3
		(21)	(13)	(33)	(22)	(15)	(38)	(21)	(16)	(35)
		[4]	[2]	[11]	[5]	[2]	[15]	[5]	[2]	[12]
	V_2	2	-1	-3	0	-1	-1	2	0	-1
		(22)	(13)	(43)	(45)	(33)	(45)	(19)	(10)	(28)
		[5]	[2]	[19]	[21]	[11]	[20]	[4]	[1]	[8]
	V_3	2	1	-2	3	1	-3	2	1	-1
		(18)	(7)	(24)	(23)	(15)	(39)	(18)	(9)	(23)
		[3]	[1]	[6]	[5]	[2]	[15]	[3]	[1]	[5]
HAN	V_1	3	1	-3	3	1	-4	3	1	-3
		(21)	(13)	(33)	(22)	(14)	(36)	(22)	(18)	(40)
		[4]	[2]	[11]	[5]	[2]	[13]	[5]	[3]	[16]
	V_2	1	0	-1	3	1	-4	1	0	-1
		(19)	(9)	(27)	(23)	(12)	(37)	(20)	(11)	(30)
		[4]	[1]	[8]	[5]	[1]	[14]	[4]	[1]	[9]
	V_3	2	1	-1	3	2	-3	2	0	-2
		(18)	(7)	(24)	(21)	(11)	(33)	(18)	(9)	(26)
		[3]	[1]	[6]	[4]	[1]	[11]	[3]	[1]	[7]
RLSR	V_1	3	1	-2	-	-	-	3	1	-3
		(21)	(13)	(30)	-	-	-	(22)	(16)	(36)
		[5]	[2]	[9]	-	-	-	[5]	[3]	[13]
	V_2	0	0	1	-	-	-	1	0	0
		(30)	(8)	(29)	-	-	-	(20)	(8)	(28)
		[9]	[1]	[9]	-	-	-	[4]	[1]	[8]
	V_3	2	0	-1	-	-	-	2	0	-1
		(17)	(7)	(22)	-	-	-	(17)	(7)	(22)
		[3]	[0]	[5]	-	-	-	[3]	[0]	[5]
CLQ	V_1	3	1	-3	3	0	-1	2	1	-2
		(21)	(13)	(35)	(31)	(34)	(78)	(27)	(18)	(46)
		[5]	[2]	[12]	[9]	[11]	[62]	[7]	[3]	[21]
	V_2	0	1	1	2	0	-4	-6	-3	4
		(33)	(18)	(64)	(24)	(18)	(38)	(39)	(21)	(51)
		[11]	[3]	[41]	[6]	[3]	[15]	[15]	[5]	[26]
	V_3	3	1	-2	3	1	-2	-31	-5	1
		(21)	(8)	(33)	(21)	(11)	(34)	(21)	(10)	(32)
		[5]	[1]	[11]	[5]	[1]	[12]	[14]	[1]	[10]
QZL	V_1	2	0	-3	3	2	-4	-31	-50	-26
		(21)	(14)	(33)	(22)	(14)	(36)	(25)	(17)	(44)

	[4]	[2]	[11]	[5]	[2]	[13]	[16]	[28]	[26]
	1	0	-2	3	2	-3	-12	-13	-1
V_2	(18)	(9)	(25)	(23)	(11)	(35)	(27)	(19)	(34)
	[3]	[1]	[6]	[5]	[1]	[12]	[8]	[5]	[11]
	2	1	-1	3	2	-2	-4	-6	0
V_3	(18)	(8)	(24)	(21)	(11)	(33)	(20)	(10)	(27)
	[3]	[1]	[6]	[4]	[1]	[11]	[4]	[1]	[7]

When both $\pi(S, \boldsymbol{X})$ and $\mathrm{E}(Y|S, \boldsymbol{X})$ are correctly modeled, we have the following summary points based on Table 2.1 and Table 2.2.

(i) The total mean square error (MSE) of the IPW estimator decreases as the model for $\mathrm{Var}(Y|\boldsymbol{X})$ gets closer to the truth, from V_1 to V_3. The CEL-IPW estimator further reduces the total MSE. Even compared to the IPW estimator employing V_3, the CEL-IPW estimator still reduces the total MSE by 11% at $N = 800$.

(ii) The AIPW, HAN and RLSR estimators have similar performance by inspecting their total MSE, especially at $N = 800$, confirming their asymptotic equivalence. When V_2 is employed, those three estimators have poor numerical performance at $N = 200$, which, however, substantially improves as N increases to 800. The CEL-AIPW estimator has smaller total MSE than those three estimators, except when they use the true value of $\mathrm{Var}(Y|\boldsymbol{X})$ (i.e. V_3), in which case, however, as N increases, the difference between the total MSE diminishes. Specifically, the ratio of the total MSE of the CEL-AIPW estimator over that of the AIPW, HAN and RLSR estimators using V_3 drops from 1.21, 1.21 and 1.51 to 1.13, 1.17 and 1.35, respectively, as N increases from 200 to 800. This observation provides a numerical support to our theory that the CEL-AIPW estimator attains the semiparametric efficiency bound, hence should be more efficient, as $N \to \infty$, than the other three estimators regardless of how

$\mathrm{Var}(Y|\boldsymbol{X})$ is modeled. The CEL-AIPW estimator has smaller total MSE than the CEL-IPW estimator as expected, since correctly modeling $\mathrm{E}(Y|S,\boldsymbol{X})$ improves efficiency.

(iii) The CEL-AIPW estimator has smaller total MSE than the CLQ estimator under all three different models for $\mathrm{Var}(Y|\boldsymbol{X})$, and has similar total MSE to the QZL estimator using the true value of $\mathrm{Var}(Y|\boldsymbol{X})$. As N increases, the CEL-AIPW estimator will become more efficient than both the CLQ and QZL estimators regardless of how $\mathrm{Var}(Y|\boldsymbol{X})$ is modeled.

When only $\pi(S,\boldsymbol{X})$ is correctly modeled, we summarize the following points.

(i) Under the same model for $\mathrm{Var}(Y|\boldsymbol{X})$, the AIPW estimator has larger total MSE than the IPW estimator, mainly due to the incorrect modeling of $\mathrm{E}(Y|S,\boldsymbol{X})$. This lack of efficiency of the AIPW estimator when the optimal augmentation term is incorrectly modeled is in full agreement with findings reported in the literature (Rubin and van der Laan 2008; Cao et al. 2009). In this case, it has been shown that both the HAN and RLSR estimators achieve higher efficiency than the IPW and AIPW estimators. Since the implementation of the RLSR estimator requires that the dimension of $\boldsymbol{\gamma}$ is no smaller than that of $\boldsymbol{\beta}$, which is not satisfied in the current scenario, we only report the results of the HAN estimator. The efficiency improvement of the HAN estimator over the AIPW estimator is apparent by comparing their total MSE, and the improvement over the IPW estimator is less obvious but can still be observed at $N = 800$. Due to the incorrect modeling of $\mathrm{E}(Y|S,\boldsymbol{X})$ again, the CEL-AIPW estimator has larger total MSE than the CEL-IPW estimator.

(ii) The CEL-AIPW estimator is apparently superior to the AIPW estimator, judg-

ing from its smaller total MSE under all three different models for $\mathrm{Var}(Y|\boldsymbol{X})$. Although hard to make theoretical comparison, the CEL-AIPW estimator appears to have comparable or better efficiency than the HAN, CLQ and QZL estimators even when they use the true value of $\mathrm{Var}(Y|\boldsymbol{X})$, as the ratio of the total MSE of the CEL-AIPW estimator over that of the latter three estimators using V_3 is 0.98, 0.94 and 0.99, respectively, at $N = 800$.

When only $\mathrm{E}(Y|S, \boldsymbol{X})$ is correctly modeled, we summarize the following points.

(i) The IPW, CEL-IPW, CLQ and QZL estimators are clearly biased. In contrast, the AIPW, CEL-AIPW, HAN and RLSR estimators have ignorable bias, due to their consistency guaranteed by the double robustness property.

(ii) Theoretical comparison of the efficiency between the CEL-AIPW estimator and the other three doubly robust estimators does not provide any definitive conclusions. However, judging from the total MSE, the CEL-AIPW estimator is superior to the other three estimators unless they use V_3, corresponding to their best case scenario. Even in that case, the CEL-AIPW estimator still has comparable total MSE. Specifically, the ratio of the total MSE of the CEL-AIPW estimator over that of the AIPW, HAN and RLSR estimators using V_3 is 1.07, 0.94 and 1.20, respectively, at $N = 800$.

In addition to all of the above points, it is also of interest to observe the numerical evidence on the convergence of the asymptotic variance estimators given in Theorem II.3, Corollary II.3 and Theorem II.4. The convergence is well demonstrated by the comparison across different sample sizes. When $N = 200$, these estimators tend to have slight underestimation, but this underestimation disappears when the sample size increases to $N = 800$.

Since our proposed CEL estimation procedure involves nonparametric calculation of weight w_{ij}, one important question is whether increasing the number of covariates would substantially affect the estimation of regression parameters. To assess such an impact, we conduct the following simulation experiment. The simulation model now involves four covariates, $X_1 \sim \mathcal{N}(0, 2^2)$, $X_2 \sim$ Bernoulli(0.5), $X_3 \sim \mathcal{N}(0, 1^2)$ and $X_4 \sim \mathcal{N}(0, 1^2)$. The surrogate variable is given by $S = 1 + X_1 + X_2 + X_3 + X_4 + \epsilon_S$ with $\epsilon_S \sim \mathcal{N}(0, 2^2)$, and the outcome Y is generated by $Y = 1 + S + 0.6X_1 + 2X_2 + 0.5X_3 + 0.5X_4 + \epsilon_Y$, where $\epsilon_Y \sim \mathcal{N}(0, \sigma_Y^2)$ with $\sigma_Y^2 = \exp(0.92 + 0.8X_1 + 0.4X_2)$. Here X_1, X_2, X_3, X_4 and ϵ_S are independently sampled. For this model, $Y|\boldsymbol{X}$ has a normal distribution with mean $\mathrm{E}(Y|\boldsymbol{X}) = 2 + 1.6X_1 + 3X_2 + 1.5X_3 + 1.5X_4$ and variance $\mathrm{Var}(Y|\boldsymbol{X}) = 4 + \exp(0.92 + 0.8X_1 + 0.4X_2)$. The missingness mechanism is set as logit $\{\pi(S, \boldsymbol{X})\} = 0.5 - 0.2S + 0.6X_1 - 0.2X_2 + 0.2X_3 + 0.2X_4$, under which approximately 48% of subjects have missing Y in the generated data. Compared to the previous simulation model, this new model has two extra continuous covariates X_3 and X_4, and has $\boldsymbol{\beta}_0 = (\beta_1, \cdots, \beta_5)^T = (2, 1.6, 3, 1.5, 1.5)^T$, $\boldsymbol{\alpha}_0 = (0.5, -0.2, 0.6, -0.2, 0.2, 0.2)^T$ and $\boldsymbol{\gamma}_0 = (1, 1, 0.6, 2, 0.5, 0.5)^T$. When $\pi(S, \boldsymbol{X})$ or $\mathrm{E}(Y|S, \boldsymbol{X})$ is incorrectly modeled, they are incorrectly modeled as before. The numerical performance of our proposed CEL-AIPW estimator based on the new simulation model is summarized in Table 2.3 using 500 simulations. In Table 2.3, the IDEAL estimator and the AIPW estimator based on V_3 are also included for the sake of comparison. Product Gaussian kernel is used for weight calculation, where the continuous covariates are standardized to have mean 0 and variance 1 for the bandwidth selection via (2.9).

Table 2.3: The effect of an increased dimension in covariates on the performance of CEL-AIPW estimator. The numbers have been multiplied by 100. For each estimator presented, three statistics are reported: bias, empirical standard error (the number in ( )), and mean square error (the number in [ ]). For CEL-AIPW estimator, the number in { } and the number in $\langle\,\rangle$ are the mean of estimated standard error based on Theorem II.3 and Theorem II.4, respectively. AIPW$_{opt}$ is the AIPW estimator using the true value of Var$(Y|\boldsymbol{X})$.

N	method	β_1	β_2	β_3	β_4	β_5
200	IDEAL	-1	0	2	1	0
		(30)	(13)	(40)	(19)	(21)
		[9]	[2]	[16]	[4]	[4]
	both models correct					
	AIPW$_{opt}$	-1	0	1	3	2
		(38)	(22)	(59)	(28)	(29)
		[14]	[5]	[35]	[8]	[9]
	CEL-AIPW	-1	0	0	2	2
		(37)	(21)	(60)	(30)	(30)
		[14]	[4]	[36]	[9]	[9]
		{39}	{15}	{57}	{28}	{28}
		$\langle 34\rangle$	$\langle 16\rangle$	$\langle 51\rangle$	$\langle 25\rangle$	$\langle 25\rangle$
	correct $\pi(\boldsymbol{S},\boldsymbol{X})$					
	AIPW$_{opt}$	2	-1	-8	0	-4
		(54)	(40)	(101)	(55)	(55)
		[29]	[16]	[102]	[30]	[30]
	CEL-AIPW	2	1	-6	-3	-1
		(43)	(26)	(76)	(40)	(39)
		[19]	[7]	[58]	[16]	[15]
		{42}	{23}	{73}	{37}	{37}
		$\langle 44\rangle$	$\langle 24\rangle$	$\langle 76\rangle$	$\langle 39\rangle$	$\langle 39\rangle$
	correct E$(Y|\boldsymbol{S},\boldsymbol{X})$					
	AIPW$_{opt}$	-1	0	1	1	1
		(37)	(21)	(57)	(28)	(28)
		[14]	[5]	[32]	[8]	[8]
	CEL-AIPW	-1	-1	0	1	2
		(37)	(23)	(62)	(33)	(31)
		[14]	[5]	[39]	[11]	[10]
		{35}	{16}	{50}	{24}	{24}
		$\langle 35\rangle$	$\langle 16\rangle$	$\langle 50\rangle$	$\langle 24\rangle$	$\langle 24\rangle$
800	IDEAL	0	0	0	0	0
		(15)	(6)	(20)	(10)	(9)
		[2]	[0]	[4]	[1]	[1]
	both models correct					
	AIPW$_{opt}$	0	0	0	0	0
		(17)	(7)	(24)	(11)	(11)
		[3]	[1]	[6]	[1]	[1]
	CEL-AIPW	0	0	0	1	1
		(18)	(9)	(25)	(13)	(13)
		[3]	[1]	[6]	[2]	[2]
		{19}	{7}	{27}	{13}	{13}
		$\langle 17\rangle$	$\langle 8\rangle$	$\langle 24\rangle$	$\langle 12\rangle$	$\langle 12\rangle$
	correct $\pi(\boldsymbol{S},\boldsymbol{X})$					
	AIPW$_{opt}$	0	1	-2	-1	0
		(25)	(17)	(45)	(25)	(26)
		[6]	[3]	[20]	[6]	[7]
	CEL-AIPW	0	1	-2	-1	0
		(21)	(12)	(36)	(19)	(20)
		[5]	[2]	[13]	[4]	[4]
		{21}	{12}	{37}	{19}	{19}
		$\langle 22\rangle$	$\langle 12\rangle$	$\langle 38\rangle$	$\langle 20\rangle$	$\langle 20\rangle$
	correct E$(Y|\boldsymbol{S},\boldsymbol{X})$					

	-1	0	0	0	0
AIPW$_{opt}$	(18)	(10)	(25)	(12)	(12)
	[3]	[1]	[6]	[1]	[2]
	0	0	0	0	1
	(19)	(10)	(27)	(15)	(15)
CEL-AIPW	[4]	[1]	[7]	[2]	[2]
	{17}	{8}	{24}	{12}	{12}
	⟨17⟩	⟨8⟩	⟨24⟩	⟨12⟩	⟨12⟩

From Table 2.3, no substantial effect on the estimation of regression coefficients resulting from the inclusion of two extra continuous covariates has been observed. The reason that an increased dimension in covariates does not have a dramatic impact on our proposed method might be that, from (2.8) and the Newton-Raphson algorithm described in Section 2.2, the calculation of our proposed estimators implicitly utilizes the conditional empirical probability p_{ij}, which may be regarded as an "upgraded" variant of the weight w_{ij} after incorporating extra information implied by the fact that $\mathrm{E}\{g(\boldsymbol{\beta}_0)|\boldsymbol{X}\} = 0$. As a result, the impact of an increased dimension in covariates on parameter estimation may be mitigated by the utilization of p_{ij}.

2.5 Data Application

We now apply the proposed CEL method to an intervention study for adolescent children of parents with HIV (Rotheram-Borus et al. 2004). In this study, a total of 307 parents having HIV with adolescent children were recruited from the Division of AIDS Services in New York City, and 423 adolescents from these families were eligible for study participation. After recruitment, each parent and each adolescent received a baseline interview, which collected information on background characteristics as well as the measurements for adolescent assessment, such as emotional distress and somatic symptoms. At the end of the baseline interview, participant families were randomly assigned either to the intervention arm or to the control arm. The

intervention in this study was designed using social learning theory and cognitive-behavioral principles (Bandura 1994). Depending on the parents' phase of illness, families received the intervention in 3 different modules, which cover different aspects of information on the tasks for either parents or adolescents. The researchers followed up on the participants every 3 months for the first 2 years and every 6 months thereafter, until the end of 6 years. At each follow-up, measurements for adolescent assessment were collected.

In our analysis we use a subset of the data that contains assessments on adolescents' emotional distress, which were collected using the Brief Symptom Inventory (BSI). BSI is a commonly used psychological survey consisting of 53 items that belong to 9 sub-groups. Each item is associated with a psychiatric symptom and has a 0 to 4 rate scale. Subjects report values to each item according to the level that they have been troubled by the corresponding symptom in the past week, with 0 meaning "having not been troubled at all" and 4 meaning "having been troubled a lot". One scientifically interesting question is whether having parents with HIV has disparate impacts on the emotional distress between boys and girls during the delivery of intervention. Such a gender disparity, if it exists, may suggest the need for the development of gender-specific interventions that could result in more beneficial achievement. We try to answer this question using data collected at the end of the first year of intervention. The data are downloaded from "http://rem.ph.ucla.edu/rob/mld/data/tabdelimiteddata/bsitotal.txt", and detailed description about the data can be found in Weiss (2005).

The outcome variable is the global severity index, which is the average rating score over all 53 items. Due to the skewed distribution of the global severity index and the possibility of occurrence of value 0, following the analysis instruction in Weiss

(2005), we create a new outcome gsi by adding a small constant 1/53 to the global severity index and then taking the log-transformation with base 2. The following model is assumed:

$$gsi = \beta_1 + \beta_2 age + \beta_3 girl + \beta_4 int + \epsilon,$$

where age is the age of adolescent at the end of the first year of intervention, $girl$ is gender indicator with $girl = 0$ for boys and $girl = 1$ for girls, int is the intervention indicator with $int = 0$ for control arm and $int = 1$ for intervention arm, and ϵ is the error term that has mean 0 conditional on all three covariates. However, scores on gsi were only available for about half of the adolescents at the end of the first year of intervention. On the other hand, almost all adolescents had their baseline gsi score observed. Therefore, we treat the baseline gsi, denoted by $bgsi$, as a surrogate variable. To better model the missing mechanism, we create two dummy variables, namely $winter$ and $summer$, as indicators for the season ($winter$ indicates November through February, $summer$ indicates July through October, and the rest time of the calendar year is treated as reference) when the measurements at the end of the first year of intervention were taken. These two dummy variables are considered as extra surrogate variables. After removing adolescents who did not have scores on $bgsi$, we end up with $N = 420$ subjects, among which 204 did not have score on gsi (the missing data proportion is 49%). There are in total 221 girls and 199 boys, and 211 are in the intervention arm and 209 are in the control arm. The average age is 16 years old, with a standard deviation 2 years. Note that some parents contributed more than one adolescents, therefore measurements from these adolescents are correlated. In our illustration we ignore such correlation and treat all adolescents as independent.

Table 2.4: Results of modeling the missingness mechanism for the intervention study data ($N = 420$).

	est	se	z-value	p-value
constant	0.473	0.819	0.578	0.564
bgsi	-0.148	0.064	-2.316	0.021
winter	0.830	0.244	3.396	0.001
summer	0.513	0.245	2.089	0.037
age	-0.067	0.049	-1.353	0.176
girl	0.042	0.204	0.206	0.837
int	0.165	0.201	0.818	0.413

est: estimated value; se: estimated standard error.

Table 2.5: Estimation results for the intervention study data ($N = 420$).

	CEL-AIPW			CEL-IPW			complete-case analysis		
	est	se	p-value	est	se	p-value	est	se	p-value
constant	-5.531	1.098	0.000	-5.393	1.208	0.000	-5.231	1.195	0.000
age	0.164	0.066	0.013	0.164	0.074	0.027	0.139	0.071	0.053
girl	0.745	0.265	0.005	0.651	0.290	0.025	0.634	0.294	0.032
int	0.277	0.266	0.297	0.186	0.289	0.520	0.200	0.296	0.500

est: estimated value; se: estimated standard error. For CEL-AIPW method, the standard error is estimated based on Theorem II.4.

To model the missingness mechanism, we fit a logistic regression model, and the results are presented in Table 2.4. It is seen that having higher score on *bgsi* significantly increases the probability of missing the interview conducted at the end of the first year of intervention. The season when the interview was conducted also plays a significant role, in the sense that subjects were more likely to take the interview during *winter* and *summer* seasons compared to the rest time of year. A linear regression model is employed to model $\mathrm{E}(Y|\boldsymbol{S}, \boldsymbol{X})$ in the augmentation term.

Table 2.5 contains the parameter estimates and corresponding p-values based on the CEL-AIPW estimator, the CEL-IPW estimator and the complete case analysis. The disparity between the CEL-AIPW estimate and the other two estimates indicates that the missingness mechanism may not be adequately modeled, but the relationship between the global severity index and the auxiliary variables is well exploited in the the CEL-AIPW estimation. Indeed, through model building process, we found that it is reasonable to use linear regression to model $\mathrm{E}(Y|\boldsymbol{S}, \boldsymbol{X})$, where all the auxiliary variables have significant effects (p-value for *bgsi* is less than 0.0001, p-value for *winter* is 0.0003, and p-value for *summer* is 0.0480). All three estimation methods in Table 2.5 conclude that gender has a significant effect on the global severity index, whereas the effect of intervention is not significant. The age effect is significant based on both the CEL-AIPW and the CEL-IPW methods, but is only marginally significant based on the complete case analysis. Due to the possible inadequacy of modeling the missingness mechanism and the strong relationship between the global severity index and the auxiliary variables, parameter values estimated by the CEL-AIPW method seem to be more reliable. Based on the CEL-AIPW estimates, on average, one year increase in age leads to roughly 12% increase in the global severity index, and girls have their global severity index roughly 68% higher than boys, where each effect is interpreted by holding the others fixed. Therefore, having parents with HIV does have different impacts on the emotional distress of boys and girls during the delivery of intervention, at least after one year of the delivery.

2.6 Conclusions

In this chapter we proposed a CEL method for analyzing missing outcome data when some surrogate (or auxiliary) variables exist. We studied the asymptotic prop-

erties of both the CEL-IPW and CEL-AIPW estimators. The innovation of our proposed method is that our estimators are not based on any estimating functions. The circumvention of the construction of estimating functions enables our estimators to possibly achieve the maximum amount of efficiency. As has been illustrated by our simulation experiments, the efficiency of existing estimating-function-based methods relies heavily on how the estimating functions are constructed. Therefore, the circumvention of constructing estimating functions by our proposed CEL method may be highly advantageous.

Some additional findings obtained during the process of the development and the application of the CEL method are worth mentioning. First, the proposed CEL method enjoys high estimation efficiency when moderate to high level of heteroscedasticity exists, especially when such heteroscedasticity needs to be modeled. In other words, when homoscedasticity is a more reasonable assumption, our CEL estimators may not outperform some of the existing estimators, such as the IPW or AIPW estimators, due to the nonparametric calculation of the weights. Second, we employ a modified cross-validation criterion (2.9) along the lines suggested by Newey (1993) for the bandwidth selection. Although this criterion works reasonably well in our simulation experiments, it has not been theoretically justified yet. Future work on better criteria and related theory is needed. Third, when the number of covariates is large and the covariates vary in scales and/or types, the kernel-based weight calculation is challenged. Although through simulation experiments we have found that the inclusion of additional covariates may not have a dramatic impact on the performance of our proposed procedure, such an impact will become influential and can not be ignored as the number of covariates keeps increasing. Therefore, it is worthwhile to explore more flexible ways to calculate the weights. Last, but not least, the

CEL estimation procedure has a relatively heavy computational burden compared to the estimating-function-based methods. Future work on the development of fast algorithms to search for the CEL estimates is highly desired.

2.7 Technical Assumptions and Proofs

Let $\mathcal{B}$, $\mathcal{A}$ and $\mathcal{G}$ denote the domain of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, respectively. Let $\mathcal{B}_0 \subseteq \mathcal{B}$ be some closed ball around $\boldsymbol{\beta}_0$. Denote Kitamura et al. (2004) by KTA. Denote $\boldsymbol{G}_{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \mathrm{E}\{\partial g(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})/\partial \boldsymbol{\alpha} | \boldsymbol{X}\}$. For a matrix $\boldsymbol{H}$ with elements H_{ij}, define $\| \boldsymbol{H} \| = \sqrt{\sum_{i,j} H_{ij}^2}$.

Assumptions 1. *(i) For any $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, there exists $\mathcal{X}_{\boldsymbol{\beta},\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*}$ in the domain of $\boldsymbol{X}$, such that $P(\boldsymbol{x} \in \mathcal{X}_{\boldsymbol{\beta},\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*}) > 0$ and $E\{g(\boldsymbol{\beta}, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*)|\boldsymbol{X} = \boldsymbol{x}\} \neq 0$ for every $\boldsymbol{x} \in \mathcal{X}_{\boldsymbol{\beta},\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*}$. (ii) $E\{\sup_{\boldsymbol{\beta}\in\mathcal{B}} |g(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})|^m\} < \infty$ for some $m \geq 8$ for any $\boldsymbol{\alpha} \in \mathcal{A}$ and $\boldsymbol{\gamma} \in \mathcal{G}$. (iii) $\mathcal{B}$, $\mathcal{A}$ and $\mathcal{G}$ are compact. (iv) $\mu(\cdot)$ and $h(\cdot)$ are continuously differentiable. (v) $\pi(\boldsymbol{\alpha}; \boldsymbol{S}, \boldsymbol{X})$ is continuously differentiable with respect to $\boldsymbol{\alpha}$ for every $\boldsymbol{S}$ and $\boldsymbol{X}$, and $\pi(\boldsymbol{\alpha}; \boldsymbol{S}, \boldsymbol{X}) > \sigma > 0$ for all $\boldsymbol{\alpha} \in \mathcal{A}$ for some σ. (vi) $0 < \inf_{\boldsymbol{X},\boldsymbol{\beta}\in\mathcal{B}_0} V_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \leq \sup_{\boldsymbol{X},\boldsymbol{\beta}\in\mathcal{B}_0} V_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) < \infty$ for any $\boldsymbol{\alpha} \in \mathcal{A}$ and $\boldsymbol{\gamma} \in \mathcal{G}$, where $V_{AIPW}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = E\{g(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})^2|\boldsymbol{X}\}$. (vii) The domain of $\boldsymbol{X}^c$ is compact. (viii) $b_N \to 0$, $N^{1-2\nu-2/\delta} b_N^{2q} \to \infty$ and $N^{1-2\nu} b_N^{5q/2} \to \infty$ as $N \to \infty$, where $\nu \in (0, 1/2)$ and $\delta \geq 8$. (ix) $\hat{\lambda}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \in \{\lambda_i \in \mathbb{R} : | \lambda_i | \leq cN^{-1/m}\}$ for some $c > 0$, $i = 1, \cdots, N$.*

Remark: Assumptions 1 (vi) guarantees that the conditional variance of AIPW residual is invertible. The restrictions on b_N in Assumptions 1 (viii) follow that in Smith (2007). Assumption 1 (ix) is similar to Assumption 3.6 in KTA, and is only needed when the focus is on some neighborhood around $\boldsymbol{\beta}_0$. For example, it is needed in establishing the asymptotic normality of $\hat{\boldsymbol{\beta}}_{AIPW}$, but not needed in establishing

the consistency. Assumptions 1 (ix) is legitimate if AIPW residual has mean zero. Numerical implementation does not require Assumptions 1 (ix).

