AN ABSTRACT OF THE DISSERTATION OF

Jianfei Zheng for the degree of Doctor of Philosophy in Statistics presented on August 28, 2017.

Title: Inference about Missing Mechanisms in Longitudinal Studies with a Refreshment Sample

Abstract approved: _____

Lan Xue

Missing data is one of the major methodological problems in longitudinal studies. It not only reduces the sample size, but also can result in biased estimation and inference. It is crucial to correctly understand the missing mechanism and appropriately incorporate it into the estimation and inference procedures. Traditional methods, such as the complete case analysis and imputation methods, are designed to deal with missing data under unverifiable assumptions of MCAR and MAR. The purpose of this dissertation is to provide an overview of procedures dealing with missing data. We especially focus on identifying and estimating attrition (missing) parameters under the non-ignorable missingness assumption using the refreshment sample in two-wave panel data. We propose a full-likelihood parametric approach which sets benchmarks for the performance of estimators in this setting. We also propose a semi-parametric method to estimate the attrition parameters by marginal density estimates with the help of two constraints from

Hirano et al. (2001) and the additional information provided by the refreshment sample. We derive asymptotic properties of the semi-parametric estimators and illustrate their performance with simulations. Inference based on bootstrapping is proposed and verified through simulations. A real data application is attempted in the Netherlands Mobility Panel. Finally, we extend the semi-parametric method to incorporate a time-invariant binary covariate and evaluate its large-sample performance with simulations.

Inference about Missing Mechanisms in Longitudinal Studies with a Refreshment Sample

by

Jianfei Zheng

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented August 28, 2017
Commencement June 2018

Doctor of Philosophy dissertation of Jianfei Zheng presented on August 28, 2017.

APPROVED:

Major Professor, representing Statistics

Chair of the Department of Statistics

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Jianfei Zheng, Author

ACKNOWLEDGEMENTS

To my beloved parents, Chenglin Li and Yi Zheng. You are the best parents I could have asked for. I could not have such a beautiful and happy life without your endless love and wisdom.

To my lovely wife, He Gao. You truly are an angel. I am lucky to have you by my side. You chose to leave everything behind and come with me to a foreign land, take care of me and start a new family with me. You are a new chapter and my future.

I sincerely thank my PhD advisor Dr. Lan Xue for your guidance. You help me push myself. You have faith in me even when I do not believe in myself. I am always inspired from our meetings, and your persistence truly encourages me. I have learned tremendously from the way you think and tackle new problems. I really appreciate everything we have been through together during my three years PhD life.

I would like to specially thank Dr. Lisa Madsen. You believed in me and granted me a teaching assistantship at the very beginning when I had just transferred to the Statistics department as an outsider, otherwise I would not have been able to continue my studies at that time. You are the key to the start of my wonderful journey in Statistics.

I would like to extend my gratitude and appreciation to my committee members who made this Ph.D. dissertation possible. Dr. Dan Edge, Dr. Yuan Jiang, Dr. Jay Kim, Dr. Virginia Lesser, Dr. Lisa Madsen and Dr. Lan Xue, thank you all for serving as my committee members. I am grateful to have you on board and advise me through my dissertation.

I have had a great learning experience in the Statistics department at Oregon State

University. Dr. Yanming Di, Dr. Sarah Emerson, Dr. Alix Giltelman, Dr. Yuan Jiang, Dr. Lisa Madsen, Dr. Charlotte Wickham and Dr. Lan Xue, thank you all for your insightful and inspiring teaching. Thanks for your patience and wisdom.

To all my best friends: Thank you all for the laughter and companionship during my time at Oregon State University. I will miss you all.

It has been seven years since I started my studies in Corvallis. I consider this place as my second home. I have a lot of great memories here, and I really appreciate what this small college town has given me. This is the place where I started finding my path and exploring my passion. I have challenged myself, and I made it! This is one of the best memories and gifts for my 30th birthday.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF TABLES

LIST OF TABLES (Continued)

Inference about Missing Mechanisms in Longitudinal Studies with a Refreshment Sample

1 Introduction

Compared to traditional cross-sectional data, panel or longitudinal data provide richer information (Hirano et al., 2001). In a longitudinal study, each unit or object is observed repeatedly over a period of time so that each unit yields a series of observations. If designed properly, a longitudinal study often is more efficient and requires a smaller sample to achieve the same power as a cross-sectional study with the same number of subjects. This is because repeated measurements from the same subject are often correlated, and subjects serve as natural blocks that reduce the variability of random errors (Liang and Zeger, 1986; Hirano et al., 2001; Fitzmaurice et al., 2008; Si et al., 2014). For this reason, longitudinal studies have been widely used in many scientific areas such as clinical trials, psychology and economics.

However, longitudinal studies often suffer from attrition where some of the subjects are unable to provide a response in the follow-up portion of the study. This results in incomplete panel data to which traditional statistical methods cannot be applied directly. For example, the Netherlands Institute for Transport Policy Analysis has been conducting the Netherlands Mobility Panel since 2013 (Hoogendoorn-Lanser et al., 2015). This panel currently involves two waves of data collection completes two-wave of data. The initial wave consists of 2380 households as experimental units. With almost 30% dropping out, there are only 1685 households remaining in the study at the time of the sec-

ond wave. Bias can be introduced in statistical inferences if attrition is ignored and the missingness is in fact systematically related to responses. It is important to understand the missing mechanism before making any statistical inferences about the population. Unfortunately, partially observed panel data alone are insufficient for distinguishing between missing mechanisms (Rubin, 1976; Hirano et al., 2001; Fitzmaurice et al., 2008; Deng et al., 2013; Si et al., 2014).

Researchers have proposed different models to explain the attrition process (Rubin, 2004). Statistical methods are then proposed to handle missingness and make valid inferences about the population, under these missingness assumptions.

Three different models have been proposed to explain the attrition process (Rubin, 2004). The first model is *Missing Completely at Random* (MCAR). This model assumes that the missingness is independent of both *variables that are always observed* and *variables that are potentially missing*. MCAR implies that the complete set, which consists of subjects who provide complete information, can be considered as a representative random sample from the population. Therefore, traditional methods for longitudinal data applied to the complete set can provide valid inferences.

The second model is *Missing At Random* (MAR), or the ignorable missingness model. This model assumes that the missing mechanism depends solely on *variables that are always observed*. For example, in a drug study, individuals might tend to drop out of the study and refuse to provide future response due to pre-treatment examination results that make them feel worried or embarrassed; the missingness of responses depends only on examination results which are always observed for every individual in the study. This missing mechanism makes the complete set no longer a representative sam-

ple of the population. Any inferences drawn from the complete set are subject to bias. Extensive efforts have been made to develop methods that can properly analyze MAR data. Imputation is the primary technique. The key implication of the MAR assumption is that the relationship between observed variables and missing variables is the same among subjects who provide complete information (the complete set) and those who do not (the incomplete set). Usually, a model for this relationship is built based on the complete set and used to impute missing values in the incomplete set. Single imputation is an approach that replaces each missing value with a reasonable guess. Statistical methods can be applied to the imputed panel as if it were fully observed (Rubin, 2004). A clear drawback of single imputation is that it is unable to take into account the uncertainty associated with the imputed values. An improvement is offered by multiple imputation which replaces each missing value with multiple guesses drawn from an imputation model (Rubin, 2004; Fitzmaurice et al., 2008). As a result, multiple panels are produced to give multiple estimates for parameters, which will be properly combined at the end to give the final estimates. The uncertainty of imputation is incorporated in the computation of standard errors, which accounts for both the variation in estimation and the variation in multiple imputations. Imputation techniques rely on the MAR assumption. Violations of the assumption can lead to biased estimation and inference (Deng et al., 2013).

The third model is *Missing Not At Random* (MNAR), or the non-ignorable missing model. It further relaxes the assumption for the missing mechanism and allows the missingness to depend on both the *observed variables* and *variables that are potentially missing*. This model has identification issues as the information provided in the panel

data is not enough to make inferences about the population (Rubin, 1976, 2004; Hirano et al., 2001; Fitzmaurice et al., 2008). Hausman and Wise (1979) developed a special case of the MNAR model that assumes the missingness depends solely on the *variables that are potentially missing*. This restriction on the missing mechanism makes the model identifiable, and it is usually referred to as the HW model in literature. Hausman and Wise (1979) showed that parameters can be estimated through a maximum likelihood approach for this special attrition model. Hirano et al. (2001) pointed out the pitfalls of assuming the MAR model or the HW model in panel studies. Even though they are both theoretically plausible and identifiable given the panel data, they can lead to very different inferences, and one is unable to distinguish between them.

Hirano et al. (2001) first proposed the use of additional information from refreshment samples in longitudinal studies to make the MNAR model identifiable. A refreshment sample is a new random sample taken from the target population during follow-up waves when attrition starts to occur. Many large panel studies now routinely include refreshment samples (Deng et al., 2013). For instance, many of the longer longitudinal studies of the National Center for Education Statistics, including the Early Childhood Longitudinal Study and the National Educational Longitudinal Study, refill their samples once or multiple times during the study. The National Educational Longitudinal Study, for example, followed 21,500 eighth graders every two years from 1988 until 2000 and included refreshment samples in years 1990 and 1992. The Netherlands Mobility Panel completed its initial survey data in 2013. A follow-up survey was administrated in 2014 and a refreshment sample was considered and incorporated.

Methods have been developed to analyze MNAR data with the use of a refresh-

ment sample. Hirano et al. (2001) proposed an additive non-ignorable model that takes MCAR, MAR and HW models as special cases to gain insights and make inferences for the attrition process. They provided and proved the fundamental identification theory and developed an estimation procedure for a two-wave binary response with no covariate. Nevo (2003) exploited the refreshment sample to compute weights that apply to the panel data. Parameters of interest were estimated through the method of moments, comparing between moments of the weighted panel and those of the refreshment sample. Bhattacharya (2008) converted Hirano's fundamental identification theory into conditional moment restrictions to make inferences about the attrition process. Compared to the method of Hirano et al. (2001), Bhattacharya's offers a simpler proof under weaker conditions. Kim et al. (2009) extended the model to account for sample attrition in the presence of population attrition. They proposed using a generated counter factual sample and the refreshment sample to identify both attrition processes. Deng (2012) and Deng et al. (2013) extended the additive non-ignorable model by including two sets of refreshment samples to handle three-wave binary response data. They took a fully Bayesian approach and used Markov chain Monte Carlo for estimation. Si et al. (2014) presented a semi-parametric additive non-ignorable model to analyze multivariate categorical responses in a two-wave panel with one refreshment sample. This approach adopted the additive non-ignorable model for the attrition process and modeled the multinomial survey responses with a Dirichlet process mixture.

In this dissertation, we present two new approaches that handle MNAR data in a two-wave panel with one refreshment sample. The first method is a fully parametric method based on likelihood. Inferences for the population are made through maximum

likelihood estimators. Adaptive Gaussian quadrature is used to overcome the integration difficulty introduced by the missing data in the construction of the likelihood. The second method is a semi-parametric approach where the kernel density estimator serves as the non-parametric component of the method and the additive non-ignorable attrition model (Hirano et al., 2001) is adopted as the parametric component. When the likelihood is correctly specified, the full-likelihood approach gives the most efficient estimators and acts as the benchmark for comparing different methods that analyze MNAR data in a two-wave panel. However, when the likelihood is misspecified, the full-likelihood method can result in bias and invalid inferences. The semi-parametric method, on the other hand, drops the requirement of a distributional specification and provides consistent inferences for the attrition process under different population conditions. In simulations, the kernel density based semi-parametric estimators perform better in terms of the mean square errors than the method of Bhattacharya (2008).

We first provide a more thorough literature review in chapter 2. In particular, we use a two-wave study to illustrate the three main missing mechanisms and common methods for analyzing the corresponding data. In chapter 3, we introduce the refreshment sample as a source of important supportive information for analyzing missing data. We then formally set up the scenario which is the primary focus of this dissertation: two-wave panel data with one refreshment sample. In chapter 4, we present the first method, a full-likelihood parametric method, and provide details on constructing the likelihood with adaptive Gaussian quadrature. The kernel density based semi-parametric model is introduced as the second method. Its asymptotic properties are investigated. Extensive simulation results are given in chapter 5 to understand the finite-sample performance of

the proposed methods and to verify and confirm our theoretical findings. A real data application using the Netherlands Mobility Panel is then attempted. In chapter 6, the semi-parametric method is extended to incorporate a time-invariant binary covariate. Corresponding asymptotic properties are assessed by simulations. Chapter 7 summarizes the present research and discusses future research.

2 Literature Review

In this chapter, we focus on simple two-wave panel data where only responses are measured with no covariate. We use these data to illustrate the implications of various missing mechanisms and the corresponding consequences on statistical inferences. We also describe common methods for dealing with the various types of missing data.

2.1 Two-Wave Panel Data

Let $\underline{Y}_i = (Y_{i1}, Y_{i2}),\ i = 1, 2, ..., N$, be bivariate responses obtained in a two-wave longitudinal study, with i indexing subjects. It is assumed that the responses in the first wave Y_{i1} are always observed, while responses in the second wave Y_{i2} are potentially missing. Let W_i be the indicator of missingness for Y_{i2} with $W_i = 1$ if Y_{i2} is observed and $W_i = 0$ otherwise. For a given sample from the population, we assume there are n_c subjects that have both observations in $\underline{Y}$, which are referred as the "completers" in literature. The data set associated with completers is called the complete set. The remaining $N - n_c$ subjects ("incompleters") have Y_2 missing, and the corresponding data set is referred to as the incomplete set. Table 2.1 shows the structure of the data.

We assume observations from different subjects are independent and identically distributed. The joint distribution of $\underline{Y}_i$ and the attrition (missingness) model are denoted

	Obs	Y_1	Y_2	W
Complete set	1	Y_{11}	Y_{12}	W_1=1
	⋮	⋮	⋮	⋮
	n_c	Y_{n_c1}	Y_{n_c2}	W_{n_c}=1
Incomplete set	n_c+1	$Y_{(n_c+1)1}$		W_{N_c+1}=0
	⋮	⋮		⋮
	N	Y_{N1}		W_N=0

Table 2.1: Case study data set

as follows:

$$
\begin{aligned}
&(\underline{Y}_i \mid \underline{\theta}) \sim f(\underline{y} \mid \underline{\theta}),\\
&(W_i = 1 \mid y_{i1}, y_{i2}, \underline{\beta}) \sim \text{Bernoulli}(\pi(y_{i1}, y_{i2}, \underline{\beta})),\\
&\pi(y_{i1}, y_{i2}, \underline{\beta}) = P(W_i = 1 \mid y_{i1}, y_{i2}, \underline{\beta}),
\end{aligned} \tag{2.1}
$$

where $\underline{\theta}$ are parameters of the population distribution and $\underline{\beta}$ are the attrition parameters for the attrition model denoted by $P(W_i = 1 \mid y_{i1}, y_{i2}, \underline{\beta})$. Then the observed likelihood of a given data set is as follows:

$$
\begin{aligned}
L_{obs}(\underline{\theta}, \underline{\beta} \mid \underline{y}, \underline{w}) &= \prod_{i=1}^{N} f(w_i, y_{i1}, y_{i2} \mid \underline{\theta}, \underline{\beta})\\
&= L(\text{completers}) L(\text{incompleters})\\
&= \prod_{i=1}^{n_c} f(W_i = 1, y_{i1}, y_{i2} \mid \underline{\theta}, \underline{\beta}) \prod_{i=n_c+1}^{N} f(W_i = 0, y_{i1} \mid \underline{\theta}, \underline{\beta})\\
&= \prod_{i=1}^{n_c} f(y_{i1}, y_{i2} \mid \underline{\theta}) P(W_i = 1 \mid y_{i1}, y_{i2}, \underline{\beta})
\end{aligned}
$$

$$\times \prod_{i=n_c+1}^{N} \int f(y_{i1}, y_2 \mid \underline{\theta}) P(W_i = 0 \mid y_{i1}, y_2, \underline{\beta}) dy_2, \tag{2.2}$$

where L_{obs} denotes the observed likelihood of the data. Since we do not observe Y_2 for the incomplete set, the likelihood for the observed part of the incomplete set is obtained by integrating the full likelihood $f(y_1, y_2 \mid \underline{\theta}) P(W = 0 \mid y_1, y_2, \underline{\beta})$ with respect to y_2. Three different classes of models have been considered for the missing mechanism which determines the form of $P(W = 1 \mid y_1, y_2, \underline{\beta})$.

2.2 MCAR

The first type of missing mechanism is *Missing Completely At Random*. It assumes that the missingness is completely independent of both variables that are always observed and variables that are potentially missing. In our simple two-wave scenario, the MCAR assumption implies that missingness is independent of both Y_1 and Y_2 so that

$$P(W_i = 1 \mid y_{i1}, y_{i2}, \underline{\beta}) = P(W_i = 1 \mid \underline{\beta}).$$

The likelihood (2.2) can then be factorized into

$$L_{obs}(\underline{\theta}, \underline{\beta} \mid \underline{y}, w) = \prod_{i=1}^{n_c} f(\underline{y}_i \mid \underline{\theta}) \prod_{i=1}^{n_c} P(W_i = 1 \mid \underline{\beta}) \prod_{i=n_c+1}^{N} f(y_{i1} \mid \underline{\theta}) \prod_{i=n_c+1}^{N} P(W_i = 0 \mid \underline{\beta}).$$

The MCAR assumption removes the dependency of missingness on responses. The attrition model moves outside of the integral, which extremely simplifies the likelihood. Inferences about $\underline{\theta}$ can then be based upon the information provided by observed $\underline{Y}_i$'s alone, ignoring the part related to the attrition model, because the two sets of parameters spaces $\underline{\theta}$ and $\underline{\beta}$ are separable. That is, we can ignore the likelihood of missingness and use the likelihood of observed responses only, $\prod_{i=1}^{n_c} f(\underline{y}_i \mid \underline{\theta}) \prod_{i=n_c+1}^{N} f(y_{i1} \mid \underline{\theta})$, to estimate parameters of interest, provided the joint distribution is correctly specified. This missing mechanism model is **ignorable** as one can separate inferences between the $\underline{\theta}$ and $\underline{\beta}$ in the likelihood. The MCAR assumption also implies that the completers can be treated as a random representative sample from the population. So if one deletes the incomplete set and uses only the complete set to analyze the panel data, valid inferences can also be obtained but with the efficiency being compromised.

2.3 MAR and Imputation

The second type of missing mechanism is *Missing At Random*. In terms of the simple case study, we say Y_2 is MAR if the probability of observing Y_2 depends on the value of Y_1, but not on the value of itself:

$$P(W_i = 1 \mid y_{i1}, y_{i2}, \underline{\beta}) = P(W_i = 1 \mid y_{i1}, \underline{\beta}).$$

In other words, if complete and incomplete cases with exactly the same value of Y_1 have a systematic difference in the value of Y_2, then the data do not satisfy the MAR

assumption. Under this model, the observed likelihood (2.2) can be factorized into

$$L_{obs}(\underline{\theta},\underline{\beta} \mid \underline{y},w) = \prod_{i=1}^{n_c} f(\underline{y}_i \mid \underline{\theta}) \prod_{i=1}^{n_c} P(W_i = 1 \mid y_{i1},\underline{\beta}) \prod_{i=n_c+1}^{N} f(y_{i1} \mid \underline{\theta}) \prod_{i=n_c+1}^{N} P(W_i = 0 \mid y_{i1},\underline{\beta}).$$

This factorization is due to the fact that the missingness is independent of Y_2 given Y_1. The attrition model can also be moved outside of the integral. With the complete separation in likelihoods between $\underline{\theta}$ and $\underline{\beta}$, the MAR is also an ignorable model. If the attrition parameters $\underline{\beta}$ are not the primary interest, we can ignore the attrition model's contribution to the likelihood and obtain the maximum likelihood estimates for $\underline{\theta}$ by maximizing the likelihood of observed responses, $\prod_{i=1}^{n_c} f(\underline{y}_i \mid \underline{\theta}) \prod_{i=n_c+1}^{N} f(y_{i1} \mid \underline{\theta})$. In contrast to the MCAR model, the complete set can no longer be considered as a representative sample from the population when data are MAR. As a result, inferences based on only the complete data are invalid under MAR.

The fact that missingness is independent of Y_2 given Y_1 implies that the conditional distribution of Y_2 given Y_1 is the same in both completers and incompleters. That is,

$$f(y_{i2} \mid y_{i1}) = f(y_{i2} \mid y_{i1}, W_i = 1) = f(y_{i2} \mid y_{i1}, W_i = 0). \tag{2.3}$$

This provides us the idea of building an imputation model as the conditional distribution of Y_2 given Y_1 from completers and make predictions for the missing Y_2 values of the incompleters. Once the imputation for every missing value is completed, we can pretend that we have fully observed panel data. Rubin (1976); Rubin and Schafer (1990) proved the validity of imputation in the analysis of longitudinal MAR data. The relationship among conditional distributions in (2.3) not only provides the foundation for making

imputations but also allows us to ignore the specification of a specific attrition model when analyzing MAR data.

In single imputation, one replaces each missing value with an imputed value to create complete panel data. Traditional statistical procedures for complete data analysis can then be applied. For instance, one can impute each missing value in a variable by the mean of its own observed values or by the conditional mean given other variables. Single imputation is easy to use, but it implicitly treats missing values as known and fixed, which fails to account for the variability of these missing values. As a consequence, the resulting standard errors of estimators will be biased downward.

Improving upon single imputation, Rubin (1987) introduced multiple imputation to account for the uncertainty of the imputed value and obtain valid statistical inferences. The idea of multiple imputation is that one imputes the missing data from an imputation model multiple times. Since we assume MAR, the conditional distribution of Y_2 given Y_1 is equivalent to the conditional distribution of Y_2 given Y_1 in the complete set, that is, $f(y_{i2} \mid y_{i1}) = f(y_{i2} \mid y_{i1}, W_i = 1)$. Therefore, an imputation model $f(y_{i2} \mid y_{i1})$ can be estimated directly from the completers. The imputation process starts by filling up missing data with random draws from this conditional distribution. For each imputed data set, a traditional method can be applied to obtain a set of estimates and standard errors for parameters of interest. Repeat this process several times, and one is able to produce multiple estimates and corresponding standard errors. Final estimates and standard errors are given by aggregating these multiple results. It is important to understand that the multiple imputation not only estimates missing values, but also preserves the variation among them. Three main approaches have been developed to implement imputation.

There are the joint model approach, the conditional model approach and the likelihood based model approach, which will be reviewed in the following subsections.

2.3.1 Joint Model

Consider our simple case study and assume that the conditional distribution of Y_2 given Y_1 can be modeled under the normality assumption as

$$Y_{i2} \mid Y_{i1} \sim N(\alpha_0 + \alpha_1 Y_{i1}, \sigma^2).$$

This model is then fitted to completers who have observations on both Y_1 and Y_2 to obtain estimates $\underline{\widehat{\alpha}} = (\widehat{\alpha}_0, \widehat{\alpha}_1)$ and $\widehat{\sigma}^2$. The imputation process for each missing value in Y_2 proceeds by obtaining a new set of parameters $(\alpha_0^*, \alpha_1^*, \sigma^*)$ from the posterior predictive distribution (Yuan, 2010):

$$\frac{(r-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{r-2},$$
$$\underline{\alpha} \sim N(\underline{\widehat{\alpha}}, \widehat{\sigma}^2 V^{-1}).$$

where $V = [\underline{1} \quad \underline{Y}_1]^T [\underline{1} \quad \underline{Y}_1]$. First, the σ^* is drawn as

$$\sigma^{*2} = \frac{(r-2)\widehat{\sigma}^2}{g},$$

where g is a random draw from $\chi^2_{n_c-2}$, and n_c is the number of completers. Then $\underline{\alpha}^* = (\alpha_0^*, \alpha_1^*)$ can be drawn as

$$\underline{\alpha}^* = \underline{\widehat{\alpha}} + \sigma^* V^{-\frac{1}{2}} \underline{Z},$$

where $V^{-\frac{1}{2}}$ represents the squared root of V^{-1}, which can be obtained by the Cholesky decomposition, and $\underline{Z}$ is a vector of two independent standard normal variables. The missing value is imputed as

$$\alpha_0^* + \alpha_1^* y_1 + \sigma^* z,$$

where z is a realization from the standard normal distribution. One imputation set becomes complete after this whole process is repeated for every missing Y_2 value. The standard analysis can be applied to this imputation set to obtain estimates as if the data were fully observed. Multiple imputation sets are produced in the same manner and estimates are combined by rules introduced by Rubin (2004).

If one can assume the monotone missing pattern where subjects who , this method is naturally extended to data with multiple waves of responses by creating a sequence of imputation models. Let Y_{i2}^* be the complete data at the second wave after imputation. The distribution of the third wave Y_3 is then modeled in the same manner as

$$Y_{i3} \mid Y_{i1}, Y_{i2}^* \sim N(\alpha_{30} + \alpha_{31} Y_{i1} + \alpha_{32} Y_{i2}^*, \sigma_3^2),$$

where α_{30}, α_{31} and α_{32} are coefficients for building the imputation model for Y_3. The process continues until all missing values have been imputed, producing one complete data set. Multiple imputation requires repeating this process to produce several complete

data sets. The advantage of this approach is its simplicity. However, this model is subject to the assumptions of the monotone missing pattern and the normality of the responses, and it can fail when any of these additional assumptions is not met.

2.3.2 Conditional Model

When the monotone missing pattern is not present, one can specify a sequence of conditional imputation distributions for each variable. For example, if we have p variables for each subject in the panel data, a set of full conditional models can be specified as follows:

$$f_k(y_k \mid y_1, y_2, ..., y_{k-1}, y_{k+1}, ..., y_p), \quad k = 1, 2, ..., p.$$

This is an attempt to define the joint distribution by specifying a conditional distribution for every variable in the data. These conditional distributions are used to impute missing values. Starting from simple guessed values, imputation is done by iterating over all conditionally specified imputation models until convergence. One iteration consists of one cycle through all p models (Rubin and Schafer, 1990; Van Buuren, 2007). The advantage is that this method is able to model different types of variables naturally – using, for example, a multiple regression model for a continuous variable, logistic regression model for binary variable and log-linear regression for nominal categorical variable. And it is convenient to incorporate interaction terms or non-linear terms if necessary for each individual model. However, there is no guarantee that the distribution of draws will converge to a valid posterior distribution (Fitzmaurice et al., 2008).

2.3.3 Likelihood Based Model

Alternative to direct imputation as shown in the joint model and conditional model, likelihood-based method is also use to estimate parameters of interest when data is subjected to missing. Recall in (2.2), if the following two conditions are satisfied,

- the data is MAR,
- the parameters for missingness $\underline{\beta}$ and the parameters of interest $\underline{\theta}$ are separable,

then the missing model becomes ignorable. The EM algorithm is an example of making imputation through the likelihood, provided the likelihood is correctly specified. In a sense, the EM algorithm imputes the missing values by the conditional mean given both the observed variables and the parameters from the previous iteration (the expectation or the E-step). The likelihood of the filled-in data is then maximized to produce a new set of estimates of parameters (the maximization or the M-step) (Fitzmaurice et al., 2008; Little and Rubin, 2014).

In our simple case study, let us assume that Y_1 and Y_2 have a bivariate normal distribution with means μ_1 and μ_2, standard deviations σ_1 and σ_2, and correlation coefficient ρ; and that Y_2 is potentially missing. At the tth step, the expectation step imputes each missing Y_2 value with the conditional mean of Y_2, given Y_1 and the current parameter values $\underline{\theta}^t = (\mu_1^t, \mu_2^t, \sigma_1^t, \sigma_2^t, \rho^t)$:

$$Y_2^{t+1} = E(Y_2 \mid Y_1, \underline{\theta}^t) = \mu_2^t + sign(\rho^t)\frac{\sigma_2^t}{\sigma_1^t}(Y_1 - \mu_1^t),$$

Let D^{t+1} denote the complete data for iteration $t+1$ after the missing Y_2 values have been replaced by the conditional means. The maximization step then update the parameters as

$$\underline{\theta}^{t+1} = \arg\max_{\underline{\theta}\in\underline{\Theta}} L(\underline{\theta} \mid D^{t+1}),$$

where the $\underline{\theta}^{t+1}$ is the updated parameters of interest at $(t+1)$th step. We repeat these two steps until changes in the updated parameters are smaller than some pre-determined thresholds. There is no need to specify a model for the missingness, but we do need a correctly specified full joint distribution of (Y_1, Y_2), and any misspecification may lead to biased estimates.

2.4 MNAR

Missing Not At Random is the third type of missing mechanism where the probability of being missing depends not only on variables that are always observed but also on the variables that are potentially missing. In our simple two-wave scenario, MNAR implies that there is a systematic difference in Y_2 between completer and incompleter even when they have the same Y_1 value. We no longer have the conditional distribution relationship that we do in the MAR model, and

$$f(y_{i2} \mid y_{i1}, W_i = 1) \neq f(y_{i2} \mid y_{i1}, W_i = 0).$$

That is, the conditional distribution of Y_2 given Y_1 for the completers is different from that for incompleters. Therefore, it is no longer valid for one to recover the missing

information by imputing them from an imputation model based on the complete data. From the likelihood perspective, (2.2) can no longer be simplified as in the case of MCAR or MAR, and the computation of the integral is inevitable:

$$L_{obs}(\underline{\theta},\underline{\beta} \mid \underline{y},w) = \prod_{i=1}^{n_c} f(\underline{y}_i \mid \underline{\theta}) P(W_i = 1 \mid \underline{y}_i,\underline{\beta}) \prod_{i=n_c+1}^{N} \int f(\underline{y}_i \mid \underline{\theta}) P(W_i = 0 \mid \underline{y}_i,\underline{\beta}) dy_2.$$

This fact complicates the computation of standard estimators, such as the MLE. More importantly, identification of parameters becomes an issue when the inferences are based on the panel data alone. Fitzmaurice et al. (2008) pointed out that the lack of identifiability leads to not only inferential problems such as estimators having high variability, but also computational problems such as the EM algorithm failing to converge or converging slowly. Furthermore, when data are MNAR, most standard analyses lead to invalid estimates for the parameters of interest.