Assumptions 2. *(i) For any* $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$*, we can find* $\mathcal{X}_{\boldsymbol{\beta},\boldsymbol{\alpha}_*}$ *in the domain of* $\boldsymbol{X}$*, such that* $P(\boldsymbol{x} \in \mathcal{X}_{\boldsymbol{\beta},\boldsymbol{\alpha}_*}) > 0$ *and* $E\{f(\boldsymbol{\beta},\boldsymbol{\alpha}_*)|\boldsymbol{X}=\boldsymbol{x}\} \neq 0$ *for every* $\boldsymbol{x} \in \mathcal{X}_{\boldsymbol{\beta},\boldsymbol{\alpha}_*}$*. (ii)* $E\left\{\sup_{\boldsymbol{\beta}\in\mathcal{B}}|f(\boldsymbol{\beta},\boldsymbol{\alpha})|^m\right\} < \infty$ *for some* $m \geq 8$ *for any* $\boldsymbol{\alpha} \in \mathcal{A}$*. (iii)* $\mathcal{B}$ *and* $\mathcal{A}$ *are compact. (iv)* $\mu(\cdot)$ *is continuously differentiable. (v)* $\pi(\boldsymbol{\alpha};\boldsymbol{S},\boldsymbol{X})$ *is continuously differentiable with respect to* $\boldsymbol{\alpha}$ *for every* $\boldsymbol{S}$ *and* $\boldsymbol{X}$*, and* $\pi(\boldsymbol{\alpha};\boldsymbol{S},\boldsymbol{X}) > \sigma > 0$ *for all* $\boldsymbol{\alpha} \in \mathcal{A}$ *for some* σ*. (vi)* $0 < \inf_{\boldsymbol{X},\boldsymbol{\beta}\in\mathcal{B}_0} V_{IPW}(\boldsymbol{\beta},\boldsymbol{\alpha}) \leq \sup_{\boldsymbol{X},\boldsymbol{\beta}\in\mathcal{B}_0} V_{IPW}(\boldsymbol{\beta},\boldsymbol{\alpha}) < \infty$ *for any* $\boldsymbol{\alpha} \in \mathcal{A}$*, where* $V_{IPW}(\boldsymbol{\beta},\boldsymbol{\alpha}) = E\{f(\boldsymbol{\beta},\boldsymbol{\alpha})^2|\boldsymbol{X}\}$*. (vii) The domain of* $\boldsymbol{X}^c$ *is compact. (viii)* $b_N \to 0$*,* $N^{1-2\nu-2/\delta}b_N^{2q} \to \infty$ *and* $N^{1-2\nu}b_N^{5q/2} \to \infty$ *as* $N \to \infty$*, where* $\nu \in (0,1/2)$ *and* $\delta \geq 8$*. (ix)* $\hat{\lambda}_i(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}}) \in \left\{\lambda_i \in \mathbb{R} : |\ \lambda_i\ | \leq cN^{-1/m}\right\}$ *for some* $c > 0$*,* $i = 1,\cdots,N$*.*

In the following proofs, without causing any confusion, we suppress the subscript "AIPW" (except for that in $\hat{\boldsymbol{\beta}}_{AIPW}$) to save notations.

Proof of Theorem II.1. Under Assumptions 1, the same arguments as in the proof of Theorem 1 in KTA yields that $N^{1/m}L(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}}) \leq F(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}}) + o_p(1)$ for any $\boldsymbol{\beta} \in \mathcal{B}$, where

$$F(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}) = -\mathrm{E}\left[\frac{|\ \mathrm{E}\{g(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})|\boldsymbol{X}\}\ |^2}{1+|\ \mathrm{E}\{g(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})|\boldsymbol{X}\}\ |}\right]$$

is continuous with respect to $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. Therefore, we have

$$N^{1/m}L(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}}) \leq F(\boldsymbol{\beta},\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*) + o_p(1), \quad \text{for any } \boldsymbol{\beta} \in \mathcal{B}. \tag{2.14}$$

From Assumptions 1, for any $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, we have

$$F(\boldsymbol{\beta},\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*) \leq -\mathrm{E}\left[\mathbb{I}(\boldsymbol{X} \in \mathcal{X}_{\boldsymbol{\beta},\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*})\frac{|\ \mathrm{E}\{g(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})|\boldsymbol{X}\}\ |^2}{1+|\ \mathrm{E}\{g(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})|\boldsymbol{X}\}\ |}\right],$$

and the right-hand side of this inequality is strictly negative. Therefore, from (2.14), the continuity of $F(\boldsymbol{\beta}, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*)$ and the compactness of $\mathcal{B}$, for any $\delta > 0$, there exists $H(\delta) > 0$, such that

(2.15)

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}/B(\boldsymbol{\beta}_0, \delta)} N^{1/m} L(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \leq \sup_{\boldsymbol{\beta} \in \mathcal{B}/B(\boldsymbol{\beta}_0, \delta)} F(\boldsymbol{\beta}, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*) + o_p(1) \leq -H(\delta) + o_p(1),$$

where $B(\boldsymbol{\beta}_0, \delta)$ is the ball centering at $\boldsymbol{\beta}_0$ with radius δ.

On the other hand, under Assumptions 1, the same arguments as in the proof of (B.4) in KTA yields that $\max_{1 \leq i \leq N} \hat{\lambda}_i(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) = o_p(N^{-1/m})$ if either $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$ or $\boldsymbol{\gamma}_* = \boldsymbol{\gamma}_0$. Therefore, from

$$\begin{aligned} L(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) &= -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \log \left\{ 1 + \hat{\lambda}_i(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) g_j(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \right\} \\ &\geq -\frac{1}{N} \sum_{i=1}^{N} \left\{ \hat{\lambda}_i(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \sum_{j=1}^{N} w_{ij} g_j(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \right\} \end{aligned}$$

we have that $N^{1/m} L(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \geq o_p(1)$. This, together with (2.15) gives the desired result. □

To prove Theorem II.2, we first prove the following lemma.

Lemma.

$$\boldsymbol{G}_{\boldsymbol{\alpha}}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*) = -E\left\{ g(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*) \boldsymbol{\psi}(\boldsymbol{\alpha}_0)^T | \boldsymbol{X} \right\}.$$

Proof of Lemma. Given Y, $\boldsymbol{S}$ and $\boldsymbol{X}$, the binomial likelihood in (2.5) and $g(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ in (2.7) only depend on R, which we redenote as $p(R, \boldsymbol{\alpha}) = \pi(\boldsymbol{\alpha})^R \{1 - \pi(\boldsymbol{\alpha})\}^{1-R}$ and $g_R(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = g(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$, respectively. Simple calculation gives that, for any $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, $\mathrm{E}\{g_R(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) | Y, \boldsymbol{S}, \boldsymbol{X}\} = Y - \mu(\boldsymbol{X}^T \boldsymbol{\beta})$. On the other hand, we have $\mathrm{E}\{g_R(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) | Y, \boldsymbol{S}, \boldsymbol{X}\} = \sum_R g_R(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) p(R, \boldsymbol{\alpha})$. Therefore, for all Y, $\boldsymbol{S}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, we have $\sum_R g_R(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) p(R, \boldsymbol{\alpha}) = Y - \mu(\boldsymbol{X}^T \boldsymbol{\beta})$. Taking partial derivative

with respect to $\boldsymbol{\alpha}$ on both sides and setting $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$ and $\boldsymbol{\gamma} = \boldsymbol{\gamma}_*$ yields

$$\sum_R \frac{\partial g_R(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)}{\partial \boldsymbol{\alpha}} p(R, \boldsymbol{\alpha}_0) + \sum_R g_R(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*) \frac{\partial p(R, \boldsymbol{\alpha}_0)/\partial \boldsymbol{\alpha}}{p(R, \boldsymbol{\alpha}_0)} p(R, \boldsymbol{\alpha}_0) = \mathbf{0},$$

which is

$$\mathrm{E}\left\{ \frac{\partial g(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)}{\partial \boldsymbol{\alpha}} | Y, \boldsymbol{S}, \boldsymbol{X} \right\} = -\mathrm{E}\left\{ g(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*) \boldsymbol{\psi}(\boldsymbol{\alpha}_0)^T | Y, \boldsymbol{S}, \boldsymbol{X} \right\}.$$

Taking expectation conditional on $\boldsymbol{X}$ on both sides then gives the desired result. □

Proof of Theorem II.2. Since $\hat{\boldsymbol{\beta}}_{AIPW}$ satisfies the equation $\partial L(\hat{\boldsymbol{\beta}}_{AIPW}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})/\partial \boldsymbol{\beta}^T = 0$, by Taylor expansion, we have

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{AIPW} - \boldsymbol{\beta}_0) = \left\{ -\frac{1}{N} \frac{\partial^2 L(\tilde{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right\}^{-1} \left\{ \frac{1}{\sqrt{N}} \frac{\partial L(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\beta}^T} \right\}, \tag{2.16}$$

where $\tilde{\boldsymbol{\beta}}$ is some point between $\hat{\boldsymbol{\beta}}_{AIPW}$ and $\boldsymbol{\beta}_0$. Under Assumptions 1, using the same arguments as in the proof of Lemma C.1 in KTA, together with Theorem II.1, we have that

$$-\frac{1}{N} \frac{\partial^2 L(\tilde{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \xrightarrow{p} I(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*). \tag{2.17}$$

On the other hand, the same argument as the proof of (A.14) in KTA gives

$$\frac{1}{\sqrt{N}} \frac{\partial L(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\beta}^T} = \frac{1}{\sqrt{N}} \boldsymbol{A}(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) + o_p\left\{ \frac{1}{\sqrt{N}} \boldsymbol{A}(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \right\},$$

where

$$\frac{1}{\sqrt{N}} \boldsymbol{A}(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) =$$
$$\frac{1}{\sqrt{N}} \sum_{i=1}^N \left[\left\{ \sum_{j=1}^N w_{ij} \frac{\partial \mu(\boldsymbol{X}_j^T \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T} \right\} \left\{ \sum_{j=1}^N w_{ij} g_j(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})^2 \right\}^{-1} \left\{ \sum_{j=1}^N w_{ij} g_j(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \right\} \right].$$

A Taylor expansion around $\boldsymbol{\alpha}_*$ and $\boldsymbol{\gamma}_*$ then gives that

$$
\begin{aligned}
&\frac{1}{\sqrt{N}}\boldsymbol{A}(\boldsymbol{\beta}_0,\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}}) \\
= \;& \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[\left\{\sum_{j=1}^{N} w_{ij}\frac{\partial\mu(\boldsymbol{X}_j^T\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}^T}\right\}\left\{\sum_{j=1}^{N} w_{ij}g_j(\boldsymbol{\beta}_0,\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})^2\right\}^{-1}\left\{\sum_{j=1}^{N} w_{ij}g_j(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)\right\}\right] \\
&+\frac{1}{N}\sum_{i=1}^{N}\left[\left\{\sum_{j=1}^{N} w_{ij}\frac{\partial\mu(\boldsymbol{X}_j^T\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}^T}\right\}\left\{\sum_{j=1}^{N} w_{ij}g_j(\boldsymbol{\beta}_0,\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})^2\right\}^{-1}\left\{\sum_{j=1}^{N} w_{ij}\frac{\partial g_j(\boldsymbol{\beta}_0,\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\gamma}})}{\partial\boldsymbol{\alpha}}\right\}\right] \\
&\sqrt{N}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_*) \\
&+\frac{1}{N}\sum_{i=1}^{N}\left[\left\{\sum_{j=1}^{N} w_{ij}\frac{\partial\mu(\boldsymbol{X}_j^T\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}^T}\right\}\left\{\sum_{j=1}^{N} w_{ij}g_j(\boldsymbol{\beta}_0,\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})^2\right\}^{-1}\left\{\sum_{j=1}^{N} w_{ij}\frac{\partial g_j(\boldsymbol{\beta}_0,\tilde{\boldsymbol{\alpha}},\tilde{\boldsymbol{\gamma}})}{\partial\boldsymbol{\gamma}}\right\}\right] \\
&\sqrt{N}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_*),
\end{aligned}
$$

where $\tilde{\boldsymbol{\alpha}}$ is some point between $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}_*$, and $\tilde{\boldsymbol{\gamma}}$ is some point between $\hat{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}_*$.

Under our assumptions,

$$
\begin{aligned}
\frac{1}{\sqrt{N}}\boldsymbol{A}(\boldsymbol{\beta}_0,\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}}) \;=\; & \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{Q}_i(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*) \\
&+\mathrm{E}\left\{\frac{\partial\mu(\boldsymbol{X}_i^T\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}^T}V_i(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)^{-1}\boldsymbol{G}_{i,\boldsymbol{\alpha}}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)\right\}\sqrt{N}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_*) \\
&+\mathrm{E}\left\{\frac{\partial\mu(\boldsymbol{X}_i^T\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}^T}V_i(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)^{-1}\boldsymbol{G}_{i,\boldsymbol{\gamma}}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)\right\}\sqrt{N}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_*) \\
&+o_p(1).
\end{aligned}
$$

When $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$, we have that

$$
\boldsymbol{G}_{i,\boldsymbol{\gamma}}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*) = \mathrm{E}\left[\mathrm{E}\left\{-\frac{R_i-\pi_i(\boldsymbol{\alpha}_0)}{\pi_i(\boldsymbol{\alpha}_0)}\frac{\partial h_i(\boldsymbol{\gamma}_*)}{\partial\boldsymbol{\gamma}}|\boldsymbol{X}_i,\boldsymbol{S}_i\right\}|\boldsymbol{X}_i\right] = 0.
$$

Therefore from the linear expansion (2.10) for $\sqrt{N}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0)$, the boundedness of $\sqrt{N}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_*)$, and using Lemma, we have

$\frac{1}{\sqrt{N}}\boldsymbol{A}(\boldsymbol{\beta}_0,\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}}) =$

$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[\boldsymbol{Q}_i(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*) - \mathrm{E}\left\{\boldsymbol{Q}_i(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*)\boldsymbol{\psi}_i(\boldsymbol{\alpha}_0)^T\right\}\left[\mathrm{E}\left\{\boldsymbol{\psi}_i(\boldsymbol{\alpha}_0)\boldsymbol{\psi}_i(\boldsymbol{\alpha}_0)^T\right\}\right]^{-1}\boldsymbol{\psi}_i(\boldsymbol{\alpha}_0)\right]$

$+o_p(1)$,

and this leads to

$$\frac{1}{\sqrt{N}}\frac{\partial L(\boldsymbol{\beta}_0,\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\beta}^T} \xrightarrow{d} N\left(0, \boldsymbol{V}_{\boldsymbol{\alpha}}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*)\right). \tag{2.18}$$

When $\boldsymbol{\gamma}_* = \boldsymbol{\gamma}_0$, we have that

$$\boldsymbol{G}_{i,\boldsymbol{\alpha}}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*) = \mathrm{E}\left\{\mathrm{E}\left[-\frac{R_i}{\pi_i(\boldsymbol{\alpha}_*)^2}\frac{\partial \pi_i(\boldsymbol{\alpha}_*)}{\partial \boldsymbol{\alpha}}\left\{Y_i - h_i(\boldsymbol{\gamma}_0)\right\}|\boldsymbol{X}_i,\boldsymbol{S}_i\right]|\boldsymbol{X}_i\right\} = 0.$$

Therefore from the linear expansion (2.11) for $\sqrt{N}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)$ and the boundedness of $\sqrt{N}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_*)$,

$\frac{1}{\sqrt{N}}\boldsymbol{A}(\boldsymbol{\beta}_0,\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}}) =$

$\frac{1}{\sqrt{N}}\sum_{i=1}^N\left[\boldsymbol{Q}_i(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_0) + \mathrm{E}\left\{\frac{\partial \mu(\boldsymbol{X}_i^T\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}V_i(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_0)^{-1}\boldsymbol{G}_{i,\boldsymbol{\gamma}}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_0)\right\}\boldsymbol{\phi}_i(\boldsymbol{\gamma}_0)\right]$

$+o_p(1)$,

and this leads to

$$\frac{1}{\sqrt{N}}\frac{\partial L(\boldsymbol{\beta}_0,\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\beta}^T} \xrightarrow{d} N\left(0, \boldsymbol{V}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_0)\right). \tag{2.19}$$

When both $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$ and $\boldsymbol{\gamma}_* = \boldsymbol{\gamma}_0$, we have

$$\frac{1}{\sqrt{N}}\boldsymbol{A}(\boldsymbol{\beta}_0,\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}}) = \frac{1}{\sqrt{N}}\sum_{i=1}^N \boldsymbol{Q}_i(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0) + o_p(1),$$

and this leads to

$$\frac{1}{\sqrt{N}}\frac{\partial L(\boldsymbol{\beta}_0,\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\beta}^T} \xrightarrow{d} N\left(0, \boldsymbol{I}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0)\right). \tag{2.20}$$

The desired results follow from (2.16) to (2.20). □

Proof of Theorem II.3. Under Assumptions 1, from Lemma D.2 in KTA, we have $\max_{1\le i,j\le N}\sup_{\boldsymbol{\beta}\in\mathcal{B}}|\hat{\lambda}_i(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})g_j(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})| = o_p(1)$. Therefore $p_{ij}(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}}) = w_{ij}\left\{1+o_p(1)\right\}$, and the $o_p(1)$ term is independent of i, j and $\boldsymbol{\beta}$. This, together with Assumptions 1, the consistency of $\hat{\boldsymbol{\beta}}_{AIPW}$, $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$, and the continuity of $g(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})$, implies that

for any $1 \leq i \leq N$, $\hat{V}_i(\hat{\boldsymbol{\beta}}_{AIPW}) = V_i(\hat{\boldsymbol{\beta}}_{AIPW}, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*) + o_p(1)$ and $\hat{\boldsymbol{G}}_{i,\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}_{AIPW}) = \boldsymbol{G}_{i,\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}_{AIPW}, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*) + o_p(1)$. Assumptions 1 guarantees that $V_i(\hat{\boldsymbol{\beta}}_{AIPW}, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*)$ is invertible with probability approaching 1. Then the weak law of large numbers gives the results. □

Proof of Theorem II.4. This follows from the weak law of large numbers and the arguments in the proof of Theorem II.2 and Theorem II.3. □

CHAPTER III

Locally Efficient and Doubly Robust Estimation with Missing Outcome: Longitudinal Data with Dropout

3.1 Background and Literature Review

In longitudinal studies, repeated measurements are collected from the subjects over certain time period. Dropout is commonly seen in longitudinal studies, where dropout means that some subjects leave the study in the middle of the follow-up and do not return. The missing data caused by dropout often complicate statistical estimation and inference. Unless the dropout is completely at random (Little 1993, 1994, 1995, 2008; Little and Rubin 2002), analysis based on a direct application of the generalized estimating equations (GEE) method (Liang and Zeger 1986) leads to biased estimation.

To correct for the selection bias due to dropout, Robins et al. (1995) and Robins and Rotnitzky (1995) proposed the inverse probability weighted GEE method. Under the assumption of missing at random (MAR) (Little and Rubin 2002), their estimator is consistent if the missingness probabilities are correctly modeled. According to Robins et al. (1994), an augmentation term that extracts more information from subjects with incomplete measurements can be incorporated to improve estimation efficiency. Along this line, Tsiatis (2006) presented a detailed study of the augmented inverse probability weighted (AIPW) complete-case GEE method. In addition to

the potential efficiency improvement, this method yields an estimator that is doubly robust (Scharfstein et al. 1999), in the sense that the estimator is consistent if either the missingness probabilities or the conditional expectations of certain functions of the full data given the observed data at each level of missingness are correctly modeled. See also Bang and Robins (2005), Rotnitzky (2008), Seaman and Copas (2009), Tsiatis et al. (2011), and Rotnitzky et al. (2012).

For the semiparametric model defined by (i) the conditional mean structure of longitudinal outcomes given covariates and (ii) the MAR mechanism, Robins and Rotnitzky (1995) derived the efficiency bound, which is the highest level of estimation efficiency achievable by any regular and asymptotically linear estimator under this semiparametric model. Obtaining an estimator that achieves the efficiency bound is not easy. One such success was given by Robins and Rotnitzky (1995), who proposed to model the following quantities: (i) the missingness probabilities, (ii) the conditional expectations of longitudinal outcomes given the observed data at each level of missingness, and (iii) certain second moments of the data. Their estimator achieves the efficiency bound when all those quantities are correctly modeled, and thus is locally efficient. Refer to Tsiatis (2006) for more details on how to obtain locally efficient estimators. However, the second moments required by Robins and Rotnitzky (1995) are beyond the variance-covariance of longitudinal outcomes. Therefore, even if the variance-covariance is correctly modeled, existing estimators are still unlikely to achieve the efficiency bound. In addition, modeling those second moments may be very difficulty due to their complex forms and the unknown data distribution. Hence, a method that avoids the modeling of any second moments would be appealing.

In this chapter, following the spirit of the AIPW approach, we propose an estimation method based on the conditional empirical likelihood (CEL) (Owen 1988, 2001;

Qin and Lawless 1994; Zhang and Gijbels 2003; Kitamura et al. 2004). Unlike most of the existing methods, our method does not need to model any second moments, including the variance-covariance of longitudinal outcomes. The proposed estimator is doubly robust; that is, the estimator is consistent if either the missingness probabilities or the conditional expectations of longitudinal outcomes given the observed data at each level of missingness are correctly modeled. When both quantities are correctly modeled, our estimator achieves the semiparametric efficiency bound. The procedure developed here is a generalization of that in Chapter II.

This chapter is organized as follows. Section 3.2 describes the data and model on which the developments are based. Section 3.3 details the proposed CEL estimation procedure. Section 3.4 concerns the numerical implementation. Section 3.5 contains the large sample properties. Section 3.6 presents the results of simulation studies. Section 3.7 consists of some concluding remarks. Technical assumptions and proofs are provided in Section 3.8.

3.2 Data and Model

Let Y_{ik} and $\boldsymbol{X}_{ik}$ denote the outcome and a vector of covariates collected from subject i ($i = 1, \cdots, N$) at time k ($k = 0, \cdots, K$), respectively, where time 0 denotes the baseline. In many practical studies, a certain set of auxiliary variables $\boldsymbol{S}_{ik}$ may also be collected at each visit k. Although they are not of direct statistical interest, these auxiliary variables can usually help explain the missingness mechanism and improve estimation efficiency. Therefore, our development in this chapter takes their possible presence into account. Write $\boldsymbol{Y} = (Y_{0,} \cdots, Y_K)^{\mathrm{T}}$, $\boldsymbol{X} = (\boldsymbol{X}_0^{\mathrm{T}}, \cdots, \boldsymbol{X}_K^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{S} = (\boldsymbol{S}_0^{\mathrm{T}}, \cdots, \boldsymbol{S}_K^{\mathrm{T}})^{\mathrm{T}}$. Our interest is to estimate the unknown p-dimensional vector

$\boldsymbol{\beta}_0$ in the following mean regression model:

$$E(Y_k \mid \boldsymbol{X}) = \mu_k(\boldsymbol{X}, \boldsymbol{\beta}_0) \quad (k = 1, \cdots, K), \tag{3.1}$$

where μ_k are user-specified link functions depending on the nature of the outcome. For example, the identity link may be used for continuous outcome, and the logit link may be used for binary outcome. The relationship between $\boldsymbol{Y}$ and $\boldsymbol{S}$ is not of direct interest, and thus $\boldsymbol{S}$ is not included in the regression model (3.1).

Define R_{ik} to be the indicator of observing subject i at time k; that is, $R_{ik} = 1$ if subject i is still in the study at time k, and $R_{ik} = 0$ otherwise. Without loss of generality, assume that data at the baseline are always observed; that is, $R_{i0} = 1$. Due to the fact that dropout leads to monotone missingness, we have that $R_{ik} = 0$ implies $R_{i(k+1)} = 0$ $(k = 1, \cdots, K-1)$. Write $\boldsymbol{R} = (R_0, \cdots, R_K)^{\mathrm{T}}$. In this chapter, we allow the auxiliary variables to be missing together with the outcome, but assume the covariates to be fully observed. Such scenario occurs, for example, when the covariates are external time-dependent variables or deterministic functions of time and baseline covariates. Therefore, our observed data are N independently and identically distributed copies of $(\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{R}^{\mathrm{T}}, \boldsymbol{R}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}, \boldsymbol{R}^{\mathrm{T}}\boldsymbol{S}^{\mathrm{T}})^{\mathrm{T}}$. The missing data caused by dropout are assumed to be MAR, in the sense that for any $k = 1, \cdots, K$,

$$\mathrm{pr}(R_k = 1 \mid R_{k-1} = 1, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{S}) = \mathrm{pr}(R_k = 1 \mid R_{k-1} = 1, \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1}), \tag{3.2}$$

where $\bar{\boldsymbol{Y}}_{k-1} = (Y_0, \cdots, Y_{k-1})^{\mathrm{T}}$ and $\bar{\boldsymbol{S}}_{k-1} = (\boldsymbol{S}_0^{\mathrm{T}}, \cdots, \boldsymbol{S}_{k-1}^{\mathrm{T}})^{\mathrm{T}}$. In other words, the probability of observing a subject at the current scheduled visit, given the fact that the subject was observed at the previous visit, does not depend on the current or future unobserved data, but only depends on the observed history. Denote the probability in (3.2) by $\pi_k(\boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1})$. As usual, the probability of observing the

complete data is assumed to be bounded away from zero, or equivalently,

$$\pi_k = \pi_k(\boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1}) > c > 0 \quad (k = 1, \cdots, K) \tag{3.3}$$

for some constant c. The semiparametric model for longitudinal data with dropout considered in this chapter is defined by (3.1), (3.2) and (3.3). This model is among the semiparametric models considered by Robins et al. (1995) and Robins and Rotnitzky (1995).