Two modeling approaches for tackling MNAR data are the selection model and the pattern mixture model (Fitzmaurice et al., 2008). We briefly review these two modeling approaches, which differ in how the joint distribution of Y and W is factored, in the following subsections.

2.4.1 Selection Model

The selection model factors the joint distribution of $\underline{Y}$ and W into the joint distribution of responses $\underline{Y}$ and the conditional distribution of W given $\underline{Y}$:

$$f(\underline{y}_i, w_i \mid \underline{\theta},\underline{\beta}) = f(\underline{y}_i \mid \underline{\theta}) f(w_i \mid \underline{y}_i,\underline{\beta}).$$

This factorization is natural in a sense that we first specify a probability model for responses followed by a model for the missingness given the responses. The two sets of parameters are separated naturally with $\underline{\theta}$ being of primary interest and $\underline{\beta}$ pertaining to missingness. From this perspective, our simple case study in (2.1) belongs to the selection model where

$$
\begin{aligned}
&(\underline{Y}_i \mid \underline{\theta}) \sim f(\underline{y} \mid \underline{\theta}),\\
&(W_i \mid \underline{y}_i, \underline{\beta}) \sim \text{Bernoulli}(\pi(\underline{y}_i, \underline{\beta})),\\
&\pi(\underline{y}_i, \underline{\beta}) = P(W_i = 1 \mid \underline{y}_i, \underline{\beta}).
\end{aligned}
$$

The corresponding observed likelihood is

$$
L_{obs}(\underline{\theta}, \underline{\beta} \mid \underline{y}, w) = \prod_{i=1}^{n_c} f(\underline{y}_i \mid \underline{\theta}) P(W_i = 1 \mid \underline{y}_i, \underline{\beta}) \prod_{i=n_c+1}^{N} \int f(\underline{y}_i \mid \underline{\theta}) P(W_i = 0 \mid \underline{y}_i, \underline{\beta}) dy_{i2}.
$$

Maximum likelihood estimation under MNAR requires iterative techniques such as the EM algorithm. However, the model is not identifiable based on the panel data alone. Constraints have to be made on the attrition model to resolve the identification problem, which usually changes the original MNAR assumption into what is effectively an MAR assumption (Rubin, 1976; Fitzmaurice et al., 2008; Little and Rubin, 2014).

2.4.2 Pattern Mixture Model

The pattern-mixture model specifies a model for the missing mechanism first and then models the conditional joint distribution of responses given the missingness indi-

cator. It assumes

$$f(\underline{y}_i, w_i \mid \underline{\gamma}, p) = f(w_i \mid p) f(\underline{y}_i \mid w_i, \underline{\gamma}),$$

where p is the parameter of the marginal distribution of the missingness indicator and $\underline{\gamma}$ contains the parameters of primary interest. In our simple case study, the pattern-mixture model can be specified as follows

$$\begin{aligned}
(W_i = 1 \mid p) &\sim \text{Bernoulli}(p), \\
(\underline{Y}_i \mid W_i = 1, \underline{\gamma}) &\sim N_2(\underline{\mu}^{(1)}, \Sigma^{(1)}), \\
(\underline{Y}_i \mid W_i = 0, \underline{\gamma}) &\sim N_2(\underline{\mu}^{(0)}, \Sigma^{(0)}),
\end{aligned}$$

where $\underline{\mu}^{(0)}$, $\Sigma^{(0)}$ and $\underline{\mu}^{(1)}$, $\Sigma^{(1)}$ are means and variance-covariance matrices of the respective conditional distributions. Fitzmaurice et al. (2008) points out that the pattern-mixture model also cannot be identified from panel data alone. Some constraints have to be placed on the model to make valid inferences on parameters. Little and Rubin (2014) give a comprehensive review of the application of the normal pattern-mixture model to missing data. They also show that additional restrictions on parameters are necessary for model identification.

This chapter illustrates three main missing mechanisms and common methods used to address the problem. The validity of statistical methods depends on correct assumptions about the missing mechanism. However, panel data alone cannot distinguish between these mechanisms. One has to make an untestable assumption about the mechanism before analyzing missing data. Panel data provides sufficient information to identify parameters of interest only in MCAR and MAR scenarios as the missing model

can be ignored. However, the estimation problem becomes intractable when data are MNAR.

3 Refreshment Sample

In chapter 2, we discussed how the presence of missing data requires careful consideration of assumptions about the missing mechanism. These assumptions are untestable given the panel data alone. Moreover, the attrition model becomes unidentifiable if data are MNAR. Additional restrictions on parameters are needed, which in turn alters the MNAR problem to either a MCAR or a MAR problem. Hirano et al. (2001) proposed to exploit the refreshment sample, a random sample from the same population, to not only resolve the identification problem in the MNAR model but also make the missing mechanism testable. In this chapter, we introduce the use of a refreshment sample in addressing the MNAR problem in section 3.1. In section 3.2, we revisit the two-wave panel data with a logistic attrition model. A closely related method proposed by Bhattacharya (2008) is then introduced in section 3.3, and it will be compared with our methods in a later chapter.

3.1 Introduction

The refreshment sample is an additional independent random sample from the same population and has fully observed data for variables that are partially missing in the panel. Many large panel studies include refreshment samples (Deng et al., 2013). For instance, the longitudinal studies of the National Center for Education Statistics, includ-

ing the Early Childhood Longitudinal Study and the National Educational Longitudinal Study, added new samples of panelists at certain points in the studies. The National Educational Longitudinal Study surveyed 21,500 eighth graders every two years from 1988 until 2000 and included refreshment samples in 1990 and 1992. The 2007-2008 Associated Press - Yahoo! News Poll (APYN) consisted of an 11-wave survey with 3 refreshment samples aimed to measure attitudes about the 2008 U.S. Presidential election and politics (Hirano et al., 2001; Deng, 2012). Starting in 2013, Hoogendoorn-Lanser et al. (2015) conducted the Netherlands Mobility Panel (MPN) study, a multiple wave longitudinal study with a refreshment sample, to understand changes in travel behavior.

Hirano et al. (2001) introduced an idea of using a refreshment sample to identify the missing mechanism. They proposed an additive non-ignorable model for two-wave data with the attrition model specified as

$$P(W = 1 \mid y_1, y_2, x) = g(\kappa_0(x) + \kappa_1(y_1, x) + \kappa_2(y_2, x)), \tag{3.1}$$

where g is a monotone function bounded in $(0,1)$, and $\kappa_1(\cdot)$, $\kappa_2(\cdot)$, $\kappa_3(\cdot)$ are arbitrary functions. The covariates are denoted by x. This additive non-ignorable model includes the MCAR and the MAR models as special cases. In particular, it leads to the MCAR model if both κ_1 and κ_2 are identically 0, and to the MAR model if only κ_2 is identically 0. When κ_2 is nonzero, the data are MNAR. It also provides an approach to test for MCAR or MAR mechanisms through testing for non-zero κ's. This model still has an untestable assumption that the missingness depends on the two responses in an additive way without any interactions. Hirano et al. (2001) showed that this additive

non-ignorable model is the weakest assumption that is identifiable and estimable using a refreshment sample, in the sense that more complex models are no longer identifiable. For two-wave binary response data with no covariates, they provided two fundamental constraints which make the attrition parameters identifiable. They proposed to estimate the parameters using the method of moments. An implementation of the additive non-ignorable model was not given for continuous responses.

With the refreshment sample showing promise in the analysis of missing data (Hirano et al., 2001), researchers began to explore different ways to use its information. Nevo (2003) exploited the refreshment sample to compute weights for adjusting the panel data. The method of moments was used to estimate parameters by comparing moments between the weighted panel and the refreshment sample. Bhattacharya (2008) converted Hirano's fundamental identification theory into conditional moment restrictions, and a set of non-parametric regressions with B-splines were used to construct the objective function for estimation. This method is closely related to Hirano's and handles missing data in the two-wave continuous response scenario which is the focus of our methods. Details of Bhattacharya's method are discussed in a later section, and the comparison of its performance with our methods will be demonstrated in a later chapter. Kim et al. (2009) extended the model to account for sample attrition in the presence of population attrition. They used a generated counter factual sample and the refreshment sample to identify both attrition processes. Deng (2012); Deng et al. (2013) extended the additive non-ignorable model by including multiple refreshment samples to handle three-wave binary response data. They took a fully Bayesian approach using Markov chain Monte Carlo for estimation. Si et al. (2014) presented a semi-parametric additive

non-ignorable model to analyze multivariate categorical responses in a two-wave panel with one refreshment sample. Their approach adopted an additive non-ignorable model for the attrition process and modeled the categorical survey responses with a Dirichlet process mixture.

In the following section, we revisit the joint model for two-wave data and a logistic attrition model to represent the missing mechanism. In section 3.3, we then describe the conditional moment restriction model proposed by Bhattacharya (2008), which is an indirect implementation of Hirano's two constraints.

3.2 Two-Wave Panel with Logistic Attrition Model

Let $\underline{Y}_i = (Y_{i1}, Y_{i2})$, $i = 1, 2, ..., N$, be i.i.d. bivariate responses observed on N subjects. We assume that the responses in the first wave $\{Y_{i1}\}_{i=1}^{N}$ are always observed, while responses in the second wave $\{Y_{i2}\}_{i=1}^{N}$ are potentially missing. Let W_i be the indicator of missingness of Y_{i2} with $W_i = 1$ if Y_{i2} is observed and $W_i = 0$ otherwise.

We assume non-ignorable missingness and allow the conditional distribution of the attrition process W to depend on the values of both Y_1 and Y_2. In particular, we assume an additive non-ignorable attrition model with a logistic regression form:

$$P(W = 1 \mid y_1, y_2) = \frac{\exp(\beta_0 + \beta_1 y_1 + \beta_2 y_2)}{1 + \exp(\beta_0 + \beta_1 y_1 + \beta_2 y_2)}, \tag{3.2}$$

where β_0, β_1, β_2 are attrition parameters. The logistic regression model is a popular parametric form used in the literature to describe the missing mechanism (Rubin, 1976;

Hirano et al., 2001; Nevo, 2003; Bhattacharya, 2008; Kim et al., 2009; Little and Rubin, 2014). It can be extended to a more flexible regression model with either a nonparametric link function instead of the logistic form, or an additive function of y_1 and y_2 instead of the linear functional forms of y_1 and y_2.

The refreshment sample provides additional information in the form of an independent sample from the second wave population. Let $\left\{Y_{i2}^r\right\}_{i=1}^n$ be the refreshment sample with sample size n. Appending the refreshment sample to the original data, we have the data structure shown in Table 3.1.

	Obs	Y_1	Y_2	W
Complete set	1	Y_{11}	Y_{12}	W_1=1
	⋮	⋮	⋮	⋮
	n_c	Y_{n_c1}	Y_{n_c2}	W_c=1
Incomplete set	n_c+1	$Y_{(n_c+1)1}$		W_{n_c+1}=0
	⋮	⋮		⋮
	N	Y_{N1}		W_N=0
Refreshment sample	1		Y_{12}^r	
	⋮		⋮	
	n		Y_{n2}^r	

Table 3.1: Case study data set with refreshment sample

The goal is to estimate the attrition parameters $\underline{\beta} = (\beta_0, \beta_1, \beta_2)$ given the data observed in Table 3.1.

3.3 Conditional Moment Restriction Model

Bhattacharya (2008) adopted Hirano's constraint equations and transformed the con-

straints into two conditional moment equations. The unknown parameters are estimated by minimizing these conditional moment equations. Bhattacharya (2008) showed that under the additive model assumption (3.1), the following two conditional expectation equations make the attrition parameters identifiable:

$$m_1(Y_1;\underline{\beta}) = E\left\{\frac{W}{g(Y_1,Y_2;\underline{\beta})} - 1 \mid Y_1\right\} = 0,$$
$$m_2(Y_2;\underline{\beta}) = E\left\{\frac{W}{g(Y_1,Y_2;\underline{\beta})} - 1 \mid Y_2\right\} = 0.$$

Estimation proceeds with non-parametric regressions of the new variable

$$U - 1 = \left\{\frac{W}{g(Y_1,Y_2;\underline{\beta})} - 1\right\}$$

onto the B-spline spaces expanded by Y_1 and Y_2 respectively.

Recall the data in Table 3.1 consist of three parts: the complete set, the incomplete set and the refreshment sample. Let B_1, $B_2^{n_c}$ and B_2^* be three B-spline bases expanded at complete Y_1, complete Y_2 and the refreshment sample Y_2^r, respectively. Notice that both $B_2^{n_c}$ and B_2^* share the same bases since Y_2 and the refreshment sample come from the same second wave. Let $\widehat{m}_1(Y_1;\underline{\beta})$ and $\widehat{m}_2(Y_2^r;\underline{\beta})$ be the sample analogues to $m_1(Y_1;\underline{\beta})$ and $m_2(Y_2;\underline{\beta})$, formulated as follows:

$$\widehat{\underline{m}}_1(Y_1;\underline{\beta}) = B_1(B_1^T B_1)^{-1} B_1^T(\underline{U} - \underline{j}_N), \tag{3.3}$$
$$\widehat{\underline{m}}_2(Y_2^r;\underline{\beta}) = B_2^*\left(\frac{B_2^{*T}B_2^*}{n}\right)^{-1}\left\{\frac{1}{N}B_2^{n_c T}\underline{U}^{n_c} - \frac{1}{n}B_2^{*T}\underline{j}_n\right\}, \tag{3.4}$$

where $\underline{U} = \{U_i\}_{i=1}^{N} = \left\{ \frac{W_i}{g(Y_{i1}, Y_{i2}; \underline{\beta})} \right\}_{i=1}^{N}$ with N being the total number of observations in the panel. Furthermore, n_c is the number of observations in the complete panel, $\underline{U}^{n_c}$ denotes a sub-vector of $\underline{U}$ for complete observations, and n is the number of observations in the refreshment sample. Here $\underline{j}_N$ (or $\underline{j}_n$) is a vector of 1's of length N (or n).

Since $U_i = 0$ when $W_i = 0$, $\underline{U}$ can always be calculated in the panel even when Y_2 is missing. The regression of $\underline{U} - \underline{j}_N$ on Y_1, in Equation (3.3), is obtained by projecting $\underline{U} - \underline{j}_N$ onto the B-spline basis expanded at Y_1. The regression of $\underline{U} - \underline{j}_N$ on Y_2, in Equation (3.4), however, is complicated due to the potential for missing Y_2 values. If Y_2 is fully observed with no missing values, then one can construct an estimator for m_2 similar to (3.3) by replacing B_1 with B_2, which is the B-spline basis expanded at full panel data points:

$$\widetilde{\underline{m}}_2(Y_2; \underline{\beta}) = B_2 \left(\frac{B_2^T B_2}{N} \right)^{-1} \left\{ \frac{1}{N} B_2^T \underline{U} - \frac{1}{N} B_2^T \underline{j}_N \right\}. \tag{3.5}$$

When there are missing Y_2 values, Bhattacharya (2008) proposed to use the refreshment sample to span the B-spline basis instead and $\widehat{\underline{m}}_2(Y_2^r; \underline{\beta})$ in (3.4) uses data from different sources to estimate the components in (3.5). The parameters $\underline{\beta}$ are estimated by minimizing the following sum of squares:

$$\widehat{\underline{\beta}} = \underset{\underline{\beta}}{\text{minimize}} \quad \left\{ \frac{1}{N} \sum \widehat{m}_1^2 + \frac{1}{n} \sum \widehat{m}_2^2 \right\}.$$

This method transforms Hirano's constraints to conditional moment restrictions and constructs an objective function through spline regressions. The B-spline basis is used

to approximate the expectation of $U - 1$ conditional on either Y_1 or Y_2. This method has the advantage of easily incorporating covariates, but it can suffer from the curse of dimensionality when too many covariates are included.

4 The Proposed Methods

In chapter 3, we reviewed the use of a refreshment sample to resolve problems posed by MNAR data. In particular, Hirano et al. (2001) proposed the use of a refreshment sample to identify an additive non-ignorable attrition model when data are MNAR in a two-wave binary response study.

We aim to develop methods to handle two-wave MNAR data with continuous responses, instead of binary responses. In this chapter, we introduce two new methods which use a refreshment sample. First, we describe a likelihood-based, fully parametric model in section 4.1. Then in section 4.2 we introduce a kernel density based semi-parametric method to estimate attrition parameters directly from Hirano's constraints. The asymptotic theory of the semi-parametric estimators is developed in section 4.3. Finally, we describe hypothesis tests for the attrition parameters and estimation of the corresponding power functions in section 4.4.

4.1 Full-Likelihood Parametric Method

In this section, we propose to estimate the attrition parameters $\underline{\beta}$ by maximizing the full likelihood function. To construct the full likelihood function, we now assume our data are normally distributed. In particular, we assume the joint distribution of the first and second wave responses, Y_1 and Y_2, are bivariate normal. The population parameters $\underline{\theta}$ consist of μ_1, the marginal mean of Y_1; μ_2, the marginal mean of Y_2; σ_{11}^2, the variance

of Y_1; σ_{22}^2, the variance of Y_2; and $\rho = \frac{\sigma_{12}}{\sigma_{11}\sigma_{22}}$, the correlation coefficient between Y_1 and Y_2. Then

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim N_2 \left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} \right).$$

We have mentioned that the parameters $(\underline{\theta}, \underline{\beta})$ in the observed likelihood (2.2) are unidentifiable in general when the missingness is non-ignorable. However, the existence of a refreshment sample resolves the issue. The three subsets of the data contribute to the likelihood independently. In the complete set, both responses been observed. The likelihood for the complete data is

$$\begin{aligned} L_{comp}(\underline{\theta}, \underline{\beta}) &= \prod_{i=1}^{n_c} f(\underline{y}_i, W_i = 1 \mid \underline{\theta}, \underline{\beta}) \\ &= \prod_{i=1}^{n_c} f(y_{i1}, y_{i2}) P(W_i = 1 \mid y_{i1}, y_{i2}) \\ &= \prod_{i=1}^{n_c} \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho^2)}} exp\left\{ -\frac{z_i}{2(1-\rho^2)} \right\} \\ &\quad \times \frac{exp(\beta_0 + \beta_1 y_{i1} + \beta_2 y_{i2})}{1 + exp(\beta_0 + \beta_1 y_{i1} + \beta_2 y_{i2})}, \end{aligned}$$

where $z_i = \frac{(y_{i1}-\mu_1)^2}{\sigma_{11}} - \frac{2\rho(y_{i1}-\mu_1)(y_{i2}-\mu_2)}{\sqrt{\sigma_{11}\sigma_{22}}} + \frac{(y_{i2}-\mu_2)^2}{\sigma_{22}}$ and $\rho = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}$.

In the incomplete panel, we only observe the first wave responses. The likelihood contributed from these observations can be expressed as

$$\begin{aligned} L_{incmp}(\underline{\theta}, \underline{\beta}) &= \prod_{i=n_c+1}^{N} f(y_{i1}, W_i = 0 \mid \underline{\theta}, \underline{\beta}) \\ &= \prod_{i=n_c+1}^{N} \int f(y_{i1}, y_2) P(W_i = 0 \mid y_{i1}, y_2) dy_2 \end{aligned}$$

$$= \prod_{i=n_c+1}^{N} \int \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho^2)}} exp\left\{-\frac{z_i}{2(1-\rho^2)}\right\} \times \frac{1}{1+exp(\beta_0+\beta_1 y_{i1}+\beta_2 y_2)} dy_2.$$

In the refreshment sample, we only observe the second wave observations whose likelihood contribution is

$$L_{refresh}(\underline{\theta}) = \prod_{i=1}^{n} f(y_{i2}^r) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_{22}}} exp\left\{-\frac{(y_{i2}^r-\mu_2)^2}{2\sigma_{22}}\right\}.$$

Then the full likelihood is the product of above three pieces:

$$L(\underline{\theta},\underline{\beta}) = L_{comp}(\underline{\theta},\underline{\beta}) \times L_{incmp}(\underline{\theta},\underline{\beta}) \times L_{refresh}(\underline{\theta}).$$

The maximum likelihood estimates $(\widehat{\underline{\theta}}_{MLE}, \widehat{\underline{\beta}}_{MLE})$ can be obtained by maximizing this full likelihood with respect to all parameters.

The calculation of the likelihood for the incomplete set is the most challenging part since it involves the integral of a joint density over Y_2, and this integration needs to be evaluated for each incomplete data point. In addition, there is no closed form solution to this integral problem. We use the Gaussian-Hermite quadrature to numerically approximate the integration. The Gaussian-Hermite quadrature has been used in the generalized linear mixed models (Molenberghs and Verbeke (2005), section 14.5). For any function

$f(x)$, the Gaussian-Hermite quadrature provides an approximation to the integration as

$$\int f(x)\phi(x)dx \approx \sum_{q=1}^{Q} \frac{w_q}{\sqrt{\pi}} f\left(\sqrt{2}x_q\right),$$

where $\phi(\cdot)$ is the standard normal density function, x_q are solutions to the Qth order Hermite polynomial and w_q are corresponding weights.

The drawback of this method is that the approximation accuracy of the integral heavily depends on the location of the mode of the function $f(x)\phi(x)$. It works well only when the mode of $f(x)\phi(x)$ is located near 0. The integral is not well approximated if the mode of $f(x)\phi(x)$ deviates from 0. Due to this drawback, the adaptive Gaussian-Hermite quadrature (Skrondal and Rabe-Hesketh, 2004; Rabe-Hesketh et al., 2005; Skrondal and Rabe-Hesketh, 2009) has been adopted to accommodate such situation. The idea of adaptive Gaussian quadrature is to treat $f(x)\phi(x)$ as a normal density function by applying the Laplace approximation. Then move nodes to the center of approximated $f(x)\phi(x)$ and scale weights correspondingly. This method works better than the Gaussian-Hermite quadrature when the integrated function $f(x)\phi(x)$ is not centered around 0. The drawback is that it requires additional computation. In particular, one has to find the mode of $f(x)\phi(x)$, which can be difficult.

Let $\widehat{x}$ be the mode of $f(x)\phi(x)$ or $\ln[f(x)\phi(x)]$ and $\widehat{\sigma}_{GQ} = \left[-\frac{\partial^2}{\partial x^2}\ln[f(x)\phi(x)]\mid_{x=\widehat{x}}\right]^{-\frac{1}{2}}$. Then the adaptive Gaussian-Hermite quadrature gives

$$\int f(x)\phi(x)dx \approx \sum_{q=1}^{Q} w_q^+ f(x_q^+),$$

where

$$x_q^+ = \widehat{x} + \sqrt{2}\widehat{\sigma}_{GQ} x_q,$$
$$w_q^+ = \widehat{\sigma}_{GQ} \frac{\phi(x_q^+)}{\phi(\sqrt{2}x_q)} \frac{w_q}{\sqrt{\pi}}.$$

We apply the adaptive Gaussian-Hermite quadrature to the integration involved in calculating the joint distribution of $(Y_1, W = 0)$. In the incomplete set, only the first wave responses are observed. The likelihood function is as follow

$$L_{incmp}(\underline{\theta}) = f(y_1, W = 0 \mid \underline{\theta}) = \int f(y_1, y_2, w = 0 \mid \underline{\theta}) dy_2$$
$$= \int f(y_1, y_2, w = 0) \phi^{-1}(y_2) \phi(y_2) dy_2,$$

where

$$f(y_1, y_2, w = 0 \mid \underline{\theta}) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho^2)}} exp\left\{-\frac{z}{2(1-\rho^2)}\right\} \frac{1}{1 + exp(\beta_0 + \beta_1 y_1 + \beta_2 y_2)}.$$

Let $g(y_2) = f(y_1, y_2, w = 0)\phi^{-1}(y_2)$. The adaptive Gaussian-Hermite quadrature is calculated as follows:

$$L_{incmp}(\underline{\theta}, \underline{\beta}) = \int g(y_2)\phi(y_2) dy_2 \approx \sum_{q=1}^{Q} w_q^+ g(x_q^+).$$

We need to compute the mode $\widehat{y_2}$ of $g(y_2)\phi(y_2)$. Since the logarithm transformation is monotone, the mode of $g(y_2)\phi(y_2)$ is the same to the mode of $\ln[g(y_2)\phi(y_2)] = \ln f(y_1, y_2, w = 0)$. In addition, we also need to compute the second derivative of

$\ln[g(y_2)\phi(y_2)]$ in order to obtain the standard deviation of $g(y_2)\phi(y_2)$ based on the Laplace approximation.

The first order derivative of $\ln[g(y_2)\phi(y_2)]$ can be found as

$$\begin{aligned}\frac{\partial}{\partial y_2}\ln f(y_1,y_2,w=0) =& \frac{\partial}{\partial y_2}\left[-\log\left(2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho^2)}\right) - \frac{z}{2(1-\rho^2)}\right.\\ & \left. -\log(1+exp(\beta_0+\beta_1 y_1+\beta_2 y_2))\right]\\ =& \frac{\rho(y_1-\mu_1)}{(1-\rho^2)\sqrt{\sigma_{11}\sigma_{22}}} - \frac{(y_2-\mu_2)}{(1-\rho^2)\sigma_{22}} - \frac{\beta_2 exp(\beta_0+\beta_1 y_1+\beta_2 y_2)}{1+exp(\beta_0+\beta_1 y_1+\beta_2 y_2)}.\end{aligned}$$

Then the mode $\widehat{y}_2$ can be found as the zero solution of the first order derivative

$$\frac{\partial}{\partial y_2}\ln f(y_1,y_2,w=0)\mid_{y_2=\widehat{y}_2}=0.$$

In addition, the second order derivative of $\ln[g(y_2)\phi(y_2)]$ is given by

$$\begin{aligned}\frac{\partial^2}{\partial y_2^2}\ln f(y_1,y_2,w=0) =& -\frac{1}{(1-\rho^2)\sigma_{22}} - \left[\frac{\beta_2^2 exp(\beta_0+\beta_1 y_1+\beta_2 y_2)}{1+exp(\beta_0+\beta_1 y_1+\beta_2 y_2)}\right.\\ & \left. -\frac{\beta_2^2 exp^2(\beta_0+\beta_1 y_1+\beta_2 y_2)}{(1+exp(\beta_0+\beta_1 y_1+\beta_2 y_2))^2}\right].\end{aligned}$$

Then we obtain

$$\widehat{\sigma}_{GQ} = \left[-\frac{\partial^2}{\partial y_2^2}\ln[g(y_2)\phi(y_2)]\mid_{y_2=\widehat{y}_2}\right]^{-\frac{1}{2}}.$$

Finally the likelihood $L_{incmp}(\underline{\theta},\underline{\beta})$ is approximated

$$L_{incmp}(\underline{\theta},\underline{\beta}) = \int g(y_2)\phi(y_2)dy_2 \approx \sum_{q=1}^{n} w_q^+ g(x_q^+)$$

$$= \sum_{q=1}^{Q} \widehat{\sigma}_{GQ} \frac{\phi(x_q^+)}{\phi(\sqrt{2}x_q)} \frac{w_q}{\sqrt{\pi}} f(y_1, x_q^+, w=0) \phi^{-1}(x_q^+)$$
$$= \sum_{q=1}^{Q} \frac{\widehat{\sigma}_{GQ} \cdot w_q \cdot \sqrt{2}}{exp(-x_q^2)} f(y_1, x_q^+, w=0),$$

where $x_q^+ = \widehat{y}_2 + \sqrt{2} \cdot \widehat{\sigma}_{GQ} \cdot x_q$.

This is the likelihood based parametric method. The refreshment sample helps to identify the population parameters $\underline{\theta}$ and attrition parameters $\underline{\beta}$ in the observed likelihood. Without the refreshment sample, this parametric method is infeasible in general non-ignorable missingness scenarios. The likelihood of the incomplete set is obtained by integrating out the missing variable Y_2, which is accomplished through adaptive Gaussian Quadrature. The maximum likelihood estimators are the most efficient if the underlying population and attrition model are correctly specified. However, misspecification of either the population or attrition models can lead to biased estimation and inference. In the next section, we introduce a semi-parametric method which does not require specifying the population density and extends Hirano's constraints to the continuous response setting. The parametric method is a useful benchmark with which to assess the performance of the proposed semi-parametric method in our simulation studies.

4.2 Kernel Density Based Semi-parametric Method

The main idea of this approach is to estimate attrition parameters $\underline{\beta}$ by using the two identification equations provided by Hirano et al. (2001) along with the refreshment

sample:

$$\int \frac{P(W=1)}{\text{logistic}(\beta_0+\beta_1 y_1+\beta_2 y_2)} f(y_1,y_2 \mid W=1)dy_2 = f_1(y_1),$$
$$\int \frac{P(W=1)}{\text{logistic}(\beta_0+\beta_1 y_1+\beta_2 y_2)} f(y_1,y_2 \mid W=1)dy_1 = f_2(y_2). \quad (4.1)$$

The integrand which is common to the two equations constructs the joint density $f(y_1,y_2)$ from the observed part $f(y_1,y_2 \mid W=1)$ by re-weighting each observed likelihood with a factor of $\frac{P(W=1)}{\text{logistic}(\beta_0+\beta_1 y_1+\beta_2 y_2)}$. If the attrition model is correctly specified, the integrand becomes the marginal joint density $f(y_1,y_2)$ with true attrition parameters $\underline{\beta}^0$:

$$\begin{aligned}\frac{P(W=1)}{\text{logistic}(\beta_0^0+\beta_1^0 y_1+\beta_2^0 y_2)} f(y_1,y_2 \mid W=1) &= \frac{f(y_1,y_2,W=1)}{P(W=1 \mid y_1,y_2)} \\ &= \frac{f(y_1,y_2,W=1)}{f(y_1,y_2,W=1)/f(y_1,y_2)} \\ &= f(y_1,y_2).\end{aligned}$$

By taking the integral, the left hand side of Equation (4.1) produces the marginal density functions. Figure 4.1 provides a visual understanding of the main idea of estimating the attrition parameters. The right hand side of Figure 4.1 shows the joint distribution of Y_1 and Y_2 in the population. This is the population we should have seen if there were no missing data. The color of each data point represents the probability of Y_2 being observed. The lighter the color, the higher the probability. This particular missingness pattern is exactly specified by the logistic additive attrition model $P(W=1 \mid y_1,y_2) = \text{logistic}(\beta_0+\beta_1 y_1+\beta_2 y_2)$. This attrition model can be seen as laying a logistic function over the joint distribution of Y_1 and Y_2. The red line represents a 50% missing rate line

on which the equation $\beta_0 + \beta_1 y_1 + \beta_2 y_2$ holds and there is a 50% probability of Y_2 being missing.

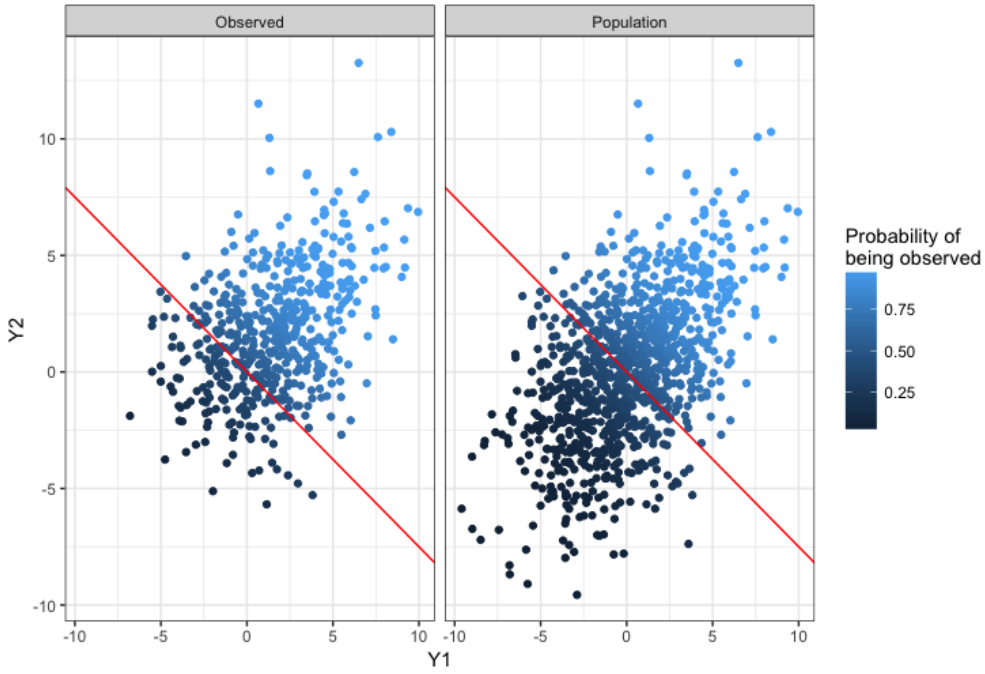


Figure 4.1: Illustration of the logistic additive attrition model.

The left hand side of Figure 4.1 shows the joint distribution we actually see due to the missing data in the second wave. The scatter plot represents the joint distribution of Y_1 and Y_2 in the complete set $f(y_1, y_2 \mid W = 1)$. If the attrition model is correctly specified, the idea of estimating attrition parameters is to find the values of $\underline{\beta}$ that transform the joint density in the complete set (left Figure 4.1) back into the joint density in the population (right Figure 4.1). When we examine the marginal density functions from this transformed joint density by taking the integrals (left hand side of Equation (4.1)), they should match with the true marginal density functions (right hand side of Equation

(4.1)).

The estimation starts with a 2-dimensional kernel density estimator for the observed joint density $f(y_1, y_2 \mid W = 1)$. For any given $\underline{y} = (y_1, y_2)^T$, the kernel density estimator is defined as

$$\widehat{f}_H(y_1, y_2 \mid W = 1) = \widehat{f}_H(\underline{y} \mid W = 1) = \frac{1}{n_c} \sum_{i=1}^{n_c} K_H(\underline{y} - \underline{Y}_i),$$

where $\underline{Y}_i = (Y_{i1}, Y_{i2})^T$, $i = 1, 2, ..., n_c$ are data points in the complete set; H is a 2×2 bandwidth matrix which is symmetric and positive definite; and $K_H(\underline{y}) = |H|^{-1/2} K(H^{-1/2}\underline{y})$, where K is the bivariate normal kernel function defined as $K(\underline{y}) = (2\pi)^{-1} exp(-\underline{y}^T \underline{y}/2)$. The kernel density estimator can viewed as setting bivariate normal densities with centers at each data point (Y_{i1}, Y_{i2}). For any given $\underline{y}$, the kernel density estimator $\widehat{f}_H$ is obtained by taking the average of these bivariate normal densities evaluated at $\underline{y}$. Next, $P(W = 1)$ can be consistently estimated by $\widehat{P}(W = 1) = n_c/N$, the proportion of complete data in the panel. For a given $\underline{\beta} = (\beta_0, \beta_1, \beta_2)^T$, we can construct an estimator for the joint density $f(y_1, y_2)$ as,

$$\widetilde{f}(y_1, y_2 \mid \underline{\beta}) = \frac{\widehat{P}(W = 1)}{\text{logistic}(\beta_0 + \beta_1 y_1 + \beta_2 y_2)} \widehat{f}_H(y_1, y_2 \mid W = 1).$$

Then we can compute the marginal densities of Y_1 and Y_2 by numerically integrating the joint distribution $\widetilde{f}(y_1, y_2 \mid \underline{\beta})$. In particular, for a given grid point y_1, the marginal

density of Y_1 can be computed by

$$\widetilde{f}_1(y_1 \mid \underline{\beta}) = \int \widetilde{f}(y_1, y_2 \mid \underline{\beta}) dy_2 \approx \sum_{i=1}^{n_{grid}} \widetilde{f}(y_1, y_{2i} \mid \underline{\beta}) \times \Delta y_2$$
$$= \sum_{i=1}^{n_{grid}} \widetilde{f}(y_1, y_{2i} \mid \underline{\beta}) \times \frac{range(y_2)}{n_{grid}},$$

where y_{2i} is the ith grid point on Y_2, and n_{grid} denotes the number of grid points in the 2-dimensional kernel density estimator. We use the same number of grid points for Y_1 and Y_2 in our application. Here the integration is approximated by the summation over the grid points expanded in the range of the second wave data. For a given y_2, $\widetilde{f}_2(y_2 \mid \underline{\beta})$ can be defined similarly. It is worth mentioning here that the range of Y_2 cannot be used to expand the the grid points due to missing data in the second wave. To get a representative range of Y_2, we use the range of the refreshment sample to expand the grid points. The resulting marginal density estimates $\widetilde{f}_1(y_1 \mid \underline{\beta})$ and $\widetilde{f}_2(y_2 \mid \underline{\beta})$ are the semi-parametric estimators, which rely on the attrition model. They can consistently estimate the true marginal densities only when the attrition model is correctly specified.

On the right hand side of Equation (4.1), we are able to obtain estimates of both marginal densities directly from the first wave and the refreshment sample. Let $\{y_{i1}\}_{i=1}^{N}$ be the data from the first wave and $\{y_{i2}^r\}_{i=1}^{n}$ be the refreshment sample. We define the following one-dimensional kernel density estimators:

$$\widehat{f}_1(y_1) = \frac{1}{N}\sum_{i=1}^{N} K_{h_1}(y_1 - y_{i1}), \qquad \widehat{f}_2(y_2) = \frac{1}{n}\sum_{i=1}^{n} K_{h_2}(y_2 - y_{i2}^r),$$

where K is the univariate normal density function, and $K_{h_i}(y) = h_i^{-1}K(y/h_i)$ with h_i

being the corresponding bandwidth for $i = 1, 2$.

The H and h_i are bandwidth selectors, and they play an important role in estimating and evaluating kernel density estimators. Several bandwidth selectors are available, and they can be divided into two classes based on the complexity of derivation (Wand and Jones, 1994). The first class is called *quick and simple* selectors; they have simple computable formulas with the goal of providing reasonable bandwidth for a wide range of situations, but without mathematical guarantees of being close to the optimal bandwidth. The second class is called *hi-tech* selectors; they commonly involve more mathematical arguments and computational effort and specially aim to minimize the Mean Integrated Square Error (MISE).

In the quick and simple class, the normal scale bandwidth selector, also known as the "rule of thumb" for choosing the bandwidth of a Gaussian kernel density estimator, is the minimizer of asymptotic MISE for a normal density with the same scale as that estimated for the underlying density (Silverman, 1986; Bowman, 1985). This bandwidth selector provides reasonable estimates for optimal bandwidth when the data are close to normal. It comes with the risk of over-smoothing and masking important multimodality features of a multimodal distribution.

In the hi-tech class, there are four commonly used bandwidth selectors, namely least squares cross-validation (LSCV) selector (Rudemo, 1982; Bowman, 1984), biased cross-validation (BCV) selector (Scott and Terrell, 1987), direct plug-in (DPI) selector (Scott et al., 1977) and smoothed cross-validation (SCV) selector (Hall et al., 1992). Wand and Jones (1994) conducted comprehensive simulations to compare the performance of these hi-tech bandwidth selectors. LSCV selector is relatively unbiased in

estimating the optimal bandwidth, but it has the largest variation. Both BCV and SCV selectors have more accuracy and less variation in estimating the optimal bandwidth, but a notable bias is present. The DPI selector shows a better combined performance in both bias and variation. In our application, the simple normal scale bandwidth selector is used due to the normal population setting and its computational efficiency.

The estimators of the attrition parameters $\underline{\beta}$ are the minimizer of the objective function $M_{N,n}(\underline{\beta})$, which is defined as,

$$\begin{aligned}
\widehat{\underline{\beta}} &= \arg\min_{\underline{\beta}} M_{N,n}(\underline{\beta}) \\
&= \arg\min_{\underline{\beta}} \left\{ M_N(\underline{\beta}) + M_n(\underline{\beta}) \right\} \\
&= \arg\min_{\underline{\beta}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left\{ e_1(y_{i1}) \left[\widetilde{f}_1(y_{i1} \mid \underline{\beta}) - \widehat{f}_1(y_{i1}) \right] \right\}^2 + \right. \\
&\qquad \left. \frac{1}{n} \sum_{i=1}^{n} \left\{ e_2(y^r_{i2}) \left[\widetilde{f}_2(y^r_{i2} \mid \underline{\beta}) - \widehat{f}_2(y^r_{i2}) \right] \right\}^2 \right\}, \qquad (4.2)
\end{aligned}$$

where $e_1(\cdot)$ and $e_2(\cdot)$ are given weight functions. Intuitively, $M_N(\underline{\beta})$ and $M_n(\underline{\beta})$ in (4.2) measure the differences between two types of marginal density estimators: the semi-parameteric estimators constructed based on the attrition model and the nonparametric kernel estimators using the fully observed data in the first wave or the refreshment sample. Only for the true attrition parameters $\underline{\beta}$ do the semi-parametric estimators provide consistent estimates of the marginals and make the objective function $M_{N,n}$ be close to zero. Therefore, our estimator $\widehat{\underline{\beta}}$ is the one such that $M_{N,n}(\widehat{\underline{\beta}})$ is as close to zero as possible.

In (4.2), the additional weight functions e_1 and e_2 enable us to adaptively compare the differences between two types of estimators of the marginal densities. For example, it is well known that the performance of kernel density estimators is less satisfactory at the boundary due to the edge effect. Therefore, the comparison at the boundary may add unnecessary noise into the objective function and lead to numerically unstable estimates of $\underline{\beta}$. In our simulation, we have chosen the weight functions $e_1(\cdot)$ and $e_2(\cdot)$ such that the summation in $M_N(\underline{\beta})$ focuses on the middle 70% of the first wave panel data and the middle 70% of the refreshment sample.

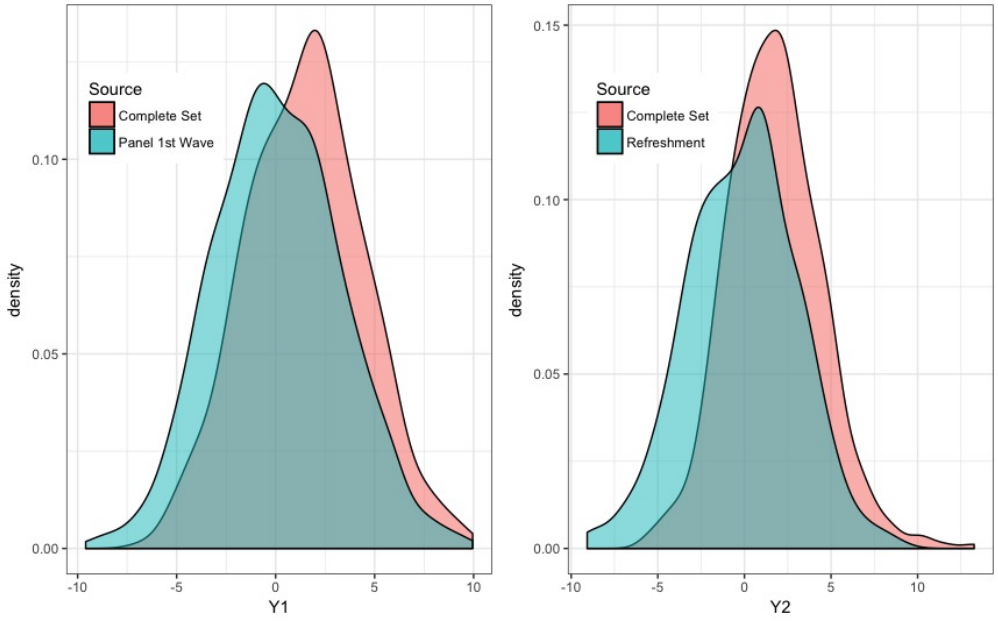


Figure 4.2: Marginal comparison

Figure 4.2 shows the intuition behind the estimation process from the marginal perspective. Green density functions represent the estimates of the true marginals. On the left hand side, the green density function is the true marginal density estimate of

Y_1, $\widehat{f_1}(y_1)$, based on first wave panel data. On the right hand side, the green density function is the true marginal estimate of Y_2, $\widehat{f_2}(y_2)$, based on the refreshment sample. These are the density estimates on the right hand side of Equation (4.1). Red density functions are the marginal density functions of the complete set. They are obtained by integrating the joint density of the complete set without applying the re-weighting factor $\frac{P(W=1)}{\text{logistic}(\beta_0+\beta_1 y_1+\beta_2 y_2)}$ on the left hand side of Equation (4.1). The difference in the two density functions is due to missing data. Thus, the idea behind the proposed method of estimating the attrition parameters is to find the values of $\underline{\beta}$ which re-weight the joint density of the complete set so that the red density functions shift to match the green density functions as closely as possible.

4.3 Asymptotic Theory for Kernel Density Based Semi-parametric Estimators

In previous sections, we introduced two new methods in analyzing a two-wave MNAR continuous response data with the help of the refreshment sample. The focus of these methods is primarily on identifying the missing mechanism in order to provide support and validation for the main data analysis. In the following section, we develop the asymptotic theory for the kernel density based semi-parametric estimators.

4.3.1 Preliminary Notation and Conditions

Let the joint distribution of (Y_1, Y_2) and the conditional distribution of (Y_1, Y_2) given $W = 1$ respectively have density functions $f(y_1, y_2)$ and $f(y_1, y_2 \mid W = 1)$, with respect to the (product) Lebesgue measure. Let $P(W = 1)$ be the marginal probability of Y_2 being observed. We assume the attrition model $P(W = 1 \mid y_1, y_2) = \text{logistic}(\beta_0^0 + \beta_1^0 y_1 + \beta_2^0 y_2)$, with $\underline{\beta}^0 = (\beta_0^0, \beta_1^0, \beta_2^0)$ being the true parameters. Let $M_{N,n}(\underline{\beta})$ be the objective function for estimating $\underline{\beta}$ as defined in Equation (4.2). Let $\widehat{P}(W = 1)$ be the estimate of $P(W = 1)$ and $\widehat{f}_H(y_1, y_2 \mid W = 1)$ be the two-dimensional kernel density estimate of $f(y_1, y_2 \mid W = 1)$ respectively. Let P_n be the empirical expectation operator defined as $P_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$ and P be the expectation operator defined as $Pf = \int f dP = E(f(X))$.

Within the objective function $M_{N,n}(\underline{\beta})$, let

$$\begin{aligned} f_1(y_1 \mid \underline{\beta}) &= \int \frac{f(y_1, y_2 \mid W = 1) P(W = 1)}{\text{logistic}(\beta_0 + \beta_1 y_1 + \beta_2 y_2)} dy_2 \\ &= \int \frac{f(y_1, y_2 \mid W = 1) P(W = 1)}{1/(1 + \exp(-\beta_0 - \beta_1 y_1 - \beta_2 y_2))} dy_2 \end{aligned}$$

denote the proposed density function of y_1 given $\underline{\beta}$. Let

$$\begin{aligned} \widetilde{f}_1(y_1 \mid \underline{\beta}) &= \int \frac{\widehat{f}_H(y_1, y_2 \mid W = 1) \widehat{P}(W = 1)}{\text{logistic}(\beta_0 + \beta_1 y_1 + \beta_2 y_2)} dy_2 \\ &= \int \frac{\frac{1}{N}\sum_{j=1}^{N} w_j K_{h_1}(y_1 - y_{j1}) K_{h_2}(y_2 - y_{j2})}{1/(1 + \exp(-\beta_0 - \beta_1 y_1 - \beta_2 y_2))} dy_2 \end{aligned}$$

be the semi-parametric estimate of $f_1(y_1 \mid \underline{\beta})$, where w_j is the missingness indicator, and $K(\cdot)$ is the kernel function with bandwidth h_1 or h_2. Let $f_1(y_1)$ be the first wave marginal

density function and $\widehat{f_1}(y_1) = \frac{1}{N}\sum_{j=1}^{N} K_{h_1}\left(y_1 - y_{j1}\right)$ be the kernel density estimate of $f_1(y_1)$. Let

$$\begin{aligned} A_{1\underline{\beta}}(y_1) =& f_1(y_1 \mid \underline{\beta}) - f_1(y_1), \\ B_{1\underline{\beta}}(y_1) =& \widetilde{f_1}(y_1 \mid \underline{\beta}) - f_1(y_1 \mid \underline{\beta}), \\ C_1(y_1) =& f_1(y_1) - \widehat{f_1}(y_1). \end{aligned} \tag{4.3}$$

Let $e_1(y_1)$ be the weight function when constructing the empirical objective function on first wave.

In the same manner, let

$$\begin{aligned} f_2(y_2 \mid \underline{\beta}) &= \int \frac{f(y_1, y_2 \mid W=1)P(W=1)}{\text{logistic}(\beta_0 + \beta_1 y_1 + \beta_2 y_2)} dy_1 \\ &= \int \frac{f(y_1, y_2 \mid W=1)P(W=1)}{1/\left(1+\exp\left(-\beta_0 - \beta_1 y_1 - \beta_2 y_2\right)\right)} dy_1 \end{aligned}$$

denote the proposed density function of y_2 given $\underline{\beta}$. Let

$$\begin{aligned} \widetilde{f_2}(y_2 \mid \underline{\beta}) &= \int \frac{\widehat{f_H}(y_1, y_2 \mid W=1)\widehat{P}(W=1)}{\text{logistic}(\beta_0 + \beta_1 y_1 + \beta_2 y_2)} dy_1 \\ &= \int \frac{\frac{1}{N}\sum_{j=1}^{N} w_j K_{h_1}\left(y_1 - y_{j1}\right) K_{h_2}\left(y_2 - y_{j2}\right)}{1/\left(1+\exp\left(-\beta_0 - \beta_1 y_1 - \beta_2 y_2\right)\right)} dy_1 \end{aligned}$$

be the semi-parametric estimate of $f_2(y_2 \mid \underline{\beta})$. Let $f_2(y_2)$ be the second wave marginal density function and $\widehat{f_2}(y_2) = \frac{1}{n}\sum_{j=1}^{n} K_{h_2}\left(y_2 - y_{j2}\right)$ be the kernel density estimate of

$f_2(y_2)$. Let

$$\begin{aligned} A_{2\underline{\beta}}(y_2) =& f_2(y_2 \mid \underline{\beta}) - f_2(y_2), \\ B_{2\underline{\beta}}(y_2) =& \widetilde{f}_2(y_2 \mid \underline{\beta}) - f_2(y_2 \mid \underline{\beta}), \\ C_2(y_2) =& f_2(y_2) - \widehat{f}_2(y_2). \end{aligned} \tag{4.4}$$

Let $e_2(y_2)$ be the weight function when constructing the empirical objective function on the second wave.

To establish our asymptotic results, we need the following conditions:

(A1) Let $S = \{(y_1, y_2) : f(y_1, y_2) > 0\}$ be the support of the joint density function of (Y_1, Y_2). Without loss of generality, assume $S = [-t, t] \times [-u, u]$ is compact. The support of $f(y_1, y_2 \mid W = 1)$ coincides with S;

(A2) the density functions $f(y_1, y_2)$ and $f(y_1, y_2 \mid W = 1)$ are both continous;

(A3) the parameters $\underline{\beta} = (\beta_0, \beta_1, \beta_2)$ belong to a compact set Θ, and without loss of generality, we assume $\beta_0 \in [-b_0, b_0]$, $\beta_1 \in [-b_1, b_1]$ and $\beta_2 \in [-b_2, b_2]$;

(A4) $K(y)$ is the kernel function with

(a) $\int K(y)dy = 1$,

(b) $\int |K(y)|\, dy < +\infty$,

(c) $K(y) \to 0$ as $|y| \to +\infty$,

(d) $\int |y \log |y||^{1/2} |dK(y)| < +\infty$;

(A5) $H = H_{n_c}$ is the bandwidth matrix for the 2-dimensional kernel, where n_c is the complete panel size, and

(a) $H \to 0$ as $n_c \to +\infty$,

(b) $n_c H^4 \to +\infty$ as $n_c \to +\infty$;

(A6) h_1 and h_2 are the bandwidths for the 1-dimensional kernel with

(a) $h_1 \to 0$ and $h_2 \to 0$,

(b) $(Nh_1)^{-1} \log N \to 0$ as $N \to +\infty$ and $(nh_2)^{-1} \log n \to 0$ as $n \to +\infty$, where N is panel size and n is the refreshment sample size.

4.3.2 Identifiability

In this section, we show that the attrition parameters $\underline{\beta} = (\beta_0, \beta_1, \beta_2)$ are well identified based on the two marginal distributions of Y_1 and Y_2.

Lemma 4.1. *Suppose conditions (A1) and (A2) are satisfied, then for almost all* $(y_1, y_2) \in S$, *there is a unique set of parameters* $(\beta_0, \beta_1, \beta_2)$ *satisfying*

$$\begin{aligned} \int \frac{P(W=1)}{logistic(\beta_0 + \beta_1 y_1 + \beta_2 y_2)} f(y_1, y_2 \mid W=1) dy_2 &= f_1(y_1), \\ \int \frac{P(W=1)}{logistic(\beta_0 + \beta_1 y_1 + \beta_2 y_2)} f(y_1, y_2 \mid W=1) dy_1 &= f_2(y_2). \end{aligned} \tag{4.5}$$

Proof of Lemma 4.1. The proof follows directly from Theorem 1 of Hirano et al. (2001)

■

Theorem 4.1. *The two constraints in Equation (4.5) are uniquely satisfied by the true parameters* $\underline{\beta}^0 = (\beta_0^0, \beta_1^0, \beta_2^0)$.

Proof of Theorem 4.1. Given the attrition model as $P(W = 1 \mid y_1, y_2) = \text{logistic}(\beta_0^0 + \beta_1^0 y_1 + \beta_2^0 y_2)$, it is sufficient to show that $\underline{\beta}^0$ satisfies Equation (4.5) with

$$\frac{P(W=1)}{\text{logistic}(\beta_0^0 + \beta_1^0 y_1 + \beta_2^0 y_2)} f(y_1, y_2 \mid W = 1) = f(y_1, y_2). \qquad \blacksquare$$

4.3.3 Consistency

The estimator of the parameters $\underline{\beta}$ is defined as the minimizer of the objective function

$$\begin{aligned} M_{N,n}(\underline{\beta}) &= M_N(\underline{\beta}) + M_n(\underline{\beta}) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left[\widetilde{f}_1(y_{i1} \mid \underline{\beta}) - \widehat{f}_1(y_{i1})\right]^2 + \frac{1}{n}\sum_{i=1}^{n}\left[\widetilde{f}_2(y_{i2} \mid \underline{\beta}) - \widehat{f}_2(y_{i2})\right]^2 \\ &= \frac{1}{N}\sum_{i=1}^{N} m_{\underline{\beta}}(y_{i1}) + \frac{1}{n}\sum_{i=1}^{n} m_{\underline{\beta}}(y_{i2}). \end{aligned}$$

Notice that we drop the concentration weights $e_1(y_{i1})$ and $e_2(y_{i2})$ from the objective function for simplicity. To show consistency of the minimizer $\widehat{\underline{\beta}}$, we proceed in two steps. First, we show the uniform convergence of $M_{N,n}(\underline{\beta})$ to its probability limit. Second, we show that this probability limit has an unique minimizer $\underline{\beta}^0$. Then the consistency follows from Theorem 5.7 of Van der Vaart (2000).

Focusing on the first part of $M_{N,n}(\underline{\beta})$, using notations defined in Equation (4.3), we

have

$$
\begin{aligned}
M_N(\underline{\beta}) =& \frac{1}{N}\sum_{i=1}^{N}\left[\widetilde{f}_1(y_{i1} \mid \underline{\beta}) - \widehat{f}_1(y_{i1})\right]^2 \\
=& \frac{1}{N}\sum_{i=1}^{N}\left\{\left[f_1(y_{i1} \mid \underline{\beta}) - f_1(y_{i1})\right] + \left[\widetilde{f}_1(y_{i1} \mid \underline{\beta}) - f_1(y_{i1} \mid \underline{\beta})\right]\right. \\
&\left. + \left[f_1(y_{i1}) - \widehat{f}_1(y_{i1})\right]\right\}^2 \\
=& \frac{1}{N}\sum_{i=1}^{N}\left\{A_{1\underline{\beta}}(y_{i1}) + B_{1\underline{\beta}}(y_{i1}) + C_1(y_{i1})\right\}^2 \\
=& \frac{1}{N}\sum_{i=1}^{N} A_{1\underline{\beta}}(y_{i1})^2 + \frac{1}{N}\sum_{i=1}^{N}\left[2A_{1\underline{\beta}}(y_{i1})B_{1\underline{\beta}}(y_{i1}) + A_{\underline{\beta}}(y_{i1})C_1(y_{i1})\right. \\
&\left. + B_{1\underline{\beta}}(y_{i1})C_1(y_{i1}) + B_{1\underline{\beta}}(y_{i1})^2 + C_1(y_{i1})^2\right].
\end{aligned}
$$

First we show that the leading term $\frac{1}{N}\sum_{i=1}^{N} A_{1\underline{\beta}}(y_{i1})^2$ uniformly converges to its probability limit under mild conditions.

Lemma 4.2. *For any θ in a compact set Θ, let $x \mapsto f_\theta(x)$ be a given measurable function. Suppose $\theta \mapsto f_\theta(x)$ is continuous for every x and suppose that there exists a function F such that $|f_\theta| \leq F$ for every $\theta \in \Theta$, and $PF < +\infty$, then*

$$\sup_{\theta\in\Theta} |P_n f_\theta - P f_\theta| \xrightarrow{P} 0.$$

This result is shown in section 19.2 of Van der Vaart (2000).