3.3 CEL Estimation

Our estimation procedure starts by defining the following AIPW residual vector:

$$\boldsymbol{g}(\boldsymbol{\beta}) = \boldsymbol{\Delta}\boldsymbol{\epsilon}(\boldsymbol{\beta}) - \sum_{k=1}^{K} \frac{R_k - \pi_k R_{k-1}}{\pi_1 \times \cdots \times \pi_k} \boldsymbol{I}_{k-1} E\left\{\boldsymbol{\epsilon}(\boldsymbol{\beta}) \mid \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1}\right\}. \tag{3.4}$$

Here $\boldsymbol{\Delta}$ is the $K \times K$ diagonal matrix with the kth diagonal element $R_k/(\pi_1 \times \cdots \times \pi_k)$, $\boldsymbol{\epsilon}(\boldsymbol{\beta}) = \{Y_1 - \mu_1(\boldsymbol{X}, \boldsymbol{\beta}), \cdots, Y_K - \mu_K(\boldsymbol{X}, \boldsymbol{\beta})\}^{\mathrm{T}}$, and $\boldsymbol{I}_{k-1}$ is the $K \times K$ diagonal matrix with the first $k-1$ diagonal elements 0 and the rest diagonal elements 1. Clearly, $\boldsymbol{I}_0$ is the $K \times K$ identify matrix. In the term $\boldsymbol{\Delta}\boldsymbol{\epsilon}(\boldsymbol{\beta})$, each available ordinary residual is weighted by the inverse probability of observing that residual. The conditional expectations $E\left\{\boldsymbol{\epsilon}(\boldsymbol{\beta}) \mid \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1}\right\}$ $(k = 1, \cdots, K)$ in the second term, the augmentation term, contain the predictive information of historic data on the current and future mean values of the longitudinal outcomes. The formulation of $\boldsymbol{g}(\boldsymbol{\beta})$ is inspired by the derivation of the semiparametric efficiency bound in Robins and Rotnitzky (1995). It is easy to see that $\boldsymbol{g}(\boldsymbol{\beta})$ only depends on the observed data. In Section 3.9, we show that $E\{\boldsymbol{g}(\boldsymbol{\beta}_0) \mid \boldsymbol{X}\} = \boldsymbol{0}$.

Using the idea of CEL (Zhang and Gijbels 2003; Kitamura et al. 2004), conditional on each $\boldsymbol{X}_i$ $(i = 1, \cdots, N)$, we consider the empirical probabilities p_{ij} defined by a discrete distribution that has support on $\{\boldsymbol{g}_j(\boldsymbol{\beta}) : j = 1, \cdots, N\}$. Our CEL

based AIPW (CEL-AIPW) estimator is defined through the following constrained optimization problem:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \max_{p_{ij}} \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\log p_{ij} \quad \text{subject to}$$

$$p_{ij} \geq 0,\ \sum_{j=1}^{N} p_{ij} = 1,\ \sum_{j=1}^{N} p_{ij}\boldsymbol{g}_j(\boldsymbol{\beta}) = \mathbf{0} \quad (i,\, j = 1, \cdots, N). \tag{3.5}$$

Here the first two constraints ensure that conditional on each $i = 1, \cdots, N$, p_{ij} are well-defined probabilities, and the last constraint is the empirical version of $E\{\boldsymbol{g}(\boldsymbol{\beta}_0) \mid \boldsymbol{X}\} = \mathbf{0}$. The objective function in (3.5) is the sum of N localized empirical log-likelihoods, where the localization is carried out by w_{ij}, which are certain non-negative weights satisfying $\sum_{j=1}^{N} w_{ij} = 1$ for each i. A technique to fulfill the purpose of localization is the nonparametric kernel method. Let $\boldsymbol{X}^c$ and $\boldsymbol{X}^d$ denote the continuous and categorical components of $\boldsymbol{X}$, respectively. Then one way to calculate w_{ij} is

$$w_{ij} = \frac{\mathcal{K}\{(\boldsymbol{X}_i^c - \boldsymbol{X}_j^c)/b_N\}\mathcal{I}(\boldsymbol{X}_i^d = \boldsymbol{X}_j^d)}{\sum_{j=1}^{N}\mathcal{K}\{(\boldsymbol{X}_i^c - \boldsymbol{X}_j^c)/b_N\}\mathcal{I}(\boldsymbol{X}_i^d = \boldsymbol{X}_j^d)}, \quad (i,\, j = 1, \cdots, N)$$

where $\mathcal{K}(\cdot)$ is a multivariate kernel function, b_N is the bandwidth parameter, and $\mathcal{I}(\cdot)$ is the indicator function. For many longitudinal clinical trial studies, the main interest is the effect of certain fixed treatment on the outcome adjusted by functions of time and some baseline covariates. In this case, $\boldsymbol{X}^c$ and $\boldsymbol{X}^d$ are respectively the continuous and categorical components of the baseline covariates and treatment, whose dimension is usually not large. Simulation studies, such as that in Chapter II and in Han (2013), have also demonstrated that the CEL method is moderately resistent to the curse of dimensionality, which is known to be a problem for most nonparametric methods.

In $\boldsymbol{g}(\boldsymbol{\beta})$ there are additional unknown quantities other than $\boldsymbol{\beta}$, which are the missingness probabilities π_k and the conditional expectations $H_{kl} = E(Y_l \mid \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1})$

$(k = 1, \cdots, K;\ l = k, \cdots, K)$. These quantities need to be estimated before we can proceed to calculate our estimator $\hat{\boldsymbol{\beta}}$. Following the current literature (Robins et al. 1995; Robins and Rotnitzky 1995; Tsiatis 2006), we postulate parametric models $\pi_k(\boldsymbol{\alpha})$ for π_k and $h_{kl}(\boldsymbol{\gamma})$ for H_{kl}, where $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are the corresponding unknown finite-dimensional parameters. The true values of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are denoted by $\boldsymbol{\alpha}_0$ and $\boldsymbol{\gamma}_0$, respectively. A commonly used estimator $\hat{\boldsymbol{\alpha}}$ for $\boldsymbol{\alpha}_0$ is the maximizer of the partial likelihood

$$\prod_{i=1}^{N}\prod_{k=1}^{K}\left[\{\pi_{ik}(\boldsymbol{\alpha})\}^{R_{ik}}\{1-\pi_{ik}(\boldsymbol{\alpha})\}^{1-R_{ik}}\right]^{R_{i(k-1)}}, \tag{3.6}$$

which is based on fully observed data. The estimation of $\boldsymbol{\gamma}_0$ is less straightforward. This is because under assumption (3.2), we have that $H_{kl} = E(Y_l \mid R_{k-1} = 1, \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1})$, which, however, is not equal to $E(Y_l \mid R_l = 1, \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1})$ for $l > k$. Therefore, estimating $\boldsymbol{\gamma}_0$ is a missing data problem that cannot be solved by complete-case analysis. In the literature there are several methods available. Using the fact that $H_{kl} = E\{Y_l/(\pi_k \times \cdots \times \pi_l) \mid R_l = 1, \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1}\}\mathrm{pr}(R_l = 1 \mid R_{k-1} = 1, \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1})$, Robins and Rotnitzky (1995) proposed to model $E\{Y_l/(\pi_k \times \cdots \times \pi_l) \mid R_l = 1, \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1}\}$ and $\mathrm{pr}(R_l = 1 \mid R_{k-1} = 1, \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1})$ separately. However, since correctly modeling the first quantity depends on a correctly specified $\pi_k(\boldsymbol{\alpha})$, this method will not yield doubly robust estimation of $\boldsymbol{\beta}_0$. Tsiatis (2006) and Tsiatis et al. (2011) suggested specifying a working model for the joint distribution of the full data $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{S})$ and then deriving $h_{kl}(\boldsymbol{\gamma})$ accordingly. In this case, $\boldsymbol{\gamma}$ parametrizes the working distribution and can be estimated by maximum likelihood theory. This method guarantees that $h_{kl}(\boldsymbol{\gamma})$ are compatible in the sense that $E\{h_{(k+1)l}(\boldsymbol{\gamma}) \mid \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1}\} = h_{kl}(\boldsymbol{\gamma})$. However, depending on the specific working distribution, finding the maximum likelihood estimator and deriving $h_{kl}(\boldsymbol{\gamma})$ may be difficult due to multiple integrals. A more flexible method was proposed

by Paik (1997), who utilized the fact that $H_{kl} = E(Y_l \mid R_k = 1, \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1})$ under assumption (3.2), and suggested fitting $h_{kl}(\boldsymbol{\gamma})$ in a sequential way so that the models to be fitted use both the observed outcomes and the imputed outcomes from previously fitted models. Although this method may yield incompatible $h_{kl}(\boldsymbol{\gamma})$, it is useful in practice when the specification of a working model for the joint distribution is not easy, especially in the situation where the number of repeated measurements is large.

After obtaining estimators $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$, we solve the constrained optimization in (3.5) with $\boldsymbol{g}(\boldsymbol{\beta})$ substituted by $\boldsymbol{g}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$, where $\boldsymbol{g}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ is defined by (3.4), but with π_k replaced by $\pi_k(\boldsymbol{\alpha})$ and H_{kl} replaced by $h_{kl}(\boldsymbol{\gamma})$. For convenience, we still denote our estimator by $\hat{\beta}$ after the substitution, but it should be clear that now $\hat{\beta}$ depends on the two nuisance parameters $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$.

It is worth pointing out that, $\boldsymbol{g}(\boldsymbol{\beta})$ defined by (3.4) has another formulation in agreement with the widely used formulation of the the AIPW complete-case GEE method (Tsiatis, 2006; Seaman and Copas, 2009). It is easy to check that the following equation holds:

$$\boldsymbol{\Delta} = \frac{R_K}{\pi_1 \times \cdots \times \pi_K} \boldsymbol{I}_0 + \sum_{k=1}^{K} \frac{C_{k-1} - (1-\pi_k) R_{k-1}}{\pi_1 \times \cdots \times \pi_k} (\boldsymbol{I}_0 - \boldsymbol{I}_{k-1}),$$

where C_k $(k = 0, \cdots, K-1)$ indicates if time k is the last visit; that is, $C_k = 1$ if $R_k = 1$ and $R_{k+1} = 0$, and $C_k = 0$ otherwise. In addition, it is easy to see that $-(R_k - \pi_k R_{k-1}) = C_{k-1} - (1-\pi_k) R_{k-1}$. Therefore, $\boldsymbol{g}(\boldsymbol{\beta})$ can be equivalently written

as follows:

$$\begin{aligned}
\boldsymbol{g}(\boldsymbol{\beta}) &= \frac{R_K}{\prod_{l'=1}^{K}\pi_{l'}}\boldsymbol{\epsilon}(\boldsymbol{\beta}) + \sum_{k=1}^{K}\frac{C_{k-1}-(1-\pi_k)R_{k-1}}{\prod_{l'=1}^{k}\pi_{l'}}(\boldsymbol{I}_0 - \boldsymbol{I}_{k-1})\boldsymbol{\epsilon}(\boldsymbol{\beta}) \\
&\qquad + \sum_{k=1}^{K}\frac{C_{k-1}-(1-\pi_k)R_{k-1}}{\prod_{l'=1}^{k}\pi_{l'}}\boldsymbol{I}_{k-1}E\left\{\boldsymbol{\epsilon}(\boldsymbol{\beta}) \mid \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1}\right\} \\
&= \frac{R_K}{\prod_{l'=1}^{K}\pi_{l'}}\boldsymbol{\epsilon}(\boldsymbol{\beta}) + \sum_{k=1}^{K}\frac{C_{k-1}-(1-\pi_k)R_{k-1}}{\prod_{l'=1}^{k}\pi_{l'}}E\left\{\boldsymbol{\epsilon}(\boldsymbol{\beta}) \mid \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1}\right\}.
\end{aligned}$$

In the last expression, the first term $R_K\boldsymbol{\epsilon}(\boldsymbol{\beta})/\prod_{l'=1}^{K}\pi_{l'}$ is the completely observed residual vector weighted by the inverse probability of observing the complete data. The second term is the augmentation term.

3.4 Numerical Implementation

Using Lagrange multipliers method, the Lagrangian of (3.5) is given by

$$\mathcal{L} = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\log p_{ij} - \sum_{i=1}^{N}\varpi_i\left(\sum_{j=1}^{N}p_{ij} - 1\right) - \sum_{i=1}^{N}\boldsymbol{\lambda}_i^{\mathrm{T}}\left\{\sum_{j=1}^{N}p_{ij}\boldsymbol{g}_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\right\},$$

where ϖ_i and $\boldsymbol{\lambda}_i$ are the Lagrange multipliers. With $\partial\mathcal{L}/\partial p_{ij} = 0$ and the constraints in (3.5), for a fixed $\boldsymbol{\beta}$, we have that

$$p_{ij}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) = \frac{w_{ij}}{1 + \hat{\boldsymbol{\lambda}}_i^{\mathrm{T}}\boldsymbol{g}_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})} \quad (i,\ j = 1, \cdots, N),$$

where $\hat{\boldsymbol{\lambda}}_i$ is the solution to the equation $\sum_{j=1}^{N}p_{ij}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\boldsymbol{g}_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) = \mathbf{0}$. It is easy to see that

$$\hat{\boldsymbol{\lambda}}_i = \arg\min_{\boldsymbol{\lambda}_i}\left[-\sum_{j=1}^{N} w_{ij}\log\left\{1 + \boldsymbol{\lambda}_i^{\mathrm{T}}\boldsymbol{g}_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\right\}\right]. \tag{3.7}$$

Therefore, the proposed estimator $\hat{\boldsymbol{\beta}}$ can be equivalently defined through the following nested optimization:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}}\sum_{i=1}^{N}\left(\min_{\boldsymbol{\lambda}_i}\left[-\sum_{j=1}^{N} w_{ij}\log\left\{1 + \boldsymbol{\lambda}_i^{\mathrm{T}}\boldsymbol{g}_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\right\}\right]\right).$$

This definition of $\hat{\boldsymbol{\beta}}$ suggests a way for numerical implementation, which can be carried out via nested optimization routines. We employ the Newton–Raphson algorithm in our implementation. For convenience, we suppress $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ in the following presentation of the algorithm.

For a fixed $\boldsymbol{\beta}$, given $\boldsymbol{\lambda}_i^{\text{old}}$, the inner optimization updates $\boldsymbol{\lambda}_i$ by

$$\boldsymbol{\lambda}_i^{\text{new}} = \boldsymbol{\lambda}_i^{\text{old}} - \boldsymbol{\Lambda}_{i,\boldsymbol{\lambda}\boldsymbol{\lambda}}^{-1}\left(\boldsymbol{\lambda}_i^{\text{old}}, \boldsymbol{\beta}\right)\boldsymbol{\Lambda}_{i,\boldsymbol{\lambda}}\left(\boldsymbol{\lambda}_i^{\text{old}}, \boldsymbol{\beta}\right),$$

where

$$\boldsymbol{\Lambda}_{i,\boldsymbol{\lambda}}\left(\boldsymbol{\lambda}, \boldsymbol{\beta}\right) = -\sum_{j=1}^{N} w_{ij}\frac{\boldsymbol{g}_j(\boldsymbol{\beta})}{1+\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{g}_j(\boldsymbol{\beta})}, \quad \boldsymbol{\Lambda}_{i,\boldsymbol{\lambda}\boldsymbol{\lambda}}\left(\boldsymbol{\lambda}, \boldsymbol{\beta}\right) = \sum_{j=1}^{N} w_{ij}\frac{\boldsymbol{g}_j(\boldsymbol{\beta})\boldsymbol{g}_j^{\mathrm{T}}(\boldsymbol{\beta})}{\{1+\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{g}_j(\boldsymbol{\beta})\}^2}.$$

An initial value can be taken as $\boldsymbol{\lambda}_i = \mathbf{0}$. To guarantee the positivity of p_{ij}, the updates should be restricted on the legitimate region $\{\boldsymbol{\lambda}_i : 1 + \boldsymbol{\lambda}_i\boldsymbol{g}_j(\boldsymbol{\beta}) \geq w_{ij}\}$.

Given $\boldsymbol{\beta}^{\text{old}}$ and the converged values $\hat{\boldsymbol{\lambda}}_i$ from the inner optimization, the outer optimization updates $\boldsymbol{\beta}$ by

$$\boldsymbol{\beta}^{\text{new}} = \boldsymbol{\beta}^{\text{old}} - \left\{\sum_{i=1}^{N}\boldsymbol{L}_{i,\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}^{\text{old}})\right\}^{-1}\left\{\sum_{i=1}^{N}\boldsymbol{L}_{i,\boldsymbol{\beta}}(\boldsymbol{\beta}^{\text{old}})\right\},$$

where

$$\begin{aligned}
\boldsymbol{L}_{i,\boldsymbol{\beta}}(\boldsymbol{\beta}) &= -\sum_{j=1}^{N} w_{ij}\frac{\boldsymbol{G}_j^{\mathrm{T}}(\boldsymbol{\beta})}{1+\hat{\boldsymbol{\lambda}}_i^{\mathrm{T}}\boldsymbol{g}_j(\boldsymbol{\beta})}\hat{\boldsymbol{\lambda}}_i, \\
\boldsymbol{L}_{i,\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}) &= -\boldsymbol{\Lambda}_{i,\boldsymbol{\lambda}\boldsymbol{\beta}}^{\mathrm{T}}(\hat{\boldsymbol{\lambda}}_i, \boldsymbol{\beta})\boldsymbol{\Lambda}_{i,\boldsymbol{\lambda}\boldsymbol{\lambda}}^{-1}(\hat{\boldsymbol{\lambda}}_i, \boldsymbol{\beta})\boldsymbol{\Lambda}_{i,\boldsymbol{\lambda}\boldsymbol{\beta}}(\hat{\boldsymbol{\lambda}}_i, \boldsymbol{\beta}), \\
\boldsymbol{\Lambda}_{i,\boldsymbol{\lambda}\boldsymbol{\beta}}(\boldsymbol{\lambda}, \boldsymbol{\beta}) &= \sum_{j=1}^{N} w_{ij}\frac{\boldsymbol{g}_j(\boldsymbol{\beta})\hat{\boldsymbol{\lambda}}_i^{\mathrm{T}}\boldsymbol{G}_j(\boldsymbol{\beta})}{\left\{1+\hat{\boldsymbol{\lambda}}_i^{\mathrm{T}}\boldsymbol{g}_j(\boldsymbol{\beta})\right\}^2} - \sum_{j=1}^{N} w_{ij}\frac{\boldsymbol{G}_j(\boldsymbol{\beta})}{1+\hat{\boldsymbol{\lambda}}_i^{\mathrm{T}}\boldsymbol{g}_j(\boldsymbol{\beta})},
\end{aligned}$$

with $\boldsymbol{G}(\boldsymbol{\beta}) = \partial\boldsymbol{g}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$.

The inner and outer optimizations should be iterated until a certain convergence criterion is satisfied. It is worth noting that (4.6) is a convex minimization problem. Therefore, for a fixed $\boldsymbol{\beta}$, the inner loop almost always converges to the global minimizer. A rigorous proof of the convergence can be established by following Chen

et al. (2002). The maximization in the outer loop is more complicated, and the convergence of the Newton-Raphson algorithm to the global maximizer may not be guaranteed. Refer to Owen (2001) for some detailed discussion on related issues in the setting of unconditional moment restrictions. Nonetheless, the nested optimization has been widely used in the literature to implement the empirical likelihood method. See, for example, Owen (2001), Kitamura (2007) and Hansen (2013), among others. According to Kitamura (2007), the nested optimization appears to be the most stable way to compute the empirical likelihood estimator.

In practice, to select the bandwidth parameter b_N, we can employ the cross-validation criterion suggested by Newey (1993) for models with conditional moment restrictions, with some modifications to suit for the case of longitudinal data with dropout. Specifically, define

$$CV(b_N) = \mathrm{tr}\left\{\sum_{i=1}^{N} \hat{\boldsymbol{G}}_i^{\mathrm{T}} \hat{\boldsymbol{V}}_{-i}^{-1} \left(\hat{\boldsymbol{g}}_i \hat{\boldsymbol{g}}_i^{\mathrm{T}} - \hat{\boldsymbol{V}}_{-i}\right) \hat{\boldsymbol{V}}_{-i}^{-1} \left(\hat{\boldsymbol{g}}_i \hat{\boldsymbol{g}}_i^{\mathrm{T}} - \hat{\boldsymbol{V}}_{-i}\right) \hat{\boldsymbol{V}}_{-i}^{-1} \hat{\boldsymbol{G}}_i\right\},$$

where $\mathrm{tr}(\cdot)$ is the trace of a matrix, $\hat{\boldsymbol{G}}_i = \boldsymbol{G}_i(\hat{\boldsymbol{\beta}}_b)$, $\hat{\boldsymbol{g}}_i = \boldsymbol{g}_i(\hat{\boldsymbol{\beta}}_b)$, $\hat{\boldsymbol{V}}_{-i} = \sum_{j \neq i} \hat{w}_{ij} \hat{\boldsymbol{g}}_j \hat{\boldsymbol{g}}_j^{\mathrm{T}}$, $\hat{\boldsymbol{\beta}}_b = \hat{\boldsymbol{\beta}}(b_N)$ is our proposed estimator obtained with a given b_N, and

$$\hat{w}_{ij} = \frac{\mathcal{K}\{(\boldsymbol{X}_i^c - \boldsymbol{X}_j^c)/b_N\}\mathcal{I}(\boldsymbol{X}_i^d = \boldsymbol{X}_j^d)}{\sum_{j \neq i} \mathcal{K}\{(\boldsymbol{X}_i^c - \boldsymbol{X}_j^c)/b_N\}\mathcal{I}(\boldsymbol{X}_i^d = \boldsymbol{X}_j^d)} \qquad (j \neq i).$$

The optimal bandwidth is chosen as the minimizer of $CV(b_N)$.

3.5 Large Sample Properties

Using the results of White (1982), we know that, as $N \to \infty$, $\hat{\boldsymbol{\alpha}} \xrightarrow{p} \boldsymbol{\alpha}_*$ and $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}_*$, and $N^{1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_*)$ and $N^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*)$ are bounded in probability, where $\boldsymbol{\alpha}_*$ and $\boldsymbol{\gamma}_*$ minimize the corresponding Kullback–Leibler distance between the probability distribution based on the postulated model and that generating the data. In general,

$\boldsymbol{\alpha}_* \neq \boldsymbol{\alpha}_0$ unless $\pi_k(\boldsymbol{\alpha})$ are correctly specified, and $\boldsymbol{\gamma}_* \neq \boldsymbol{\gamma}_0$ unless $h_{kl}(\boldsymbol{\gamma})$ are correctly specified.

When $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$, similar argument to that given in (3.8) in Section 3.9 yields $E\{\boldsymbol{g}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*) \mid \boldsymbol{X}\} = \mathbf{0}$ regardless if $\boldsymbol{\gamma}_* = \boldsymbol{\gamma}_0$. On the other hand, it is easy to check that

$$\boldsymbol{\Delta} = \boldsymbol{I}_0 + \sum_{k=1}^{K} \frac{R_k - \pi_k R_{k-1}}{\pi_1 \times \cdots \times \pi_k} \boldsymbol{I}_{k-1},$$

which leads to

$$\boldsymbol{g}(\boldsymbol{\beta}) = \boldsymbol{\epsilon}(\boldsymbol{\beta}) + \sum_{k=1}^{K} \frac{R_k - \pi_k R_{k-1}}{\pi_1 \times \cdots \times \pi_k} \boldsymbol{I}_{k-1} \left[\boldsymbol{\epsilon}(\boldsymbol{\beta}) - E\left\{\boldsymbol{\epsilon}(\boldsymbol{\beta}) \mid \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1}\right\}\right],$$

and thus we have

$$\boldsymbol{g}(\boldsymbol{\beta}, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*) = \boldsymbol{\epsilon}(\boldsymbol{\beta}) + \sum_{k=1}^{K} \frac{R_k - \pi_k(\boldsymbol{\alpha}_*) R_{k-1}}{\pi_1(\boldsymbol{\alpha}_*) \times \cdots \times \pi_k(\boldsymbol{\alpha}_*)} \left\{\mathbf{0}_{k-1}^{\mathrm{T}}, Y_k - h_{kk}(\boldsymbol{\gamma}_*), \cdots, Y_K - h_{kK}(\boldsymbol{\gamma}_*)\right\}^{\mathrm{T}},$$

where $\mathbf{0}_{k-1}$ is the $(k-1)$-dimensional zero vector. When $\boldsymbol{\gamma}_* = \boldsymbol{\gamma}_0$, based on the above expression, in Section 3.9 we show that $E\{\boldsymbol{g}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_0) \mid \boldsymbol{X}\} = \mathbf{0}$ regardless if $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$. Therefore, we have the following Theorem.

Theorem III.1. *For the semiparametric model defined by (3.1), (3.2) and (3.3), if either* $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$ *or* $\boldsymbol{\gamma}_* = \boldsymbol{\gamma}_0$*, we have* $E\{\boldsymbol{g}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*) \mid \boldsymbol{X}\} = \mathbf{0}$.

The conditional mean zero property of $\boldsymbol{g}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ given by Theorem III.1 leads to the double robustness property of the proposed estimator $\hat{\boldsymbol{\beta}}$, which is stated by the following Theorem.

Theorem III.2. *For the semiparametric model defined by (3.1), (3.2) and (3.3), under the assumptions given in Section 3.9, if either* $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$ *or* $\boldsymbol{\gamma}_* = \boldsymbol{\gamma}_0$*, we have* $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ *as* $N \to \infty$.

From Theorem III.2, $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}_0$ if either π_k or H_{kl} are

correctly modeled. This property provides double protection on consistency against model misspecification. A sketched proof of Theorem III.2 is given in Section 3.9.

To derive the asymptotic distribution of $\hat{\boldsymbol{\beta}}$, write $\boldsymbol{V}(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}) = E\{\boldsymbol{g}(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})^{\otimes 2} \mid \boldsymbol{X}\}$, where $\boldsymbol{D}^{\otimes 2} = \boldsymbol{D}\boldsymbol{D}^{\mathrm{T}}$ for any matrix $\boldsymbol{D}$, $\boldsymbol{G}_{\boldsymbol{\alpha}}(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}) = E\{\partial \boldsymbol{g}(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})/\partial\boldsymbol{\alpha} \mid \boldsymbol{X}\}$, $\boldsymbol{G}_{\boldsymbol{\gamma}}(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}) = E\{\partial \boldsymbol{g}(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})/\partial\boldsymbol{\gamma} \mid \boldsymbol{X}\}$, $\boldsymbol{\mu}(\boldsymbol{\beta}) = \{\mu_1(\boldsymbol{X},\boldsymbol{\beta}),\cdots,\mu_K(\boldsymbol{X},\boldsymbol{\beta})\}^{\mathrm{T}}$, $\boldsymbol{\mu}_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \partial\boldsymbol{\mu}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$, $\boldsymbol{Q}(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}) = \boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{V}^{-1}(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})\boldsymbol{g}(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})$,
$\boldsymbol{Q}_{\boldsymbol{\alpha}}(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}) = \boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{V}^{-1}(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})\boldsymbol{G}_{\boldsymbol{\alpha}}(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})$,
$\boldsymbol{Q}_{\boldsymbol{\gamma}}(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}) = \boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{\beta})\boldsymbol{V}^{-1}(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})\boldsymbol{G}_{\boldsymbol{\gamma}}(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})$, and $\boldsymbol{J}(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}) = E\{\boldsymbol{Q}(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})^{\otimes 2}\}$. Since $\hat{\boldsymbol{\alpha}}$ maximizes (3.6), we know that

$$N^{1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_*) = -\left[E\{\partial\boldsymbol{\psi}(\boldsymbol{\alpha}_*)/\partial\boldsymbol{\alpha}\}\right]^{-1} N^{-1/2}\sum_{i=1}^{N}\boldsymbol{\psi}_i(\boldsymbol{\alpha}_*) + o_p(1)$$

from White (1982), where $\boldsymbol{\psi}(\boldsymbol{\alpha})$ is the score function of (3.6). Let $\boldsymbol{\phi}(\boldsymbol{\gamma})$ denote the influence function of $\hat{\boldsymbol{\gamma}}$; that is, $N^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*) = N^{-1/2}\sum_{i=1}^{N}\boldsymbol{\phi}_i(\boldsymbol{\gamma}_*) + o_p(1)$. The following Theorem gives the asymptotic distribution of $\hat{\boldsymbol{\beta}}$, and a sketched proof is given in Section 3.9.

Theorem III.3. *For the semiparametric model defined by (3.1), (3.2) and (3.3), under the assumptions given in Section 3.9, when either $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$ or $\boldsymbol{\gamma}_* = \boldsymbol{\gamma}_0$, as $N \to \infty$, $N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ converges to a normal distribution with mean $\mathbf{0}$ and variance $\boldsymbol{U} = \boldsymbol{J}^{-1}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)\boldsymbol{M}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)\boldsymbol{J}^{-1}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)$, where*

$$\begin{aligned}\boldsymbol{M}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*) \;=\; var\Bigg(&\boldsymbol{Q}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*) - E\{\boldsymbol{Q}_{\boldsymbol{\alpha}}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)\}\left[E\left\{\frac{\partial\boldsymbol{\psi}(\boldsymbol{\alpha}_*)}{\partial\boldsymbol{\alpha}}\right\}\right]^{-1}\boldsymbol{\psi}(\boldsymbol{\alpha}_*)\\ &+E\{\boldsymbol{Q}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*)\}\boldsymbol{\phi}(\boldsymbol{\gamma}_*)\Bigg).\end{aligned}$$

Following Theorem III.3, there are several important simplifications. When π_k are correctly modeled, or equivalently, when $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$, we have that (i) $\boldsymbol{G}_{\boldsymbol{\alpha}}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*) = -E\{\boldsymbol{g}(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_*)\boldsymbol{\psi}^{\mathrm{T}}(\boldsymbol{\alpha}_0) \mid \boldsymbol{X}\}$ (Lemma 9.1 in Tsiatis 2006), (ii) the information

equality $-E\{\partial\boldsymbol{\psi}(\boldsymbol{\alpha}_0)/\partial\boldsymbol{\alpha}\} = E\{\boldsymbol{\psi}(\boldsymbol{\alpha}_0)^{\otimes 2}\}$, and (iii) $\boldsymbol{G}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*) = \mathbf{0}$, the proof of which is given in Section 3.9. These facts lead to the simplification in the following Corollary.

Corollary III.1. *Under the same conditions as in Theorem III.3 but with* $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$, *the asymptotic variance of* $N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ *becomes*
$\boldsymbol{U} = \boldsymbol{J}^{-1}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)\boldsymbol{M}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)\boldsymbol{J}^{-1}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)$, *where*

$$\boldsymbol{M}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*) = var\left[\boldsymbol{Q}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*) - E\{\boldsymbol{Q}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)\boldsymbol{\psi}^T(\boldsymbol{\alpha}_0)\}\{E\boldsymbol{\psi}(\boldsymbol{\alpha}_0)^{\otimes 2}\}^{-1}\boldsymbol{\psi}(\boldsymbol{\alpha}_0)\right].$$

It is easy to see that $\boldsymbol{M}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)$ is the variance of the residual after taking the least square regression of $\boldsymbol{Q}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)$ on $\boldsymbol{\psi}(\boldsymbol{\alpha}_0)$. Hence, when $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$, an augmentation of $\pi_k(\boldsymbol{\alpha})$ by adding interaction or higher order terms of the variables that are already in $\pi_k(\boldsymbol{\alpha})$ increases the dimension of $\boldsymbol{\alpha}$, and thus the dimension of $\boldsymbol{\psi}(\boldsymbol{\alpha})$, which is likely to make $\boldsymbol{M}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)$ smaller in the positive-definite sense. The reduction in $\boldsymbol{M}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*)$ leads to reduction in $\boldsymbol{U}$, and thus efficiency improvement in estimating $\boldsymbol{\beta}_0$. See Robins et al. (1995) for more discussion on this observation.

Now suppose that π_k may not be correctly modeled, but H_{kl} are. In other words, we have $\boldsymbol{\gamma}_* = \boldsymbol{\gamma}_0$. In Section 3.9, we show that $\boldsymbol{G}_{\boldsymbol{\alpha}}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_0) = \mathbf{0}$, which leads to the simplification given in the following Corollary.

Corollary III.2. *Under the same conditions as in Theorem III.3 but with* $\boldsymbol{\gamma}_* = \boldsymbol{\gamma}_0$, *the asymptotic variance of* $N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ *becomes*
$\boldsymbol{U} = \boldsymbol{J}^{-1}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_0)\boldsymbol{M}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_0)\boldsymbol{J}^{-1}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_0)$, *where*

$$\boldsymbol{M}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_0) = var\left[\boldsymbol{Q}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_0) + E\{\boldsymbol{Q}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_0)\}\boldsymbol{\phi}(\boldsymbol{\gamma}_0)\right].$$

When both $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$ and $\boldsymbol{\gamma}_* = \boldsymbol{\gamma}_0$, we have $\boldsymbol{M}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0) = \text{var}\{\boldsymbol{Q}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0)\} = \boldsymbol{J}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0)$, which leads to the following Corollary.

Corollary III.3. *Under the same conditions as in Theorem III.3 but with* $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$ *and* $\boldsymbol{\gamma}_* = \boldsymbol{\gamma}_0$*, the asymptotic variance of* $N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ *becomes* $\boldsymbol{U} = \boldsymbol{J}^{-1}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0)$.

As shown in Robins and Rotnitzky (1995), $\boldsymbol{J}^{-1}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0)$ is the efficiency bound for the semiparametric model defined by (3.1), (3.2) and (3.3). Therefore, our estimator $\hat{\boldsymbol{\beta}}$ attains the semiparametric efficiency bound when both π_k and H_{kl} are correctly modeled. In other words, $\hat{\boldsymbol{\beta}}$ is locally efficient. It is worth pointing out that, the local efficiency of existing estimators, such as the ones in Robins et al. (1995) and Tsiatis (2006), requires correctly modeling certain second moments of the data in addition to π_k and H_{kl}. These second moments are usually difficulty to model in practice due to their complex forms and the unknown data distribution. On the contrary, the local efficiency of $\hat{\boldsymbol{\beta}}$ only requires correctly modeling π_k and H_{kl}, which is easier to achieve.

To be complete, we provide a consistent estimator of the asymptotic variance matrix $\boldsymbol{U}$. Write $\hat{\boldsymbol{g}}_i(\boldsymbol{\beta}) = \boldsymbol{g}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$, $\hat{\boldsymbol{V}}_i(\boldsymbol{\beta}) = \sum_{j=1}^N p_{ij}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \hat{\boldsymbol{g}}_j(\boldsymbol{\beta})^{\otimes 2}$, $\hat{\boldsymbol{Q}}_i(\boldsymbol{\beta}) = \boldsymbol{\mu}_{i,\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{\beta}) \hat{\boldsymbol{V}}_i^{-1}(\boldsymbol{\beta}) \hat{\boldsymbol{g}}_i(\boldsymbol{\beta})$, $\hat{\boldsymbol{J}}(\boldsymbol{\beta}) = N^{-1} \sum_{i=1}^N \hat{\boldsymbol{Q}}_i(\boldsymbol{\beta})^{\otimes 2}$,

$\hat{\boldsymbol{G}}_{i,\boldsymbol{\alpha}}(\boldsymbol{\beta}) = \sum_{j=1}^N p_{ij}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \{\partial \boldsymbol{g}_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) / \partial \boldsymbol{\alpha}\}$,

$\hat{\boldsymbol{G}}_{i,\boldsymbol{\gamma}}(\boldsymbol{\beta}) = \sum_{j=1}^N p_{ij}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \{\partial \boldsymbol{g}_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) / \partial \boldsymbol{\gamma}\}$,

$$\begin{aligned} \hat{\boldsymbol{M}}_i(\boldsymbol{\beta}) \;\; = \;\; & \hat{\boldsymbol{Q}}_i(\boldsymbol{\beta}) - \left\{ \frac{1}{N} \sum_{i=1}^N \boldsymbol{\mu}_{i,\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{\beta}) \hat{\boldsymbol{V}}_i(\boldsymbol{\beta})^{-1} \hat{\boldsymbol{G}}_{i,\boldsymbol{\alpha}}(\boldsymbol{\beta}) \right\} \left\{ \frac{1}{N} \sum_{i=1}^N \frac{\partial \boldsymbol{\psi}_i(\hat{\boldsymbol{\alpha}})}{\partial \boldsymbol{\alpha}} \right\}^{-1} \boldsymbol{\psi}_i(\hat{\boldsymbol{\alpha}}) \\ & + \left\{ \frac{1}{N} \sum_{i=1}^N \boldsymbol{\mu}_{i,\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{\beta}) \hat{\boldsymbol{V}}_i(\boldsymbol{\beta})^{-1} \hat{\boldsymbol{G}}_{i,\boldsymbol{\gamma}}(\boldsymbol{\beta}) \right\} \boldsymbol{\phi}_i(\hat{\boldsymbol{\gamma}}). \end{aligned}$$

Using the Weak Law of Large Numbers and arguments similar to that in the proof of Theorem II.3, $\hat{\boldsymbol{J}}^{-1}(\hat{\boldsymbol{\beta}}) \left\{ N^{-1} \sum_{i=1}^N \hat{\boldsymbol{M}}_i(\hat{\boldsymbol{\beta}})^{\otimes 2} \right\} \hat{\boldsymbol{J}}^{-1}(\hat{\boldsymbol{\beta}}) \xrightarrow{p} \boldsymbol{U}$, as $N \to \infty$.

3.6 Simulation Experiments

In this section we evaluate the finite sample performance of our proposed estimator and make comparison with the AIPW estimator. We use the same simulation model as that considered in Tsiatis et al. (2011). For each subject i ($i = 1, \cdots, N$), the model has one time-dependent covariate $X_{ik}^{(1)} = k$ ($k = 0, 1, 2$) and two time-independent covariates $X_i^{(2)} \sim \mathcal{N}(5, 1)$ and $X_i^{(3)} \sim \text{Ber}(0.5)$. The longitudinal outcomes are generated via the following linear mixed model:

$$Y_{ik} = \eta_{i1} + \eta_{i2} X_{ik}^{(1)} + X_i^{(2)} - X_i^{(3)} + e_{ik} \quad (i = 1, \cdots, N;\ k = 0, 1, 2),$$

where $(\eta_{i1}, \eta_{i2})^{\mathrm{T}} \sim \mathcal{N}\{(1, 2.5)^{\mathrm{T}}, \boldsymbol{\Sigma}\}$, $\boldsymbol{\Sigma}$ is a 2×2 diagonal matrix with diagonal elements 0.3 and 0.2 and off-diagonal elements 0.1, and $e_{ik} \sim \mathcal{N}(0, 1)$. Let $\tilde{Y}_{i0} = \mathcal{I}(Y_{i0} > 5.8)$ and $\tilde{Y}_{i1} = \mathcal{I}(Y_{i1} > 6.2)$, the missingness mechanism for dropout is specified as $\pi_{i1} = [1 + \exp\{\alpha_{0,1}^{(0)} + \alpha_{0,1}^{(1)} \tilde{Y}_{i0}\}]^{-1}$ and $\pi_{i2} = [1 + \exp\{\alpha_{0,2}^{(0)} + \alpha_{0,2}^{(1)} \tilde{Y}_{i0} + \alpha_{0,2}^{(2)} \tilde{Y}_{i1}\}]^{-1}$, where $\boldsymbol{\alpha}_0 = \{\alpha_{0,1}^{(0)}, \alpha_{0,1}^{(1)}, \alpha_{0,2}^{(0)}, \alpha_{0,2}^{(1)}, \alpha_{0,2}^{(2)}\}^{\mathrm{T}} = (-2.0, 2.5, -2.0, 2.0, 2.5)^{\mathrm{T}}$. Under this missingness mechanism, there are approximately 33% and 74% of the subjects with missing Y_1 and Y_2, respectively. It is easy to see that $\boldsymbol{\beta}_0 = \{\beta_0^{(0)}, \cdots, \beta_0^{(3)}\}^{\mathrm{T}} = (1.0, 2.5, 1.0, -1.0)^{\mathrm{T}}$.

To make comparison, we also calculate the AIPW complete-case GEE estimator (Tsiatis 2006; Seaman and Copas 2009). This estimator is given by the solution to the equation

$$\sum_{i=1}^{N} \left[\frac{R_{i2}}{\pi_{i1}\pi_{i2}} \boldsymbol{D}_i(\boldsymbol{\beta}) + \frac{C_{i0} - (1 - \pi_{i1})}{\pi_{i1}} E\left\{\boldsymbol{D}_i(\boldsymbol{\beta}) \mid \boldsymbol{X}_i, Y_{i0}\right\} \right.$$
$$\left. + \frac{C_{i1} - (1 - \pi_{i2}) R_{i1}}{\pi_{i1}\pi_{i2}} E\left\{\boldsymbol{D}_i(\boldsymbol{\beta}) \mid \boldsymbol{X}_i, Y_{i0}, Y_{i1}\right\} \right] = \mathbf{0},$$

where $C_{ik} = 1$ if subject i's last observed measurement is at time k and $C_{ik} = 0$ otherwise ($k = 0, 1$), $\boldsymbol{X}_i = \{X_{i1}^{(1)}, X_{i2}^{(1)}, X_i^{(2)}, X_i^{(3)}\}$, and $\boldsymbol{D}_i(\boldsymbol{\beta})$ is a set of full-data-

based estimating functions. In our simulation study we let $\boldsymbol{D}_i(\boldsymbol{\beta})$ be the quasi-score function (Wedderburn, 1974)

$$\left\{\begin{array}{cc} 1 & 1 \\ X_{i1}^{(1)} & X_{i2}^{(1)} \\ X_i^{(2)} & X_i^{(2)} \\ X_i^{(3)} & X_i^{(3)} \end{array}\right\} \boldsymbol{\Omega}^{-1} \left\{\begin{array}{l} Y_{i1} - \beta_0^{(0)} - \beta_0^{(1)} X_{i1}^{(1)} - \beta_0^{(2)} X_i^{(2)} - \beta_0^{(3)} X_i^{(3)} \\ Y_{i2} - \beta_0^{(0)} - \beta_0^{(1)} X_{i2}^{(1)} - \beta_0^{(2)} X_i^{(2)} - \beta_0^{(3)} X_i^{(3)} \end{array}\right\},$$

where the weighting matrix $\boldsymbol{\Omega}$ takes two different values, the 2×2 identity matrix $\boldsymbol{I}_{2\times 2}$ and the 2×2 matrix with diagonal elements 1.7 and 2.5 and off-diagonal elements 1.0. The latter matrix is the true variance-covariance $\text{var}\{(Y_{i1}, Y_{i2})^{\mathrm{T}} \mid \boldsymbol{X}_i\}$.

We consider three scenarios to make the comparison: (i) only π_k are correctly modeled; (ii) only H_{kl} are correctly modeled; (iii) both π_k and H_{kl} are correctly modeled. It is clear that the partial likelihood (3.6) can be re-written as

$$\left[\prod_{i=1}^{N} \{\pi_{i1}(\boldsymbol{\alpha})\}^{R_{i1}} \{1-\pi_{i1}(\boldsymbol{\alpha})\}^{1-R_{i1}}\right] \left[\prod_{i:R_{i1}=1} \{\pi_{i2}(\boldsymbol{\alpha})\}^{R_{i2}} \{1-\pi_{i2}(\boldsymbol{\alpha})\}^{1-R_{i2}}\right].$$

Therefore, $\{\alpha_{0,1}^{(0)}, \alpha_{0,1}^{(1)}\}^{\mathrm{T}}$ and $\{\alpha_{0,2}^{(0)}, \alpha_{0,2}^{(1)}, \alpha_{0,2}^{(2)}\}^{\mathrm{T}}$ can be estimated separately by maximizing the terms inside the first and the second brackets in the above expression, respectively. Due to the data generating model we are considering, a correctly specified model $h_{kl}(\boldsymbol{\gamma})$ for H_{kl} should be a linear regression of Y_l on $\boldsymbol{X}$ and $\bar{\boldsymbol{Y}}_{k-1}$, and we use the method in Paik (1997) to estimate $\boldsymbol{\gamma}_0$. Following the simulation set-up in Tsiatis et al. (2011), incorrect model for π_k is specified by replacing $(\tilde{Y}_{i0}, \tilde{Y}_{i1})$ in the correct model by (Y_{i0}, Y_{i1}), and incorrect model for H_{kl} is specified by eliminating the regressors $\boldsymbol{X}$ from the correct linear regression model and replacing (Y_{i0}, Y_{i1}) by $[\exp\{(Y_{i0}/9)^2\}, (Y_{i0}+3)/\{1+\exp(Y_{i1})\}+1]$. We consider sample sizes $N = 200, 800$, and use 1000 Monte Carlo replications to summarize the results.

Table 3.1: Simulation results based on 1000 Monte Carlo replications. The numbers have been multiplied by 100

		correct π_k			correct H_{kl}			both correct		
		Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE
						$N = 200$				
AIPWCC-I	$\beta^{(0)}$	-28	186	105	20	553	81	-3	145	85
	$\beta^{(1)}$	7	54	32	-3	92	22	0	43	25
	$\beta^{(2)}$	4	30	17	-4	101	14	0	24	14
	$\beta^{(3)}$	1	52	30	2	77	25	4	47	29
AIPWCC-V	$\beta^{(0)}$	-20	143	85	13	386	65	-1	113	69
	$\beta^{(1)}$	7	54	32	-3	92	22	0	43	25
	$\beta^{(2)}$	3	22	13	-2	65	12	0	18	11
	$\beta^{(3)}$	1	38	25	1	52	22	3	35	22
CEL-AIPW	$\beta^{(0)}$	-8	96	62	14	384	51	0	83	56
	$\beta^{(1)}$	6	53	31	-2	60	21	0	44	25
	$\beta^{(2)}$	1	15	10	-2	63	8	0	13	9
	$\beta^{(3)}$	0	26	16	-1	61	17	1	25	16
						$N = 800$				
AIPWCC-I	$\beta^{(0)}$	-7	88	53	22	441	50	-1	70	41
	$\beta^{(1)}$	1	25	15	-3	64	14	0	20	13
	$\beta^{(2)}$	1	15	9	-4	82	9	0	12	7
	$\beta^{(3)}$	-1	26	17	4	66	15	0	22	14
AIPWCC-V	$\beta^{(0)}$	-5	67	41	15	302	40	-1	54	32
	$\beta^{(1)}$	1	25	15	-3	64	14	0	20	13
	$\beta^{(2)}$	1	11	7	-3	52	7	0	9	5
	$\beta^{(3)}$	-1	19	12	3	42	12	0	16	11
CEL-AIPW	$\beta^{(0)}$	-1	42	27	12	248	27	-1	39	25
	$\beta^{(1)}$	1	24	14	-1	30	12	0	21	13
	$\beta^{(2)}$	0	7	5	-2	48	5	0	6	4
	$\beta^{(3)}$	0	12	8	1	18	8	0	11	8

RMSE: root mean square error. MAE: median absolute error. AIPWCC-I: augmented inverse probability weighted complete-case estimator with $\boldsymbol{\Omega}$ the identity matrix. AIPWCC-V: augmented inverse probability weighted complete-case estimator with $\boldsymbol{\Omega} = \text{var}\{(Y_{i1}, Y_{i2})^{\mathrm{T}} \mid \boldsymbol{X}_i\}$. CEL-AIPW: conditional empirical likelihood based augmented inverse probability weighted estimator.

Due to the data generating process, it is easy to check that the two augmented inverse probability weighted complete-case estimators of $\beta_0^{(1)}$ with $\boldsymbol{\Omega} = \boldsymbol{I}_{2\times 2}$ and with $\boldsymbol{\Omega} = \text{var}\{(Y_{i1}, Y_{i2})^{\text{T}} \mid \boldsymbol{X}_i\}$ are identical, and both are given by the solution to the equation

$$\sum_{i=1}^{N}\left[\frac{R_{i2}}{\pi_{i1}\pi_{i2}}\{Y_{i2}-Y_{i1}-\beta^{(1)}\}+\frac{C_{i0}-(1-\pi_{i1})}{\pi_{i1}}\{H_{i12}-H_{i11}-\beta^{(1)}\}\right.$$
$$\left.+\frac{C_{i1}-(1-\pi_{i2})R_{i1}}{\pi_{i1}\pi_{i2}}\{H_{i22}-Y_{i1}-\beta^{(1)}\}\right]=0.$$

This fact is confirmed by our simulation results in Table 3.1. From Table 3.1, the three estimators under comparison have ignorable bias in most scenarios, confirming their double robustness property. The relatively large bias and root mean square error of the three estimators of $\beta_0^{(0)}$ when only H_{kl} are correctly modeled are due to the fact that, as already pointed out in Tsiatis et al. (2011), the incorrect models for π_k occasionally produce near-zero estimated values, which lead to large inverse probability weights and destabilize the numerical behavior of all estimators considered here. It is clearly seen that, the augmented inverse probability weighted complete-case estimator using $\boldsymbol{\Omega} = \text{var}\{(Y_{i1}, Y_{i2})^{\text{T}} \mid \boldsymbol{X}_i\}$ has smaller root mean square error and smaller median absolute error compared to the estimator using $\boldsymbol{\Omega} = \boldsymbol{I}_{2\times 2}$, except for the case of estimating $\beta_0^{(1)}$. Therefore, using the true variance-covariance matrix to construct the estimating function $\boldsymbol{D}_i(\boldsymbol{\beta})$ helps improve the efficiency. Our proposed estimator further reduces the root mean square error and the median absolute error in general. When both π_k and H_{kl} are correctly modeled, our estimator attains the semiparametric efficiency bound asymptotically, whereas the two augmented inverse probability weighted complete-case estimators do not, even if the true variance-covariance matrix is used to derive them. This fact is well demonstrated by the smaller root mean square error of our estimator in the simulation

results.

3.7 Data Application

In this section we analyze the data collected from the National Cooperative Gallstone Study (NCGS). This is a longitudinal study, of which one major interest is the treatment effect of the drug chenodiol on cholesterol gallstones. There are three treatment arms, the high-dose group (750 mg per day), the low-dose group (375 mg per day), and the placebo group. After randomization prior to baseline, the three groups have 305, 306 and 305 patients, respectively, and each patient is administratively scheduled to be followed for as long as 28 months. Following the analysis in Wei and Lachin (1984), we will restrict our consideration to those patients who are assigned to the high dose and the placebo groups and those who have floating stones.

It was suggested in this study that the drug chenodiol dissolves gallstones to reduce cholesterol secretion into gallbladder bile. As the gallstones dissolves, they might pass into the biliary tree and cause an exacerbation of gallbladder symptoms. In addition, the reduction of cholesterol secretion into gallbladder bile might increase the level of serum cholesterol, which is a known risk factor for atherosclerotic disease. Therefore, serum cholesterol level is measured for all patients at the baseline and at 6, 12, 20 and 24 months of follow-up. An important question is whether there is a significant effect of the high dose of chenodiol on the progression of patients' serum cholesterol level during follow-up, adjusting for the baseline level (Wei and Lachin, 1984).

Because of the termination of follow-up and the missing visits of some patients, many serum cholesterol measurements are unobserved. Wei and Lachin's (1984) analysis assumed that the causes for the missing observations are completely inde-

pendent of the nature or values of the serum cholesterol level, namely the missing completely at random mechanism (Little and Rubin 2002). A close inspection of the data reveals that, there are 2 patients having intermittent missing visits. According to Robins and Rotnitzky (1995), when the mechanism (3.2) is assumed, the measurements after a missing visit do not add further information to improve estimation efficiency. Therefore, for those 2 patients, we remove their measurements after their first missing visit. After such adjustment, the missing pattern in the observed data becomes monotone for each patient. In other words, we have longitudinal data with dropout. The total number of patients is 103, with 62 in the high-dose group and 41 in the placebo group. The missingness rate at 6, 12, 20, and 24 months of follow-up is 0%, 10%, 24% and 35%, respectively.

In order to study the effect of chenodiol on the progression of serum cholesterol level adjusting for the baseline, we fit the model

$$Y_{ik} = \beta_1 + \beta_2 bsc_i + \beta_3 k + \beta_4 trt_i \times k + \epsilon_{ik},$$

where Y_{ik} is the serum cholesterol level measured at time k, $k = 6, 12, 20, 24$, bsc_i is the baseline serum cholesterol level, trt_i is the indicator of treatment with $trt_i = 1$ for the high-dose group and $trt_i = 0$ for the placebo group, and ϵ_{ik} is the error term with $E(\epsilon_{ik} \mid bsc_i, k) = 0$. Because of the randomization prior to the baseline, there is no need to include the main effect of treatment into the above model. Our main interest is β_4. When $\beta_4 = 0$, there is no effect of high-dose chenodiol on the progression of serum cholesterol level adjusting for the baseline. The missingness mechanism is modeled by fitting separate logistic regression models at each follow-up visit, similar to what has been done in our simulation study. Paik's (1997) sequential imputation method is used to calculate the expectations in the augmentation term.

Table 3.2: Analysis results for the data from the National Cooperative Gallstone Study

	est	se	p-value
intercept	97.122	12.851	< 0.000
bsc	0.627	0.053	< 0.000
k	0.624	0.210	0.003
$bsc \times k$	0.097	0.288	0.736

est: estimated value. se: standard error.

The results of our analysis are summarized in Table 3.2. It is seen that, adjusting for the baseline, the serum cholesterol level increases significantly with time, but the high-dose group does not have a different rate of increase compared to the placebo group. Therefore, there is no effect of high-dose chenodiol on the progression of serum cholesterol level adjusting for the baseline. This conclusion is in agreement with that in Wei and Lachin (1984).

3.8 Conclusions

In this chapter we proposed the CEL method for analyzing longitudinal data with dropout. We studied the asymptotic properties of the CEL-AIPW estimator. The innovation of our proposed method is that our estimator is locally efficient with no need to model any second moments of the data. In order to achieve the semiparametric efficiency bound, the second moments that are required to be correctly modeled by existing estimators have complex forms, and thus are difficulty to model, especially with an unknown data distribution in practical studies. Therefore, the circumvention of modeling the second moments by our proposed CEL method may be highly

advantageous. The issues associated with the CEL method discussed in Section 2.6 also apply to the development in this chapter. Refer to Section 2.6 for these issues.

3.9 Technical Assumptions and Proofs

This section contains the proofs of the theoretical results. Technical details of the proofs for Theorem III.2 and Theorem III.3 can be filled in following the proofs in Chapter II and Kitamura et al. (2004).