Lemma 4.3. *Suppose (A1)–(A4) are satisfied. The set of functions* $\left[f_1(y_1 \mid \underline{\beta}) - f_1(y_1)\right]^2 =$

$A_{1\underline{\beta}}(y_1)^2$ indexed by $\underline{\beta}$ is in the Glivenko-Cantelli class and

$$\sup_{\underline{\beta}\in\Theta}\left|P_N A_{1\underline{\beta}}^2 - PA_{1\underline{\beta}}^2\right| \xrightarrow{P} 0.$$

Proof of Lemma 4.3. The proof is given in the appendix. ■

Lemma 4.4. *Suppose conditions (A4), (A5) and (A6) are satisfied. Then*

$$\sup_{\underline{\beta}}\left\{\frac{1}{N}\sum_{i=1}^{N}\Big[2A_{1\underline{\beta}}(y_{i1})B_{1\underline{\beta}}(y_{i1}) + A_{1\underline{\beta}}(y_{i1})C_1(y_{i1}) + B_{1\underline{\beta}}(y_{i1})C_1(y_{i1}) + B_{1\underline{\beta}}(y_{i1})^2 + C_1(y_{i1})^2\Big]\right\} = o_p(1).$$

Proof of Lemma 4.4. The proof is given in the appendix. ■

Lemma 4.5. *Under (A1) – (A6), $M_N(\underline{\beta})$ uniformly converges to its probability limit $E\left[f_1(Y_1 \mid \underline{\beta}) - f_1(Y_1)\right]^2$. That is,*

$$\sup_{\underline{\beta}\in\Theta}\left|M_N(\underline{\beta}) - E\left[f_1(Y_1 \mid \underline{\beta}) - f_1(Y_1)\right]^2\right| \xrightarrow{P} 0.$$

Proof of Lemma 4.5. The proof follows from Lemmas 4.3 and 4.4. ■

Lemma 4.6. *Under (A1) – (A6), $M_{N,n}(\underline{\beta})$ uniformly converges to its probability limit $E\left[f_1(Y_1 \mid \underline{\beta}) - f_1(Y_1)\right]^2 + E\left[f_2(Y_2 \mid \underline{\beta}) - f_2(Y_2)\right]^2$.*

Proof of Lemma 4.6. Similar to Lemma 4.5, one can show that $M_n(\underline{\beta})$ uniformly con-

verges to its probability limit $E\left[f_2(Y_2 \mid \underline{\beta}) - f_2(Y_2)\right]^2$, where

$$\begin{aligned} M_n(\underline{\beta}) =& \frac{1}{n}\sum_{i=1}^{n}\left[\widetilde{f}_2(y_{i2} \mid \underline{\beta}) - \widehat{f}_2(y_{i2})\right]^2 \\ =& \frac{1}{n}\sum_{i=1}^{n}\left\{\left[f_2(y_{i2} \mid \underline{\beta}) - f_2(y_{i2})\right] + \left[\widetilde{f}_2(y_{i2} \mid \underline{\beta}) - f_2(y_{i2} \mid \underline{\beta})\right]\right. \\ & \left. + \left[f_2(y_{i2}) - \widehat{f}_2(y_{i2})\right]\right\}^2. \end{aligned}$$

Together with Lemma 4.5, this proves Lemma 4.6. ■

Theorem 4.2. *The minimizer $\widehat{\underline{\beta}}$ of $M_{N,n}(\underline{\beta})$ converges in probability to $\underline{\beta}^0$, the unique minimizer of $E\left[f_1(Y_1 \mid \underline{\beta}) - f_1(Y_1)\right]^2 + E\left[f_2(Y_2 \mid \underline{\beta}) - f_2(Y_2)\right]^2$.*

Proof of Theorem 4.2. By Lemma 4.1, if the attrition model is correctly specified, then $\underline{\beta} = \underline{\beta}^0$ is the unique set of parameters that satisfies Equation (4.5),

$$\begin{aligned} \int \frac{P(W=1)}{\text{logistic}(\beta_0^0 + \beta_1^0 y_1 + \beta_2^0 y_2)} f(y_1, y_2 \mid W=1) dy_2 = f_1(y_1), \\ \int \frac{P(W=1)}{\text{logistic}(\beta_0^0 + \beta_1^0 y_1 + \beta_2^0 y_2)} f(y_1, y_2 \mid W=1) dy_1 = f_2(y_2). \end{aligned}$$

That is, for almost all $(y_1, y_2) \in S$, $\underline{\beta} = \underline{\beta}^0$ is the unique set of parameters that have

$$f_1(y_1 \mid \underline{\beta}^0) - f_1(y_1) = 0 \quad \text{and} \quad f_2(y_2 \mid \underline{\beta}^0) - f_2(y_2) = 0.$$

Thus $\underline{\beta}_0$ is the unique minimizer of

$$E\left[f_1(Y_1 \mid \underline{\beta}) - f_1(Y_1)\right]^2 + E\left[f_2(Y_2 \mid \underline{\beta}) - f_2(Y_2)\right]^2.$$

Combining with Lemma 4.6, the consistency of the minimizer $\widehat{\underline{\beta}}$ to $\underline{\beta}^0$ follows from Theorem 5.7 of Van der Vaart (2000). ■

4.3.4 Asymptotic Normality

The estimator $\widehat{\underline{\beta}}$ based on the objective function $M_{N,n}\left(\underline{\beta}\right)$ is an M-estimator. The asymptotic properties of $\widehat{\underline{\beta}}$ can be evaluated through the form of a Z-estimator by taking the derivative of $M_{N,n}\left(\underline{\beta}\right)$. There are two parts in $M_{N,n}\left(\underline{\beta}\right)$, namely $M_N\left(\underline{\beta}\right)$ and $M_n\left(\underline{\beta}\right)$. In the following we will tackle each part separately and put them back together at the end to obtain the asymptotic properties of $\widehat{\underline{\beta}}$.

4.3.4.1 The First Part, $M_N\left(\underline{\beta}\right)$

Using the notation in Equation (4.3), the first part of the objective function $M_{N,n}\left(\underline{\beta}\right)$ is

$$\begin{aligned} M_N\left(\underline{\beta}\right) &= \frac{1}{N}\sum_{i=1}^{N}\left\{e_1\left(y_{i1}\right)\left[\widetilde{f}_1(y_{i1}\mid\underline{\beta})-\widehat{f}_1(y_{i1})\right]\right\}^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}\left\{e_1\left(y_{i1}\right)\left[A_{1\underline{\beta}}(y_{i1})+B_{1\underline{\beta}}(y_{i1})+C_1(y_{i1})\right]\right\}^2. \end{aligned}$$

Then the first order derivative of $M_N\left(\underline{\beta}\right)$ is

$$\varphi_N\left(\underline{\beta}\right) = \frac{\partial}{\partial \underline{\beta}} M_N\left(\underline{\beta}\right)$$
$$= \frac{2}{N}\sum_{i=1}^{N} e_1^2(y_{i1})\left[A_{1\underline{\beta}}(y_{i1}) + B_{1\underline{\beta}}(y_{i1}) + C_1(y_{i1})\right]\left[\frac{\partial}{\partial \underline{\beta}} A_{1\underline{\beta}}(y_{i1}) + \frac{\partial}{\partial \underline{\beta}} B_{1\underline{\beta}}(y_{i1})\right].$$

When $\varphi_N\left(\underline{\beta}\right)$ is evaluated at the truth $\underline{\beta}^0$, $f_1(y_{i1} \mid \underline{\beta}^0) = f_1(y_{i1})$, and $A_{1\underline{\beta}^0}(y_{i1}) = 0$. Then

$$\begin{aligned}\varphi_N\left(\underline{\beta}^0\right) =& \frac{2}{N}\sum_{i=1}^{N}\left[e_1^2(y_{i1})\frac{\partial}{\partial \underline{\beta}} A_{1\underline{\beta}^0}(y_{i1})\left(B_{1\underline{\beta}_0}(y_{i1}) + C_1(y_{i1})\right)\right.\\ &\left. + e_1^2(y_{i1})\frac{\partial}{\partial \underline{\beta}} B_{1\underline{\beta}^0}(y_{i1})\left(B_{1\underline{\beta}^0}(y_{i1}) + C_1(y_{i1})\right)\right]\\ \approx& \frac{1}{N}\sum_{i=1}^{N}\left[2e_1^2(y_{i1})\frac{\partial}{\partial \underline{\beta}} A_{1\underline{\beta}^0}(y_{i1})\left(B_{1\underline{\beta}^0}(y_{i1}) + C_1(y_{i1})\right)\right].\end{aligned}$$

The approximation is due to the fact that $B_{1\underline{\beta}^0}(y_{i1})$, $C_1(y_{i1})$ and $\frac{\partial}{\partial \underline{\beta}} B_{1\underline{\beta}^0}(y_{i1})$ are $o_p(1)$. Thus, the first term is the dominant term. Define $\frac{\partial}{\partial \underline{\beta}} A_{1\underline{\beta}^0}(y_{i1}) = \underline{g}(y_{i1}) = [g_1(y_{i1}), g_2(y_{i1}), g_3(y_{i1})]^T$ with

$$\begin{aligned}g_1(y_{i1}) &= \int f(y_{i1}, y_2 \mid W=1) P(W=1) \exp\left(-\beta_0^0 - \beta_1^0 y_{i1} - \beta_2^0 y_2\right)(-1)\, dy_2\\ &= \int \frac{f(y_{i1}, y_2)}{1 + \exp\left(\beta_0^0 + \beta_1^0 y_{i1} + \beta_2^0 y_2\right)}(-1)\, dy_2,\\ g_2(y_{i1}) &= \int \frac{f(y_{i1}, y_2)}{1 + \exp\left(\beta_0^0 + \beta_1^0 y_{i1} + \beta_2^0 y_2\right)}(-y_{i1})\, dy_2,\\ g_3(y_{i1}) &= \int \frac{f(y_{i1}, y_2)}{1 + \exp\left(\beta_0^0 + \beta_1^0 y_{i1} + \beta_2^0 y_2\right)}(-y_2)\, dy_2.\end{aligned}$$

Define a function

$$T_1(x,y,z,w) = \int \frac{wK_{h_1}(z-x)K_{h_2}(y_2-y)}{1/\left(1+\exp\left(-\beta_0^0-\beta_1^0 z-\beta_2^0 y_2\right)\right)}dy_2 - K_{h_1}(z-x).$$

Then we have

$$\begin{aligned}
\varphi_N\left(\underline{\beta}^0\right) \approx & \frac{1}{N}\sum_{i=1}^{N}\left[2e_1^2(y_{i1})\frac{\partial}{\partial\underline{\beta}}A_{1\underline{\beta}^0}(y_{i1})\left(B_{1\underline{\beta}^0}(y_{i1})+C_1(y_{i1})\right)\right] \\
= & \frac{1}{N}\sum_{i=1}^{N}\left[2e_1^2(y_{i1})\underline{g}(y_{i1})\left(\int \frac{\frac{1}{N}\sum_{j=1}^{N} w_j K_{h_1}(y_{i1}-y_{j1})K_{h_2}(y_2-y_{j2})}{1/\left(1+\exp\left(-\beta_0^0-\beta_1^0 y_{i1}-\beta_2^0 y_2\right)\right)}dy_2\right.\right. \\
& \left.\left. -\frac{1}{N}\sum_{j=1}^{N}K_{h_1}(y_{i1}-y_{j1})\right)\right] \\
= & \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[2e_1^2(y_{i1})\underline{g}(y_{i1})T_1(y_{j1},y_{j2},y_{i1},w_j)\right].
\end{aligned}$$

Let $\underline{X}_i = [Y_{i1}, Y_{i2}, W_i]^T$ and $\underline{X}_j = \left[Y_{j1}, Y_{j2}, W_j\right]^T$ be independent samples from the panel. Define a symmetric function

$$h\left(\underline{x}_i,\underline{x}_j\right) = e_1^2(y_{i1})\underline{g}(y_{i1})T_1\left(y_{j1},y_{j2},y_{i1},w_j\right) + e_1^2\left(y_{j1}\right)\underline{g}\left(y_{j1}\right)T_1\left(y_{i1},y_{i2},y_{j1},w_i\right).$$

Then

$$\varphi_N\left(\underline{\beta}^0\right) \approx \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} h\left(\underline{x}_i,\underline{x}_j\right)$$

is a V-statistic.

Lemma 4.7. *Assume (A1) - (A6). Define* $h_1\left(\underline{X}_i\right) = E\left[h\left(\underline{X}_i,\underline{X}_j\right) \mid \underline{X}_i\right]$ *and* $\Sigma_1 =$

Var $[h_1(\underline{X})]$. *Then one has* $E\left[h\left(\underline{X}_i, \underline{X}_j\right)\right] \approx \underline{0}$, *and*

$$\sqrt{N}\varphi_N\left(\underline{\beta}^0\right) \sim N\left(\underline{0}, 4\Sigma_1\right).$$

Proof of Lemma 4.7. The proof is given in the appendix. ■

Lemma 4.8. *The probability limit of the second derivative of* $M_N\left(\underline{\beta}\right)$ *is*

$$E\left[\frac{\partial^2}{\partial \underline{\beta}^2} M_N\left(\underline{\beta}^0\right)\right] \approx 2E\left[e_1^2(Y_1)\,\underline{g}(Y_1)\,\underline{g}(Y_1)^T\right].$$

Proof of Lemma 4.8. The proof is given in the appendix. ■

4.3.4.2 Second Part, $M_n\left(\underline{\beta}\right)$

Using the notation in Equation (4.4), the second part of the objective function $M_{N,n}\left(\underline{\beta}\right)$ is

$$\begin{aligned} M_n\left(\underline{\beta}\right) &= \frac{1}{n}\sum_{i=1}^{n}\left[e_2(y_{i2})\left[\widetilde{f}_2(y_{i2} \mid \underline{\beta}) - \widehat{f}_2(y_{i2})\right]\right]^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[e_2(y_{i2})\left[A_{2\underline{\beta}}(y_{i2}) + B_{2\underline{\beta}}(y_{i2}) + C_2(y_{i2})\right]\right]^2. \end{aligned}$$

The first order derivative of $M_n\left(\underline{\beta}\right)$ is

$$\varphi_n\left(\underline{\beta}\right) = \frac{\partial}{\partial \underline{\beta}} M_n\left(\underline{\beta}\right)$$
$$= \frac{2}{n}\sum_{i=1}^{n} e_2^2(y_{i2})\left[A_{2\underline{\beta}}(y_{i2}) + B_{2\underline{\beta}}(y_{i2}) + C_2(y_{i2})\right]\left[\frac{\partial}{\partial \underline{\beta}} A_{2\underline{\beta}}(y_{i2}) + \frac{\partial}{\partial \underline{\beta}} B_{2\underline{\beta}}(y_{i2})\right].$$

When $\varphi_n\left(\underline{\beta}\right)$ is evaluated at the truth $\underline{\beta}^0$, $f_2(y_{i2} \mid \underline{\beta}^0) = f_2(y_{i2})$, and $A_{2\underline{\beta}^0}(y_{i2}) = 0$. Then

$$\begin{aligned}\varphi_n\left(\underline{\beta}^0\right) =& \frac{2}{n}\sum_{i=1}^{n}\left[e_2^2(y_{i2})\frac{\partial}{\partial \underline{\beta}} A_{2\underline{\beta}^0}(y_{i2})\left(B_{2\underline{\beta}^0}(y_{i2}) + C_2(y_{i2})\right)\right.\\ &\left. + e_2^2(y_{i2})\frac{\partial}{\partial \underline{\beta}} B_{2\underline{\beta}^0}(y_{i2})\left(B_{2\underline{\beta}^0}(y_{i2}) + C_2(y_{i2})\right)\right]\\ \approx& \frac{1}{n}\sum_{i=1}^{n}\left[2e_2^2(y_{i2})\frac{\partial}{\partial \underline{\beta}} A_{2\underline{\beta}^0}(y_{i2})\left(B_{2\underline{\beta}^0}(y_{i2}) + C_2(y_{i2})\right)\right].\end{aligned}$$

The approximation is due to the fact that $B_{2\underline{\beta}^0}(y_{i2})$, $C_2(y_{i2})$ and $\frac{\partial}{\partial \underline{\beta}} B_{2\underline{\beta}^0}(y_{i2})$ are $o_p(1)$. Thus, the first term is the dominant term. Define $\frac{\partial}{\partial \underline{\beta}} A_{2\underline{\beta}^0}(y_{i2}) = \underline{k}(y_{i2}) = [k_1(y_{i2}), k_2(y_{i2}), k_3(y_{i2})]^T$ with

$$\begin{aligned}k_1(y_{i2}) &= \int f(y_1, y_{i2} \mid W = 1) P(W = 1) \exp\left(-\beta_0^0 - \beta_1^0 y_1 - \beta_2^0 y_{i2}\right)(-1)\, dy_1\\ &= \int \frac{f(y_1, y_{i2})}{1 + \exp\left(\beta_0^0 + \beta_1^0 y_1 + \beta_2^0 y_{i2}\right)}(-1)\, dy_1,\\ k_2(y_{i2}) &= \int \frac{f(y_1, y_{i2})}{1 + \exp\left(\beta_0^0 + \beta_1^0 y_1 + \beta_2^0 y_{i2}\right)}(-y_1)\, dy_1,\\ k_3(y_{i2}) &= \int \frac{f(y_1, y_{i2})}{1 + \exp\left(\beta_0^0 + \beta_1^0 y_1 + \beta_2^0 y_{i2}\right)}(-y_{i2})\, dy_1.\end{aligned}$$

Define a function

$$T_2(x,y,z,w) = \int \frac{wK_{h_1}(y_1-x)K_{h_2}(z-y)}{1/\left(1+\exp\left(-\beta_0^0-\beta_1^0 y_1-\beta_2^0 z\right)\right)}dy_1 - f_2(z).$$

Then we have

$$\begin{aligned}
\varphi_n\left(\underline{\beta}^0\right) \approx & \frac{1}{n}\sum_{i=1}^{n}\left[2e_2^2(y_{i2})\frac{\partial}{\partial\underline{\beta}}A_{2\underline{\beta}^0}(y_{i2})\left(B_{2\underline{\beta}^0}(y_{i2})+C_2(y_{i2})\right)\right] \\
= & \frac{1}{n}\sum_{i=1}^{n}\left[2e_2^2(y_{i2})\underline{k}(y_{i2})\left(\int \frac{\frac{1}{N}\sum_{j=1}^{N} w_j K_{h_1}\left(y_1-y_{j1}\right)K_{h_2}\left(y_{i2}-y_{j2}\right)}{1/\left(1+\exp\left(-\beta_0^0-\beta_1^0 y_1-\beta_2^0 y_{i2}\right)\right)}dy_1\right.\right. \\
& \left.\left. -f_2(y_{i2})+f_2(y_{i2})-\frac{1}{n}\sum_{l=1}^{n}K_{h_2}(y_{i2}-y_{l2})\right)\right] \\
= & \frac{1}{nN}\sum_{i=1}^{n}\sum_{j=1}^{N}\left[2e_2^2(y_{i2})\underline{k}(y_{i2})T_2\left(y_{j1},y_{j2},y_{i2},w_j\right)\right] \\
& +\frac{1}{n^2}\sum_{i=1}^{n}\sum_{l=1}^{n}\left[2e_2^2(y_{i2})\underline{k}(y_{i2})\left(f_2(y_{i2})-K_{h_2}(y_{i2}-y_{l2})\right)\right] \\
= & \varphi_n^{(1)}\left(\underline{\beta}^0\right)+\varphi_n^{(2)}\left(\underline{\beta}^0\right).
\end{aligned}$$

Lemma 4.9. *Define* $h_1^{(1)}\left(\underline{X}_j\right) = E\left[2e_2^2(Y_{i2})\underline{k}(Y_{i2})T_2\left(Y_{j1},Y_{j2},Y_{i2},W_j\right) \mid \underline{X}_j\right]$ *and* $\Sigma_2^{(1)} = Var\left[h_1^{(1)}(\underline{X})\right]$. *Then*

$$\sqrt{N}\varphi_n^{(1)}\left(\underline{\beta}^0\right) \sim N\left(\underline{0},\Sigma_2^{(1)}\right).$$

Proof of Lemma 4.9. The proof is given in the appendix. ■

Note that $\varphi_n^{(2)}\left(\underline{\beta}^0\right)$ is a V-statistic. Let

$$\begin{aligned} h^{(2)}(y_{i2}, y_{l2}) =& e_2^2(y_{i2})\,\underline{k}(y_{i2})\left(f_2(y_{i2}) - K_{h_2}(y_{i2} - y_{l2})\right) \\ &+ e_2^2(y_{l2})\,\underline{k}(y_{l2})\left(f_2(y_{l2}) - K_{h_2}(y_{l2} - y_{i2})\right), \end{aligned}$$

where y_{i2} and y_{j2} represent independent refreshment samples. Then

$$\varphi_n^{(2)}\left(\underline{\beta}^0\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{l=1}^{n} h^{(2)}(y_{i2}, y_{l2}). \tag{4.6}$$

Lemma 4.10. *Define* $h_1^{(2)}(Y_{i2}) = E\left(h^{(2)}(Y_{i2}, Y_{l2}) \mid Y_{i2}\right)$ *and* $\Sigma_2^{(2)} = Var\left[h_1^{(2)}(Y)\right]$. *Then* $E\left[h^{(2)}(Y_{i2}, Y_{l2})\right] \approx \underline{0}$ *and*

$$\sqrt{n}\varphi_n^{(2)}\left(\underline{\beta}^0\right) \sim N\left(\underline{0}, 4\Sigma_2^{(2)}\right).$$

Proof of Lemma 4.10. The proof is given in the appendix. ■

Lemma 4.11. *The probablity limit of the second derivative of* $M_n\left(\underline{\beta}\right)$ *is*

$$E\left[\frac{\partial^2}{\partial \underline{\beta}^2} M_n\left(\underline{\beta}^0\right)\right] \approx 2E\left[e_2^2(Y_2)\,\underline{k}(Y_2)\,\underline{k}(Y_2)^T\right].$$

Proof of Lemma 4.11. The proof is given in the appendix. ■

Theorem 4.3. *Let* $N = rn$, *where r is the ratio between N and n, and define*

$$\Sigma_{cov} = Cov\left[h_1(\underline{X}), h_1^{(1)}(\underline{X})\right].$$

Then we have the following asymptotic property for $\widehat{\underline{\beta}}$

$$\sqrt{N}\left(\widehat{\underline{\beta}}-\widehat{\underline{\beta}}_0\right)\sim N\left(\underline{0},\left(V^{-1}\right)\Sigma\left(V^{-1}\right)^T\right),$$

where $\Sigma=4\Sigma_1+\Sigma_2^{(1)}+4r\Sigma_2^{(2)}+4\Sigma_{cov}$ *and* $V=E\left[\frac{\partial^2}{\partial\underline{\beta}^2}M_N\left(\underline{\beta}^0\right)\right]+E\left[\frac{\partial^2}{\partial\underline{\beta}^2}M_n\left(\underline{\beta}^0\right)\right]$.

Proof of Theorem 4.3. The proof is given in the appendix. ■

4.4 Hypothesis Testing

Our primary goal of using the asymptotic theory, which was developed in the previous section, is to establish hypothesis tests for missing mechanisms. Testing missing mechanisms can be accomplished by testing attrition parameters β_1 and β_2 in the additive non-ignorable model as follows:

$$H_0:\ \text{Data are MCAR} \iff H_0:\ \beta_1=0 \text{ and } \beta_2=0.$$

$$H_0:\ \text{Data are MAR} \iff H_0:\ \beta_2=0.$$

$$H_0:\ \text{Data are MCAR} \iff H_0:\ \beta_2\neq 0.$$

A Wald-type test statistic can be constructed given the asymptotic normality of the semi-parametric estimators:

$$Z=\frac{\widehat{\beta}_i-\beta_{i0}}{SE_{\widehat{\beta}_i}}=\frac{\widehat{\beta}_i}{SE_{\widehat{\beta}_i}}\sim N(0,1),\quad \text{for } i=1,2, \tag{4.7}$$

where $\widehat{\beta}_i$'s are semi-parametric estimators and $SE_{\widehat{\beta}_i}$'s are corresponding standard errors. The $100(1-\alpha)\%$ confidence interval can also be defined as

$$CI: \quad \widehat{\beta}_i \pm z_{1-\alpha/2} SE_{\widehat{\beta}_i}, \quad \text{for } i = 1,2,$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)^{\text{th}}$ quantile of the standard normal distribution. The asymptotic theory of the semi-parametric estimators gives the asymptotic formula for computing the standard errors. However, this computation requires both the true population density functions and the true attrition parameters, which makes the asymptotic formula an unrealistic approach to obtaining standard errors for the construction of both test statistics and confidence intervals in real data applications. We propose to use the bootstrap technique to numerically approximate standard errors. Bootstrapping treats the empirical distribution of the observed data as the population distribution. Then bootstrapped samples from this empirical distribution are treated as if they were repeated samples from the population. The sampling distribution based on bootstrapped samples is taken as an approximation to the true sampling distribution, and its standard deviation SE_{boot} is the bootstrap standard error of the semi-parametric estimator. The accuracy of this bootstrap SE can be assessed numerically by comparing to the empirical SE based on the simulation. In particular, the comparison is made through power functions of hypothesis tests based on the test statistic defined in (4.7). Let the level α test function

$\phi(Y_1, Y_2, W)$ be defined as

$$\phi(Y_1, Y_2, W) = \begin{cases} 1, & \text{reject null if} \quad |Z| \geq z_{1-\alpha/2} \\ 0, & \text{fail to reject} \quad \text{otherwise} \end{cases}. \tag{4.8}$$

And let $Q_1(\beta_1)$ and $Q_2(\beta_2)$ be the power functions for β_1 and β_2 respectively, defined as

$$\begin{aligned} Q_i(\beta_i) =& P(\text{Reject the null} \mid \beta_i \text{ is the true parameter}) \\ =& P_{\beta_i}(\phi(Y_1, Y_2, W) = 1), \quad \text{for } i = 1, 2. \end{aligned} \tag{4.9}$$

The test statistic in (4.7) depends on the choice of the SE, so the power function in (4.9) does also. For each attrition parameter β_i, power functions based on different SE's can be formed. The accuracy of the bootstrap SE is assessed by seeing how well the bootstrap-based power function reproduces the simulation-based power function.

In this chapter, we introduced two new methods in analyzing MNAR continuous response data with the help of a refreshment sample. The focus of these methods is to estimate attrition parameters and reveal missing mechanisms. The first method, the full-likelihood parametric method, serves as a performance benchmark to which the second method, the kernel density based semi-parametric method, will be compared. The asymptotic theory for the semi-parametric estimators was developed to investigate their large sample behaviors. Finally, hypothesis tests were established to fulfill the goal of making inferences about the missing mechanism.

5 Numerical Results

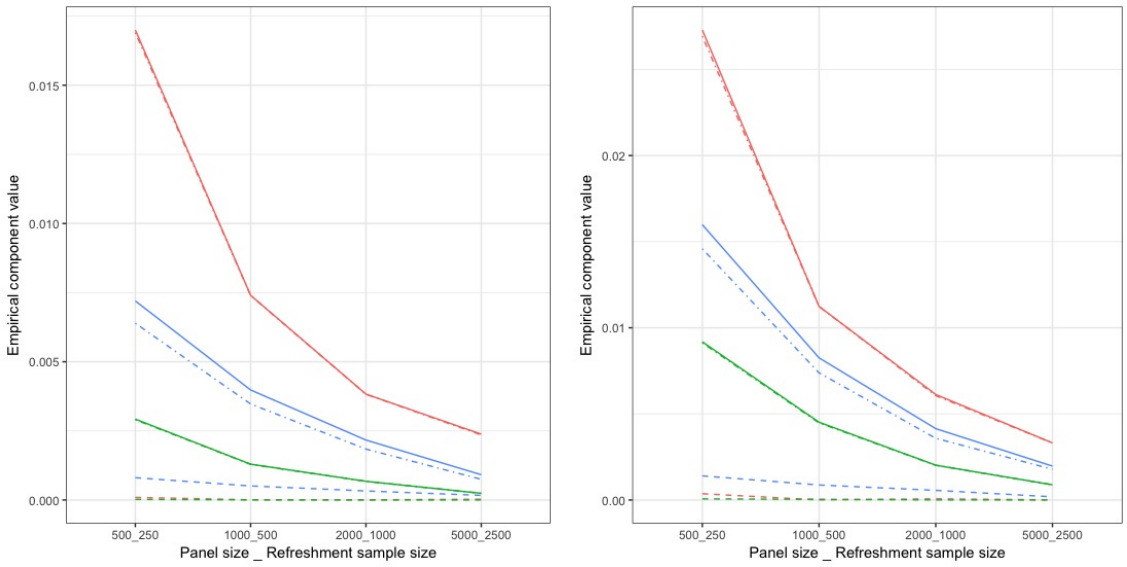
In this chapter, we demonstrate the numerical performance of the proposed methods. We compare finite-sample performance of different methods in section 5.1. In section 5.2, we consider some insights from the asymptotic variance formula for the proposed semi-parametric estimator and use simulations to validate properties of the asymptotic variance. In section 5.3, we illustrate the application of bootstrapping to the previously described hypothesis tests. Finally, we show an application of the semi-parametric method to real data from the Netherlands Mobility Panel in section 5.4.

5.1 Finite-Sample Performance

We use simulation studies to compare finite-sample performance of the parametric and semi-parametric methods proposed in chapter 4. We also compare the proposed methods with Bhattacharya's conditional moment restriction method. The numerical performance of different estimators is evaluated in terms of mean square errors. A total of 1000 data sets is generated, and different methods are applied to yield 1000 estimates from which the empirical squared bias, variance and mean square error are calculated. This process is repeated with different panel sizes and refreshment sample sizes. Our focus is on the performance of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ since they are the parameters that determine the underlying missing mechanism.

5.1.1 Bivariate Normal Population

We generate data from a bivariate normal distribution with mean $\underline{0}$, marginal variances of 10 and correlation coefficient of 0.5. The true attrition model has the form of a logistic regression with true attrition parameters of $\beta_0 = 0$, $\beta_1 = 0.3$, $\beta_2 = 0.4$. Three methods are applied to obtain estimates of attrition parameters. Figure 5.1 compares the finite-sample performances in terms of empirical squared bias, variance and MSE for $\widehat{\beta}_1$ and $\widehat{\beta}_2$ respectively. On the x-axis are different panel size and refreshment sample size combinations with both sample sizes increasing along the x-axis. The parametric method is plotted in green, the semi-parametric method is in blue and Bhattacharya's conditional moment restriction method is in red (denoted as CMR). In addition, for all three methods, dash, dotted dash and solid lines stand for empirical squared bias, variance and MSE respectively. Figure 5.1 clearly shows that MSEs of both parametric and semi-parametric methods decrease as the sample sizes increase, which supports our asymptotic results. In addition, we see that both the parametric and the semi-parameter methods perform better than Bhattacharya's method, with the CMR method having the largest MSE for all sample size combinations.



(a) Finite-sample performance of $\widehat{\beta}_1$

(b) Finite-sample performance of $\widehat{\beta}_2$

Figure 5.1: Comparison of finite-sample performance with normal responses. (True parameters: $\beta_0 = 0$, $\beta_1 = 0.3$, $\beta_2 = 0.4$). Parametric method is plotted in green, semi-parametric method is in blue and Bhattacharya's conditional moment restriction method is in red. For all three methods, dash, dotted dash and solid lines stand for empirical squared bias, variance and MSE, respectively.

Tables 5.1 and 5.2 give results for a panel size of 5000 and a refreshment size of 2500. Most of the MSE is due to the variance of the estimators. The parametric estimator of β_1 has about one third the variance of the semi-parametric estimator, which in turn has about one third the variance of the CMR estimator. Due to the attrition in the second wave, we do not have as much information to estimate β_2 as we do to estimate β_1. This might be one reason why the variances of $\widehat{\beta}_2$ are larger for all three methods. The parametric estimator of β_2 has about half the variance of the semi-parametric estimator, which in turn has about half the variance of the CMR estimator. Therefore, in

the following we will focus on the comparison between parametric and semi-parametric methods to get a better understanding of their finite-sample performance.

	Squared Bias (10^{-5})	Variance (10^{-5})	MSE (10^{-5})
CMR	2.89	235.25	238.15
Semi-parametric	16.68	74.72	91.41
Parametric	0.06	24.08	24.14

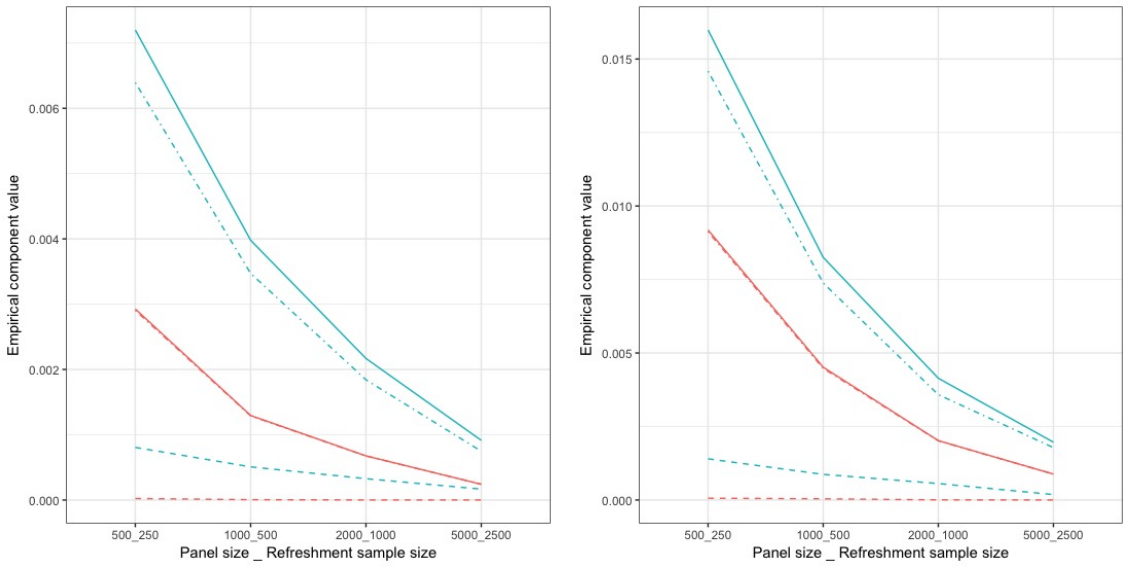
Table 5.1: Empirical squared bias, variance and MSE of $\widehat{\beta}_1$ for three different methods with panel size of 5000, and refreshment sample size of 2500.

	Squared Bias (10^{-5})	Variance (10^{-5})	MSE (10^{-5})
CMR	0.82	330.59	331.40
Semi-parametric	18.36	178.23	196.68
Parametric	0.03	88.12	88.15

Table 5.2: Empirical squared bias, variance and MSE of $\widehat{\beta}_2$ for three different methods with panel size of 5000, and refreshment sample size of 2500.

In Figure 5.2, the parametric method is plotted in red and the semi-parametric method is in cyan. The parametric method under the normality assumption performs better than the semi-parametric method. This is expected since the parametric method makes full use of the density functions to construct the likelihood. The parametric method is most efficient in this scenario and can be treated as the benchmark. The semi-parametric estimators give reasonable performance with both squared bias and variance approaching 0 as sample size increases. In addition, as sample size increases, the difference between

the two methods decreases.



(a) Finite-sample performance of $\widehat{\beta}_1$

(b) Finite-sample performance of $\widehat{\beta}_2$

Figure 5.2: Comparison of finite-sample performance with normal responses (True parameters: $\beta_0 = 0$, $\beta_1 = 0.3$, $\beta_2 = 0.4$). Parametric method is plotted in red and semi-parametric method is in cyan. For both methods, dash, dotted dash and solid lines stand for empirical squared bias, variance and MSE, respectively.

5.1.2 Gamma-t population

In real data applications, we often encounter non-normal data. It is of interest to compare the performance of the parametric and semi-parametric methods when the distribution is misspecified.

Copulas are useful for creating a non-normal joint density with given marginals and controlled correlation (Yan et al., 2007). For the following simulations we created a

Gamma-t joint distribution with the first wave following a $Gamma(3,2)$ marginal distribution and the second wave following a t distribution with degree of freedom of 6. In order to make this non-normal joint distribution comparable to the previous bivariate normal case, the *Gamma* distribution is shifted to center at 0, and the t distribution is scaled by 3. A correlation coefficient of 0.5 is generated by the Copula method. As a result, the non-normal joint distribution centers at zero, and the *Gamma* marginal has variance of 12 while the t distribution has variance of 13.5. Compared with the bivariate normal distribution, the non-normal distribution has the same means (0) and slightly larger marginal variances. Figure 5.3 plots this non-normal *Gamma-t* distribution, where Y_1 has the *Gamma* marginal distribution and Y_2 has the t marginal distribution. The sample correlation coefficient, in this particular sample, is 0.5.

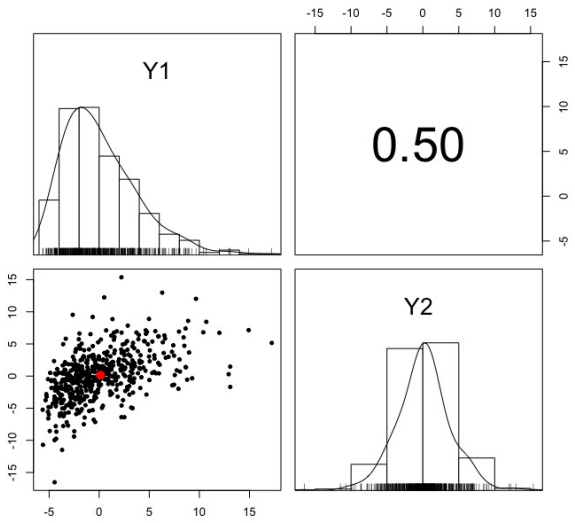


Figure 5.3: Non-normal data generated by the Copula method and modeled with $Y_1 \sim Gamma(3,2)$ centering at 0 and $\frac{1}{3}Y_2 \sim t_6$. The sample correlation coefficient is 0.5.

Figure 5.4 plots the pairs (y_1, y_2) for the complete set only (left), and the full panel assuming no attrition (right). The colors of the points represent the probability of Y_2 being observed. The lighter the color, the higher the probability. The red line consists of points (y_1, y_2) that satisfy $\beta_0 + \beta_1 y_1 + \beta_2 y_2 = 0$ in the additive non-ignorable attrition model (3.2). As a result, data points on this line have 50% probability of Y_2 being observed.

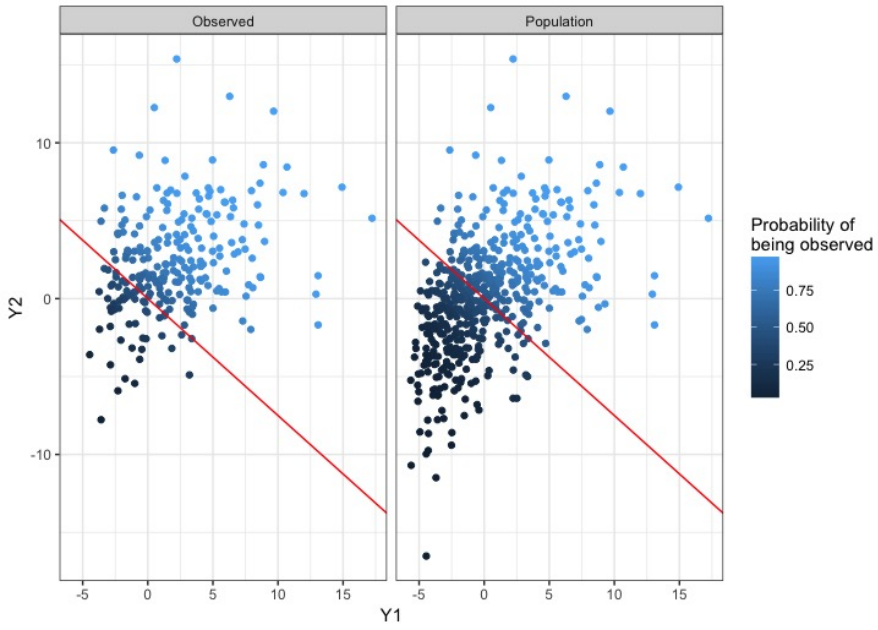
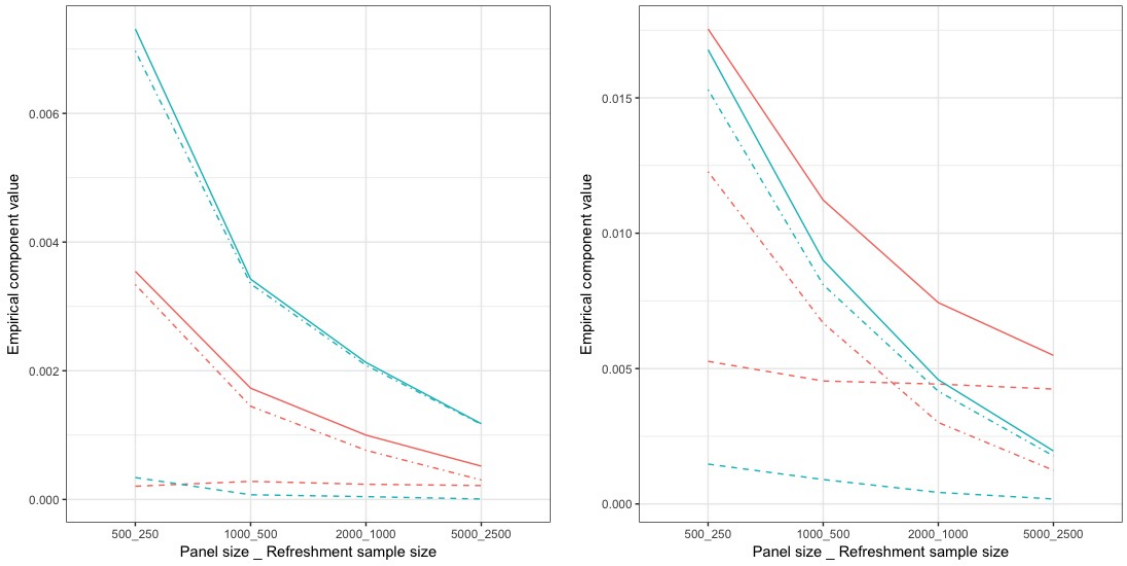


Figure 5.4: Non-normal population with attrition. Left: Data in the complete set only, with missing data deleted. Right: Full panel data assuming no attrition.

For the performance of $\widehat{\beta}_1$, Figure 5.5 shows that the parametric method still has better performance in terms of MSE. However, as sample size increases, the parametric method has non-decreasing bias while the semi-parametric method has decreasing bias. The variance of the semi-parametric estimator $\widehat{\beta}_1$, though, is still larger than the

parametric one. Overall, the departure from normality does not penalize the numerical performance of $\widehat{\beta}_1$ much in terms of MSE, but it does introduce a small bias into the parametric estimator. Table 5.3 shows numerical results comparing between semi-parametric and parametric estimators for panel size of 5000 and refreshment sample size of 2500.



(a) Finite-sample performance of $\widehat{\beta}_1$

(b) Finite-sample performance of $\widehat{\beta}_2$

Figure 5.5: Comparison of finite-sample performance with gamma-t responses (True parameters: $\beta_0 = 0$, $\beta_1 = 0.3$, $\beta_2 = 0.4$). Parametric method is plotted in red and semi-parametric method is in cyan. For both methods, dash, dotted dash and solid lines stand for empirical squared bias, variance and MSE, respectively.

	Squared Bias (10^{-5})	Variance (10^{-5})	MSE (10^{-5})
Semi-parametric	0.74	116.92	117.66
Parametric	21.61	30.26	51.83

Table 5.3: Non-normal Gamma-t population scenario. Empirical squared bias, variance and MSE of $\widehat{\beta}_1$ for both parametric and semi-parametric methods with panel size of 5000, and refreshment sample size of 2500.

For $\widehat{\beta}_2$, the semi-parametric method performs much better than the parametric method. The departure from normality introduces a much larger bias to the parametric $\widehat{\beta}_2$ than it does for $\widehat{\beta}_1$. Table 5.4 compares results between semi-parametric and parametric estimators for panel size of 5000 and refreshment sample size of 2500. In this large sample case, the much larger bias of the parametric estimator overwhelms its advantage in the variance, which results in a larger empirical MSE compared to the semi-parametric estimator. This demonstrates that the semi-parametric method is quite robust against non-normality in the response distribution.

	Squared Bias (10^{-5})	Variance (10^{-5})	MSE (10^{-5})
Semi-parametric	18.38	177.55	195.94
Parametric	424.40	124.51	548.91

Table 5.4: Non-normal Gamma-t population scenario. Empirical squared bias, variance and MSE of $\widehat{\beta}_2$ for both parametric and semi-parametric methods with panel size of 5000, and refreshment sample size of 2500.

In summary, we have made comparisons of the finite-sample performance among different methods. Our proposed parametric and semi-parametric methods have better

performance than CMR. When the joint distribution is correctly specified, the parametric method performs the best. When the distribution is misspecified, however, there will be bias in the parametric estimator, while the semi-parametric estimator, being free of distributional assumptions, gives consistent performance in the presence of non-normal populations. In the next section, we will take a closer look at the asymptotic variance of the semi-parametric estimator. In particular, we want to investigate effects of different population parameters and data transformation on the asymptotic variance of the semi-parametric estimator.

5.2 Understanding of Asymptotic Variance of Semi-parametric Estimator

The asymptotic variance formula in Theorem 4.3 helps us to better understand the large sample performance of the proposed semi-parametric method under different scenarios. In particular, it provides guidance on how different population parameters affect the performance of $\widehat{\beta}$. In this subsection, several scenarios have been considered, based on which theoretical results are calculated through asymptotic variance formula along with the verification through simulations.

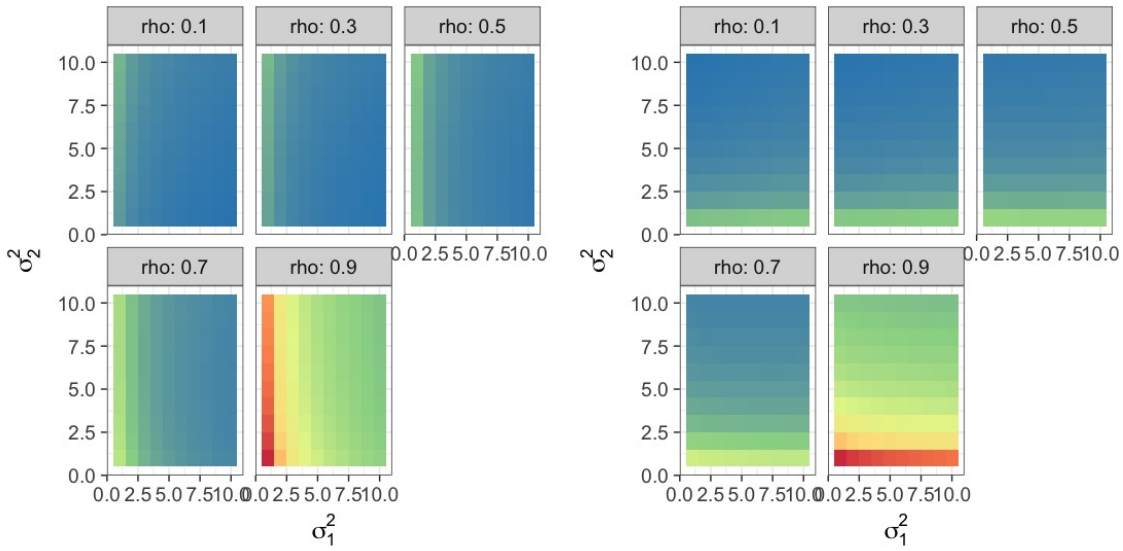
5.2.1 Effect of Marginal Variation

First we investigate the effect of the variances of Y_1 and Y_2 and the correlation coefficient ρ on the variability of $\underline{\widehat{\beta}}$ with both waves' marginal means, panel size and refreshment sample size fixed. We consider five correlation coefficient settings: $\rho =$

$0.1, 0.2, ..., 0.5$. For each setting the standard errors of $\underline{\widehat{\beta}}$ are calculated through the asymptotic formula at different combinations of the two waves' marginal variances, with both marginal variances ranging from 0 to 10.

Figure 5.6 shows the effect of the marginal variances σ_1^2 and σ_2^2 on the standard errors of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ computed through the asymptotic formula. Both plots show the trend that the asymptotic standard errors increase as the correlation coefficient increases across different marginal variance combinations. The left plot shows that the asymptotic standard error of $\underline{\widehat{\beta}}_1$ decreases as the first wave marginal variance σ_1^2 increases. The right plot shows a similar relationship between the asymptotic standard error of $\underline{\widehat{\beta}}_2$ and the second wave marginal variance σ_2^2. Less correlation and more variability in the marginal distributions give more stable estimates of attrition parameters.

Table 5.5 compares standard errors between simulation results and the asymptotic formula for three settings. The simulation results verify the asymptotic findings above except that the simulation produces systematically smaller standard errors than the asymptotic formula. This disagreement is the result of many Taylor expansion approximations involved in the development of the asymptotic formula. Therefore, the standard errors computed from the asymptotic formula can be considered as conservative estimates of the true standard errors.



(a) $SE_{\widehat{\beta}_1}$ (b) $SE_{\widehat{\beta}_2}$

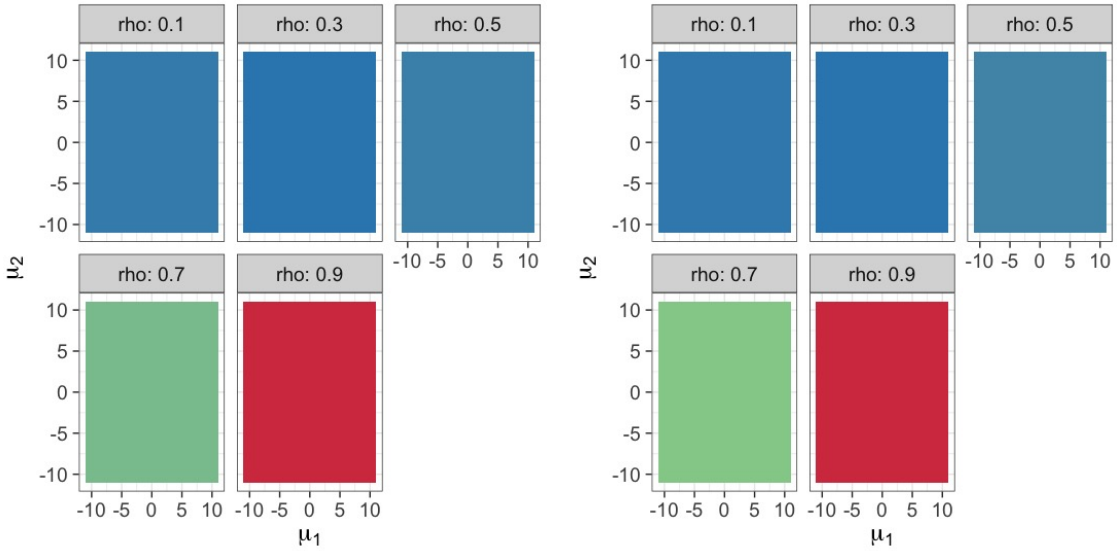
Figure 5.6: The effect of marginal variances σ_1^2 and σ_2^2 on standard errors of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ computed from the asymptotic formula. The color represents the value of standard error. The value is larger in the red direction and smaller in the blue direction. Plots are separated by the levels of correlation coefficient ρ. The population is bivariate normal with both marginal means of 0. The panel size is 5000 and refreshment sample size is 2500. True values of attrition parameters β_1 and β_2 are 0.3 and 0.4 respectively.

		Asymptotic Formula		Simulation	
σ_1^2	σ_2^2	$SE_{\widehat{\beta}_1}$	$SE_{\widehat{\beta}_2}$	$SE_{\widehat{\beta}_1}$	$SE_{\widehat{\beta}_2}$
1	1	0.105	0.161	0.081	0.131
5	5	0.043	0.066	0.036	0.055
10	10	0.035	0.048	0.028	0.041

Table 5.5: The effect of marginal variances σ_1^2 and σ_2^2 on standard errors of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ computed from both the asymptotic formula and simulation. The population is bivariate normal with both marginal means of 0 and correlation coefficient of 0.5. The panel size is 5000 and refreshment sample size is 2500. True values of attrition parameters β_1 and β_2 are 0.3 and 0.4 respectively.

5.2.2 The Effect of Marginal Mean

Now we want to verify our postulation that the standard errors of $\underline{\widehat{\beta}}_1$ and $\underline{\widehat{\beta}}_2$ are invariant to the data location. Standard errors are calculated again through the asymptotic formula under five correlation settings at different combinations of two marginal means with both means ranging from -10 to 10. Both marginal variances are set to be 10 and all other parameters are fixed as previously. Figure 5.7 confirms our postulation and shows again that less correlation stabilizes the estimates. Table 5.6 gives numerical results for five particular population centers, and the same conclusion can be made. This invariance property implies that the same testing results for $\underline{\widehat{\beta}}_1$ and $\underline{\widehat{\beta}}_2$ can be obtained after centering the data to zero mean.



(a) $SE_{\widehat{\beta}_1}$ (b) $SE_{\widehat{\beta}_2}$

Figure 5.7: The effect of marginal means μ_1 and μ_2 on standard errors of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ computed from the asymptotic formula. The color represents the value of standard error. The value is larger in the red direction and smaller in the blue direction. Plots are separated by the levels of correlation coefficient ρ. The population is bivariate normal with both marginal variances of 10. The panel size is 5000 and refreshment sample size is 2500. True values of attrition parameters β_1 and β_2 are 0.3 and 0.4 respectively.

		Asymptotic Formula		Simulation	
μ_1	μ_2	$SE_{\widehat{\beta}_1}$	$SE_{\widehat{\beta}_2}$	$SE_{\widehat{\beta}_1}$	$SE_{\widehat{\beta}_2}$
3	4	0.035	0.048	0.028	0.041
-5	2	0.035	0.048	0.028	0.041
-2	-4	0.035	0.048	0.028	0.041
3	-1	0.035	0.048	0.028	0.041
0	0	0.035	0.048	0.028	0.041

Table 5.6: The effect of marginal means μ_1 and μ_2 on standard errors of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ computed from both the asymptotic formula and simulation. The population is bivariate normal with both marginal variances of 10 and correlation coefficient of 0.5. The panel size is 5000 and refreshment sample size is 2500. True values of attrition parameters β_1 and β_2 are 0.3 and 0.4 respectively.

5.2.3 Effect of The Rotation in Missing Direction

First, we introduce notation and concepts used in this subsection. The attrition model consists of the logistic function $P(W = 1 \mid y_1, y_2) = logistic(\beta_0 + \beta_1 y_1 + \beta_2 y_2)$, which can be thought of as a surface over the (y_1, y_2) plane. This surface gives the probability of Y_2 being observed for every data point, and the missingness probability is constant along any line where $\beta_0 + \beta_1 y_1 + \beta_2 y_2$ is constant. For example, in Figure 5.8, $\beta_0 + \beta_1 y_1 + \beta_2 y_2 = 0$ corresponds to a line of data points that have 50% of having Y_2 missing, which we refer to as the reference line. For all simulations, we set this reference line to go through the population mean, which results in about 50% of the Y_2 data missing if the population has a symmetric joint distribution such as the bivariate normal

distribution we used. The vector $(\beta_1, \beta_2)^T$ gives a direction that is perpendicular to the reference line. We refer to this vector as the normal vector in the following context. The probability of Y_2 being observed increases in the direction of the normal vector. The length of the normal vector, $\sqrt{\beta_1^2 + \beta_2^2}$, indicates how fast the probability of Y_2 being observed increases. The surface of the logistic function increases gradually along the direction of the normal vector if the length of the vector is relatively small, and it increases dramatically otherwise. For a very large length, we will find that almost all Y_2 data are observed on one side of the reference line, and hardly any Y_2 data are observed on the other side.

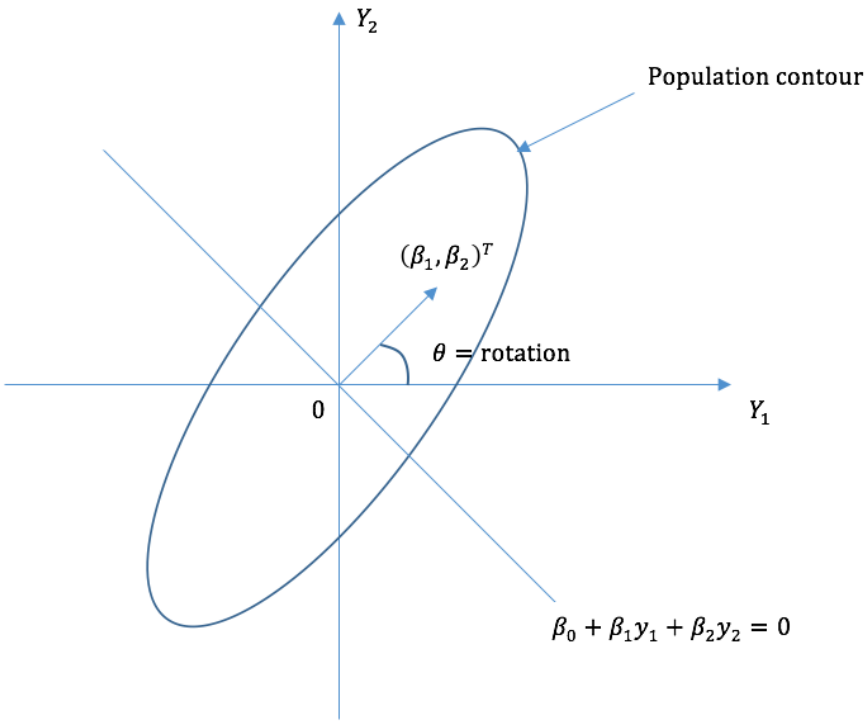
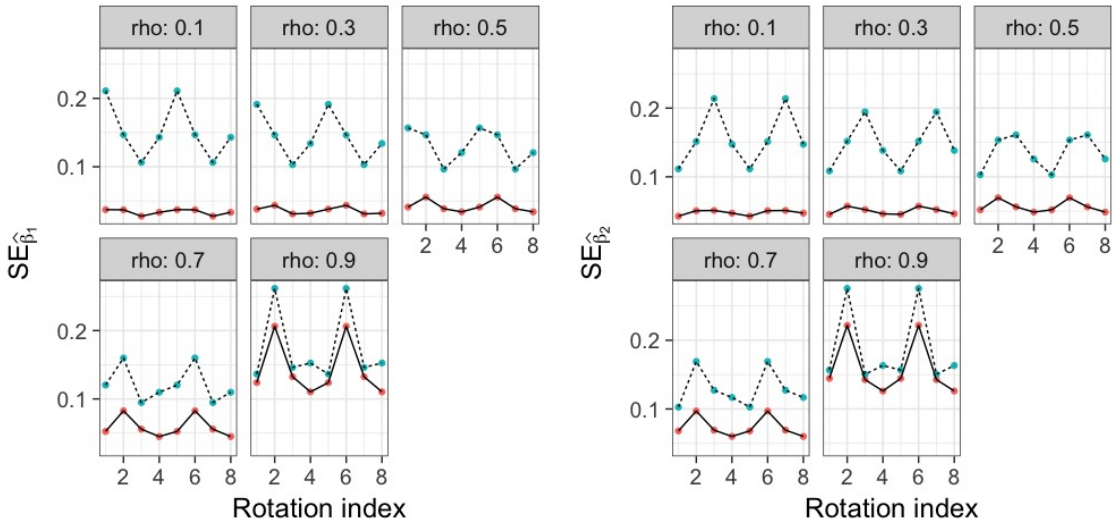


Figure 5.8: The definitions of the reference line, normal vector and rotation.