Proof of $E\{\boldsymbol{g}(\boldsymbol{\beta}_0) \mid \boldsymbol{X}\} = \mathbf{0}$.

$$
\begin{aligned}
& E\{\boldsymbol{g}(\boldsymbol{\beta}_0) \mid \boldsymbol{X}\} \\
= \; & E\left[E\{\boldsymbol{g}(\boldsymbol{\beta}_0) \mid \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{S}\} \mid \boldsymbol{X}\right] \\
= \; & E\left\{E(\boldsymbol{\Delta} \mid \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{S})\boldsymbol{\epsilon}(\boldsymbol{\beta}_0) \mid \boldsymbol{X}\right\} \\
& -E\left[\sum_{k=1}^{K} \frac{E(R_k - \pi_k R_{k-1} \mid \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{S})}{\pi_1 \times \cdots \times \pi_k} \boldsymbol{I}_{k-1} E\left\{\boldsymbol{\epsilon}(\boldsymbol{\beta}_0) \mid \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1}\right\} \mid \boldsymbol{X}\right] \\
= \; & \mathbf{0}. \qquad (3.8)
\end{aligned}
$$

The last equality follows from the facts that, under monotone missingness and (3.2), $E(R_k \mid \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{S}) = E(\pi_k R_{k-1} \mid \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{S}) = \pi_k \mathrm{pr}(R_{k-1} = 1 \mid \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{S}) = \pi_1 \times \cdots \times \pi_k$. □

Proof of $E\{\boldsymbol{g}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_0) \mid \boldsymbol{X}\} = \mathbf{0}$.

$$
\begin{aligned}
& E\{\boldsymbol{g}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_0) \mid \boldsymbol{X}\} \\
= & \sum_{k=1}^{K} E\Bigg(E\Bigg[\frac{R_k - \pi_k(\boldsymbol{\alpha}_*) R_{k-1}}{\prod_{l'=1}^{k} \pi_{l'}(\boldsymbol{\alpha}_*)} \left\{ 0_{k-1}^{\mathrm{T}}, Y_k - h_{kk}(\boldsymbol{\gamma}_0), \cdots, Y_K - h_{kK}(\boldsymbol{\gamma}_0) \right\}^{\mathrm{T}} \\
& \qquad \mid \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1} \Bigg] \mid \boldsymbol{X} \Bigg) \\
= & \sum_{k=1}^{K} E\Bigg(E\Bigg[\frac{R_k - \pi_k(\boldsymbol{\alpha}_*) R_{k-1}}{\prod_{l'=1}^{k} \pi_{l'}(\boldsymbol{\alpha}_*)} \left\{ 0_{k-1}^{\mathrm{T}}, Y_k - h_{kk}(\boldsymbol{\gamma}_0), \cdots, Y_K - h_{kK}(\boldsymbol{\gamma}_0) \right\}^{\mathrm{T}} \\
& \qquad \mid R_{k-1} = 1, \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1} \Bigg] \times \mathrm{pr}(R_{k-1} = 1 \mid \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1}) \mid \boldsymbol{X} \Bigg) \\
= & \sum_{k=1}^{K} E\Bigg(\frac{\pi_k - \pi_k(\boldsymbol{\alpha}_*)}{\prod_{l'=1}^{k} \pi_{l'}(\boldsymbol{\alpha}_*)} E\left[\left\{ 0_{k-1}^{\mathrm{T}}, Y_k - h_{kk}(\boldsymbol{\gamma}_0), \cdots, Y_K - h_{kK}(\boldsymbol{\gamma}_0) \right\}^{\mathrm{T}} \mid \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1} \right] \\
& \qquad \times \prod_{l'=1}^{k-1} \pi_{l'} \mid \boldsymbol{X} \Bigg) \\
= & \ \mathbf{0}
\end{aligned}
$$

The third equality uses the facts that, under (3.2), $R_k \perp Y_l \mid (R_{k-1} = 1, \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1})$ and $H_{kl} = E(Y_l \mid R_{k-1} = 1, \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1})$ $(k = 1, \cdots, K;\ l = k, \cdots, K)$. □

Assumptions (Assumptions used to prove Theorem III.2 and Theorem III.3)**.** *Let* $\mathcal{B}$, $\mathcal{A}$, $\mathcal{G}$ *and* $\mathcal{X}$ *denote the domain of* $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ *and* $\boldsymbol{X}$, *respectively. Let* $\mathcal{B}_0 \subseteq \mathcal{B}$ *be some closed ball around* $\boldsymbol{\beta}_0$. *For a matrix* $\boldsymbol{D}$ *with elements* D_{st}, *define* $\| \boldsymbol{D} \| = (\sum_{s,t} D_{st}^2)^{1/2}$. *We make the following assumptions. (i)* $\mathcal{B}$, $\mathcal{A}$, $\mathcal{G}$ *and* $\mathcal{X}$ *are compact. (ii) For any* $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, *there exists* $\mathcal{X}_{\boldsymbol{\beta},\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*} \subseteq \mathcal{X}$ *such that* $pr(\boldsymbol{X} \in \mathcal{X}_{\boldsymbol{\beta},\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*}) > 0$ *and* $E\{g(\boldsymbol{\beta}, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*) \mid \boldsymbol{X}\} \neq 0$ *for every* $\boldsymbol{X} \in \mathcal{X}_{\boldsymbol{\beta},\boldsymbol{\alpha}_*,\boldsymbol{\gamma}_*}$. *(iii)* $E\{\sup_{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}} \| \boldsymbol{g}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \|^m\} < \infty$ *for some* $m \geq 8$. *(iv)* $0 < \inf_{\boldsymbol{X},\boldsymbol{\beta} \in \mathcal{B}_0,\boldsymbol{\alpha},\boldsymbol{\gamma}} \boldsymbol{\xi}^T \boldsymbol{V}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \boldsymbol{\xi} \leq \sup_{\boldsymbol{X},\boldsymbol{\beta} \in \mathcal{B}_0,\boldsymbol{\alpha},\boldsymbol{\gamma}} \boldsymbol{\xi}^T \boldsymbol{V}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \boldsymbol{\xi} < \infty$ *for any* $\| \boldsymbol{\xi} \| = 1$. *(v)* $\inf_{\boldsymbol{\alpha}} \pi_k(\boldsymbol{\alpha}) > 0$. *(vi)* $\mu_k(\boldsymbol{X}, \boldsymbol{\beta})$, $\pi_k(\boldsymbol{\alpha})$ *and* $h_{kl}(\boldsymbol{\gamma})$ *are continuously differentiable. (vii)* $b_N \to 0$, $N^{1-2\nu-2/m} b_N^{2q} \to \infty$ *and* $N^{1-2\nu} b_N^{5q/2} \to \infty$ *as* $N \to \infty$, *where* $\nu \in (0, 1/2)$, $m \geq 8$ *and* q *is the dimension of* $\boldsymbol{X}^c$. *(viii)* $\| \hat{\boldsymbol{\lambda}}_i \| \leq \delta N^{-1/m}$ *for some* $\delta > 0$.

Proof of Theorem III.2. Define

$$L(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} \log\left\{1 + \hat{\boldsymbol{\lambda}}_i^{\mathrm{T}} \boldsymbol{g}_j(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})\right\}.$$

It is clear that $\hat{\boldsymbol{\beta}}$ is the maximizer of $L(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$. First, we have $N^{1/m} L(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \leq F(\boldsymbol{\beta}, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*) + o_p(1)$ for any $\boldsymbol{\beta} \in \mathcal{B}$, where

$$F(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = -E\left[\frac{\| E\{\boldsymbol{g}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \mid \boldsymbol{X}\}\|^2}{1 + \| E\{\boldsymbol{g}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \mid \boldsymbol{X}\}\|}\right].$$

This, together with Assumptions (i) and (ii), yields that, for any $\delta > 0$, there exists $c(\delta) > 0$, such that

$$\sup_{\boldsymbol{\beta} \in \mathcal{B}/B(\boldsymbol{\beta}_0, \delta)} N^{1/m} L(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \leq \sup_{\boldsymbol{\beta} \in \mathcal{B}/B(\boldsymbol{\beta}_0, \delta)} F(\boldsymbol{\beta}, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*) + o_p(1) \leq -c(\delta) + o_p(1),$$

where $B(\boldsymbol{\beta}_0, \delta)$ is the ball centering at $\boldsymbol{\beta}_0$ with radius δ. Second, we have

$$L(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \geq -\frac{1}{N}\sum_{i=1}^{N} \hat{\boldsymbol{\lambda}}_i^{\mathrm{T}}(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \sum_{j=1}^{N} w_{ij} \boldsymbol{g}_j(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$$

and $\max_{1 \leq i \leq N} \hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) = o_p(N^{-1/m})$ if either $\boldsymbol{\alpha}_* = \boldsymbol{\alpha}_0$ or $\boldsymbol{\gamma}_* = \boldsymbol{\gamma}_0$, which leads to that $N^{1/m} L(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) \geq o_p(1)$. These two facts give the consistency of $\hat{\boldsymbol{\beta}}$. □

Proof of Theorem III.3. Taking the Taylor expansion of $\partial L(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})/\partial \boldsymbol{\beta} = \mathbf{0}$ around $\boldsymbol{\beta}_0$ gives that

$$N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \left\{-\frac{\partial^2 L(\tilde{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right\}^{-1} \left\{N^{1/2} \frac{\partial L(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\beta}}\right\},$$

where $\tilde{\boldsymbol{\beta}}$ is some point between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$. We have

$$-\frac{\partial^2 L(\tilde{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \xrightarrow{p} \boldsymbol{J}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*)$$

and

$$\begin{aligned} N^{1/2} \frac{\partial L(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\beta}} &= N^{-1/2} \sum_{i=1}^{N} \boldsymbol{Q}_i(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*) \\ &\quad + E\left\{\boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{\beta}_0) \boldsymbol{V}^{-1}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*) \boldsymbol{G}_{\boldsymbol{\alpha}}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*)\right\} N^{1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_*) \\ &\quad + E\left\{\boldsymbol{\mu}_{\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{\beta}_0) \boldsymbol{V}^{-1}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*) \boldsymbol{G}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_*)\right\} N^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*) + o_p(1). \end{aligned}$$

The desired result then follows. □

Proof of $\boldsymbol{G}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*) = \mathbf{0}$.

$$
\begin{aligned}
& \boldsymbol{G}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_*) \\
= & -\sum_{k=1}^{K} E\left(E\left[\frac{R_k - \pi_k(\boldsymbol{\alpha}_0) R_{k-1}}{\pi_1(\boldsymbol{\alpha}_0) \times \cdots \times \pi_k(\boldsymbol{\alpha}_0)} \left\{ \begin{array}{c} \mathbf{0}_{(k-1)\times d} \\ \partial h_{kk}(\boldsymbol{\gamma}_*)/\partial \boldsymbol{\gamma}^{\mathrm{T}} \\ \vdots \\ \partial h_{kK}(\boldsymbol{\gamma}_*)/\partial \boldsymbol{\gamma}^{\mathrm{T}} \end{array} \right\} \mid \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{S} \right] \mid \boldsymbol{X} \right) = \mathbf{0},
\end{aligned}
$$

where d denotes the dimension of $\boldsymbol{\gamma}$. □

Proof of $\boldsymbol{G}_{\boldsymbol{\alpha}}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_0) = \mathbf{0}$.

$$
\begin{aligned}
& \boldsymbol{G}_{\boldsymbol{\alpha}}(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_*, \boldsymbol{\gamma}_0) \\
= & \sum_{k=1}^{K} E\left(E\left[\left\{ \begin{array}{c} \mathbf{0}_{k-1} \\ Y_k - h_{kk}(\boldsymbol{\gamma}_0) \\ \vdots \\ Y_K - h_{kK}(\boldsymbol{\gamma}_0) \end{array} \right\} \frac{\partial \left\{ \frac{R_k - \pi_k(\boldsymbol{\alpha}_*) R_{k-1}}{\pi_1(\boldsymbol{\alpha}_*) \times \cdots \times \pi_k(\boldsymbol{\alpha}_*)} \right\}}{\partial \boldsymbol{\alpha}^{\mathrm{T}}} \mid \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1} \right] \mid \boldsymbol{X} \right) \\
= & \sum_{k=1}^{K} E\Bigg(E\Bigg[\left\{ \begin{array}{c} \mathbf{0}_{k-1} \\ Y_k - h_{kk}(\boldsymbol{\gamma}_0) \\ \vdots \\ Y_K - h_{kK}(\boldsymbol{\gamma}_0) \end{array} \right\} \frac{\partial \left\{ \frac{R_k - \pi_k(\boldsymbol{\alpha}_*) R_{k-1}}{\pi_1(\boldsymbol{\alpha}_*) \times \cdots \times \pi_k(\boldsymbol{\alpha}_*)} \right\}}{\partial \boldsymbol{\alpha}^{\mathrm{T}}} \mid R_{k-1} = 1, \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1} \Bigg] \\
& \qquad \times \mathrm{pr}(R_{k-1} = 1 \mid \boldsymbol{X}, \bar{\boldsymbol{Y}}_{k-1}, \bar{\boldsymbol{S}}_{k-1}) \mid \boldsymbol{X} \Bigg) \\
= & \mathbf{0}.
\end{aligned}
$$

□

CHAPTER IV

Analysis of Unbalanced Longitudinal Data

4.1 Background and Literature Review

Longitudinal data are repeated measurements collected from the study subjects over certain time period. To efficiently evaluate the relationship between the mean of longitudinal outcomes and the covariates, the variance-covariance of these outcomes needs to be taken into account. Marginal models for longitudinal data that specify model structures for both the marginal mean and the variance-covariance have been discussed widely in the literature (e.g. Diggle et al. 2002; Song 2007). Estimation and inference for marginal models may be based on, for example, the quasi-likelihood approach (Wedderburn 1974; Heyde 1997) or the inference functions approach (Godambe 1960, 1991). Liang and Zeger (1986) proposed the popular generalized estimating equations (GEE) method, which assumes a working model for the within-subject correlation of longitudinal outcomes. When the correlation is incorrectly modeled, the GEE estimator preserves consistency, but may lose substantial estimation efficiency (Wang and Carey 2003). To improve efficiency, Qu et al. (2000) proposed the quadratic inference functions (QIF) estimator, which has been shown to be equally efficient to the GEE estimator when the correlation is correctly modeled, and more efficient when the correlation is incorrectly modeled. Wang and Lin

(2005) pointed out that, to avoid losing efficiency, not only the correlation, but also the marginal variance of longitudinal outcomes need to be correctly modeled. Some strategies for modeling the second moments can be found in Pan and MacKenzie (2003) and Ye and Pan (2006). However, the introduction of more models beyond that for the marginal mean brings more risk of model misspecification, which can lead to poor efficiency and misleading conclusions. Therefore, many researchers proposed to model the variance-covariance nonparametrically; see, for example, Jiang et al. (2007) and Li (2011).

Conditional empirical likelihood (CEL) method (Zhang and Gijbels 2003; Kitamura et al. 2004) is a generalization of empirical likelihood (EL) method (Owen 1988, 1990, 2001; Qin and Lawless 1994) to the setting where the model is defined by conditional moment restrictions. Despite the success of EL method in various statistical research areas, its application in longitudinal data analysis has not been studied adequately. Some existing works include You et al. (2006) and Xue and Zhu (2007), who studied partially linear models for longitudinal data. But those authors did not account for the within-subject correlation. For data with continuous outcome, Wang et al. (2010) proposed a generalized empirical likelihood (GEL) method, which requires a pre-estimated variance-covariance matrix. To the best of our knowledge, all existing applications of EL method to longitudinal data analysis follow the setting of Qin and Lawless (1994), where the model is defined by unconditional moment restrictions.

In this paper, we propose to model longitudinal data using the CEL method, which only requires to specify a model for the marginal mean of the outcomes, and thus eliminates the dependence of data analysis on modeling the variance-covariance. Possible unbalanced follow-up visits are dealt with via stratification according to

distinct follow-up patterns. We show that the CEL method is closely connected to the GEE method, in the sense that the proposed estimator can be regarded as the solution to a set of estimating equations analogous to GEE. The difference is that, the CEL method implicitly incorporates a consistently estimated variance-covariance matrix, which, in theory, endows the proposed estimator the same efficiency as that of the GEE estimator employing the true variance-covariance. We also derive the asymptotic distribution of the proposed estimator, and provide two ways to estimate the asymptotic variance.

This article is organized as follows. In Section 4.2, we discuss longitudinal data analysis using the CEL method. Section 4.3 contains large sample properties. Section 4.4 presents numerical implementation and some discussion on computational issues. In Section 4.5, we conduct simulation experiments to study the finite sample performance of the CEL method. Sections 4.6 contains a data application. Section 4.7 consists of concluding remarks. All the technical details are provided in Section 4.8.

4.2 Methodology

4.2.1 Model and estimation

For each subject i, $i = 1, \ldots, N$, let Y_{it} denote the outcome and $\boldsymbol{X}_{it}$ denote a p-element vector of covariates measured at time $t = 1, \ldots, n_i$. Write $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\top$, $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}^\top, \ldots, \boldsymbol{X}_{in_i}^\top)^\top$. We only specify a model for the marginal mean of Y_{it} given $\boldsymbol{X}_i$. Specifically, we assume the marginal mean regression model

$$E(Y_{it} \mid \boldsymbol{X}_i) = \mu(\boldsymbol{X}_{it}^\top \boldsymbol{\beta}) \quad \text{for some } \boldsymbol{\beta} = \boldsymbol{\beta}_0 \in \mathcal{R}^p,$$

where $\mu(\cdot)$ is a known link function and $\boldsymbol{\beta}$ is the vector of regression coefficients. Define the residual vector to be $\boldsymbol{g}_i(\boldsymbol{\beta}) = \{Y_{i1} - \mu(\boldsymbol{X}_{i1}^\top \boldsymbol{\beta}), \ldots, Y_{in_i} - \mu(\boldsymbol{X}_{in_i}^\top \boldsymbol{\beta})\}^\top$.

Then, our model assumption becomes

$$E\{\boldsymbol{g}_i(\boldsymbol{\beta}_0) \mid \boldsymbol{X}_i\} = \mathbf{0}, \quad i = 1, \ldots, N. \tag{4.1}$$

The above setting accomodates unbalanced longitudinal data, which arise from many practical studies. For example, when an original clinical trial study is planned to use a balanced design, namely all subjects are pre-scheduled to visit the clinic at the same pre-fixed time points, unbalanced data result from subjects' missing visits. Therefore, unbalanced data can be treated within the framework of missing data analysis. In this case, similar to the GEE method, our above formulation implicitly made the missing completely at random (MCAR) assumption (Little and Rubin 2002). The MCAR assumption is also widely adopted for longitudinal observational studies where unbalanced data are collected at available visits, with no clear underlying missingness mechanism. Following Liang and Zeger (1986), we focus our discussion on available-data analysis without modeling the missingness mechanism. Such an approach is also taken as the default by many popular statistical softwares in analyzing unbalanced longitudinal data. To deal with the unbalanced data structure, we stratify subjects according to their follow-up patterns. Let $\tau_i = \{1, \ldots, n_i\}$ denote the (intermittent) follow-up pattern for subject i, and let $S_i = \{j : 1 \leq j \leq N \text{ and } \tau_j = \tau_i\}$ be the stratum where subject i belongs. Let Ω denote the collection of all distinct follow-up patterns observed in the data, and for $\omega \in \Omega$, let S^ω denote the set of indices for subjects having pattern ω. Thus, $\{S^\omega\}_{\omega\in\Omega}$ constitutes the collection of all distinct strata in the data. Subjects in the same stratum may be thought to be sampled from a balanced study. The model defined by (4.1) has a stratum-wise interpretation: for any $\omega \in \Omega$, $E\{\boldsymbol{g}_i(\boldsymbol{\beta}_0) \mid \boldsymbol{X}_i\} = \mathbf{0}$ for $i \in S^\omega$. Clearly, when the whole data are balanced, all subjects belong to one unique stratum.

Given each subject i, we consider the empirical probabilities p_{ij} defined by a discrete distribution that has support on $\{\boldsymbol{g}_j(\boldsymbol{\beta}) : j \in S_i\}$. Our CEL estimator of $\boldsymbol{\beta}_0$ is defined through the following constrained optimization problem.

$$\hat{\boldsymbol{\beta}}_{CEL} = \arg\max_{\boldsymbol{\beta}} \max_{p_{ij}} \sum_{i=1}^{N} \sum_{j \in S_i} w_{ij} \log p_{ij} \quad \text{subject to}$$

$$p_{ij} \geq 0 \ (j \in S_i,\, i = 1, \cdots, N),\ \sum_{j \in S_i} p_{ij} = 1 \ (i = 1, \cdots, N),$$

$$\sum_{j \in S_i} p_{ij} \boldsymbol{g}_j(\boldsymbol{\beta}) = \mathbf{0} \ (i = 1, \cdots, N). \tag{4.2}$$

Here the first two constraints make sure that, conditional on each $i = 1, \cdots, N$, p_{ij} are well-defined probabilities, while the last constraint is the empirical version of the model assumption in (4.1). The objective function in (4.2) is the sum of N localized empirical log-likelihoods, where the localization is carried out by certain non-negative weights w_{ij}, satisfying $\sum_{j \in S_i} w_{ij} = 1$ for each i. A technique to fulfill the purpose of localization is the nonparametric kernel method. Write $\boldsymbol{X}_i = (\boldsymbol{X}_i^c, \boldsymbol{X}_i^d)$, where $\boldsymbol{X}_i^c$ consists of continuous components with dimension q_i and $\boldsymbol{X}_i^d$ consists of categorical components, respectively. Time itself as a covariate may be excluded in the calculation of the weights, since subjects from the same stratum have a common set of visit times. The weights w_{ij} are calculated as

$$w_{ij} = \frac{\mathcal{K}\left(\frac{\boldsymbol{X}_i^c - \boldsymbol{X}_j^c}{b_N}\right) I(\boldsymbol{X}_i^d = \boldsymbol{X}_j^d)}{\sum_{j \in S_i} \mathcal{K}\left(\frac{\boldsymbol{X}_i^c - \boldsymbol{X}_j^c}{b_N}\right) I(\boldsymbol{X}_i^d = \boldsymbol{X}_j^d)} \quad j \in S_i, \tag{4.3}$$

where $\mathcal{K}\{(\boldsymbol{X}_i^c - \boldsymbol{X}_j^c)/b_N\} = \prod_{l=1}^{q_i} K\{(X_i^{(l)} - X_j^{(l)})/b_N\}$, $K(\cdot)$ is a second order kernel function and symmetric around zero (e.g. standard Gaussian kernel), b_N is the bandwidth parameter, and $X_i^{(l)}$ denotes the l-th component of $\boldsymbol{X}_i^c$. When the model assumption (4.1) is true and the parameter space for $\boldsymbol{\beta}$ is compact, $\hat{\boldsymbol{\beta}}_{CEL}$ is well-defined according to the EL and CEL theory (Owen 2001; Kitamura et al. 2004).

A common bandwidth parameter b_N for different strata as in (4.3) is appropriate,

since from our assumption in Section 4.3, each stratum has size proportional to N. Results from Smith (2007) may serve as a rule of thumb for the order of b_N. That is, $b_N \to 0$, $N^{1-2\nu-2/\delta} b_N^{2q} \to \infty$ and $N^{1-2\nu} b_N^{5q/2} \to \infty$ as $N \to \infty$, where $\nu \in (0, 1/2)$, $\delta \geq 8$ and $q = \max_i q_i$. In data analysis, however, a data-driven bandwidth selection procedure is often desired. In this paper we implement a cross-validation criterion suggested by Newey (1993) for models with conditional moment restrictions, with some modifications to suit for the context of possibly unbalanced longitudinal data. Specifically, define

$$CV(b_N) = \mathrm{tr}\left\{ \sum_{i=1}^{N} \hat{\boldsymbol{G}}_i^\top \hat{\boldsymbol{V}}_{-i}^{-1} \left(\boldsymbol{g}_i \boldsymbol{g}_i^\top - \hat{\boldsymbol{V}}_{-i} \right) \hat{\boldsymbol{V}}_{-i}^{-1} \left(\boldsymbol{g}_i \boldsymbol{g}_i^\top - \hat{\boldsymbol{V}}_{-i} \right) \hat{\boldsymbol{V}}_{-i}^{-1} \hat{\boldsymbol{G}}_i \right\}, \tag{4.4}$$

where $\hat{\boldsymbol{G}}_i = \partial \boldsymbol{g}_i(\hat{\boldsymbol{\beta}})/\partial \boldsymbol{\beta}$, $\hat{\boldsymbol{V}}_{-i} = \sum_{j \in S_i} \hat{w}_{ij} \boldsymbol{g}_j(\hat{\boldsymbol{\beta}}) \boldsymbol{g}_j(\hat{\boldsymbol{\beta}})^\top$, $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(b_N)$ is the CEL estimator obtained with a given b_N, and

$$\hat{w}_{ii} = 0, \quad \hat{w}_{ij} = \frac{\mathcal{K}\left(\frac{\boldsymbol{X}_i^c - \boldsymbol{X}_j^c}{b_N} \right) I(\boldsymbol{X}_i^d = \boldsymbol{X}_j^d)}{\sum_{j \in S_i, j \neq i} \mathcal{K}\left(\frac{\boldsymbol{X}_i^c - \boldsymbol{X}_j^c}{b_N} \right) I(\boldsymbol{X}_i^d = \boldsymbol{X}_j^d)} \quad \text{for } j \in S_i \text{ and } j \neq i.$$

The optimal bandwidth b_N is chosen as the minimizer of $CV(b_N)$.

4.2.2 Reformulation and connection to GEE

Calculating $\hat{\boldsymbol{\beta}}_{CEL}$ requires us to solve a constrained optimization problem. Using Lagrange multipliers, define the Lagrangian as

$$\mathcal{L} = \sum_{i=1}^{N} \left(\sum_{j \in S_i} w_{ij} \log p_{ij} \right) - \sum_{i=1}^{N} \varpi_i \left(\sum_{j \in S_i} p_{ij} - 1 \right) - \sum_{i=1}^{N} \boldsymbol{\lambda}_i^\top \left\{ \sum_{j \in S_i} p_{ij} \boldsymbol{g}_j(\boldsymbol{\beta}) \right\},$$

where scalar $\varpi_i \in \mathcal{R}$ and vector $\boldsymbol{\lambda}_i \in \mathcal{R}^{n_i}$ are Lagrange multipliers associated with the second and third constraints in (4.2), respectively. Taking $\partial \mathcal{L}/\partial p_{ij} = 0$, together with the constraints in (4.2), it can be easily shown that, for a fixed $\boldsymbol{\beta}$,

$$p_{ij}(\boldsymbol{\beta}) = \frac{w_{ij}}{1 + \hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta})^\top \boldsymbol{g}_j(\boldsymbol{\beta})}, \quad j \in S_i, \quad i = 1, \ldots, N, \tag{4.5}$$

where $\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta})$ is the root of equation $\sum_{j\in S_i} p_{ij}(\boldsymbol{\beta})\boldsymbol{g}_j(\boldsymbol{\beta}) = \mathbf{0}$. It is easy to see that

$$\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\lambda}_i\in\mathcal{R}^{n_i}} \left[-\sum_{j\in S_i} w_{ij}\log\left\{1+\boldsymbol{\lambda}_i^\top \boldsymbol{g}_j(\boldsymbol{\beta})\right\}\right] \stackrel{\text{def}}{=} \arg\min_{\boldsymbol{\lambda}_i\in\mathcal{R}^{n_i}} \Lambda_i(\boldsymbol{\lambda}_i,\boldsymbol{\beta}). \tag{4.6}$$

where $\Lambda_i(\boldsymbol{\lambda}_i,\boldsymbol{\beta}) = -\sum_{j\in S_i} w_{ij}\log\left\{1+\boldsymbol{\lambda}_i^\top \boldsymbol{g}_j(\boldsymbol{\beta})\right\}$. Thus the objective function in (4.2) can be rewritten as a function of $\boldsymbol{\beta}$ only, namely

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{N}\sum_{j\in S_i} w_{ij}\log p_{ij} = \sum_{i=1}^{N} L_i(\boldsymbol{\beta}) + \sum_{i=1}^{N}\sum_{j\in S_i} w_{ij}\log w_{ij}, \tag{4.7}$$

where $L_i(\boldsymbol{\beta}) = \Lambda_i\{\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta}),\boldsymbol{\beta}\}$. Therefore, the CEL estimator can be equivalently defined as

$$\hat{\boldsymbol{\beta}}_{CEL} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{N}\left\{\min_{\boldsymbol{\lambda}_i\in\mathcal{R}^{n_i}} \Lambda_i(\boldsymbol{\lambda}_i,\boldsymbol{\beta})\right\}. \tag{4.8}$$

Equation (4.8) provides a way to implement the CEL method, which will be detailed in Section 4.4.