In Figure 5.9, we consider two different lengths of normal vectors, 0.5 and 1. A length of 0.5 results in a gradual missingness pattern and the probability of being observed is bounded from 0 and 1 for almost all data points. A length of 1 shows a dramatic missing pattern where part of data is almost always observed and the other part is missing most of times. The X-axis represents different rotation θ values. There are 8 equally spaced rotations with the angular coordinate turning from $\frac{3}{2}\pi$ to $\frac{5}{4}\pi$ counterclockwise.



(a) $SE_{\widehat{\beta}_1}$ (b) $SE_{\widehat{\beta}_2}$

Figure 5.9: The effect of the normal vector rotation on standard errors of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ computed from the asymptotic formula. The y-axis shows the value of the standard error. The solid and dashed lines correspond to normal vector length of 0.5 and 1 respectively. Plots are separated by the levels of correlation coefficient ρ. The population is bivariate normal with both marginal means of 0 and variances of 10. The panel size is 5000 and refreshment sample size is 2500.

Figure 5.9 shows that the semi-parametric estimator has larger variance when data are dramatically missing (the length of normal vector = 1). When data are gradually

missing, less correlation in the data results in less variability of the estimator. Furthermore, Figure 5.9 reveals that minimum asymptotic standard errors are achieved when the normal vector parallels the major axis of the population contour. In our bivariate normal case, the joint distribution has a positive correlation coefficient and both marginal variances are equal, which results in a population contour with a 45 degree major axis. Two rotation scenarios give the parallel relationship between the normal vector and this 45 degree major axis, namely rotation scenarios of 4 and 8 (i.e. $\frac{1}{4}\pi$ and $\frac{5}{4}\pi$ respectively). When the normal vector is perpendicular to the major axis, asymptotic standard errors are at the maximum (i.e. $\frac{7}{4}\pi$ and $\frac{3}{4}\pi$ corresponding to rotation scenarios 2 and 6). These results are verified in the numerical results for four particular cases given in Table 5.7. Again we see that the asymptotic formula gives conservative estimates for standard errors.

		Polar Coordinate		Asymptotic Formula		Simulation	
β_1	β_2	radial	angular	$SE_{\widehat{\beta}_1}$	$SE_{\widehat{\beta}_2}$	$SE_{\widehat{\beta}_1}$	$SE_{\widehat{\beta}_2}$
0.354	0.354	0.5	$\pi/4$	0.0340	0.049	0.029	0.041
0.354	-0.354	0.5	$7\pi/4$	0.056	0.070	0.041	0.052
0.707	0.707	1	$\pi/4$	0.121	0.126	0.070	0.074
0.707	-0.707	1	$7\pi/4$	0.147	0.154	0.070	0.076

Table 5.7: The effect of the normal vector rotation on standard errors of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ computed from both the asymptotic formula and simulation. The population is bivariate normal with both marginal means of 0, variances of 10 and correlation coefficient of 0.5. The panel size is 5000 and refreshment sample size is 2500.

5.2.4 Effect of Sample Size

Table 5.8 shows the effect of sample sizes on the standard error of the semi-parametric estimators. As expected, the standard errors decrease as both sample sizes increase. In addition, if the panel size is fixed, standard errors decrease with increasing refreshment sample size.

Sample Size		Asymptotic Formula		Simulation	
Panel	Refreshment	$SE_{\widehat{\beta}_1}$	$SE_{\widehat{\beta}_2}$	$SE_{\widehat{\beta}_1}$	$SE_{\widehat{\beta}_2}$
500	250	0.111	0.152	0.079	0.118
1000	500	0.078	0.108	0.061	0.083
2000	1000	0.055	0.076	0.043	0.062
5000	1000	0.036	0.060	0.030	0.051
5000	2500	0.035	0.048	0.028	0.041
5000	5000	0.034	0.043	0.028	0.035

Table 5.8: The effect of the sample size on standard errors of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ computed from both the asymptotic formula and simulation. The population is bivariate normal with both marginal means of 0, variances of 10 and correlation coefficient of 0.5. The true attrition parameters β_1 and β_2 are 0.3 and 0.4 respectively.

5.2.5 Effect of Transformation

It is common to use transformations on the data set when performing data analysis. We are interested in the effect that data transformation has on the asymptotic properties of the estimator $\underline{\widehat{\beta}}$. In the following, we choose to examine how transformation changes

the objective function. We will focus our attention on a common linear transformation, namely shifting and scaling. Let's assume that

$$y_1^* = \frac{1}{s}(y_1 - c_1),$$
$$y_2^* = \frac{1}{s}(y_2 - c_2),$$

where $(y_1, y_2)^T$ is original data point, $(y_1^*, y_2^*)^T$ is the transformed data point, s is the common scaling factor for both responses, and c_1 and c_2 are the respective shifting factors. Furthermore, let $\widehat{\underline{\beta}}^*$ be the minimizer of the objective function $M_{N,n}^*(\underline{\beta})$ for the transformed data. Let $\underline{\beta}^{0*}$ be the unique minimizer of $E\left[f_1^*(Y_1^* \mid \underline{\beta}) - f_1^*(Y_1^*)\right]^2 + E\left[f_2^*(Y_2^* \mid \underline{\beta}) - f_2^*(Y_2^*)\right]^2$. Here the asterisk notation indicates that corresponding terms are based on transformed data. By Theorem 4.2, $\widehat{\underline{\beta}}^*$ is consistent for $\underline{\beta}^{0*}$. Given that $\widehat{\underline{\beta}}$ converges in probability to $\underline{\beta}^0$, it is sufficient to show the relationship between $\widehat{\underline{\beta}}^*$ and $\widehat{\underline{\beta}}$ by examining the relationship between $\underline{\beta}^{0*}$ and $\underline{\beta}^0$. Then one has

$$\begin{aligned}
\underline{\beta}^{0*} &= \min_{\underline{\beta}^* \in \mathbb{R}^3} \left\{ E\left[f_1^*(Y_1^* \mid \underline{\beta}^*) - f_1^*(Y_1^*)\right]^2 + E\left[f_2^*(Y_2^* \mid \underline{\beta}^*) - f_2^*(Y_2^*)\right]^2 \right\} \\
&= \min_{\underline{\beta}^* \in \mathbb{R}^3} \left\{ E\left[\int \frac{f^*(Y_1^*, y_2^* \mid W=1)P(W=1)}{1/(1+\exp(-\beta_0^* - \beta_1^* Y_1^* - \beta_2^* y_2^*))} dy_2^* - f_1^*(Y_1^*)\right]^2 \right. \\
&\quad \left. + E\left[\int \frac{f^*(y_1^*, Y_2^* \mid W=1)P(W=1)}{1/(1+\exp(-\beta_0^* - \beta_1^* y_1^* - \beta_2^* Y_2^*))} dy_1^* - f_2^*(Y_2^*)\right]^2 \right\}.
\end{aligned}$$

By change of variables, one can show that

$$\begin{aligned}
\underline{\beta}^{0*} = &\min_{\underline{\beta}^* \in \mathbb{R}^3} \left\{ E\left[s \int \frac{f(Y_1, y_2 \mid W=1)P(W=1)}{1/(1+\exp(-\beta_0^* + \frac{\beta_1^* c_1}{s} + \frac{\beta_2^* c_2}{s} - \frac{\beta_1^*}{s} Y_1 - \frac{\beta_2^*}{s} y_2))} dy_2 - s f_1(Y_1) \right]^2 \right. \\
&\left. + E\left[s \int \frac{f(y_1, Y_2 \mid W=1)P(W=1)}{1/(1+\exp(-\beta_0^* + \frac{\beta_1^* c_1}{s} + \frac{\beta_2^* c_2}{s} - \frac{\beta_1^*}{s} y_1 - \frac{\beta_2^*}{s} Y_2))} dy_1 - s f_2(Y_2) \right]^2 \right\} \\
= &\min_{\underline{\beta} \in \mathbb{R}^3} s \left\{ E\left[\int \frac{f(Y_1, y_2 \mid W=1)P(W=1)}{1/(1+\exp(-\beta_0 - \beta_1 Y_1 - \beta_2 y_2))} dy_2 - f_1(Y_1) \right]^2 \right. \\
&\left. + E\left[\int \frac{f(y_1, Y_2 \mid W=1)P(W=1)}{1/(1+\exp(-\beta_0 - \beta_1 y_1 - \beta_2 Y_2))} dy_1 - f_2(Y_2) \right]^2 \right\} \\
= &\underline{\beta}^0,
\end{aligned}$$

where $\beta_0 = \beta_0^* - \frac{\beta_1^* c_1}{s} - \frac{\beta_2^* c_2}{s}$, $\beta_1 = \frac{\beta_1^*}{s}$ and $\beta_2 = \frac{\beta_2^*}{s}$ is a mapping from parameter space $\mathbb{R}^3$ to $\mathbb{R}^3$. As a result we should have the same relationship between $\widehat{\underline{\beta}}^*$ and $\widehat{\underline{\beta}}$ as follows:

$$\begin{aligned}
\widehat{\beta}_0^* &= \widehat{\beta}_0 + \widehat{\beta}_1 c_1 + \widehat{\beta}_2 c_2, \\
\widehat{\beta}_1^* &= s\widehat{\beta}_1, \\
\widehat{\beta}_2^* &= s\widehat{\beta}_2.
\end{aligned} \tag{5.1}$$

Therefore, centering the data will not affect the estimation of β_1 and β_2. The corresponding test procedure is invariant to centering as well. Scaling changes the point estimate, but it also alters the standard errors by the same amount. Thus, the corresponding test procedure is also invariant to scale transformation. In the following sections, we use simulation to verify the invariant property of hypothesis testing for β_1 and β_2 when a

linear transformation is applied to the responses.

5.2.5.1 Effect of Centering

We perform simulations where the original (Y_1, Y_2) population is centered at an arbitrarily chosen location, and the data are centered so that post-transformation, the population is centered at the origin. As before, we let the reference line (50% missingness probability) go through the population center. Let $\underline{\beta}$ be the original true attrition parameters and $\underline{\beta}^*$ be the post-centering true parameters. In Table 5.9, the original settings describe the original population, while the centered settings describe the population after centering. Under each setting, the asymptotic variances of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are calculated based on the asymptotic formula. We draw 1000 samples from the original population and estimate the attrition parameters using both non-centered and centered data. Standard errors of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are computed from the corresponding sampling distributions. Both asymptotic equation and simulation results show that centering data has no effect on the standard errors of $\underline{\beta}_1$ and $\underline{\beta}_2$. This invariance property means that the same statistical inferences about β_1 and β_2 will result whether the data are centered or not.

Original settings					Centered settings				
μ_1	μ_2	β_0	β_1	β_2	μ_1	μ_2	β_0^*	β_1^*	β_2^*
3	4	-2.5	0.3	0.4	0	0	0	0.3	0.4
	Asymptotic		Simulation			Asymptotic		Simulation	
$SE_{\widehat{\beta}_1}$	0.03495		0.02740		$SE_{\widehat{\beta}_1}$	0.03495		0.02740	
$SE_{\widehat{\beta}_2}$	0.04815		0.04220		$SE_{\widehat{\beta}_2}$	0.04815		0.04221	

Table 5.9: The effect of centering data on standard errors of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ computed from both the asymptotic formula and simulation. The population is bivariate normal with both marginal variances of 10 and correlation coefficient of 0.5. The panel size is 5000 and refreshment sample size is 2500.

5.2.5.2 Effect of Scaling

The goal of this subsection is to show with simulation how hypothesis testing for β_1 and β_2 is invariant to scaling of the response data. For each of three scaling factors, we generate 1000 samples from the bivariate normal distribution with both marginal means of 0, variances of 10 and correlation coefficient of 0.5. The panel size and refreshment sample size are again 5000 and 2500 respectively. The true attrition parameters β_1 and β_2 are 0.3 and 0.4 respectively. Before estimating parameters, we divide the response data by the scaling factor. Consequently, the true attrition parameters are changed after the scaling according to (5.1). For example, the attrition parameters β_1 and β_2 become 0.6 and 0.8 if the scaling factor is $s = 2$. A set of point estimates for the post-scaling attrition parameters β_1^* and β_2^* are obtained and the asymptotic variance is calculated from the asymptotic formula using parameters based on this post-scaling population.

The simulated standard errors are obtained by computing the standard deviation of the empirical sampling distribution. We generate 1000 samples from the original population, and the same scale constant is applied on these samples. Estimates of parameters are obtained to form the empirical sampling distributions whose standard deviations are calculated to represent simulated empirical SEs. Now the z-statistic can be constructed with the point estimates and two different SEs. Table 5.10 shows results for the three scaling factors. Again the simulated empirical SEs are systematically smaller than the asymptotic SEs. The z-statistics is invariant with scaling, showing that one can still obtain consistent test results (i.e. p-values) when the data are scaled.

Method	s	β_1^*	β_2^*	$SE_{\widehat{\beta}_1^*}$	$SE_{\widehat{\beta}_2^*}$	$z_{\beta_1^*}$	$z_{\beta_2^*}$
Asymptotic	2	0.6	0.8	0.062	0.096	10.38	6.69
	1	0.3	0.4	0.035	0.048	9.20	6.66
	0.25	0.075	0.1	0.009	0.012	9.20	6.65
Simulation	2	0.6	0.8	0.0548	0.084	11.74	7.59
	1	0.3	0.4	0.027	0.042	11.73	7.60
	0.25	0.075	0.1	0.007	0.011	11.75	7.61

Table 5.10: The effect of transforming data by a scaling factor (s) on standard errors of $\widehat{\beta}_1$ and $\widehat{\beta}_2$ computed from both the asymptotic formula and simulation. The population is bivariate normal with both marginal variances of 10 and correlation coefficient of 0.5. The panel size is 5000 and refreshment sample size is 2500. The prior-scaling true attrition parameters β_1 and β_2 are 0.3 and 0.4 respectively.

In this section, we used the asymptotic formula in Theorem 4.2 to get a better understanding of the semi-parametric estimator's performance. It helps us to understand

when the semi-parametric method will work and warns us when it might fail. In the next section, we will illustrate how bootstrapping can be used to obtain standard errors for the semi-parametric estimators in real data applications.

5.3 Bootstrapping in Applications

As shown in previous sections, the asymptotic formula gives a systematically larger standard errors than the empirical SEs obtained by simulation. This may be due to the fact that all higher order terms are ignored and a Taylor expansion is repeatedly used to simplify integrals involving kernel densities in the development of the asymptotic theory.

The empirical standard errors in the simulation give the best approximation of the truth, yet they are unavailable in practice. The asymptotic formula, on the other hand, not only gives higher standard errors, but also depends on true values of parameters and true population density functions. In section 4.4, we proposed the use of bootstrap technique as an alternative approach to numerically approximate the standard errors of semi-parametric estimators. We also proposed to compare different SEs by comparing their associated power functions. In the following, we first illustrate the process of using the simulation to approximate power functions based on different SEs. Then we discuss the results from a visual comparison of these power functions.

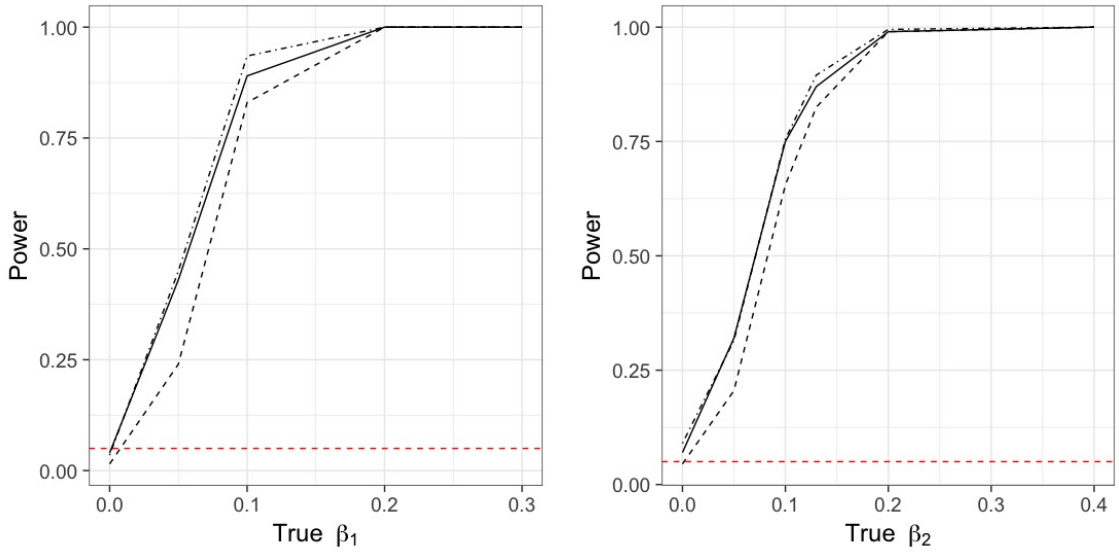
A total of 200 samples with panel size 5000 and refreshment size 2500 are drawn from a bivariate normal population with marginal means of 0, variances of 10 and correlation coefficient of 0.5. Attrition parameters are estimated by the semi-parametric

method in each sample. The SEs based on the asymptotic formula can be calculated by plugging in the true parameters and the population distribution. The empirical SEs are simulated separately as follows. First, another 1000 samples from the same population are drawn. Estimates from these 1000 samples give the empirical sampling distribution whose standard deviation gives the simulated empirical SE.

The bootstrap SE is generated as follows. For each of the original 200 samples, 500 bootstrap samples are created and attrition parameters are estimated through the semi-parametric method. Each bootstrap sample consists of two parts: a bootstrapped panel and a bootstrapped refreshment sample. The bootstrapped panel is a random sample of the same size as the original panel, drawn with replacement from the original panel data. The bootstrapped refreshment sample is a random sample of the same size as the original refreshment sample, drawn with replacement from the original refreshment data. The standard deviation of these 500 bootstrap estimates is the bootstrap SE.

The goal is to investigate whether the bootstrap method is a reasonable way to obtain standard errors, since neither asymptotic nor empirical SE is feasible. The power function in (4.9) is used to compare the three methods. The power function of β_i (for $i = 1, 2$) is calculated by holding all other parameters as fixed and changing only one β_i within a range of values. As mentioned previously, 200 samples are drawn from the population and 200 estimates $\widehat{\beta}_i$ are obtained through the semi-parametric method. Three different SEs, namely asymptotic formula SE, empirical SE and bootstrap SE, are used to construct the test statistics defined in (4.7). These test statistics are compared with the level $\alpha = 0.05$ critical value to obtain test results, either rejecting the null hypothesis that $\beta_{i0} = 0$ or not, according to the test function in (4.8). The proportion of null hypotheses

rejected among those 200 tests for each method is calculated as an estimate of the power function.



(a) Power function of β_1

(b) Power function of β_2

Figure 5.10: Power function comparison. The solid, dash and dot-dash lines represent the power functions based on the bootstrap SE, the asymptotic formula SE and the empirical SE respectively. The red dash line on the bottom is at the significance level, 0.05. The power function of β_1 is evaluated at (0, 0.05, 0.1, 0.2, 0.3). The power function of β_2 is evaluated at (0, 0.05, 0.1, 0.13, 0.2, 0.4).

In Figure 5.10, the solid, dash and dot-dash lines represent the power functions based on the bootstrap SE, the asymptotic formula SE and the empirical SE. The red dash line on the bottom is at the significance level, 0.05. The power functions evaluated at $\beta_i = 0$ give the empirical sizes of the tests. The three tests with different SE construction methods are all well calibrated in the plot. In addition, the three power functions indicate that the test is unbiased regardless of the means of constructing standard errors. More

importantly, the power functions based on bootstrap and empirical SEs are close to each other and both have overall higher power than the one based on the asymptotic formula SE. This shows a good approximation of the bootstrap SE to the empirical SE. As a result, the bootstrap method is valid to provide the SE for semi-parametric estimators when performing hypothesis tests in real data applications.

5.4 Netherlands Mobility Panel

Since 2013, the Netherlands Institute for Transport Policy Analysis (KiM) has conducted the Netherlands Mobility Panel (MPN), a multiple wave longitudinal study aimed to understand changes in travel behavior over time. Hoogendoorn-Lanser et al. (2015) provides more detailed information about the panel.

The MPN samples households as survey units and collects travel information by distributing questionnaires to members in each household. Currently, the MPN has the initial and second wave data available from years 2013 and 2014 respectively. The database consists of three main parts: household data, personal data and individual travel diary data. The household data are collected through a household questionnaire in which one person representing each participating household answers a number of questions about the household. The personal data are gathered through an individual questionnaire in which every member within the household who is aged 12 or older answers several questions about themselves. In the travel diary data, these members report their trip information during three consecutive days and answer some additional questions. All data are collected through a web-based survey.

There were 3572 households in the initial wave and 4685 households in the second wave based on the household data. A refreshment sample was included in the second wave. Since travel behavior is the primary research interest, we will focus on those households that completed their travel diary data. Not all households that agreed to participate in the study provided travel diary data. Among those 3572 households in the first wave, there were 2380 households that provided both household information and travel diary data. Among those, there were 1685 households that continued to report their travel behaviors during the second wave of data collection. The rest of the 695 households had no response on the travel dairy. A refreshment sample of 1382 households with both household information and travel diary data was identified and separated from the second wave data set. These three sets of households corresponding to the complete set, incomplete set and refreshment sample, respectively, are summarized in Table 5.11.

Wave 1	Wave 2	
2380 with travel dairy	1685 with travel dairy	Complete set
		Incomplete set
	1382 with travel dairy	Refreshment

Table 5.11: Two-wave Netherlands Mobility Panel data.

The primary use of the Netherlands Mobility Panel is to investigate travel behavior over time. Several studies have analyzed the MPN data from different perspectives to provide insights about travel behaviors. Kroesen et al. (2016) presented a research on the mutual causality between travelers' attitudes and their travel behaviors. Only the complete set was used to draw their conclusions. Hoogendoorn et al. (2016) focused

on estimating the nonresponse bias by modeling the nonresponse behavior. They used an MAR assumption to model the attrition process through a logistic regression with observed household and demographic covariates. La Paix et al. (2016) discussed the measurement of non-random attrition effects on mobility rates using trip diary data. Their analysis assumed that trip diary data were MAR as they evaluated attrition only through observed demographic data.

The above literature either assumed MCAR and used the complete set to gain insight about travel behavior or assumed MAR and incorporated an attrition model to make inferences about the population. In our analysis, we relax the assumption about missing mechanism in the Netherlands Mobility Panel, allowing for potentially MNAR data, and use the refreshment sample to estimate MNAR attrition parameters and gain useful insights for future studies. There are many variables relating to travel behavior in the travel diary data. Each respondent reports travel information including purpose of trip, distance of trip, main transportation mode and travel time. In our study, we use the total travel time as the measurement of travel fatigue, and investigate whether the missing mechanism relates to this measurement. In the travel diary, respondents reported every trip they made during the three-day survey. So we create the total travel time by summing all travel time records from all members within each household and rescale the sum by a natural log transformation. Figure 5.11 shows the comparison in marginal densities of the log transformed total travel time on each wave. Here Y_1 and Y_2 are the total travel time on the natural log scale at initial wave and second wave respectively. In the left panel of Figure 5.11, the estimated marginal density of Y_1 based on the complete set is plotted in red while the one based on the full panel Y_1 is plotted in green.

In the right panel, the estimated marginal density of Y_2 based on the complete set is in red, and the estimated density based on the refreshment sample is in green. The estimated marginal densities of Y_1 and Y_2 based on the complete set (in red) can be biased due to missingness in the data. In contrast, full panel Y_1 and the refreshment sample provide more accurate estimates (in green) for the true marginal densities of Y_1 and Y_2 respectively.

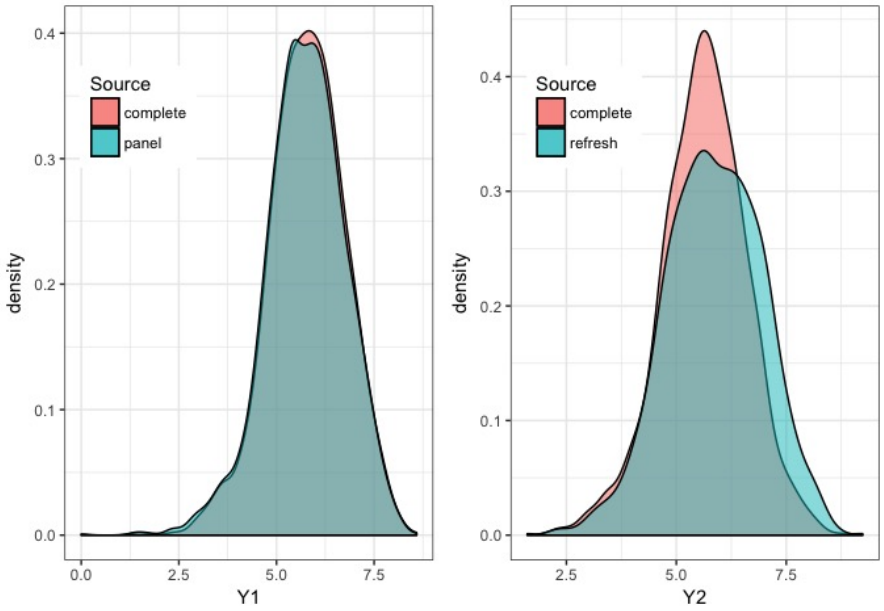


Figure 5.11: Marginal density comparison

We consider three possible attrition models corresponding to the three possible missing mechanisms, namely MCAR, MAR and MNAR. Results are shown in Table 5.12. Let W_i denote the missingness (attrition) indicator for the ith subject with $W_i = 1$ if Y_2 is observed for the subject i and $W_i = 0$ otherwise. We assume an additive logistic model

for the probability of $W_i = 1$ as

$$\pi = P(W_i = 1 \mid y_1, y_2, \underline{\beta}) = \text{logistic}(\beta_0 + \beta_1 y_1 + \beta_2 y_2). \tag{5.2}$$

Under the MCAR assumption, the probability of Y_2 being missing at the second wave is independent of both responses Y_1 and Y_2. This can be modeled by setting both β_1 and β_2 to 0 in (5.2). As a result, the attrition parameter β_0 can be estimated from an interception only logistic regression of W,

$$\pi_{\text{MCAR}} = P(W_i = 1 \mid y_1, y_2, \underline{\beta}) = \text{logistic}(\beta_0).$$

Under this MCAR assumption, subjects randomly drop out of the study and we estimate the probability of observing Y_2 at the second wave to be $logistic(0.89) = 0.71$. We have 95% confidence that this probability is from 0.69 to 0.73.

If the data are MAR – that is, if the attrition is associated with responses through the value of Y_1 only – we can specify the attrition model by setting β_2 to be 0 so that

$$\pi_{\text{MAR}} = P(W_i = 1 \mid y_1, y_2, \underline{\beta}) = \text{logistic}(\beta_0 + \beta_1 y_1).$$

A logistic regression model can be built with W as response and Y_1 as the only covariate. Since both W and Y_1 are fully observed in the panel, we can easily estimate the attrition parameters β_0 and β_1 in this logistic regression. This time, we estimate that the odds for Y_2 being observed at the second wave is $e^{0.15} = 1.16$ times greater with each unit increase in Y_1 (which, recall, is total travel time on the natural log scale). We have 95%

confidence that this multiplicative change is from 1.06 to 1.27. The significance of $\widehat{\beta}_1$ in the MAR model provides strong evidence against the MCAR assumption.

Finally, we extend our analysis by allowing for the possibility of MNAR data and using the refreshment sample to estimate the MNAR model parameters. In particular, we use (5.2) to model the attrition process. The attrition model in (5.2) allows the attrition process to depend on both Y_1 and Y_2. The MAR model can then be justified through (5.2) by testing for whether $\beta_2 = 0$. We apply the kernel density based semi-parametric method to obtain estimates of attrition parameters. Their 95% confidence intervals are constructed through bootstrapping. Figure 5.12 plots the sampling distributions of the bootstrapped semi-parametric estimators. The red vertical lines represent the point estimates from the original data.

The results in Table 5.12 indicate that there is strong evidence that the missingness is related to both Y_1 and Y_2, which implies that the data are MNAR. The positive estimate of β_1 indicates that the probability of Y_2 being observed at the second wave increases as the value of Y_1 increases, given Y_2 is fixed. The negative estimate of β_2 indicates that the probability of Y_2 being observed decreases as the value of Y_2 increases, given Y_1 is fixed. The latter finding is what we would expect from the marginal density comparison for Y_2 in Figure 5.11. The complete set in red has a density leaning toward lower values of Y_2 compared with the potential true marginal density in green. That is, it appears that the larger Y_2 values tended to go missing.