As seen, the CEL method does not require to explicitly model the variance-covariance matrix $\mathrm{var}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i)$. However, this matrix is taken into account implicitly in the optimization procedure (4.6) through the use of probabilities p_{ij}. To elaborate, note that the third constraint in (4.2) implies

$$\mathbf{0} = \sum_{j\in S_i} p_{ij}(\boldsymbol{\beta})\left\{1-\frac{w_{ij}}{p_{ij}(\boldsymbol{\beta})}\right\}\boldsymbol{g}_j(\boldsymbol{\beta}) + \sum_{j\in S_i} w_{ij}\boldsymbol{g}_j(\boldsymbol{\beta}),$$

and also from (4.5) we have

$$1-\frac{w_{ij}}{p_{ij}(\boldsymbol{\beta})} = -\boldsymbol{g}_j^\top(\boldsymbol{\beta})\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta}), \quad j\in S_i.$$

Therefore, solving for $\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta})$ leads to

$$\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta}) = \left\{\sum_{j\in S_i} p_{ij}(\boldsymbol{\beta})\boldsymbol{g}_j(\boldsymbol{\beta})\boldsymbol{g}_j^\top(\boldsymbol{\beta})\right\}^{-1}\left\{\sum_{j\in S_i} w_{ij}\boldsymbol{g}_j(\boldsymbol{\beta})\right\}. \tag{4.9}$$

From the proof of Theorem IV.3 in Section 4.3, the term $\sum_{j\in S_i} p_{ij}(\hat{\boldsymbol{\beta}}_{CEL})\boldsymbol{g}_j(\hat{\boldsymbol{\beta}}_{CEL})\boldsymbol{g}_j(\hat{\boldsymbol{\beta}}_{CEL})^\top$ is a consistent estimator of $\mathrm{var}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i)$. This implies that a consistent estimator of

$\text{var}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i)$ is involved in (4.9), which is the optimal solution to (4.6). Due to the positive definiteness of $\text{var}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i)$, $\sum_{j\in S_i} p_{ij}(\hat{\boldsymbol{\beta}}_{CEL})\boldsymbol{g}_j(\hat{\boldsymbol{\beta}}_{CEL})\boldsymbol{g}_j(\hat{\boldsymbol{\beta}}_{CEL})^\top$ is positive definite, at least when N is large. Note that $\sum_{j\in S_i} p_{ij}(\hat{\boldsymbol{\beta}}_{CEL})\boldsymbol{g}_j(\hat{\boldsymbol{\beta}}_{CEL})\boldsymbol{g}_j(\hat{\boldsymbol{\beta}}_{CEL})^\top$ is a data-driven estimate of the variance-covariance matrix, and thus the CEL method has the flexibility of allowing the variance-covariance to be different across individuals, unlike the GEE method which requires a common within-subject correlation structure, such as AR-1 or compound symmetry.

Since $\hat{\boldsymbol{\beta}}_{CEL}$ maximizes $L(\boldsymbol{\beta})$, $\hat{\boldsymbol{\beta}}_{CEL}$ is the solution to equation $\partial L(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \mathbf{0}$. From (4.7) and the fact that $\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta})$ is the root of $\sum_{j\in S_i} p_{ij}(\boldsymbol{\beta})\boldsymbol{g}_j(\boldsymbol{\beta}) = \mathbf{0}$, $\hat{\boldsymbol{\beta}}_{CEL}$ is the solution to equation

$$\sum_{i=1}^{N}\sum_{j\in S_i} p_{ij}(\boldsymbol{\beta})\boldsymbol{G}_j(\boldsymbol{\beta})^\top\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta}) = \mathbf{0},$$

where $\boldsymbol{G}_j(\boldsymbol{\beta}) = \partial\boldsymbol{g}_j(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$. Plugging $\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta})$ given by (4.9) into this equation leads to

$$\sum_{i=1}^{N}\left\{\sum_{j\in S_i} p_{ij}(\boldsymbol{\beta})\boldsymbol{G}_j(\boldsymbol{\beta})^\top\right\}\left\{\sum_{j\in S_i} p_{ij}(\boldsymbol{\beta})\boldsymbol{g}_j(\boldsymbol{\beta})\boldsymbol{g}_j(\boldsymbol{\beta})^\top\right\}^{-1}\left\{\sum_{j\in S_i} w_{ij}\boldsymbol{g}_j(\boldsymbol{\beta})\right\} = \mathbf{0}.$$

Being the solution to the above estimating equation, $\hat{\boldsymbol{\beta}}_{CEL}$ can be regarded as a GEE-type estimator, where the matrix $\text{var}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i)$ is estimated nonparametrically by using the empirical probabilities p_{ij} rather than parametrically by assuming a working model. Because $\hat{\boldsymbol{\beta}}_{CEL}$ is obtained under a consistently estimated $\text{var}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i)$, it in theory will have improved efficiency upon the GEE estimator employing an incorrectly modeled variance-covariance matrix.

4.3 Large Sample Properties

To facilitate the presentation, we introduce some extra notation. It is clear that $\boldsymbol{G}_i(\boldsymbol{\beta}) = E\{\partial\boldsymbol{g}_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta} \mid \boldsymbol{X}_i\}$. Denote $\boldsymbol{V}_i(\boldsymbol{\beta}) = E\left\{\boldsymbol{g}_i(\boldsymbol{\beta})\boldsymbol{g}_i(\boldsymbol{\beta})^\top \mid \boldsymbol{X}_i\right\}$. For any set

A, let $|A|$ denote the cardinality of A. Define $r_\omega = \lim_{N\to\infty} |S^\omega|/N$. As shown in the proofs of the following theorems, a stratum with $r_\omega = 0$ will be asymptotically negligible. Therefore, without loss of generality, we assume that $r_\omega > 0$ for all $\omega \in \Omega$. This assumption, together with the fact that for any $\omega \in \Omega$, $E\{\boldsymbol{g}_i(\boldsymbol{\beta}_0) \mid \boldsymbol{X}_i\} = \mathbf{0}$ for $i \in S^\omega$, implies that $\mathbf{0}$ is inside the convex hull of $\{\boldsymbol{g}_j(\boldsymbol{\beta}) : j \in S_i\}$ for $i = 1, \cdots, N$, at least when N is large and $\boldsymbol{\beta}$ is close to $\boldsymbol{\beta}_0$. In other words, constraints in (4.2) are satisfied asymptotically. Technical assumptions and proofs for theoretical results developed in this section are provided in Section 4.8.

Consistency of the proposed CEL estimator is established in the following theorem.

Theorem IV.1. *Under Assumptions (i)-(vii) in Section 4.8, we have* $\hat{\boldsymbol{\beta}}_{CEL} \xrightarrow{p} \boldsymbol{\beta}_0$ *as* $N \to \infty$.

The following theorem presents the asymptotic distribution of the CEL estimator.

Theorem IV.2. *Under Assumptions (i)-(viii) in Section 4.8, we have* $\sqrt{N}(\hat{\boldsymbol{\beta}}_{CEL} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}\{\mathbf{0}, \boldsymbol{J}(\boldsymbol{\beta}_0)^{-1}\}$ *as* $N \to \infty$, *where* $\boldsymbol{J}(\boldsymbol{\beta}_0) = \sum_{\omega\in\Omega} r_\omega \boldsymbol{J}_\omega(\boldsymbol{\beta}_0)$, *and for all* $i \in S^\omega$, $\boldsymbol{J}_\omega(\boldsymbol{\beta}_0) = E\left\{\boldsymbol{G}_i(\boldsymbol{\beta}_0)^\top \boldsymbol{V}_i(\boldsymbol{\beta}_0)^{-1} \boldsymbol{G}_i(\boldsymbol{\beta}_0)\right\}$.

For balanced longitudinal data where all subjects belong to a unique stratum, we have $\boldsymbol{J}(\boldsymbol{\beta}_0) = E\left\{\boldsymbol{G}(\boldsymbol{\beta}_0)^\top \boldsymbol{V}(\boldsymbol{\beta}_0)^{-1} \boldsymbol{G}(\boldsymbol{\beta}_0)\right\}$. This is the semiparametric efficiency bound for the model defined by (4.1) (Chamberlain 1987) when the data are balanced. Thus, $\hat{\boldsymbol{\beta}}_{CEL}$ has the optimal efficiency among all regular and asymptotically linear estimators under model (4.1) with balanced longitudinal data.

To consistently estimate $\boldsymbol{J}(\boldsymbol{\beta}_0)$, we consider two approaches. The first approach is based on Lemma C1 in Kitamura et al. (2004). As one can show that, for any $\omega \in \Omega$,

$$-\frac{1}{|S^\omega|}\frac{\partial^2 L^\omega}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}(\hat{\boldsymbol{\beta}}_{CEL}) \xrightarrow{p} \boldsymbol{J}_\omega(\boldsymbol{\beta}_0)$$

as $N \to \infty$, where $L^{\omega}(\boldsymbol{\beta})$ is defined by

$$L^{\omega}(\boldsymbol{\beta}) = \sum_{i \in S^{\omega}} L_i(\boldsymbol{\beta}) + \sum_{i \in S^{\omega}} \sum_{j \in S_i} w_{ij} \log w_{ij},$$

a consistent estimator of $\boldsymbol{J}(\boldsymbol{\beta}_0)$ takes the following form that is similar to an observed information matrix:

$$-\frac{1}{N} \frac{\partial^2 L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\top}} (\hat{\boldsymbol{\beta}}_{CEL}). \tag{4.10}$$

The calculation of (4.10) is straightforward, and is a byproduct of the Newton-Raphson algorithm searching for $\hat{\boldsymbol{\beta}}_{CEL}$. See more discussion in Section 4.4.

The second approach is to estimate $\boldsymbol{J}(\boldsymbol{\beta}_0)$ using the estimator given by the following theorem.

Theorem IV.3. *Under Assumptions (i)-(ix) in Section 4.8, as $N \to \infty$, we have*

$$\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{G}_i(\hat{\boldsymbol{\beta}}_{CEL})^{\top} \left\{ \sum_{j \in S_i} p_{ij}(\hat{\boldsymbol{\beta}}_{CEL}) \boldsymbol{g}_j(\hat{\boldsymbol{\beta}}_{CEL}) \boldsymbol{g}_j(\hat{\boldsymbol{\beta}}_{CEL})^{\top} \right\}^{-1} \boldsymbol{G}_i(\hat{\boldsymbol{\beta}}_{CEL}) \xrightarrow{p} \boldsymbol{J}(\boldsymbol{\beta}_0) \tag{4.11}$$

The above estimator is a simplification of the well-known "sandwich" estimator (Liang and Zeger 1986)

$$\left\{ \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{G}_i^{\top} \boldsymbol{V}_{\mathrm{w},i}^{-1} \boldsymbol{G}_i \right\}^{-1} \left\{ \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{G}_i^{\top} \boldsymbol{V}_{\mathrm{w},i}^{-1} \mathrm{var}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) \boldsymbol{V}_{\mathrm{w},i}^{-1} \boldsymbol{G}_i \right\} \left\{ \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{G}_i^{\top} \boldsymbol{V}_{\mathrm{w},i}^{-1} \boldsymbol{G}_i \right\}^{-1},$$

where $\boldsymbol{V}_{\mathrm{w}}$ is a working variance-covariance matrix. For the CEL method, $\boldsymbol{V}_{\mathrm{w},i} = \sum_{j \in S_i} p_{ij}(\boldsymbol{\beta}) \boldsymbol{g}_j(\boldsymbol{\beta}) \boldsymbol{g}_j(\boldsymbol{\beta})^{\top}$ consistently estimates $\mathrm{var}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i)$. Therefore, the variability matrix and the sensitivity matrix in the "sandwich" estimator cancel each other asymptotically, yielding the estimator given in (4.11).

Although both estimators (4.10) and (4.11) are consistent, they may perform differently under finite sample size. A numerical comparison and some practical suggestions will be given in the simulation studies in Section 4.5.

Finally, the optimality of $\hat{\boldsymbol{\beta}}_{CEL}$ is established by the following theorem.

Theorem IV.4. *Under Assumptions (i)-(viii) in Section 4.8, we have* $\sqrt{N}(\hat{\boldsymbol{\beta}}_{CEL} - \hat{\boldsymbol{\beta}}_{opt}) = o_p(1)$ *as* $N \to \infty$*, where* $\hat{\boldsymbol{\beta}}_{opt}$ *is the most efficient estimator among the class of estimators that solve estimating equations of the form* $\sum_{i=1}^{N} \boldsymbol{D}(\boldsymbol{X}_i, \boldsymbol{\beta})\boldsymbol{g}_i(\boldsymbol{\beta}) = 0$*, with* $\boldsymbol{D}(\boldsymbol{X}_i, \boldsymbol{\beta})$ *being a* $p \times n_i$ *matrix that may vary across different subjects.*

This result can be easily justified by noting that $\hat{\boldsymbol{\beta}}_{opt}$ should solve equation $\sum_{i=1}^{N} \boldsymbol{G}_i(\boldsymbol{\beta})^\top \boldsymbol{V}_i(\boldsymbol{\beta})^{-1}\boldsymbol{g}_i(\boldsymbol{\beta}) = \mathbf{0}$ based on the theory of inference functions (e.g. Godambe 1991; Heyde 1997). One special case is when the longitudinal outcomes follow a normal distribution, in which this equation is actually the score equation. In such a case, $\hat{\boldsymbol{\beta}}_{CEL}$ is asymptotically equivalent to the maximum likelihood estimator in theory.

4.4 Numerical Implementation

Numerical implementation of the proposed CEL method can be accomplished based on equation (4.8) with the invocation of nested optimization routines. Two optimization loops are required. The inner loop updates the Lagrange multipliers $\boldsymbol{\lambda}_i(\boldsymbol{\beta})$, $i = 1, \ldots, N$, for a fixed $\boldsymbol{\beta}$, and the outer loop updates $\boldsymbol{\beta}$. We use Newton-Raphson algorithm for both loops.

For a fixed $\boldsymbol{\beta}$, given $\boldsymbol{\lambda}_i^{old}$, $\boldsymbol{\lambda}_i$ is updated as

$$\boldsymbol{\lambda}_i^{new} = \boldsymbol{\lambda}_i^{old} - \boldsymbol{\Lambda}_{i,\boldsymbol{\lambda}\boldsymbol{\lambda}}^{-1}\left(\boldsymbol{\lambda}_i^{old}, \boldsymbol{\beta}\right)\boldsymbol{\Lambda}_{i,\boldsymbol{\lambda}}\left(\boldsymbol{\lambda}_i^{old}, \boldsymbol{\beta}\right),$$

where

$$\boldsymbol{\Lambda}_{i,\boldsymbol{\lambda}}(\boldsymbol{\lambda}_i, \boldsymbol{\beta}) = -\sum_{j \in S_i} w_{ij}\frac{\boldsymbol{g}_j(\boldsymbol{\beta})}{1 + \boldsymbol{\lambda}_i^\top \boldsymbol{g}_j(\boldsymbol{\beta})}, \quad \boldsymbol{\Lambda}_{i,\boldsymbol{\lambda}\boldsymbol{\lambda}}(\boldsymbol{\lambda}_i, \boldsymbol{\beta}) = \sum_{j \in S_i} w_{ij}\frac{\boldsymbol{g}_j(\boldsymbol{\beta})\boldsymbol{g}_j(\boldsymbol{\beta})^\top}{\left\{1 + \boldsymbol{\lambda}_i^\top \boldsymbol{g}_j(\boldsymbol{\beta})\right\}^2}.$$

The algorithm may start with an initial value $\boldsymbol{\lambda}_i = 0$. The converged value gives the estimated $\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta})$ for a fixed $\boldsymbol{\beta}$ as the solution to (4.6).

For a given $\boldsymbol{\beta}^{old}$, $\boldsymbol{\beta}$ is updated as

$$\boldsymbol{\beta}^{new} = \boldsymbol{\beta}^{old} - \left\{ \sum_{i=1}^{N} \boldsymbol{L}_{i,\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}^{old}) \right\}^{-1} \left\{ \sum_{i=1}^{N} \boldsymbol{L}_{i,\boldsymbol{\beta}}(\boldsymbol{\beta}^{old}) \right\},$$

where

$$\begin{aligned}
\boldsymbol{L}_{i,\boldsymbol{\beta}}(\boldsymbol{\beta}) &= -\sum_{j \in S_i} w_{ij} \frac{\boldsymbol{G}_j(\boldsymbol{\beta})^\top}{1 + \hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta})^\top \boldsymbol{g}_j(\boldsymbol{\beta})} \hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta}), \\
\boldsymbol{L}_{i,\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}) &= -\boldsymbol{\Lambda}_{i,\boldsymbol{\lambda}\boldsymbol{\beta}}^\top \left\{ \hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta}), \boldsymbol{\beta} \right\} \boldsymbol{\Lambda}_{i,\boldsymbol{\lambda}\boldsymbol{\lambda}}^{-1} \left\{ \hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta}), \boldsymbol{\beta} \right\} \boldsymbol{\Lambda}_{i,\boldsymbol{\lambda}\boldsymbol{\beta}} \left\{ \hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta}), \boldsymbol{\beta} \right\}, \\
\boldsymbol{\Lambda}_{i,\boldsymbol{\lambda}\boldsymbol{\beta}} \left\{ \hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta}), \boldsymbol{\beta} \right\} &= \sum_{j \in S_i} w_{ij} \frac{\boldsymbol{g}_j(\boldsymbol{\beta}) \hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta})^\top \boldsymbol{G}_j(\boldsymbol{\beta})}{\left\{ 1 + \hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta})^\top \boldsymbol{g}_j(\boldsymbol{\beta}) \right\}^2} - \sum_{j \in S_i} w_{ij} \frac{\boldsymbol{G}_j(\boldsymbol{\beta})}{1 + \hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta})^\top \boldsymbol{g}_j(\boldsymbol{\beta})}.
\end{aligned}$$

Iterate the inner and outer loops until certain convergence criterion is satisfied. At the convergence, the algorithm produces the CEL estimator $\hat{\boldsymbol{\beta}}_{CEL}$. It is easy to see that (4.6) is a convex minimization problem. Therefore, for a fixed $\boldsymbol{\beta}$, the inner loop almost always converges to the global minimizer. A proof of this convergence can be given following Chen et al. (2002). The maximization in the outer loop is more complicated, and the convergence of the Newton-Raphson algorithm may not be guaranteed. See Owen (2001) for some detailed discussion on related issues in the setting of unconditional moment restrictions. Nonetheless, the nested optimization is widely used by many researchers to implement the EL (CEL) method. See, for example, Owen (2001), Kitamura (2007) and Hansen (2013). According to Kitamura (2007), the nested optimization appears to be "the most stable way to compute the EL estimator".

A complication associated with the numerical implementation arises from the constraints in (4.2), in that $\mathbf{0}$ may not be in the convex hull spanned by $\{\boldsymbol{g}_j(\boldsymbol{\beta}) : j \in S_i\}$, for either all or certain values of $\boldsymbol{\beta}$. This could happen for many reasons. For example, assumption (4.1) is incorrect, or the number of follow-ups is large, or the size of a stratum is small, or the initial value of $\boldsymbol{\beta}$ is far from the true $\boldsymbol{\beta}_0$, or combinations

of the above. If this happens, the constraints in (4.2) are no longer compatible, thus no solution exists. Specifically, some probabilities provided by (4.5) may be negative. A practical suggestion to overcome this difficulty is to restrict the minimization in (4.8) over the legitimate region $\left\{\boldsymbol{\lambda}_i \in \mathcal{R}^{n_i} : 1 + \boldsymbol{\lambda}_i^\top \boldsymbol{g}_j(\boldsymbol{\beta}) \geq w_{ij},\ j \in S_i\right\}$, $i = 1, \cdots, N$. Other possible solutions may be found in, for example, Owen (2001), Chen et al. (2008) and Liu and Chen (2010).

For unbalanced longitudinal data, to avoid potential algorithmic failure caused by small stratum size, we consider a strategy of data augmentation by creating pseudo-subjects from those who have longer and more complete follow-up visits. Specifically, for subject i, we redefine $S_i = \{j : 1 \leq j \leq N \text{ and } \tau_j \supseteq \tau_i\}$ and

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{N} \min_{\boldsymbol{\lambda}_i \in \mathcal{R}^{n_i}} \left[-\sum_{j \in S_i} w_{ij} \log\left\{1 + \boldsymbol{\lambda}_i^\top \boldsymbol{g}_j^{\dagger_i}(\boldsymbol{\beta})\right\} \right] + \sum_{i=1}^{N} \sum_{j \in S_i} w_{ij} \log w_{ij},$$

where the superscript $\dagger_i$ means selecting components from the residual vector $\boldsymbol{g}_j(\boldsymbol{\beta})$ according to the set τ_i. Therefore, $\boldsymbol{g}_j^{\dagger_i}(\boldsymbol{\beta})$ may be considered as the residual vector of a pseudo-subject created from subject j whose visit times contain subject i's visit times as a subset. As N increases, the small-stratum-size problem is unlikely to occur with the creation of pseudo-subjects and with our assumption that each stratum has size proportional to N. In the extreme case that a small-size stratum still exists, for example, a stratum with only one subject, that stratum can be discarded. In the simulation studies in Section 4.5, we employ this strategy, which improves the numerical performance of $\hat{\boldsymbol{\beta}}_{CEL}$ in general.

In the case of binary outcome data, we have $\text{var}(Y_{it} \mid \boldsymbol{X}_i) = \mu_{it}(1-\mu_{it})$ where $\mu_{it} = \mu(\boldsymbol{X}_{it}^\top \boldsymbol{\beta})$. This information could be used to improve the finite sample performance

of the CEL estimator by redefining the residual vector to be

$$\boldsymbol{g}_i(\boldsymbol{\beta}) = \Big\{ Y_{i1} - \mu_{i1}, \ldots, Y_{in_i} - \mu_{in_i}, \\ (Y_{i1} - \mu_{i1})^2 - \mu_{i1}(1-\mu_{i1}), \ldots, (Y_{in_i} - \mu_{in_i})^2 - \mu_{in_i}(1-\mu_{in_i}) \Big\}^\top.$$

Model assumption (4.1) is still satisfied. Estimation and inference remain the same, except that the dimension of the Lagrange multiplier $\boldsymbol{\lambda}_i$ increases to $2n_i$.

4.5 Simulation Experiments

The setup of our simulation experiments concerns a typical longitudinal study that involves three covariates: a continuous baseline covariate b_i, a binary treatment indicator d_i, and the visit time t. Assuming that there are five designed follow-up visits, we generate the longitudinal outcomes Y_{it} from the following model:

$$Y_{it} = \beta_1 + \beta_2 b_i + \beta_3 d_i + \beta_4 t + \beta_5 d_i \times t + \epsilon_{it}, \quad t = 1, \cdots, 5, \quad i = 1, \cdots, N,$$

where $b_i \sim N(0, 3^2)$ and is truncated between -7 and 7, $d_i \sim$ Bernoulli(0.5), $d_i \times t$ is the treatment and time interaction, $(\epsilon_{i1}, \ldots, \epsilon_{i5})^\top \sim N_5(\mathbf{0}, \boldsymbol{V}_i^{1/2} \boldsymbol{R} \boldsymbol{V}_i^{1/2})$, $\boldsymbol{V}_i$ is a 5×5 diagonal matrix with the t-th diagonal element $\sigma_{it}^2 = \exp(\alpha_1 + \alpha_2 |b_i| + \alpha_3 t)$, and $\boldsymbol{R}$ takes the first-order autoregressive structure with correlation coefficient ρ. The true values are set as $(\beta_1, \ldots, \beta_5)^\top = (0.5, 0.5, 1.0, 0.3, 0.3)^\top$, $(\alpha_1, \alpha_2, \alpha_3)^\top = (0.4, 0.6, 0.4)^\top$, and $\rho = 0.5$.

To create unbalanced data, the probability that subject i misses the t-th visit is given by

$$P(Y_{it} \text{ is missing}) = \frac{1}{1 + \exp(3 - 0.1|b_i| - 0.01t)}.$$

In the missing data literature, this missingness mechanism is still MCAR (Little and Rubin 2002), despite the dependence of the missingness probability on covariates.

We only use this missingness model to generate our data. Because of the nature of MCAR, the missingness mechanism is not needed to be accounted for in the analysis. To ease the computational burden, we control the total number of strata by discarding subjects with no more than two visits. It can be verified that the final observed residuals satisfy model assumption (4.1). On average, approximately 30% of subjects have incomplete visits in the generated data, and their follow-up patterns are irregular. We conduct 500 replications to calculate the summary results. The Gaussian kernel is employed to calculate the weights, where the baseline covariate is standardized to have mean 0 and variance 1 in bandwidth selection. We compare the CEL estimator with the GEE estimator under different working correlation structures, including the true one. In addition, we compare the CEL estimator with Wang et al.'s (2010) GEL estimator, for which the variance-covariance matrix is estimated separately using various models. In effect, Wang et al.'s GEL estimator may be regarded as either a GEE2 (Prentice and Zhao, 1991) estimator, or an EL estimator based on a set of unconditional moment conditions specified by GEE2. Therefore, the comparison between CEL and GEL may be referred to either as the comparison between CEL and GEE2 or as the comparison between CEL and GEE2-based EL.