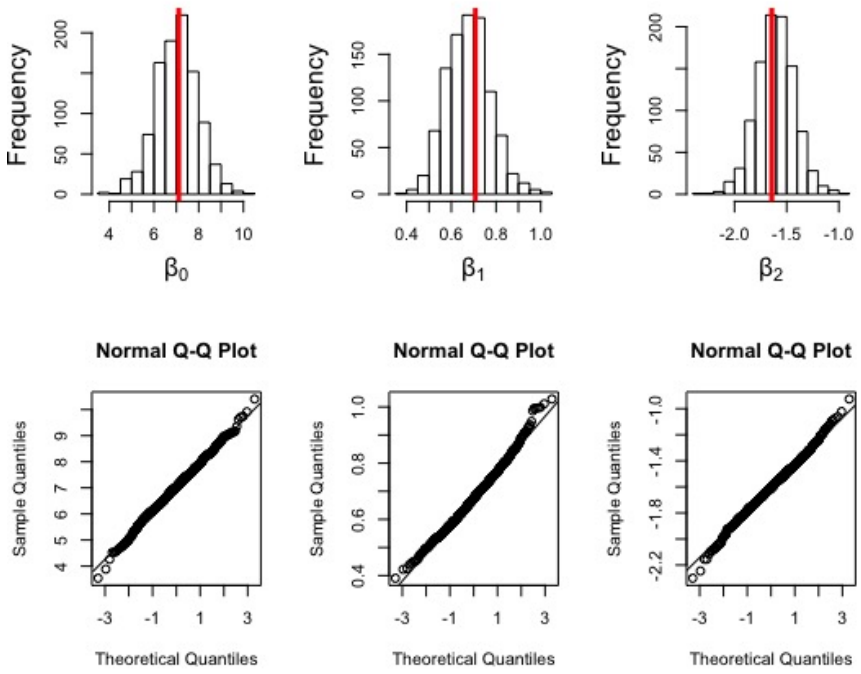


Figure 5.12: Sampling distributions of bootstrapped semi-parametric estimators in Netherlands Mobility Panel application.

Attrition model	MCAR		MAR		MNAR	
$logit(\pi) =$	β_0		$\beta_0 + \beta_1 y_1$		$\beta_0 + \beta_1 y_1 + \beta_2 y_2$	
$\widehat{\beta}_0$	0.89	(0.80, 0.97)	0.03	(-0.47, 0.54)	7.11	(5.09, 8.91)
$\widehat{\beta}_1$			0.15	(0.06, 0.24)	0.71	(0.50, 0.88)
$\widehat{\beta}_2$					-1.64	(-1.97, -1.25)

Table 5.12: Point estimates and 95% confidence intervals for attrition parameters in different attrition models for the Netherlands Mobility Panel.

We also provide a scatter plot of the complete set with the estimated reference line

(50% missingness probability) and normal vector attached in Figure 5.13. Again, the normal vector indicates the direction along which the probability of observing Y_2 increases. This result reveals that participants who have less travel time at the initial wave but more travel time at the follow-up wave tend to have higher probability of a missing second-wave response. Our attempt at explaining this finding is that the increasing burden of reporting participants' travel information as well as increasing travel fatigue they are about to experience tends to dissuade them from staying in the MPN study.

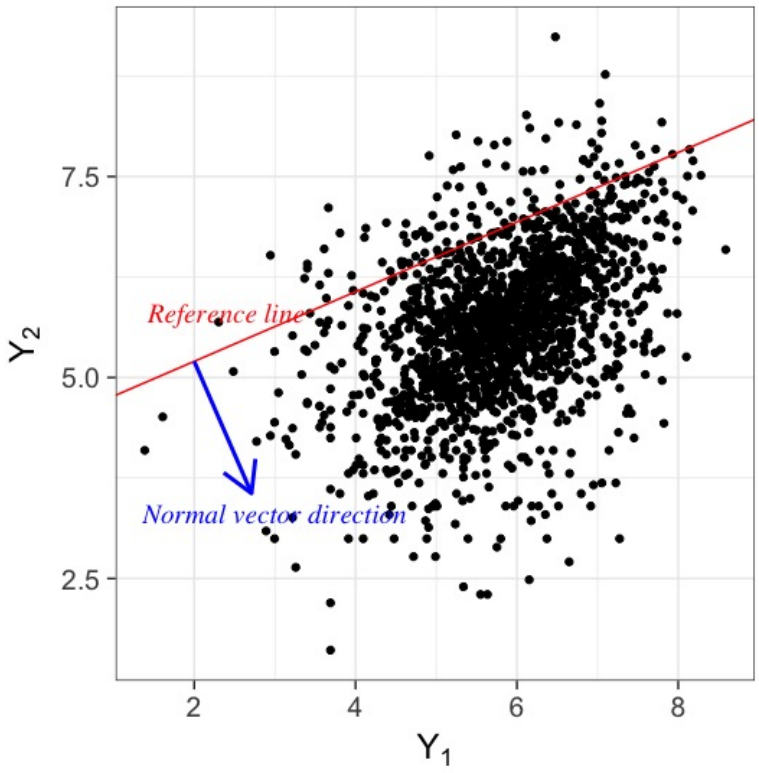


Figure 5.13: Scatter plot of MNP complete set with estimated reference line and corresponding normal vector direction. The normal vector points in the direction of a higher probability of Y_2 being observed.

In this real data application, the refreshment sample helps to build a MNAR attrition model (5.2), and attrition parameters are estimated through the kernel density based semi-parametric method. Confidence intervals based on bootstrapping provide evidence that neither MCAR nor MAR models can adequately explain the attrition process in the Netherlands Mobility Panel. This application demonstrates how different missing mechanism assumptions can lead to different understandings of the attrition process. These assumptions are untestable given the panel data alone. With the refreshment sample, however, three missing mechanisms are characterized into the additive non-ignorable attrition model (5.2) and can be estimated. As a result, we can provide more informative knowledge of the missing process instead of making untestable assumptions. As a final note, this testable attrition model still has an untestable assumption. That is, we need to assume the missing process depends on data in this additive form as shown in (5.2).

6 One Time Invariant Categorical Covariate Extension

In previous chapters, the missing mechanism is assumed to depend only on the response variables through an additive non-ignorable model. Often times, this dependency involves covariates, so it is desirable to incorporate covariates into our missing mechanism model. In this chapter, we extend our semi-parametric model to incorporate one time-invariant binary covariate. This covariate is related to both the responses and to the missing mechanism. In section 6.1, we first describe the population and give a parametric specification for the missing mechanism model in this case. In section 6.2, we explain the estimation procedure. Finally, in section 6.3, we show primary simulation results.

6.1 Models for the Population and Missing Mechanism

Let Y_{i1} and Y_{i2} denote the ith responses at the first and second waves respectively. Let X_i represent a time invariant categorical covariate for the ith observation. For simplicity, we assume that X_i only takes value in two levels, 0 and 1. We further assume a linear model for the relationship between the responses and covariate as follows:

$$\begin{aligned} Y_{i1} &= \alpha_{01} + \alpha_{11} X_i + \varepsilon_{i1}, \\ Y_{i2} &= \alpha_{02} + \alpha_{12} X_i + \varepsilon_{i2}, \end{aligned} \tag{6.1}$$

where the α's are the linear model coefficients and the ε's are the random errors. Here $(\varepsilon_{i1}, \varepsilon_{i2})$ is assumed to follow a bivariate distribution with mean $\underline{0}$ and variance covariance matrix Σ. Let W_i be an indicator variable, with $W_i = 0$ indicating that Y_{i2} is missing. An additive missing model is assumed for W_i, such that

$$P(W_i = 1 \mid y_{i1}, y_{i2}, x_i) = \text{logistic}(\beta_0 + \beta_1 y_{i1} + \beta_2 y_{i2} + \beta_3 x_i).$$

This attrition model is for demonstration purposes only. Its simplicity means that there is a straightforward interpretation for the covariate as the main effect on the odds ratio of being observed. A more complex attrition model can be specified according to Hirano et al. (2001) as shown in (3.1):

$$P(W = 1 \mid y_1, y_2, x) = g(\kappa_0(x) + \kappa_1(y_1, x) + \kappa_2(y_2, x)),$$

where g is a monotone function taking on values in the interval $(0, 1)$, and $\kappa_1(\cdot)$, $\kappa_2(\cdot)$, $\kappa_3(\cdot)$ are arbitrary functions of the responses and the covariate. It is important, however, to note that no interaction terms between y_1 and y_2 are allowed in this additive model.

In addition, a refreshment sample is also included at the second wave. Table 6.1 shows the observed data in this scenario. Similar to the no-covariate case, the observed panel data can be separated into two sets according to the values of W. The complete set consists of observations with $W = 1$, and we fully observe every variable in this set. The rest of the panel data then form the incomplete set. Again, the goal is to understand the attrition process by estimating the attrition parameters ($\underline{\beta}$) from the data that we observe in Table 6.1.

	Obs	Y_1	Y_2	X	W
Complete set	1	Y_{11}	Y_{12}	X_1	W_1=1
	⋮	⋮	⋮	⋮	⋮
	n_c	Y_{n_c1}	Y_{n_c2}	X_{n_c}	W_c=1
Incomplete set	n_c+1	$Y_{(n_c+1)1}$		X_{n_c+1}	W_{n_c+1}=0
	⋮	⋮		⋮	⋮
	N	Y_{N1}		X_N	W_N=0
Refreshment sample	1		Y^r_{12}	X^r_1	
	⋮		⋮	⋮	
	n		Y^r_{n2}	X^r_n	

Table 6.1: Observed full data set with one categorical explanatory variable.

6.2 Method

Estimates of the attrition parameters ($\underline{\beta}$) can be obtained through Hirano et al. (2001)'s two constraints on the covariate x,

$$\int \frac{P(W=1 \mid x)}{\text{logistic}(\beta_0+\beta_1 y_1+\beta_2 y_2+\beta_3 x)} f(y_1,y_2 \mid W=1,x)dy_2 = f_1(y_1 \mid x),$$
$$\int \frac{P(W=1 \mid x)}{\text{logistic}(\beta_0+\beta_1 y_1+\beta_2 y_2+\beta_3 x)} f(y_1,y_2 \mid W=1,x)dy_1 = f_2(y_2 \mid x).$$

The idea for estimating the $\underline{\beta}$ parameters in this scenario is similar to what we have done in the no-covariate case. We can consider the previous no-covariate situation as a special case where the covariate X has only one level. With one binary covariate, we can separate the data into two subsets defined by the levels of X. In each subset, we

construct two constraints as follows. For $X = 0$, we have

$$\int \frac{P(W=1 \mid X=0)}{\text{logistic}(\beta_0+\beta_1 y_1+\beta_2 y_2)} f(y_1,y_2 \mid W=1,X=0)dy_2 = f_1(y_1 \mid X=0),$$
$$\int \frac{P(W=1 \mid X=0)}{\text{logistic}(\beta_0+\beta_1 y_1+\beta_2 y_2)} f(y_1,y_2 \mid W=1,X=0)dy_1 = f_2(y_2 \mid X=0).$$

And for $X = 1$, we have

$$\int \frac{P(W=1 \mid X=1)}{\text{logistic}(\beta_0+\beta_1 y_1+\beta_2 y_2+\beta_3)} f(y_1,y_2 \mid W=1,X=1)dy_2 = f_1(y_1 \mid X=1),$$
$$\int \frac{P(W=1 \mid X=1)}{\text{logistic}(\beta_0+\beta_1 y_1+\beta_2 y_2+\beta_3)} f(y_1,y_2 \mid W=1,X=1)dy_1 = f_2(y_2 \mid X=1).$$

The true attrition parameters $\underline{\beta}^0$ are the only set of parameters that satisfy the above constraints. As a result, the estimates for these parameters can be obtained by minimizing the distance between the conditional density functions on both sides of these four constraints. The estimation procedure starts with estimating the conditional density components in the above constraints. In each subset, we estimate

$$f(y_1,y_2 \mid W=1,X=i), \qquad P(W=1 \mid X=i),$$
$$f_1(y_1 \mid X=i), \qquad f_2(y_2 \mid X=i),$$

where $i = 0$ or 1. We adopt similar notation to that of chapter 4 and consider the estimation of these density components under the constraints for $X = 1$ as an example. First,

we estimate the conditional joint distribution $f(y_1, y_2 \mid W = 1, X = 1)$ as

$$\widehat{f}_H(y_1, y_2 \mid W = 1, X = 1) = \widehat{f}_H(\underline{y} \mid W = 1, X = 1) = \frac{1}{n_{11}} \sum_{i=1}^{n_{11}} K_H(\underline{y} - \underline{Y}_i),$$

where $\underline{Y}_i = (Y_{i1}, Y_{i2})^T$, $i = 1, 2, ..., n_{11}$ indexes the data points with both $W = 1$ and $X = 1$, and H is a 2×2 bandwidth matrix that is symmetric and positive definite. Additionally, $K_H(\underline{y}) = |H|^{-1/2} K(H^{-1/2}\underline{y})$, where K is the bivariate normal kernel function defined as $K(\underline{y}) = (2\pi)^{-1} exp(-\underline{y}^T \underline{y}/2)$. Next, $P(W = 1 \mid X = 1)$ can be consistently estimated by $\widehat{P}(W = 1 \mid X = 1) = n_{11}/N_1$, where N_1 is the number of observations with $X = 1$. For a given $\underline{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$, we can construct the following estimator of the conditional joint density $f(y_1, y_2 \mid X = 1)$:

$$\widetilde{f}(y_1, y_2 \mid X = 1, \underline{\beta}) = \frac{\widehat{P}(W = 1 \mid X = 1)}{\text{logistic}(\beta_0 + \beta_1 y_1 + \beta_2 y_2 + \beta_3)} \widehat{f}_H(y_1, y_2 \mid W = 1, X = 1).$$

The conditional density of Y_1 given $X = 1$ can be computed by integrating the conditional joint distribution $\widetilde{f}(y_1, y_2 \mid X = 1, \underline{\beta})$ with respect to y_2. This can be numerically approximated as follows:

$$\begin{aligned} \widetilde{f}_1(y_1 \mid X = 1, \underline{\beta}) = \int \widetilde{f}(y_1, y_2 \mid X = 1, \underline{\beta}) dy_2 &\approx \sum_{i=1}^{n_{grid}} \widetilde{f}(y_1, y_{2i} \mid X = 1, \underline{\beta}) \times \Delta y_2 \\ &= \sum_{i=1}^{n_{grid}} \widetilde{f}(y_1, y_{2i} \mid \underline{\beta}) \times \frac{range(y_2)}{n_{grid}}, \end{aligned}$$

where y_{2i} is the ith grid point on Y_2 and n_{grid} denotes the number of grid points in the 2-dimensional kernel density estimator. Similarly, for a given y_2, the conditional density

$\widetilde{f}_2(y_2 \mid X=1,\underline{\beta})$ can be defined in the same manner. The conditional density estimates $\widetilde{f}_1(y_1 \mid X=1,\underline{\beta})$ and $\widetilde{f}_2(y_2 \mid X=1,\underline{\beta})$ are semi-parametric estimators that rely on the attrition model. They consistently estimate the true marginal densities only when the attrition model is correctly specified.

Let $\{y_{i1}\}_{i=1}^{N_1}$ be the first wave responses with $X=1$ and $\{y_{i2}^r\}_{i=1}^{n_1}$ be the refreshment sample with $X=1$. We define the following one dimensional kernel density estimators:

$$\widehat{f}_1(y_1 \mid X=1) = \frac{1}{N_1}\sum_{i=1}^{N_1} K_{h_1}(y_1 - y_{i1}), \qquad \widehat{f}_2(y_2 \mid X=1) = \frac{1}{n_1}\sum_{i=1}^{n_1} K_{h_2}(y_2 - y_{i2}^r),$$

where K is the univariate normal density function and $K_{h_i}(y) = h_i^{-1}K(y/h_i)$, with h_i being the corresponding bandwidth for $i=1,2$.

In the subset with $X=0$, a set of similar conditional density estimators can be constructed, and they are denoted as

$$\widetilde{f}_1(y_1 \mid X=0,\underline{\beta}), \qquad \widetilde{f}_2(y_2 \mid X=0,\underline{\beta}),$$
$$\widehat{f}_1(y_1 \mid X=0), \qquad \widehat{f}_2(y_2 \mid X=0).$$

The objective function $M(\underline{\beta})$ takes the form of the mean squared differences between the corresponding conditional density functions from the left and right hand sides of the

four constraints,

$$\begin{aligned}
M(\underline{\beta}) =& M_{N_0}(\underline{\beta}) + M_{n_0}(\underline{\beta}) + M_{N_1}(\underline{\beta}) + M_{n_1}(\underline{\beta}) \\
=& \frac{1}{N_0}\sum_{i=1}^{N_0}\left[\widetilde{f}_1(y_{i1} \mid X=0,\underline{\beta}) - \widehat{f}_1(y_{i1} \mid X=0)\right]^2 \\
&+ \frac{1}{n_0}\sum_{i=1}^{n_0}\left[\widetilde{f}_2(y^r_{i2} \mid X=0,\underline{\beta}) - \widehat{f}_2(y^r_{i2} \mid X=0)\right]^2 \\
&+ \frac{1}{N_1}\sum_{i=1}^{N_1}\left[\widetilde{f}_1(y_{i1} \mid X=1,\underline{\beta}) - \widehat{f}_1(y_{i1} \mid X=1)\right]^2 \\
&+ \frac{1}{n_1}\sum_{i=1}^{n_1}\left[\widetilde{f}_2(y^r_{i2} \mid X=1,\underline{\beta}) - \widehat{f}_2(y^r_{i2} \mid X=1)\right]^2.
\end{aligned}$$

Notice that there are four comparisons since there are four constraints. The vector of semi-parametric estimators of the attrition parameters ($\underline{\beta}$) is the minimizer of the objective function $M(\underline{\beta})$:

$$\widehat{\underline{\beta}} = \arg\min_{\underline{\beta}} M(\underline{\beta}).$$

6.3 Simulation Results

Instead of developing the asymptotic theory, we use simulation to demonstrate the large sample performance of the semi-parametric estimators in this one-covariate case.

We generate data from the following model:

$$
\begin{aligned}
Y_{i1} &= 2 + X_i + \varepsilon_{i1}, \\
Y_{i2} &= 1 + 3X_i + \varepsilon_{i2}, \\
\underline{\varepsilon} &= \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix} \right).
\end{aligned} \tag{6.2}
$$

The additive missing model is set up as

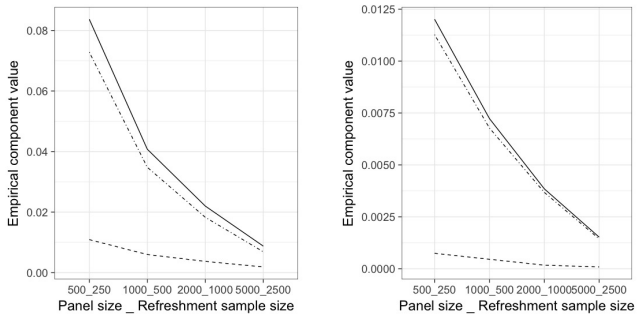
$$
P(W_i = 1 \mid Y_{i1}, Y_{i2}, X_i) = \text{logistic}(-0.8 + 0.2Y_{i1} + 0.4Y_{i2} - 1.4X_i). \tag{6.3}
$$

The attrition parameters are set up such that given any value of the covariate X, there is a non-zero probability to observe the responses Y_1 and Y_2 almost everywhere on their corresponding support. That is, given $X = x$, the support of $f(y_1, y_2 \mid W = 1, x)$ coincides with the support of $f(y_1, y_2 \mid x)$. In this particular additive missing model setting, we are able to control the probability of Y_2 being missing so that it is about 50% on average and ranges from 20% to 80% given either level of X.

6.3.1 Large Sample Performance

The large sample performance of the semi-parametric estimators is examined by computing the empirical mean squared errors (MSEs) using simulations. The true parameters are kept unchanged during the simulation; only the panel size and refreshment sample size are increased. Under each sample size setting, we draw 1000 samples from

the population. The attrition parameters are estimated by the semi-parametric method for each sample. The squared bias and variance are calculated for each estimator based on those 1000 estimates, and the corresponding MSE is computed. Figure 6.1 shows plots of the MSE versus panel and refreshment sample sizes. Again, the X-axis represents the combination of panel size and refreshment sample size. The dashed, dot-dash and solid lines represent the squared bias, variance, and MSE respectively. The decreasing trends of the empirical MSE suggest that the semi-parametric estimators are consistent.



(a) $\widehat{\beta}_0$ (b) $\widehat{\beta}_1$

(c) $\widehat{\beta}_2$ (d) $\widehat{\beta}_3$

Figure 6.1: Large sample performance of semi-parametric estimators in the one-covariate case. The dashed, dot-dash and solid lines represent the squared bias, variance, and MSE respectively.

6.3.2 Asymptotic Sampling Distribution

The asymptotic sampling distribution is simulated from the same population with a panel size of 5000 and a refreshment sample size of 2500. The red vertical lines in Figure 6.2 represent the true values of the attrition parameters. The Q-Q plots suggest the asymptotic normality of the estimators.

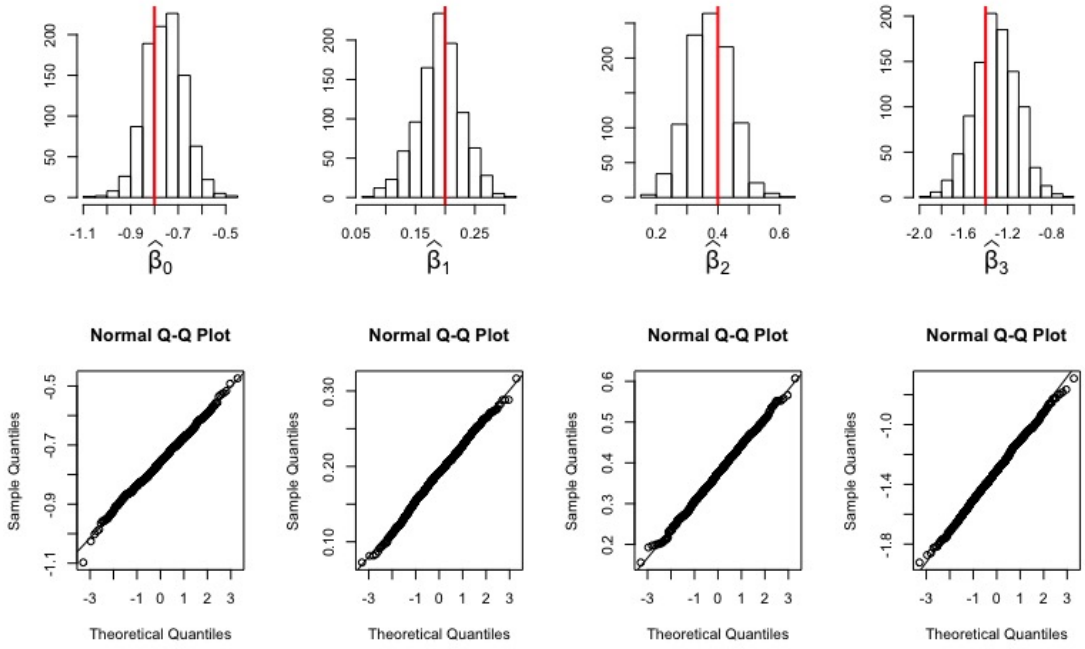


Figure 6.2: Sampling distributions of the semi-parametric estimators in the one-covariate case. These sampling distributions are based on a sample with a panel size of 5000 and a refreshment sample size of 2500.

6.3.3 Inference with Bootstrapping

In this section, we investigate inference using the bootstrap technique. The large sample performance in the previous section shows the asymptotic normality of the semi-parametric estimators, from which, an approach to inference can be developed. In the following, we illustrate the setup for hypothesis testing in the one-covariate case. We carry out the test by first computing confidence intervals and then rejecting the null if the corresponding confidence interval does not contain the null hypothesized value. The construction of confidence intervals requires the estimation of standard errors. We compute the empirical standard errors using simulation. In real data applications, however, simulation is not an option. Thus, the bootstrap technique is considered as an alternative approach to estimate standard errors. We use the power function to investigate the performance of the bootstrap SE. Again, we focus on hypothesis testing for β_1 and β_2 in the one-covariate case. The null hypothesis is

$$H_0 : \beta_{i0} = 0, \quad \text{for } i = 1, 2.$$

We define two types of confidence intervals based on two types of standard errors, namely the empirical simulation SE and the bootstrap SE,

$$\begin{aligned} CI_{sim} : &\quad \widehat{\beta}_i \pm z_{1-\alpha} SE_{sim}, \\ CI_{boot} : &\quad \widehat{\beta}_i \pm z_{1-\alpha} SE_{boot}, \end{aligned}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)^{\text{th}}$ quantile of the standard normal distribution. The level α test function $\phi(Y_1, Y_2, X, W)$ is then defined as

$$\phi(Y_1, Y_2, X, W) = \begin{cases} 1, & \text{reject null} \quad \text{if} \quad 0 \notin CI \\ 0, & \text{fail to reject} \quad \text{otherwise} \end{cases}.$$

Let $Q_1(\beta_1)$ and $Q_2(\beta_2)$ be the power functions for β_1 and β_2 respectively, defined as

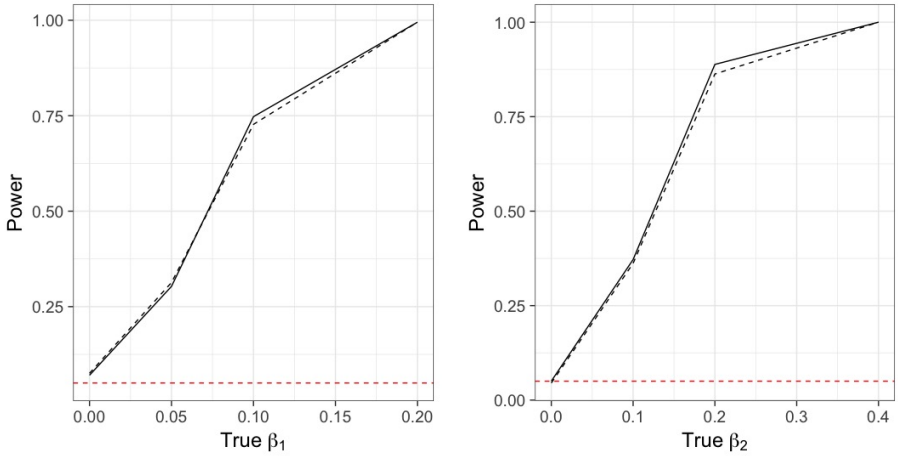
$$\begin{aligned} Q_i(\beta_i) =& P(\text{Reject the null} \mid \beta_i \text{ is the true parameter}) \\ =& P_{\beta_i}(\phi(Y_1, Y_2, X, W) = 1), \quad \text{for } i = 1, 2. \end{aligned}$$

We now consider the parametric setup for estimating these power functions. We fix all parameters appearing in (6.2) and (6.3). We only change the values of the attrition parameters β_1 and β_2 when working with their power functions. Here, we take the estimation of the power function of β_1 as an example. We change the value of β_1 from 0 to 0.2. At each value of β_1, we draw 200 samples from the corresponding population and obtain 200 estimates for β_1 through the semi-parametric method. We then compute the standard error in two different approaches. For the empirical simulation SE, we draw another 1000 samples from the same population and obtain 1000 estimates. We use these 1000 estimates to obtain an estimate of the empirical sampling distribution. The standard deviation of this distribution is the empirical SE. This SE stays the same during the construction of the confidence intervals. On the other hand, we compute different bootstrap SEs for different samples. For each of the 200 samples, we generate

500 bootstrap re-samples, from which we obtain estimates of β_1. The bootstrap SE is computed as the standard deviation of these 500 bootstrapped estimates. Now, we can build confidence intervals for β_1 with these two types of SEs. The power function is approximated by calculating the proportion of confidence intervals that exclude the null hypothesis $\beta_1 = 0$. That is,

$$\widehat{Q}_{sim}(\beta_1) = \frac{1}{200}\sum_{i=1}^{200} \mathbb{1}(0 \notin CI^i_{sim}),$$
$$\widehat{Q}_{boot}(\beta_1) = \frac{1}{200}\sum_{i=1}^{200} \mathbb{1}(0 \notin CI^i_{boot}),$$

where $\mathbb{1}(\cdot)$ is the indicator function. The power function for β_2 is created in the same manner. Figure 6.3 shows the resulting power functions. Solid and dashed lines represent the power functions based on the empirical and bootstrap SEs respectively. The horizontal dashed line on the bottom indicates the level of the test, $\alpha = 0.05$. The bootstrap method gives a close approximation to the empirical power function. These plots demonstrate promising results for using bootstrap technique in the analysis of real data.



(a) Power function of β_1 (b) Power function of β_2

Figure 6.3: Comparison of power functions in the one-covariate case. Solid and dashed lines represent the power functions based on the empirical and bootstrap SEs respectively. The power function is evaluated at 0, 0.05, 0.1, 0.15 and 0.2 for β_1 and at 0, 0.1, 0.2 and 0.4 for β_2.