Table 4.1 presents summarized results of comparisons for both $N = 100$ and $N = 500$. To speed up computing, a prefixed bandwidth $b_N = 2N^{-1/10}$ is used, and this rate of b_N converging to 0 is selected using the rule of thumb described in Section 4.2.1. When σ_{it}^2 is not correctly modeled, Table 4.1 clearly indicates that $\hat{\boldsymbol{\beta}}_{CEL}$ outperforms both the GEE estimator and the GEL estimator in terms of mean square error, regardless of whether the correlation structure is correctly specified. More specifically, when $N = 100$, the CEL estimator reduces the total

Table 4.1: Numerical comparison of different estimators for unbalanced longitudinal data analysis under different sample sizes. The numbers have been multiplied by 100

		cel	*gee.ar*	*gee.cs*	*gel*1.*ar*	*gel*1.*cs*	*gel*2.*ar*	*gel*2.*cs*	*gel*3.*ar*	*gel*3.*cs*
N=100	$\boldsymbol{\beta}_1$	3	-2	-2	-1	-1	-1	-1	2	2
		(64)	(82)	(84)	(72)	(73)	(72)	(73)	(47)	(49)
		[42]	[67]	[70]	[52]	[53]	[52]	[54]	[23]	[24]
	$\boldsymbol{\beta}_2$	-1	-1	-1	-1	-1	-1	-1	-1	-1
		(19)	(24)	(23)	(19)	(20)	(19)	(20)	(14)	(14)
		[4]	[6]	[5]	[4]	[4]	[4]	[4]	[2]	[2]
	$\boldsymbol{\beta}_3$	-5	-1	-1	-2	-2	-2	-1	-4	-4
		(92)	(113)	(118)	(100)	(101)	(100)	(102)	(65)	(67)
		[86]	[128]	[140]	[99]	[103]	[99]	[104]	[42]	[45]
	$\boldsymbol{\beta}_4$	-1	1	1	1	1	1	1	-1	-1
		(26)	(30)	(29)	(27)	(27)	(27)	(27)	(19)	(19)
		[7]	[9]	[9]	[7]	[7]	[7]	[7]	[4]	[4]
	$\boldsymbol{\beta}_5$	1	1	0	1	0	1	0	3	2
		(37)	(44)	(43)	(39)	(39)	(39)	(39)	(25)	(25)
		[14]	[19]	[19]	[16]	[15]	[16]	[15]	[6]	[6]
N=500	$\boldsymbol{\beta}_1$	4	2	2	3	3	3	3	2	2
		(28)	(36)	(38)	(31)	(33)	(31)	(33)	(21)	(22)
		[8]	[13]	[14]	[10]	[11]	[10]	[11]	[4]	[5]
	$\boldsymbol{\beta}_2$	0	-1	-1	-1	-1	-1	-1	-1	0
		(8)	(11)	(11)	(9)	(9)	(9)	(9)	(6)	(6)
		[1]	[1]	[1]	[1]	[1]	[1]	[1]	[0]	[0]
	$\boldsymbol{\beta}_3$	-7	-4	-5	-6	-6	-6	-6	-3	-2
		(39)	(51)	(55)	(44)	(47)	(45)	(48)	(30)	(32)
		[16]	[26]	[30]	[20]	[23]	[20]	[23]	[9]	[10]
	$\boldsymbol{\beta}_4$	-1	0	-1	-1	-1	-1	-1	-1	-1
		(11)	(14)	(14)	(13)	(13)	(13)	(13)	(8)	(8)
		[1]	[2]	[2]	[2]	[2]	[2]	[2]	[1]	[1]
	$\boldsymbol{\beta}_5$	3	2	2	2	2	2	2	1	1
		(16)	(21)	(21)	(18)	(19)	(18)	(19)	(11)	(12)
		[3]	[4]	[4]	[3]	[4]	[3]	[4]	[1]	[1]

cel: CEL; *gee* GEE; *gel*1, *gel*2, *gel*3: GEL with σ_{it}^2 being modeled by $\alpha_1 + \alpha_2 t$, $\exp(\alpha_1 + \alpha_2 t)$ and the truth, respectively; *ar*: first-order autoregressive; *cs*: compound symmetry.Three summary statistics are calculated: bias, empirical standard error (number in parentheses), and mean square error (number in brackets).

mean square error by 34%, 38%, 15%, 17%, 15% and 18% compared to the 2nd to the 7th estimators, respectively; and when $N = 500$, the amount of reduction becomes even higher as 40%, 46%, 22%, 30%, 23% and 31%, respectively. Such robustness against incorrect modeling of the variance-covariance matrix is a clear advantage of our proposed method. Since the data are generated from the normal distribution, $\hat{\boldsymbol{\beta}}_{CEL}$ is asymptotically equivalent to the maximum likelihood estimator, which is essentially the estimator given by the second last column. Due to the nonparametric nature of the CEL estimator, its finite-sample total mean square error is larger than that of the maximum likelihood estimator. However, this difference becomes smaller as the sample size increases. Specifically, the total mean square error of the CEL estimator is 98% higher than that of the maximum likelihood estimator when $N = 100$, and this number drops to 79% when $N = 500$. A similar trend is observed by comparing the CEL estimator with the estimator given by the last column, in which case the former has 86% higher total mean square error than the latter when $N = 100$, and this number drops to 61% when $N = 500$. From Table 4.1, an additional observation is that, under the same model for σ_{it}^2, misspecification of the correlation structure causes only marginal efficiency loss. On the other hand, different models for σ_{it}^2 lead to significant efficiency variation. This is in a full agreement with the conclusion drawn by Wang and Lin (2005).

Table 4.2 reports results concerning the influence of the bandwidth on the performance of the CEL estimator. With $N = 300$, the bandwidth is given by $b_N = cN^{-1/10}$, which varies with $c = 0.6, 0.9, \ldots, 2.7$. Table 4.2 also includes results based on the cross-validation-selected bandwidth. The value of c that gives the smallest total mean square error is 0.9. When c becomes larger (toward oversmoothing), both

Table 4.2: Effect of bandwidth parameter on the CEL estimator for unbalanced longitudinal data analysis. The numbers have been multiplied by 100

	$c = 0.6$	$c = 0.9$	$c = 1.2$	$c = 1.5$	$c = 1.8$	$c = 2.1$	$c = 2.4$	$c = 2.7$	CV
	3	3	4	4	4	4	4	4	4
$\boldsymbol{\beta}_1$	(34)	(33)	(34)	(35)	(36)	(37)	(38)	(39)	(38)
	[11]	[11]	[12]	[13]	[13]	[14]	[15]	[15]	[14]
	0	1	-1	-1	-1	-1	-1	0	0
$\boldsymbol{\beta}_2$	(10)	(10)	(10)	(10)	(10)	(10)	(11)	(11)	(11)
	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]	[1]
	-4	-4	-5	-5	-5	-5	-6	-6	-6
$\boldsymbol{\beta}_3$	(48)	(47)	(48)	(49)	(50)	(51)	(52)	(53)	(52)
	[23]	[22]	[23]	[24]	[25]	[27]	[28]	[29]	[27]
	-1	-1	-1	-1	-1	-1	-1	-1	-1
$\boldsymbol{\beta}_4$	(13)	(13)	(13)	(14)	(14)	(15)	(15)	(15)	(15)
	[2]	[2]	[2]	[2]	[2]	[2]	[2]	[2]	[2]
	2	2	2	2	2	2	2	2	2
$\boldsymbol{\beta}_5$	(19)	(18)	(19)	(19)	(20)	(20)	(21)	(21)	(20)
	[4]	[3]	[3]	[4]	[4]	[4]	[4]	[5]	[4]

$N = 300$; $b_N = cN^{-1/10}$; CV: cross-validation. Three summary statistics are calculated: bias, empirical standard error (number in parentheses), and mean square error (number in brackets).

the bias (except for β_2) and the empirical standard error increase. Overall, the impact of different bandwidths does not appear to be dramatic. The total mean square error when $c = 0.9$ is 25% smaller than that when $c = 2.7$. This empirical property has also been reported in Kitamura et al. (2004), and is of practical importance, as it could save substantial computing cost on the search for the optimal bandwidth. Taking our simulation study as an example, for smaller values of c, such as 0.6 and 0.9, the algorithm requires much longer time to converge in comparison to cases with a larger value of c.

Table 4.3 reports the numerical performances of the two asymptotic variance estimators, (4.10) and (4.11), under different sample sizes. The bandwidth is prefixed as $b_N = 2N^{-1/10}$. The empirical variance of the CEL estimator based on 500 replications is used as the true value of the variance when calculating the bias. As shown in Table 4.3, both estimators appear to be consistent, and formula (4.10) has smaller bias but slightly larger variance. Both formulas tend to underestimate the variance

Table 4.3: Numerical comparison of asymptotic variance estimators for CEL method under different sample sizes.

	$N = 100$	$N = 300$	$N = 500$	$N = 1000$
$\text{var}(\boldsymbol{\beta}_1)_{(4.10)}$	-0.033	-0.008	-0.002	0.000
	(0.134)	(0.027)	(0.014)	(0.004)
$\text{var}(\boldsymbol{\beta}_1)_{(4.11)}$	-0.071	-0.015	-0.005	-0.001
	(0.103)	(0.022)	(0.011)	(0.004)
$\text{var}(\boldsymbol{\beta}_2)_{(4.10)}$	-0.024	-0.006	-0.003	-0.002
	(0.003)	(0.001)	(0.001)	(0.000)
$\text{var}(\boldsymbol{\beta}_2)_{(4.11)}$	-0.025	-0.007	-0.003	-0.002
	(0.002)	(0.001)	(0.000)	(0.000)
$\text{var}(\boldsymbol{\beta}_3)_{(4.10)}$	-0.084	0.003	0.003	0.002
	(0.240)	(0.045)	(0.021)	(0.007)
$\text{var}(\boldsymbol{\beta}_3)_{(4.11)}$	-0.164	-0.012	-0.003	0.000
	(0.179)	(0.035)	(0.017)	(0.006)
$\text{var}(\boldsymbol{\beta}_4)_{(4.10)}$	-0.008	0.000	0.000	0.000
	(0.023)	(0.004)	(0.002)	(0.001)
$\text{var}(\boldsymbol{\beta}_4)_{(4.11)}$	-0.014	-0.001	0.000	0.000
	(0.017)	(0.004)	(0.002)	(0.001)
$\text{var}(\boldsymbol{\beta}_5)_{(4.10)}$	-0.014	0.001	0.000	0.000
	(0.040)	(0.007)	(0.003)	(0.001)
$\text{var}(\boldsymbol{\beta}_5)_{(4.11)}$	-0.027	-0.001	-0.001	0.000
	(0.030)	(0.006)	(0.003)	(0.001)

$\text{var}()_{(4.10)}$: estimator based on formula (4.10); $\text{var}()_{(4.11)}$: estimator based on formula (4.11). Two summary statistics are calculated: bias and empirical standard error (number in parentheses).

Table 4.4: Effect of bandwidth parameter on asymptotic variance estimators for CEL method.

	$c = 0.6$	$c = 0.9$	$c = 1.2$	$c = 1.5$	$c = 1.8$	$c = 2.1$	$c = 2.4$	$c = 2.7$	CV
$\text{var}(\boldsymbol{\beta}_1)_{(4.10)}$	-0.052	-0.041	-0.025	-0.028	-0.013	-0.008	-0.010	-0.013	-0.012
	(0.020)	(0.030)	(0.035)	(0.037)	(0.030)	(0.027)	(0.027)	(0.026)	(0.030)
$\text{var}(\boldsymbol{\beta}_1)_{(4.11)}$	-0.039	-0.020	-0.014	-0.015	-0.014	-0.015	-0.017	-0.019	-0.016
	(0.013)	(0.016)	(0.019)	(0.021)	(0.021)	(0.022)	(0.022)	(0.023)	(0.023)
$\text{var}(\boldsymbol{\beta}_2)_{(4.10)}$	-0.009	-0.008	-0.007	-0.008	-0.007	-0.007	-0.007	-0.008	-0.007
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$\text{var}(\boldsymbol{\beta}_2)_{(4.11)}$	-0.007	-0.005	-0.005	-0.006	-0.006	-0.007	-0.007	-0.008	-0.007
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$\text{var}(\boldsymbol{\beta}_3)_{(4.10)}$	-0.109	-0.083	-0.047	-0.040	-0.008	0.003	-0.001	-0.006	-0.003
	(0.032)	(0.047)	(0.059)	(0.062)	(0.048)	(0.045)	(0.046)	(0.047)	(0.050)
$\text{var}(\boldsymbol{\beta}_3)_{(4.11)}$	-0.082	-0.041	-0.022	-0.015	-0.013	-0.012	-0.014	-0.017	-0.013
	(0.018)	(0.023)	(0.027)	(0.030)	(0.033)	(0.036)	(0.038)	(0.040)	(0.038)
$\text{var}(\boldsymbol{\beta}_4)_{(4.10)}$	-0.008	-0.007	-0.004	-0.004	-0.002	0.000	-0.001	-0.001	-0.001
	(0.004)	(0.005)	(0.006)	(0.006)	(0.005)	(0.004)	(0.004)	(0.004)	(0.005)
$\text{var}(\boldsymbol{\beta}_4)_{(4.11)}$	-0.006	-0.003	-0.002	-0.002	-0.002	-0.001	-0.002	-0.002	-0.002
	(0.002)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
$\text{var}(\boldsymbol{\beta}_5)_{(4.10)}$	-0.017	-0.013	-0.007	-0.006	-0.001	0.001	0.001	0.000	0.000
	(0.005)	(0.008)	(0.010)	(0.010)	(0.008)	(0.007)	(0.007)	(0.008)	(0.008)
$\text{var}(\boldsymbol{\beta}_5)_{(4.11)}$	-0.012	-0.005	-0.002	-0.001	-0.001	-0.001	-0.002	-0.002	-0.001
	(0.003)	(0.004)	(0.004)	(0.005)	(0.005)	(0.006)	(0.006)	(0.007)	(0.006)

$N = 300$; $b_N = cN^{-1/10}$; CV: cross-validation; $\text{var}()_{(4.10)}$: estimator based on formula (4.10); $\text{var}()_{(4.11)}$: estimator based on formula (4.11). Two summary statistics are calculated: bias and empirical standard error (number in parentheses).

of $\hat{\boldsymbol{\beta}}_{CEL,2}$, the CEL estimator of the regression coefficient for the continuous baseline covariate. The issue that the sandwich estimator may underestimate the asymptotic variance has been observed by many researchers; see, for example, Kauermann and Carroll (2001) and references therein.

Table 4.4 reports the influence of the bandwidth on the two asymptotic variance estimators when $N = 300$ and $b_N = cN^{-1/10}$, with $c = 0.6, 0.9, \ldots, 2.7$. We can see that small bandwidth leads to underestimation of the asymptotic variance for both estimators. As the bandwidth increases, such underestimation disappears, except for the variance of $\hat{\boldsymbol{\beta}}_{CEL,2}$. For a larger bandwidth, the bias of formula (4.10) is smaller than that of formula (4.11). When the bandwidth is selected by the cross-validation, both estimators perform reasonably well, and formula (4.10) has smaller bias.

To further compare formulas (4.10) and (4.11), we consider a Wald-type test

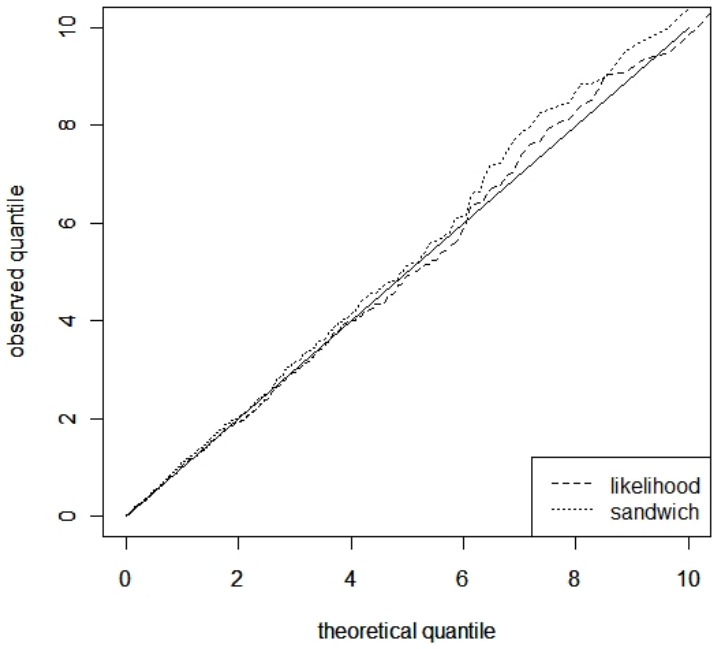

Figure 4.1: QQ-plot to compare the asymptotic χ^2_2 distribution of two Wald statistics based on "likelihood" (formula (4.10)) and "sandwich" (formula (4.11)) asymptotic variance estimators, respectively.

statistic to test for the following hypothesis of the overall treatment effect:

$$\mathrm{H}_0: \quad \beta_3 = 1 \text{ and } \beta_5 = 0.3.$$

Under H_0, both Wald statistics constructed from (4.10) and (4.11) follow the χ^2_2 distribution asymptotically. In the comparison, we take $N = 300$ and b_N is selected by the cross-validation. Figure 1 presents a QQ-plot for these two Wald statistics. It can be seen that although both statistics approximately follow the χ^2_2 distribution, the one based on formula (4.10) turns out to be closer to the 45^o line. This agrees with the conclusion drawn from Table 4 that the variance estimate given by (4.10) has smaller bias than that given by (4.11).

In summary, we recommend using the cross-validation to select the bandwidth,

and using formula (4.10) to estimate the asymptotic variance.

4.6 Data Application

We now illustrate the proposed CEL method by analyzing the data collected from the Kenya primary school nutritional intervention study (Neumann et al., 2003). This is a randomized controlled trial designed to examine the relationship of growth, cognitive development and physical activity with the intake of animal source foods, adjusted by other covariates. A total of 554 school children from 12 schools in rural Embu District, Kenya, are randomized to four nutritional intervention arms: Meat, Milk, Energy and Control. Meat, milk and energy in the form of extra vegetable oil are added into the local plant-based dish Githeri for the Meat, Milk and Energy arms, respectively. The Control arm has no feeding intervention. Lasting for over 2 years (from July 1998 to December 2000), this study includes a baseline visit during the period of July to August, 1998, and the delivery of intervention and data collection over the period of September 1998 to December 2000. The cognitive function is measured at the baseline.

We analyze a sub dataset of this study that concerns cognitive outcomes, and the data are downloaded from
http://rem.ph.ucla.edu/rob/mld/data/tabdelimiteddata/cognitive.txt. More details about the data can be found in Weiss (2005). The cognitive ability is measured by the Raven's score determined on the Raven's colored progressive matrices. Each subject has five follow-up visits, and the first (baseline) visit takes place prior to the randomization. The randomization is marked as time zero in our analysis.

Since all subjects effectively belong to the Control arm at the baseline visit, the baseline Raven's score is treated as a covariate (*braven*) in the model. Other covari-

Table 4.5: Analysis results on Kenya school nutritional intervention data ($N = 524$).

	cel			gee.ar			gee.cs			gee.un		
	est	se	p	est	se	p	est	se	p	est	se	p
intercept	11.53	0.955	0.00	11.67	1.055	0.00	11.44	1.072	0.00	11.47	1.061	0.00
age	0.119	0.075	0.11	0.111	0.082	0.18	0.149	0.082	0.07	0.137	0.082	0.09
ses	0.009	0.004	0.02	0.006	0.004	0.12	0.007	0.004	0.09	0.006	0.004	0.08
braven	0.241	0.037	0.00	0.251	0.044	0.00	0.247	0.045	0.00	0.250	0.044	0.00
boy	0.511	0.178	0.00	0.636	0.178	0.00	0.567	0.180	0.00	0.596	0.179	0.00
time	0.885	0.149	0.00	1.010	0.142	0.00	0.954	0.142	0.00	0.958	0.141	0.00
$energy \times time$	0.119	0.193	0.54	-0.127	0.189	0.50	-0.060	0.186	0.75	-0.089	0.186	0.63
$meat \times time$	0.538	0.199	0.01	0.354	0.203	0.08	0.405	0.204	0.05	0.392	0.201	0.05
$milk \times time$	-0.019	0.191	0.92	-0.273	0.188	0.15	-0.227	0.187	0.23	-0.234	0.186	0.21

cel: CEL; *gee* GEE; *ar*: first-order autoregressive; *cs*: compound symmetry; *un*: unstructured; est: estimated value; se: estimated standard error, where for CEL method, it is based on formula (4.10); p: p-value.

ates include the baseline age (age), baseline social economic status (ses) determined by an extensive survey, gender (boy), and visit time ($time$). For the intervention, three dummy variables, $meat$, $milk$ and $energy$ are created.

After removing the subjects with unmeasured covariates and the subjects who only have baseline visit, our analysis is based on 524 subjects with different patterns of follow-up. We use the average time of each visit in our analysis, and the following model is fitted for the data:

$$\begin{aligned} E(raven_{it}) &= \beta_1 + \beta_2 age_i + \beta_3 ses_i + \beta_4 braven_i + \beta_5 boy_i \\ &\quad + (\beta_6 + \beta_7 energy_i + \beta_8 meat_i + \beta_9 milk_i) \times time_t. \end{aligned}$$

Our residual analysis suggests that there is no evidence regarding the dependence of the marginal variance on any covariate. Therefore, we only compare results obtained from the CEL method and the GEE method (Table 4.5).

Both the CEL and the GEE methods find that the cognitive ability significantly improves over time, and is significantly higher for boys than for girls. It is interesting to note that our CEL method has identified social economic status as a significant factor for the cognitive ability, while the GEE method fails to detect this significance.

As for the growth rate under the three different nutritional interventions compared to the Control arm, our CEL method indicates that meat significantly improves the development of the cognitive ability, whereas the GEE method shows only a similar trend with no statistical significance. Neither the CEL nor the GEE finds significance of milk or energy intervention in helping cognitive growth.

4.7 Conclusions

In this paper we propose to model longitudinal data using the CEL method. No explicit modeling of the variance-covariance of the longitudinal outcomes is required. In the process of simulation experiments and data application, we have learned some additional insights that are worth mentioning. First, the CEL method enjoys high estimation efficiency when moderate to high level of heteroscedasticity exists, especially when it is difficult to postulate a reasonable model for such heteroscedasticity. In other words, when homoscedasticity is a more reasonable assumption, the CEL estimator may not outperform some of the existing estimators, such as the GEE estimator, due to the nonparametric weight calculation. Second, the cross-validation criterion (4.4) for bandwidth selection has not been theoretically justified yet, and it tends to produce over-smoothing. Third, when the number of covariates is large and the covariates vary in scales and/or types, the kernel-based weight calculation is challenging. Although through simulation studies Han (2013) and Han et al. (2013) demonstrated that the CEL method is not dramatically affected by moderately increasing the number of covariates, it is still worthwhile to explore more flexible methods to calculate the weights. Fourth, our stratification strategy dealing with the possible unbalancedness of longitudinal data is more appropriate for data collected from balanced study designs. Therefore, studies in which subjects visit at

irregular follow-up times call for some adjustments. For example, in analyzing the longitudinal nutrition data, the average time of each visit is used. Fifth, the stratification strategy may not perform well if one or more strata have small size. Although our suggested creation of pseudo-subjects is effective to reduce the impact of small stratum size, this issue may need some additional attention, especially when the total number of follow-up visits and the number of distinct visit patterns are both large. Last, but not least, a major hurdle for applying the CEL method is its computational burden, which is a well known open problem in the field of EL methodology. Future work on the development of fast algorithms is of great interest.

In the literature, in addition to the CEL method, there are other methods available for estimation under models defined by conditional moment restrictions, including Newey (1993) and Donald et al. (2003). Newey's (1993) method requires a preliminary estimator of the parameter of interest, and needs to explicitly estimate the variance-covariance matrix in a nonparametric fashion. However, it is often difficult to find a well-behaved estimate of this matrix in practice (Kitamura et al. 2004). The CEL method, on the contrary, achieves the same efficiency automatically without estimating the variance-covariance. Donald et al. (2003) proposed to construct an increasing number of unconditional moment restrictions to "span" the conditional moment restrictions that define the model, and use the EL method for estimation. As the sample size and the number of unconditional moment restrictions both go to infinity, Donald et al.'s (2003) estimator achieves the same efficiency as that of the CEL estimator. However, no numerical study of this method has been reported, therefore its finite sample performance is unknown.

4.8 Technical Assumptions and Proofs

This section is devoted to the technical assumptions and sketched proofs of the results stated in Section 4.3. Let $\mathcal{B}$ denote the domain of $\boldsymbol{\beta}$, and let $\mathcal{B}_0 \subseteq \mathcal{B}$ be some closed ball around $\boldsymbol{\beta}_0$. For a matrix A with elements A_{ij}, define $\| A \| = \sqrt{\sum_{i,j} A_{ij}^2}$. Denote Kitamura et al. (2004) as KTA.

Assumptions. *(i) There exists* $\omega_0 \in \Omega$*, such that for any* $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$*, we can find* $\mathcal{X}_{\boldsymbol{\beta}}$*, so that* $P(\boldsymbol{x} \in \mathcal{X}_{\boldsymbol{\beta}}) > 0$ *and* $E\{\boldsymbol{g}_i(\boldsymbol{\beta}) \mid \boldsymbol{X}_i = \boldsymbol{x}\} \neq 0$ *for every* $\boldsymbol{x} \in \mathcal{X}_{\boldsymbol{\beta}}$ *and* $i \in S^{\omega_0}$*. (ii) For any* $1 \leq i \leq N$*,* $E\left\{\sup_{\boldsymbol{\beta}\in\mathcal{B}} \|\boldsymbol{g}_i(\boldsymbol{\beta})\|^m\right\} < \infty$ *for some* $m \geq 2$*. (iii)* $\mathcal{B}$ *is compact. (iv)* $\mu(\cdot)$ *is continuously differentiable on* $\mathcal{B}_0$*. (v) For any* $\|\boldsymbol{\xi}\| = 1$*,* $0 < \inf_{\boldsymbol{X}_i, \boldsymbol{\beta}\in\mathcal{B}_0} \boldsymbol{\xi}^\top \boldsymbol{V}_i(\boldsymbol{\beta})\boldsymbol{\xi} \leq \sup_{\boldsymbol{X}_i, \boldsymbol{\beta}\in\mathcal{B}_0} \boldsymbol{\xi}^\top \boldsymbol{V}_i(\boldsymbol{\beta})\boldsymbol{\xi} < \infty$*. (vi) The domain of* $\boldsymbol{X}^c$ *is compact. (vii)* $b_N \to 0$*,* $N^{1-2\nu-2/\delta} b_N^{2q} \to \infty$ *and* $N^{1-2\nu} b_N^{5q/2} \to \infty$ *as* $N \to \infty$*, where* $\nu \in (0, 1/2)$*,* $\delta \geq 8$ *and* $q = \max_i q_i$*. (viii)* $\hat{\boldsymbol{\lambda}}_i(\boldsymbol{\beta}) \in \left\{\boldsymbol{\lambda}_i \in \mathcal{R}^{n_i} : \|\boldsymbol{\lambda}_i\| \leq c|S_i|^{-1/m}\right\}$ *for some* $c > 0$*. (ix)* $E\left\{\sup_{\boldsymbol{\beta}\in\mathcal{B}_0} \|\boldsymbol{G}_i(\boldsymbol{\beta})\|^2\right\} < \infty$*.*

Remark: Assumption (v) guarantees that the variance-covariance matrix restricted on each stratum is invertible. The restrictions on b_N in Assumption (vii) follow that in Smith 2007. Assumption (viii) is similar to Assumption 3.6 in KTA, and is only needed when the focus is on some neighborhood around $\boldsymbol{\beta}_0$. For example, it is needed in establishing the asymptotic normality of $\hat{\boldsymbol{\beta}}_{CEL}$, but not needed in establishing the consistency. Assumption (viii) is legitimate if (4.1) is correct. Our numerical implementation does not require Assumption (viii). Assumption (ix) is used to guarantee the uniform weak law of large numbers in the proof of Theorem IV.3.

Proof of Theorem IV.1. Let

$$\begin{aligned} L_N(\boldsymbol{\beta}) &\stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{j\in S_i} -w_{ij}\log\left\{1+\hat{\boldsymbol{\lambda}}_i^{\top}(\boldsymbol{\beta})\boldsymbol{g}_j(\boldsymbol{\beta})\right\}\right] \\ &= \sum_{\omega\in\Omega}\frac{|S^{\omega}|}{N}\frac{1}{|S^{\omega}|}\left[\sum_{i,j\in S^{\omega}} -w_{ij}\log\left\{1+\hat{\boldsymbol{\lambda}}_i^{\top}(\boldsymbol{\beta})\boldsymbol{g}_j(\boldsymbol{\beta})\right\}\right] \stackrel{\text{def}}{=} \sum_{\omega\in\Omega}\frac{|S^{\omega}|}{N}L_N^{\omega}(\boldsymbol{\beta}), \end{aligned}$$

then $\hat{\boldsymbol{\beta}}_{CEL} = \arg\max_{\boldsymbol{\beta}\in\mathcal{B}} L_N(\boldsymbol{\beta})$. Under Assumptions (ii)-(vii), following the same proof of (A.6) and (A.7) as in KTA, we have $\sup_{\boldsymbol{\beta}\in\mathcal{B}}|S^{\omega}|^{1/m}L_N^{\omega}(\boldsymbol{\beta}) < o_p(1)$ for all $\omega\in\Omega$. Therefore, for any $\epsilon>0$, let $B(\boldsymbol{\beta}_0,\epsilon)$ be a open ball centered at $\boldsymbol{\beta}_0$with radius ϵ, we have

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}\setminus B(\boldsymbol{\beta}_0,\epsilon)}\left(\frac{|S^{\omega}|}{N}\right)^{1-1/m}|S^{\omega}|^{1/m}L_N^{\omega}(\boldsymbol{\beta}) < o_p(1), \quad \text{for all } \omega\in\Omega.$$

In addition, for stratum ω_0 as in Assumption (i), the same proof of (A.8) as in KTA gives

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}\setminus B(\boldsymbol{\beta}_0,\epsilon)}\left(\frac{|S^{\omega_0}|}{N}\right)^{1-1/m}|S^{\omega_0}|^{1/m}L_N^{\omega_0}(\boldsymbol{\beta}) < -r_{\omega_0}^{1-1/m}H(\epsilon)+o_p(1),$$

where $H(\epsilon)>0$ is a constant depending on ϵ. Now we have

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}\setminus B(\boldsymbol{\beta}_0,\epsilon)} N^{1/m}L_N(\boldsymbol{\beta}) = \sup_{\boldsymbol{\beta}\in\mathcal{B}\setminus B(\boldsymbol{\beta}_0,\epsilon)}\sum_{\omega\in\Omega}\left(\frac{|S^{\omega}|}{N}\right)^{1-1/m}|S^{\omega}|^{1/m}L_N^{\omega}(\boldsymbol{\beta}) < -r_{\omega_0}^{1-1/m}H(\epsilon)+o_p(1).$$

On the other hand, the same proof of (A.9) as in KTA gives $|S^{\omega}|^{1/m}L_N^{\omega}(\boldsymbol{\beta}_0)\geq o_p(1)$ for all $\omega\in\Omega$, therefore $N^{1/m}L_N(\boldsymbol{\beta}_0)\geq o_p(1)$. Thus, $\hat{\boldsymbol{\beta}}_{CEL}$ must lie inside $B(\boldsymbol{\beta}_0,\epsilon)$, and consistency of $\hat{\boldsymbol{\beta}}_{CEL}$ then follows by the arbitrariness of ϵ. □

Proof of Theorem IV.2. Define

$$\boldsymbol{A}^{\omega} = \frac{1}{\sqrt{|S^{\omega}|}}\sum_{i\in S^{\omega}}\left\{\sum_{j\in S^{\omega}} w_{ij}\partial\boldsymbol{g}_j(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}\right\}^{\top}\left\{\sum_{j\in S^{\omega}} w_{ij}\boldsymbol{g}_j(\boldsymbol{\beta}_0)\boldsymbol{g}_j(\boldsymbol{\beta}_0)^{\top}\right\}^{-1}\left\{\sum_{j\in S^{\omega}} w_{ij}\boldsymbol{g}_j(\boldsymbol{\beta}_0)\right\},$$

then following the proof of (A.14) and Lemma C.1 as in KTA, one can easily show that, for any $\omega\in\Omega$,

$$-\frac{1}{\sqrt{|S^{\omega}|}}\frac{\partial L^{\omega}(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}} = \boldsymbol{A}^{\omega}+o_p(1) \text{ and } \sup_{\boldsymbol{\beta}\in\mathcal{B}_0}\|-\frac{1}{|S^{\omega}|}\frac{\partial^2 L^{\omega}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\top}}-\boldsymbol{J}_{\omega}(\boldsymbol{\beta})\| = o_p(1),$$

where $L^\omega(\boldsymbol{\beta})$ is defined in Section 4.2. Since $\hat{\boldsymbol{\beta}}_{CEL}$ maximizes $L(\boldsymbol{\beta})$, $\partial L(\hat{\boldsymbol{\beta}}_{CEL})/\partial\boldsymbol{\beta} = \mathbf{0}$. From Taylor expansion around $\boldsymbol{\beta}_0$, for some $\tilde{\boldsymbol{\beta}}$ between $\hat{\boldsymbol{\beta}}_{CEL}$ and $\boldsymbol{\beta}_0$, we have

$$\begin{aligned}\mathbf{0} &= \frac{1}{\sqrt{N}}\frac{\partial L(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}} + \frac{1}{N}\frac{\partial^2 L(\tilde{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\sqrt{N}(\hat{\boldsymbol{\beta}}_{CEL} - \boldsymbol{\beta}_0) \\ &= -\sum_{\omega\in\Omega}\frac{\sqrt{|S^\omega|}}{\sqrt{N}}\boldsymbol{A}^\omega + o_p(1) - \sum_{\omega\in\Omega}\frac{|S^\omega|}{N}\boldsymbol{J}_\omega(\tilde{\boldsymbol{\beta}})\sqrt{N}(\hat{\boldsymbol{\beta}}_{CEL} - \boldsymbol{\beta}_0) + o_p(\sqrt{N}\|\hat{\boldsymbol{\beta}}_{CEL} - \boldsymbol{\beta}_0\|).\end{aligned}$$

Solving for $\sqrt{N}(\hat{\boldsymbol{\beta}}_{CEL} - \boldsymbol{\beta}_0)$ we have

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{CEL} - \boldsymbol{\beta}_0) = -\left\{\sum_{\omega\in\Omega}\frac{|S^\omega|}{N}\boldsymbol{J}_\omega(\tilde{\boldsymbol{\beta}}) + o_p(1)\right\}^{-1}\left\{\sum_{\omega\in\Omega}\frac{\sqrt{|S^\omega|}}{\sqrt{N}}\boldsymbol{A}^\omega + o_p(1)\right\}.$$

From the Central Limit Theorem we have $\boldsymbol{A}^\omega \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{J}_\omega(\boldsymbol{\beta}_0))$. Therefore, from the continuity of $\boldsymbol{J}_\omega(\boldsymbol{\beta})$, the consistency of $\hat{\boldsymbol{\beta}}_{CEL}$ and Slutsky's Theorem, we have the desired result $\sqrt{N}(\hat{\boldsymbol{\beta}}_{CEL} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{J}(\boldsymbol{\beta}_0)^{-1})$. □

Proof of Theorem IV.3. Under Assumptions (ii) and (viii), from Lemma D.2 in KTA, we have $\max_{1\le i\le N, j\in S_i}\sup_{\boldsymbol{\beta}\in\mathcal{B}}|\hat{\boldsymbol{\lambda}}_i^\top(\boldsymbol{\beta})\boldsymbol{g}_j(\boldsymbol{\beta})| = o_p(1)$. Therefore $p_{ij}(\boldsymbol{\beta}) = w_{ij}\{1 + o_p(1)\}$ and the $o_p(1)$ term is independent of i, j and $\boldsymbol{\beta}$. This result, together with the consistency of $\hat{\boldsymbol{\beta}}_{CEL}$, implies that for any $1 \le i \le N$,

$$\sum_{j\in S_i} p_{ij}(\hat{\boldsymbol{\beta}}_{CEL})\boldsymbol{g}_j(\hat{\boldsymbol{\beta}}_{CEL})\boldsymbol{g}_j(\hat{\boldsymbol{\beta}}_{CEL})^\top = \boldsymbol{V}_i(\hat{\boldsymbol{\beta}}_{CEL}) + o_p(1).$$

Then by Assumption (v) we have

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^N \boldsymbol{G}_i(\hat{\boldsymbol{\beta}}_{CEL})^\top\left\{\sum_{j\in S_i} p_{ij}(\hat{\boldsymbol{\beta}}_{CEL})\boldsymbol{g}_j(\hat{\boldsymbol{\beta}}_{CEL})\boldsymbol{g}_j(\hat{\boldsymbol{\beta}}_{CEL})^\top\right\}^{-1}\boldsymbol{G}_i(\hat{\boldsymbol{\beta}}_{CEL}) \\ =\ &\sum_{\omega\in\Omega}\frac{|S^\omega|}{N}\frac{1}{|S^\omega|}\sum_{i\in S^\omega}\boldsymbol{G}_i(\hat{\boldsymbol{\beta}}_{CEL})^\top\left\{\boldsymbol{V}_i(\hat{\boldsymbol{\beta}}_{CEL})^{-1} + o_p(1)\right\}\boldsymbol{G}_i(\hat{\boldsymbol{\beta}}_{CEL}).\end{aligned}$$

Therefore Theorem IV.3 follows from Assumption (ix) and the uniform weak law of large numbers (e.g. Lemma 2.4 in Newey and McFadden 1994). □

CHAPTER V

Some Possible Future Works

In this dissertation we have investigated the CEL approach to statistical analysis with missing data. We considered both cases of cross-sectional data and longitudinal data. One distinctive feature of the proposed CEL method is that, the resulting estimator can achieve high estimation efficiency without modeling any second moments of the data. Because of the difficulty of modeling the second moments in practical studies, the circumvention of it by our proposed CEL method may be highly advantageous.

There are still several issues that deserve further investigation. First, in our developments we employed a modified cross-validation criterion along the lines suggested by Newey (1993) for the bandwidth selection. Although this criterion works reasonably well in our simulation experiments, it has not been theoretically justified yet. Future work on better criteria and related theory is needed.

Second, when the number of covariates is large and the covariates vary in scales and/or types, the kernel-based weight calculation is challenged. Although through simulation experiments in Chapter II we have found that the inclusion of additional covariates may not have a dramatic impact on the performance of our proposed procedure, such an impact will become influential and can not be ignored as the

number of covariates keeps increasing. Therefore, it is worthwhile to explore more flexible ways to calculate the weights.

Third, the computational burden of empirical likelihood method is known to be the main obstacle preventing its application in many practical problems. Future efforts need to be made to alleviate the computational burden while retaining the theoretical advantages of empirical likelihood method.

In addition to the above issues, there are several extensions of the current developments worth future exploration. First, our developments only concern the missing outcome problem. The problem of missing covariates is equally important, and may be more challenging, as the weights used by the CEL method to construct the localized empirical likelihood are calculated based on the covariates. Furthermore, practical studies often generate data with both missing outcome and missing covariates. The extension of the current developments to the setting of missing covariates, and more generally, to the setting where both the outcome and the covariates are subject to missingness, is an important and interesting research topic.

Second, to deal with the missing outcome problem in both cases of cross-sectional data and longitudinal data, we made the assumption that the missingness mechanism is ignorable (MAR or MCAR). Although this assumption holds for many practical studies, such as the two-stage design studies in cross-sectional data case, it does not cover the scenario where the missingness depends on the missing values, namely the nonignorable missingness mechanism. It is of interest to investigate how to generalize the CEL method to analyze data under a more complex missingness mechanism.

Third, in Chapter II we proposed both the CEL-IPW estimator and the CEL-AIPW estimator. Although the CEL-AIPW estimator attains the semiparametric efficiency bound when both the missingness mechanism and the conditional mean of

the outcome given the surrogate variables and the covariates are correctly modeled, it may lose substantial efficiency compared to the CEL-IPW estimator when the conditional mean is incorrectly modeled. It is desirable to find an estimator that is doubly robust, locally efficient, and has uniformly improved efficiency over the CEL-IPW estimator. This may be done by following the idea in Han (2012).

Fourth, it is worthwhile to study the variable selection problems in the presence of missing data using empirical likelihood. Most existing variable selection methods are based on penalizing the least square or the likelihood function. However, for semiparametric problems, the likelihood approach faces the challenge of model misspecification and the least square approach does not yield efficient estimation. Empirical likelihood seems to be a good alternative due to its robustness against model misspecification and its efficiency in estimation. In addition, the rapid development of variable selection techniques has mainly focused on dealing with fully observed data. Missing data pose great challenges, both theoretically and numerically. The least square approach is not directly applicable and the calculation of the likelihood function may become infeasible due to multiple integration. The IPW approach provides a promising solution and deserves thorough investigation. The applicability of variable selection techniques could be considerably extended.

Fifth, the current empirical likelihood methodology treats the empirical likelihood as static, despite the reality that the actual data collected may be dynamic and vary over a continuum, such as time. It is interesting to study the empirical likelihood in a dynamic fashion, where the likelihood function also varies over the continuum. Such a generalization of empirical likelihood will have a broad range of applications, such as solving problems raised by functional data analysis, image data analysis and dynamic treatment regime.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] A. Bandura. *Social cognitive theory and exercise of control over HIV infection.* Preventing AIDS: Theories and Methods of Behavioral Interventions. New York, NY: Plenum Press, 1994.

[2] H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972, 2005.

[3] P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore, 1993.

[4] G.E.P. Box. Sampling and Bayes inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A*, 143:383–430, 1980.

[5] W. Cao, A. A. Tsiatis, and M. Davidian. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96:723–734, 2009.

[6] G. Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34:305–334, 1987.

[7] J. Chen and N. E. Breslow. Semiparametric efficient estimation for the auxiliary outcome problem with conditional mean model. *Canadian Journal of Statistics*, 32:359–372, 2004.

[8] J. Chen, R. R. Sitter, and C. Wu. Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89:230–237, 2002.

[9] S. X. Chen and H. Cui. An extended empirical likelihood for generalized linear models. *Statistica Sinica*, 13:69–81, 2003.

[10] S. X. Chen and H. Cui. On the second order properties of empirical likelihood with moment restrictions. *Journal of Econometrics*, 141:492–516, 2007.

[11] S. X. Chen, D. H. Y. Leung, and J. Qin. Information recovery in a study with surrogate endpoints. *Journal of the American Statistical Association*, 98:1052–1062, 2003.

[12] S. X. Chen, D. H. Y. Leung, and J. Qin. Improving semiparametric estimation by using surrogate data. *Journal of the Royal Statistical Society, Series B*, 70:803–823, 2008.

[13] S. X. Chen and I. Van Keilegom. A review on empirical likelihood methods for regression. *Test*, 18:415–447, 2009.

[14] N. Cressie and T. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B*, 46:440–464, 1984.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39:1–22, 1977.

[16] T. DiCiccio, P. Hall, and J. Romano. Empirical likelihood is bartlett-correctable. *Annals of Statistics*, 19:1053–1061, 1991.

[17] P. Diggle, P. Heagerty, K. Y. Liang, and S. L. Zeger. *Analysis of Longitudinal Data, Second Edition*. Oxford University Press, Oxford, 2002.

[18] V. P. Godambe. An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, 31:1208–1212, 1960.

[19] V. P. Godambe. *Estimating Functions*. Oxford University Press, Oxford, 1991.

[20] P. Hall. Pseudo-likelihood theory for empirical likelihood. *Annals of Statistics*, 18:121–140, 1990.

[21] P. Hall and B. LaScala. Methodology and algorithms for empirical likelihood. *International Statistical Review*, 58:109–127, 1990.

[22] P. Han. A note on improving the efficiency of inverse probability weighted estimator using the augmentation term. *Statistics and Probability Letters*, 82:2221–2228, 2012.

[23] P. Han. Model selection with missing outcome data using conditional empirical likelihood. submitted, 2013.

[24] B. E. Hansen. *Econometrics*. draft graduate textbook, 2013.

[25] L. P. Hansen. Large sample properties of generalized methods of moments estimators. *Econometrica*, 50:1029–1054, 1982.

[26] C. C. Heyde. Fixed sample and asymptotic optimality for classes of estimating functions. *Contemporary Mathematics*, 80:241–247, 1988.

[27] C. C. Heyde. *Quasi-likelihood and Its Application*. Springer-Verlag, New York, 1997.

[28] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685, 1952.

[29] J. Jiang, Y. Luan, and Y-G. Wang. Iterative estimating equations: Linear convergence and asymptotic properties. *Annals of Statistics*, 35:2233–2260, 2007.

[30] J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. New York: Wiley, 2002.

[31] J. D. Y. Kang and J. L. Schafer. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science*, 22:523–539, 2007.

[32] G. Kauermann and R. J. Carroll. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96:1387–1396, 2001.

[33] Y. Kitamura. *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress, Vol. 3.*, chapter Empirical Likelihood Methods in Econometrics: Theory and Practice, pages 174–237. Cambridge University Press, 2007.

[34] Y. Kitamura, G. Tripathi, and H. Ahn. Empirical likelihood-based inference in conditional moment restriction models. *Econometrica*, 72:1667–1714, 2004.

[35] E. D. Kolaczyk. Empirical likelihood for generalized linear model. *Statistica Sinica*, 4:199–218, 1994.

[36] N. A. Lazar. Bayesian empirical likelihood. *Biometrika*, 90:319–326, 2003.

[37] E. L. Lehmann. *Elements of Large-Sample Theory*. Springer, 1998.

[38] G. Li, R. Li, and M. Zhou. *Empirical likelihood in survival analysis*, pages 336–350. Contemporary Multivariate Analysis and Experimental Designs, World Scientific, 2005.

[39] Y. Li. Efficient semiparametric regression for longitudinal data with nonparametric covariance estimation. *Biometrika*, 98:355–370, 2011.

[40] K. Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *biometrika*, 73(1):13–22, 1986.

[41] R. J. A. Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88:125–134, 1993.

[42] R. J. A. Little. A class of pattern mixture models for normal missing data. *Biometrika*, 81:471–483, 1994.

[43] R. J. A. Little. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90:1112–1121, 1995.

[44] R. J. A. Little. Calibrated Bayes: A Bayes/frequentist roadmap. *The American Statistician*, 60:213–223, 2006.

[45] R. J. A. Little. *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods*, chapter Selection and Pattern-Mixture Models, pages 409–431. Taylor & Francis, 2008.

[46] R. J. A. Little. Calibrated Bayes, for statistics in general, and missing data in particular. *Statistical Science*, 26:162–174, 2011.

[47] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley, New York, 2 edition, 2002.

[48] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1990.

[49] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley-Interscience, 2008.

[50] X. Meng and D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80:267278, 1993.

[51] C. G. Neumann, N. O. Bwibo, S. P. Murphy, M. Sigman, S. Whaley, L. H. Allen, D. Guthrie, R. E. Weiss, and M. W. Demment. Animal source foods improve dietary quality, micronutrient status, growth and cognitive function in kenyan school children: Background, study design and baseline findings. *The Journal of Nutrition*, 133:3941S–3949S, 2003.

[52] W. K. Newey. *Handbook of Statistics, Vol 11.*, chapter Efficient Estimation of Models with Conditional Moment Restrictions, pages 419–454. North-Holland, Amsterdam, 1993.

[53] W. K. Newey and D. L. McFadden. *Large Sample Estimation and Hypothesis Testing*. Handbook of Econometrics, Vol 4. Amsterdam, The Netherlands: Elsevier Science, 1994.

[54] W. K. Newey and R. J. Smith. Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72:219–255, 2004.

[55] T. Otsu. Penalized empirical likelihood estimation of semiparametric models. *Journal of Multivariate Analysis*, 98:1923–1954, 2007.

[56] T. Otsu. Empirical likelihood estimation of conditional moment restriction models with unknown functions. *Econometric Theory*, 27:8–46, 2011.

[57] A. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75:237–249, 1988.

[58] A. Owen. Empirical likelihood ratio confidence regions. *Annals of Statistics*, 18:90–120, 1990.

[59] A. Owen. Empirical likelihood for linear models. *Annals of Statistics*, 19:1725–1747, 1991.

[60] A. Owen. *Empirical Likelihood*. Chapman & Hall/CRC Press, New York, 2001.

[61] M. C. Paik. The generalized estimating equations approach when data are not missing completely at random. *Journal of the American Statistical Association*, 92:1320–1329, 1997.

[62] J. Pan and G. MacKenzie. Model selection for joint mean-covariance structures in longitudinal studies. *Biometrika*, 90:239–244, 2003.

[63] M. S. Pepe. Inference using surrogate outcome data and a validation sample. *Biometrika*, 79:355–365, 1992.

[64] M. S. Pepe, M. Reilly, and T. R. Fleming. Auxiliary outcome data and the mean score method. *Journal of Statistical Planning and Inference*, 42:137–160, 1994.

[65] R. L. Prentice and L. P. Zhao. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, 47:825–839, 1991.

[66] J. Qin and J. Lawless. Empirical likelihood and general estimating equations. *Annals of Statistics*, 22:300–325, 1994.

[67] J. Qin, B. Zhang, and D. H. Y. Leung. Empirical likelihood in missing data problems. *Journal of the American Statistical Association*, 104:1492–1503, 2009.

[68] A. Qu, B. Lindsay, and B. Li. Improving generalised estimating equations using quadratic inference functions. *Biometrika*, 87:823–836, 2000.

[69] J. M. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regresion models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, Mar. 1995.

[70] J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, Sep. 1994.

[71] 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, Mar. 1995.

[72] J. M. Robins, M. Sued, Q. Gomez-Lei, and A. Rotnitzky. Comment: performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science*, 22:544–559, 2007.

[73] M. J. Rotheram-Borus, M. Lee, Y. Y. Lin, and P. Lester. Six-year intervention outcomes for adolescent children of parents with the human immunodeficiency virus. *Archives of Pediatrics & Adolescent Medicine*, 158:742–748, 2004.

[74] A. Rotnitzky. *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods*, chapter Inverse Probability Weighted Methods, pages 453–476. Taylor & Francis, 2008.

[75] A. Rotnitzky, Q. Lei, M. Sued, and J. M. Robins. Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99:439–456, 2012.

[76] A. Rotnitzky, J. M. Robins, and D. O. Scharfstein. Semiparametric regression for repeated outcomes with non-ignorable non-response. *Journal of the American Statistical Association*, 93:1321–1339, 1998.

[77] D. B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

[78] D. B. Rubin. Multiple imputations in sample surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pages 20–34, 1978.

[79] D. B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12:1151–1172, 1984.

[80] D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York, 1987.

[81] D. B. Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91:473–489, 1996.

[82] D. B. Rubin and M. J. van der Laan. Empirical efficiency maximization: improved locally efficient covariate adjustment in randomized experiments and survival analysis. *International Journal of Biostatistics*, 4:article 5, 2008.

[83] D. O. Scharfstein, A. Rotnitzky, and J. M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 12 1999.

[84] S. M. Schennach. Bayesian exponentially tilted empirical likelihood. *Biometrika*, 92:31–46, 2005.

[85] S. Seaman and A. Copas. Doubly robust generalized estimating equations for longitudinal data. *Statistics in Medicine*, 28:937–955, 2009.

[86] J Shao. *Mathematical Statistics*. Springer, 2003.

[87] R. J. Smith. Efficient information theoretic inference for conditional moment restrictions. *Journal of Econometrics*, 138:430–460, 2007.

[88] P. X.-K. Song. *Correlated Data Analysis: Modeling, Analytics, and Applications*. Springer, New York, 2001.

[89] Z. Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101:1619–1637, 2006.

[90] Z. Tan. Comment: Improved local efficiency and double robustness. *The International Journal of Biostatistics*, 4:Article 10, 2008.

[91] Z. Tan. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97:661–682, 2010.

[92] G. Tripathi and Y. Kitamura. Testing conditional moment restrictions. *Annals of Statistics*, 31:2059–2095, 2003.

[93] A. A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer, New York, 2006.

[94] A. A. Tsiatis, M. Davidian, and W. Cao. Improved doubly robust estimation when data are monotonely coarsened, with application to longitudinal studies with dropout. *Biometrics*, 67:536–545, 2011.

[95] M. J. van der Laan and J. M. Robins. *Unified methods for censored longitudinal data and causality*. Springer, 2003.

[96] A. W. van der Varrt. *Asymptotic Statistics*. Cambridge University Press, 1998.

[97] D. Wang and S. X. Chen. Empirical likelihood for estimating equations with missing values. *Annals of Statistics*, 37:490–517, 2009.

[98] L. Wang, A. Rotnitzky, and X. Lin. Nonparametric regression with missing outcomes using weighted kernel estimating equations. *Journal of the American Statistical Association*, 105:1135–1146, 2010.

[99] S. Wang, L. Qian, and R. J. Carroll. Generalized empirical likelihood methods for analyzing longitudinal data. *Biometrika*, 97:79–93, 2010.

[100] Y. G. Wang and V. Carey. Working correlation structure misspecification, estimation and covariate design: Implications for generalized estimating equations performance. *Biometrika*, 90:29–41, 2003.

[101] Y. G. Wang and X. Lin. Effects of variance-function misspecification in analysis of longitudinal data. *Biometrics*, 61:413–421, 2005.

[102] R. W. M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, 61:439–447, 1974.

[103] L. J. Wei and J. M. Lachin. Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *Journal of the American Statistical Association*, 79:653–661, 1984.

[104] R. E. Weiss. *Modeling Longitudinal Data*. Springer, 2005.

[105] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25, 1982.

[106] C. F. J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11:95103, 1983.

[107] L. Xue and L. Zhu. Empirical likelihood semiparametric regression analysis for longitudinal data. *Biometrika*, 94:921–937, 2007.

[108] Y. Yang and X. He. Bayesian empirical likelihood for quantile regression. *Annals of Statistics*, to appear, 2012.

[109] H. Ye and J. Pan. Modelling of covariance structures in generalised estimating equations for longitudinal data. *biometrika*, 93:927 – 941, 2006.

[110] J. You, G. Chen, and Y. Zhou. Block empirical likelihood for longitudinal partially linear regression models. *Canadian Journal of Statistics*, 34:79–96, 2006.

[111] M. Yu and B. Nan. A revisit of semiparametric regression models with missing data. *Statistca Sinica*, 16:1193–1212, 2006.

[112] J. Zhang and I. Gijbels. Sieve empirical likelihood and extensions of the generalized least squares. *Scandinavian Journal of Statistics*, 30:1–24, 2003.