In this chapter, we have extended our semi-parametric method by including a binary time invariant covariate. The large sample performance and asymptotic sampling distribution of the semi-parametric estimators were assessed using simulation. The power function based on bootstrap re-sample was constructed and the comparison with the empirical power function showed the validity of using bootstrap technique in hypothesis testing for real data applications.

7 Discussion

Attrition in longitudinal studies is a common problem that one has to take into consideration when making statistical inferences based on incomplete data. Depending on whether the attrition process relates to variables that are always observed or variables that are subject to missingness, three main missing mechanism models have been proposed in the literature. The simplest model is missing completely at random (MCAR), which assumes that subjects randomly drop out of a study. The missing at random (MAR) model assumes that subjects drop out for reasons that only depends on variables that are always observed, such as those contained in demographic or satellite data. The missing not at random (MNAR) model allows for the attrition process to depend on variables that are subject to missingness. For MCAR and MAR models, even though many well developed methods are available, their assumptions are usually untestable and this can lead to biased estimates and inference if the assumed missing mechanism is in fact incorrect. On the other hand, given the panel data alone, the MNAR model is not identifiable in most cases.

As a result, Hirano et al. (2001) proposed an additive MNAR model and illustrated the use of a refreshment sample to better understand missing mechanisms. In particular, they showed that the additive MNAR model becomes identifiable in the presence of a refreshment sample. Moreover, this particular type of attrition model can characterize all three missing mechanisms, and it is the weakest model that can be identified using a

refreshment sample.

In this dissertation, two methods are proposed with the goal of making inferences about missing mechanisms in two-wave panel data. Our methods are able to handle more general continuous data instead of the binary data that has been considered by most of the literature. Both methods adopt the additive non-ignorable missing mechanism model proposed by Hirano et al. (2001) to characterize the attrition process. With a specification of the population distribution, our first method computes the full likelihood of the observed data, including both panel and refreshment samples. The attrition parameters are estimated by maximizing the full likelihood. The elegance of this method lies in its straightforward idea of using the likelihood function. The computation, however, is complicated by the presence of attrition. In particular, the observed likelihood for the incomplete set is constructed by integrating out missing variables. Since this integral usually lacks a closed form, adaptive Gaussian quadrature is proposed to approximate it, helping to resolve the analytical problem. The maximum likelihood estimators are asymptotically efficient estimators. Therefore, the parametric method sets the benchmark for the comparison of different estimation methods in this setting.

Our second approach is a kernel density based semi-parametric method, which provides estimates for the attrition parameters using Hirano et al. (2001)'s fundamental constraints. As the non-parametric component, kernel density estimators are used to estimate all of the densities in the constraints. The attrition model is the parametric component in this method, where the attrition probability is assumed to follow a logistic regression model. The use of the refreshment sample provides the key estimate for the second wave marginal density in the constraints. Semi-parametric estimators are

shown to be consistent and asymptotic normally distributed. As a result, a Wald-type test statistic is constructed to make inferences on the attrition parameters. A method of constructing standard errors for semi-parametric estimators using the bootstrap technique is also proposed and its validity is confirmed through simulations.

With the correct specification of the population distribution, the full-likelihood estimator has a smaller asymptotic mean squared error than that of the semi-parametric estimator, as expected. However, the full-likelihood method fails to give consistent estimates when the distribution is misspecified. This reveals an advantage of using the semi-parametric method in real data applications, where it is difficult to specify the population distribution. It always gives consistent estimators and inference, without requiring the specification of a population distribution.

In summary, both of the proposed methods are able to identify the attrition process in a more general two-wave continuous population with the help of a refreshment sample. Compared to a closely related method, our two new methods are easier to understand and enjoy better numerical performances. The finite sample performance of the full-likelihood parametric estimators can be considered as the benchmark, while the semi-parametric method is recommended for real data applications due to its distribution free feature.

However, our methods have some limitations. First, it is not straightforward to numerically carry out the computations for the proposed methods. Both methods estimate attrition parameters by optimizing an objective function that does not have a closed-form solution. Therefore, numerical optimization techniques are required to solve the optimization problems. These techniques are computationally intensive, as the objec-

tive function needs to be re-evaluated at each iteration in the optimization process. The full-likelihood method requires the re-evaluation of the objective function using adaptive Gaussian quadrature, which requires the greatest computational effort. The semi-parametric method, on the other hand, requires the calculation of both one- and two-dimensional kernel density estimates in every iteration.

Even through we have extended our semi-parametric method to incorporate an additional binary covariate, it is still far from being able to accommodate complex datasets that contain multiple covariates, which are common in real applications. The difficulty of incorporating multiple covariates lies in estimating the conditional density of the response variables given the covariates. When there is only one binary covariate, we can separate the data using that covariate and estimate the conditional densities directly from the separated data using kernel density estimators. When we add additional binary covariates, the size of each separated dataset gets smaller and the available information for estimating the conditional densities using the kernel method is reduced. Therefore, the performance of the kernel density estimator is largely compromised. As a result, the estimates of attrition parameters become unstable, with large standard errors. If the covariate is a continuous variable, it then becomes infeasible to follow the same idea of partitioning the data into every value of the covariate and using the kernel method to estimate the conditional densities.

Including covariates, both time invariant and time dependent, into the attrition model is a possible direction for future research. The curse of dimensionality and the requirement of a large sample size limit the semi-parametric method because of its kernel density-based nature. One cannot directly apply a similar method to data with a complex

covariate structure, unless careful assumptions or modifications can be made to resolve this problem. The likelihood-based method could be a good start here. Suppose that we have one continuous, time invariant explanatory variable X. With two-wave panel and refreshment samples, we could again separate the data into three sets: a complete set, an incomplete set, and a refreshment sample. The likelihood can then be constructed for these three sets separately. In particular, the observed likelihood of the complete set is

$$\begin{aligned} L_{comp} &= \prod_{i=1}^{n_c} f(y_{i1}, y_{i2}, W_i = 1 \mid x_i) \\ &= \prod_{i=1}^{n_c} f(y_{i1}, y_{i2} \mid x_i) P(W_i = 1 \mid y_{i1}, y_{i2}, x_i). \end{aligned}$$

The likelihood of the incomplete set, which requires integrating the joint distribution with respect to y_2 to compute, is

$$\begin{aligned} L_{incmp} &= \prod_{i=n_c+1}^{N} f(y_{i1}, W_i = 0 \mid x_i) \\ &= \prod_{i=n_c+1}^{N} \int f(y_{i1}, y_2 \mid x_i) P(W_i = 0 \mid y_{i1}, y_2, x_i) dy_2. \end{aligned}$$

And the refreshment sample has likelihood function

$$L_{refresh} = \prod_{i=1}^{n} f(y_{i2}^r \mid x_i).$$

The full observed likelihood is the product of these three parts. Parametric assumptions are required for both the population and the attrition process. That is, we need to specify $f(y_1, y_2, W = 1 \mid x)$, $f(y_2^r \mid x)$, and $P(W = 1 \mid y_1, y_2, x)$. Computational techniques must

be considered in order to solve the integral when constructing the likelihood of the incomplete dataset as well.

Additionally, our current methods are limited to the analysis of longitudinal data containing two waves, which restricts the extent to which our methods can be applied to data containing three or more waves. Thus, another research direction is to extend current methods to the multi-wave scenario. Deng (2012) developed methods for three-wave binary responses where multiple refreshment samples are used. He considered several missingness scenarios, and provided comprehensive discussions for each case. The three-wave problem is more complex than the two-wave problem because it has more missingness patterns. Table 7.1 shows the appearance of the panel data when missingness is monotonic, which means that if subjects are missing at the sencond wave, they cannot come back and are thus missing at the third wave as well. Table 7.2 gives missing patterns for more general cases. That is, subjects are allowed to come back at the third wave even if their observations are missing at the second wave. Considerations should also be made with regard to the refreshment sample. In a three-wave problem, two refreshment samples are usually allowed to supplement the panel data. One is collected at the second wave and the other is collected at the third wave. Depending on whether we follow up the refreshment sample taken at the second wave, the structure of the observed data at the third wave will be different. This can be seen by comparing Table 7.1 and Table 7.3.

Wave 1	Wave 2	Wave 3	
Y_1	Y_2	Y_3	Complete set
Y_1	Y_2		Incomplete at wave 3
Y_1			Incomplete at wave 2, 3
	Y_2		Refreshment 1
		Y_3	Refreshment 2

Table 7.1: The three-wave panel scenario with monotone missingness and no refreshment follow-up.

Wave 1	Wave 2	Wave 3	
Y_1	Y_2	Y_3	Complete set
Y_1	Y_2		Incomplete at wave 3
Y_1		Y_3	Incomplete at wave 2
Y_1			Incomplete at wave 2, 3
	Y_2		Refreshment 1
		Y_3	Refreshment 2

Table 7.2: The three-wave panel scenario with stochastic missingness and no refreshment follow-up.

Wave 1	Wave 2	Wave 3	
Y_1	Y_2	Y_3	Complete set
Y_1	Y_2		Incomplete at wave 3
Y_1			Incomplete at wave 2, 3
	Y_2	Y_3	Refreshment 1
	Y_2		
		Y_3	Refreshment 2

Table 7.3: The three-wave panel scenario with monotone missingness and refreshment follow-up.

One idea for modeling the missing process in a three-wave panel is to assume a Markov property for consecutive missing mechanisms. That is, we assume whether or not subjects are missing depends only on Y1 and Y2 at the second wave and only on Y2 and Y3 at the third wave. The Markov property here specifies that the conditional probability of being observed (or missing) at the next wave depends on all previous responses only through the current one. Since the goal is to make inferences on these consecutive missing mechanisms, the dataset can be divided into consecutive pairs. One subset contains all of the information in both wave 1 and wave 2, and the other subset contains all of the information in both wave 2 and wave 3. The first subset is used to model the missing mechanism from wave 1 to wave 2. The second subset can be used to build a different missing mechanism model for the attrition at wave 3. However, careful attention must be paid when one models the missing mechanism at the third wave. For instance, in Table 7.3, there are multiple sources of information about Y_2 and Y_3. For

example, the information about Y_2 comes from three different sources: the complete set, the incomplete set, and the refreshment sample. Except for the refreshment sample, none of the other sources are representative of the population at the second wave. One needs to find ways to properly incorporate these details when modeling the attrition process.

The analysis of missing data has a long history in longitudinal studies and has brought enormous challenges to researchers. For the first time, the emergence of the refreshment sample method gives researchers the ability to test missing mechanism assumptions that were untestable in the past, helping to make more informative inferences. The great potential of the refreshment sample in the analysis of missing data has opened a promising field of research, to which our methods contribute.

Bibliography

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843.

Bhattacharya, D. (2008). Inference in panel data models under attrition caused by unobservables. *Journal of Econometrics*, 144(2):430–446.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2):353–360.

Bowman, A. W. (1985). A comparative study of some kernel-based nonparametric density estimators. *Journal of Statistical Computation and Simulation*, 21(3-4):313–327.

Chen, B., Yi, G. Y., and Cook, R. J. (2010). Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random. *Journal of the American Statistical Association*, 105(489):336–353.

Deng, Y. (2012). *Modeling missing data in panel studies with multiple refreshment samples*. PhD thesis, Duke University.

Deng, Y., Hillygus, D. S., Reiter, J. P., Si, Y., Zheng, S., et al. (2013). Handling attrition in longitudinal studies: The case for refreshment samples. *Statistical Science*, 28(2):238–256.

Devroye, L. P. and Wagner, T. J. (1980). The strong uniform consistency of kernel density estimates. *Multivariate Analysis V: Proceedings of the fifth International Symposium on Multivariate Analysis*, 5:59–77.

Efromovich, S. (2011). Nonparametric regression with predictors missing at random. *Journal of the American Statistical Association*, 106(493):306–319.

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2008). *Longitudinal data analysis*. CRC Press.

Hall, P., Marron, J., and Park, B. U. (1992). Smoothed cross-validation. *Probability Theory and Related Fields*, 92(1):1–20.

Han, P. (2012). A note on improving the efficiency of inverse probability weighted estimator using the augmentation term. *Statistics & Probability Letters*, 82(12):2221–2228.

Hausman, J. A. and Wise, D. A. (1979). Attrition bias in experimental and panel data: the gary income maintenance experiment. *Econometrica: Journal of the Econometric Society*, 47(2):455–473.

Hirano, K., Imbens, G. W., Ridder, G., and Rubin, D. B. (2001). Combining panel data sets with attrition and refreshment samples. *Econometrica*, 69(6):1645–1659.

Hoogendoorn-Lanser, S., Schaap, N. T., and OldeKalter, M.-J. (2015). The netherlands mobility panel: An innovative design approach for web-based longitudinal travel data collection. *Transportation Research Procedia*, 11:311–329.

Kim, J. K. and Yu, C. L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of the American Statistical Association*, 106(493):157–165.

Kim, S. (2009). Sample attrition in the presence of population attrition. Technical report, Citeseer.

Kitamura, Y., Tripathi, G., and Ahn, H. (2004). Empirical likelihood-based inference in conditional moment restriction models. *Econometrica*, 72(6):1667–1714.

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

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer New York.

Nevo, A. (2003). Using weights to adjust for sample selection when auxiliary information is available. *Journal of Business & Economic Statistics*, 21(1):43–52.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, 128(2):301–323.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Rubin, D. B. and Schafer, J. L. (1990). Efficiently creating multiple imputations for incomplete multivariate normal data. *Proceedings of the Statistical Computing Section of the American Statistical Association*, 83:88.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9(2):65–78.

Scott, D. W., Tapia, R. A., and Thompson, J. R. (1977). Kernel density estimation revisited. *Nonlinear Analysis: Theory, Methods & Applications*, 1(4):339–372.

Scott, D. W. and Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation. *Journal of the american Statistical association*, 82(400):1131–1146.

Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*. John Wiley & Sons.

Si, Y., Reiter, J. P., and Hillygus, D. S. (2014). Semi-parametric selection models for potentially non-ignorable attrition in panel studies with refreshment samples. *Political Analysis*, 23(1):92–112.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. CRC Press.

Silverman, B. W. et al. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *The Annals of Statistics*, 6(1):177–184.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. CRC Press.

Skrondal, A. and Rabe-Hesketh, S. (2009). Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(3):659–687.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, 16(3):219–242.

Van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge University Press.

Wand, M. P. and Jones, M. C. (1994). *Kernel smoothing*. CRC Press.

Wooldridge, J. M. (2002). Inverse probability weighted m-estimators for sample selection, attrition, and stratification. *Portuguese Economic Journal*, 1(2):117–139.

Yan, J. et al. (2007). Enjoy the joy of copulas: with a package copula. *Journal of Statistical Software*, 21(4):1–21.

Yuan, Y. C. (2010). Multiple imputation for missing data: Concepts and new development (version 9.0). *SAS Institute Inc, Rockville, MD*, 49.

APPENDIX

A Lemmas and Proofs

Proof of Lemma 4.3. Under assumptions (A1), (A2) and (A3), for every $\underline{\beta}$,

$$\begin{aligned}
\left|A_{1\underline{\beta}}(y_1)\right|^2 &= \left|f(y_1 \mid \underline{\beta}) - f_1(y_1)\right|^2 \\
&= \left|\int \frac{f(y_1, y_2 \mid W=1)P(W=1)}{\text{logistic}(\underline{\beta y})} dy_2 - f_1(y_1)\right|^2 \\
&= \left|\int f(y_1, y_2 \mid W=1)P(W=1)(1+\exp(-\underline{\beta y}))dy_2 - f_1(y_1)\right|^2 \\
&\leq F(y_1),
\end{aligned}$$

for some $F(y_1)$ that only depends on y_1 and $E[F(Y_1)] < +\infty$. The inequality holds since all functions are continuous and have compact support. By Lemma 4.2 we have that

$\frac{1}{N}\sum_{i=1}^{N} A_{1\underline{\beta}}(y_{i1})^2$ uniformly converges to its probability limit $EA_{1\underline{\beta}}(y_1)^2$, and

$$\sup_{\underline{\beta}} \left| P_N A_{1\underline{\beta}}^2 - P A_{1\underline{\beta}}^2 \right| \xrightarrow{P} 0. \qquad \blacksquare$$

Proof of Lemma 4.4. By the uniform convergence of the univariate density estimator given in Theorem A of Silverman et al. (1978), we have

$$\sup_{y_1} |C_1(y_1)| = \sup_{y_1} \left| \widehat{f}_1(y_1) - f_1(y_1) \right| \xrightarrow{\text{a.s.}} 0 \quad \text{as} \quad N \to \infty.$$

As a result

$$0 \leq \frac{1}{N}\sum_{i=1}^{N} C_1^2(y_{i1}) \leq \sup_{y_1} |C_1(y_1)|^2 = o_p(1). \tag{A.1}$$

In addition, we can show that

$$\begin{aligned}
B_{1\underline{\beta}}^2(y_1) &= \left[\widetilde{f}(y_1 \mid \underline{\beta}) - f(y_1 \mid \underline{\beta}) \right]^2 \\
&= \left[\int \frac{\widehat{P}(W=1)\widehat{f}_H(y_1, y_2 \mid W=1)}{\text{logistic}(\underline{\beta y})} dy_2 - \int \frac{P(W=1) f(y_1, y_2 \mid W=1)}{\text{logistic}(\underline{\beta y})} dy_2 \right]^2 \\
&= \left[\int \frac{\widehat{P}\widehat{f}_H - Pf}{\text{logistic}(\underline{\beta y})} dy_2 \right]^2 \\
&= \left[\int \frac{\widehat{P}\widehat{f}_H - Pf + P\widehat{f}_H - P\widehat{f}_H}{\text{logistic}(\underline{\beta y})} dy_2 \right]^2 \\
&= \left[\int \frac{(\widehat{P} - P)\widehat{f}_H + P(\widehat{f}_H - f)}{\text{logistic}(\underline{\beta y})} dy_2 \right]^2 .
\end{aligned}$$

By WLLN, one has

$$\widehat{P}(W=1)-P(W=1)=\frac{1}{N}\sum_{i=1}^{N}I(W_i=1)-P(W=1)=o_p(1). \tag{A.2}$$

Furthermore, by the strong uniform convergence in multivariate case Devroye and Wagner (1980), we can show that with (A5),

$$\sup_{y_1,y_2}\left|\widehat{f}_H(y_1,y_2\mid W=1)-f(y_1,y_2\mid W=1)\right|\xrightarrow{P}0 \quad \text{as} \quad n_c\to\infty. \tag{A.3}$$

Combining (A.2) and (A.3), one has

$$\frac{1}{N}\sum_{i=1}^{N}B_{1\underline{\beta}}(y_{i1})^2=o_p(1). \tag{A.4}$$

By Cauchy–Schwarz inequality, for any $\underline{\beta}$, we have

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}\left[2A_{1\underline{\beta}}(y_{i1})B_{1\underline{\beta}}(y_{i1})\right] &\le 2\sqrt{\frac{1}{N}\sum_{i=1}^{N}A_{1\underline{\beta}}(y_{i1})^2\frac{1}{N}\sum_{i=1}^{N}B_{1\underline{\beta}}(y_{i1})^2}=o_p(1),\\
\frac{1}{N}\sum_{i=1}^{N}\left[2A_{1\underline{\beta}}(y_{i1})C_1(y_{i1})\right] &\le 2\sqrt{\frac{1}{N}\sum_{i=1}^{N}A_{1\underline{\beta}}(y_{i1})^2\frac{1}{N}\sum_{i=1}^{N}C_1(y_{i1})^2}=o_p(1),\\
\frac{1}{N}\sum_{i=1}^{N}\left[2B_{1\underline{\beta}}(y_{i1})C_1(y_{i1})\right] &\le 2\sqrt{\frac{1}{N}\sum_{i=1}^{N}B_{1\underline{\beta}}(y_{i1})^2\frac{1}{N}\sum_{i=1}^{N}C_1(y_{i1})^2}=o_p(1).
\end{aligned} \tag{A.5}$$

Therefore Lemma 4.4 follows from (A.1), (A.4) and (A.5). ■

Proof of Lemma 4.7.

$$\begin{aligned} E\left[h\left(\underline{X}_i,\underline{X}_j\right)\right] =& E\left[e_1^2\left(Y_{i1}\right)\underline{g}\left(Y_{i1}\right)T_1\left(Y_{j1},Y_{j2},Y_{i1},W_j\right)\right]+ \\ & E\left[e_1^2\left(Y_{j1}\right)\underline{g}\left(Y_{j1}\right)T_1\left(Y_{i1},Y_{i2},Y_{j1},W_i\right)\right] \\ =& I+II. \end{aligned}$$

For I, conditional on $\underline{X}_i$

$$\begin{aligned} I &= E\left\{E\left[e_1^2\left(Y_{i1}\right)\underline{g}\left(Y_{i1}\right)T_1\left(Y_{j1},Y_{j2},Y_{i1},W_j\right)\mid\underline{X}_i\right]\right\} \\ &= E\left\{e_1^2\left(Y_{i1}\right)\underline{g}\left(Y_{i1}\right)\int T_1\left(y_{j1},y_{j2},Y_{i1},w_j\right)f\left(\underline{x}_j\mid\underline{X}_i\right)d\underline{x}_j\right\}. \end{aligned}$$

Let $u_1=\frac{y_{j1}-Y_{i1}}{h_1}$, $y_{j1}=Y_{i1}+u_1h_1$, $dy_{j1}=h_1du_1$, and $u_2=\frac{y_{j2}-y_2}{h_2}$, $y_{j2}=y_2+u_2h_2$, $dy_{j2}=h_2du_2$. Note that $\underline{X}_i$ and $\underline{X}_j$ are independent and $\int w_jf\left(w_j\mid y_{j1},y_{j2}\right)dw_j=E\left(W_j\mid y_{j1},y_{j2}\right)=P\left(W_j=1\mid y_{j1},y_{j2}\right)=1/\left(1+\exp\left(-\beta_0^0-\beta_1^0y_{j1}-\beta_2^0y_{j2}\right)\right)$. With change of variable and the Taylor expansion, one has

$$\begin{aligned} I =& E\left\{e_1^2\left(Y_{i1}\right)\underline{g}\left(Y_{i1}\right)\left(\int\int\int\frac{P\left(W_j=1\mid y_{j1},y_{j2}\right)K\left(u_1\right)K\left(u_2\right)}{1/\left(1+\exp\left(-\beta_0^0-\beta_1^0Y_{i1}-\beta_2^0y_2\right)\right)}f\left(y_{j1},y_{j2}\right)\right.\right. \\ & \left.\left.\times dy_2du_1du_2-\int K\left(u_1\right)f_1\left(y_{j1}\right)du_1\right)\right\} \\ \approx& E\left\{e_1^2\left(Y_{i1}\right)\underline{g}\left(Y_{i1}\right)\left(\int\int\int K\left(u_1\right)K\left(u_2\right)du_1du_2f\left(Y_{i1},y_2\right)dy_2-f_1\left(Y_{i1}\right)\right)\right\} \\ =& E\left\{e_1^2\left(Y_{i1}\right)\underline{g}\left(Y_{i1}\right)\left(\int f\left(Y_{i1},y_2\right)dy_2-f_1\left(Y_{i1}\right)\right)\right\} \\ =& E\left\{e_1^2\left(Y_{i1}\right)\underline{g}\left(Y_{i1}\right)\left(f_1\left(Y_{i1}\right)-f_1\left(Y_{i1}\right)\right)\right\} \\ =& \underline{0}. \end{aligned}$$

Also II approximates to $\underline{0}$ conditional on $\underline{X}_j$. Thus

$$E\left[h\left(\underline{X}_i,\underline{X}_j\right)\right] \approx \underline{0}.$$

For the variance Σ_1,

$$\begin{aligned}
h_1\left(\underline{X}_i\right) =& E\left[h\left(\underline{X}_i,\underline{X}_j\right) \mid \underline{X}_i\right] \\
=& E\left[e_1^2\left(Y_{j1}\right)\underline{g}\left(Y_{j1}\right)T_1\left(Y_{i1},Y_{i2},Y_{j1},Wi\right) \mid \underline{X}_i\right] \\
=& \int\int\int e_1^2\left(y_{j1}\right)\underline{g}\left(y_{j1}\right)\frac{W_i K_{h_1}\left(y_{j1}-Y_{i1}\right)K_{h_2}\left(y_2-Y_{i2}\right)}{1/\left(1+\exp\left(-\beta_0^0-\beta_1^0 y_{j1}-\beta_2^0 y_2\right)\right)} f\left(y_{j1},y_{j2}\right) \\
& \times dy_2 dy_{j1} dy_{j2} - \int e_1^2\left(y_{j1}\right)\underline{g}\left(y_{j1}\right)K_{h_1}\left(y_{j1}-Y_{i1}\right)f_1\left(y_{j1}\right)dy_{j1}.
\end{aligned}$$

Let $u_1 = \frac{y_{j1}-Y_{i1}}{h_1}$, $y_{j1} = Y_{i1}+u_1h_1$, $dy_{j1} = h_1du_1$, and $u_2 = \frac{y_2-Y_{i2}}{h_2}$, $y_2 = Y_{i2}+u_2h_2$, $dy_2 = h_2du_2$. With change of variable, the Taylor expansion gives that

$$\begin{aligned}
h_1\left(\underline{X}_i\right) \approx& \int\int\int e_1^2\left(Y_{i1}\right)\underline{g}\left(Y_{i1}\right)\frac{W_iK\left(u_1\right)K\left(u_2\right)}{1/\left(1+\exp\left(-\beta_0^0-\beta_1^0Y_{i1}-\beta_2^0Y_{i2}\right)\right)} f\left(Y_{i1},y_{j2}\right) \\
& \times du_2du_1dy_{j2} - \int e_1^2\left(Y_{i1}\right)\underline{g}\left(Y_{i1}\right)K\left(u_1\right)f_1\left(Y_{i1}\right)du_1 \\
=& e_1^2\left(Y_{i1}\right)\underline{g}\left(Y_{i1}\right)W_if_1\left(Y_{i1}\right)\left(1+\exp\left(-\beta_0^0-\beta_1^0Y_{i1}-\beta_2^0Y_{i2}\right)\right) \\
& - e_1^2\left(Y_{i1}\right)\underline{g}\left(Y_{i1}\right)f_1\left(Y_{i1}\right).
\end{aligned}$$

Then $\Sigma_1 = Var\left[h_1\left(\underline{X}\right)\right] = E\left[h_1\left(\underline{X}\right)h_1\left(\underline{X}\right)^T\right] = \left\{E\left[h_{ij}\left(\underline{X}\right)\right]\right\}_{i,j=1}^3$, where $h_{ij}\left(\underline{X}\right)$ is the

ij^{th} element in the matrix $h_1(\underline{X})h_1(\underline{X})^T$ such that

$$\begin{aligned}
h_{ij}(\underline{X}) =& e_1^4(Y_1)g_i(Y_1)g_j(Y_1)f_1^2(Y_1)\left[W\left(1+\exp\left(-\beta_0^0-\beta_1^0Y_1-\beta_2^0Y_2\right)\right)-1\right]^2 \\
=& e_1^4(Y_1)g_i(Y_1)g_j(Y_1)f_1^2(Y_1)W^2\left(1+\exp\left(-\beta_0^0-\beta_1^0Y_1-\beta_2^0Y_2\right)\right)^2 \\
& -2e_1^4(Y_1)g_i(Y_1)g_j(Y_1)f_1^2(Y_1)w\left(1+\exp\left(-\beta_0^0-\beta_1^0Y_1-\beta_2^0Y_2\right)\right) \\
& +e_1^4(Y_1)g_i(Y_1)g_j(Y_1)f_1^2(Y_1).
\end{aligned}$$

Then we have for each element of the matrix $h_1(\underline{X})h_1(\underline{X})^T$

$$\begin{aligned}
E\left(h_{ij}(\underline{X})\right) =& E\left[e_1^4(Y_1)g_i(Y_1)g_j(Y_1)f_1^2(Y_1)\left(1+\exp\left(-\beta_0^0-\beta_1^0Y_1-\beta_2^0Y_2\right)\right)\right] \\
& -E\left[e_1^4(Y_1)g_i(Y_1)g_j(Y_1)f_1^2(Y_1)\right] \\
=& E\left[e_1^4(Y_1)g_i(Y_1)g_j(Y_1)f_1^2(Y_1)\exp\left(-\beta_0^0-\beta_1^0Y_1-\beta_2^0Y_2\right)\right].
\end{aligned}$$

By the relationship between V- and U-statistics introduced in Section 5.7.3 and Theorem A in Section 5.5.1 Serfling (2009), based on Lemma A.1 one has asymptotic normality of $\varphi_N\left(\underline{\beta}^0\right)$ as

$$\sqrt{N}\varphi_N\left(\underline{\beta}^0\right) \sim N(\underline{0}, 4\Sigma_1). \qquad \blacksquare$$

Proof of Lemma 4.8. Recall

$$\begin{aligned}
M_N\left(\underline{\beta}\right) &= \frac{1}{N}\sum_{i=1}^{N}\left[e_1(y_{i1})\left[\widetilde{f}_1(y_{i1}\mid\underline{\beta})-\widehat{f}_1(y_{i1})\right]\right]^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[e_1(y_{i1})\left[A_{1\underline{\beta}}(y_{i1})+B_{1\underline{\beta}}(y_{i1})+C_1(y_{i1})\right]\right]^2
\end{aligned}$$

and

$$\frac{\partial}{\partial \underline{\beta}} M_N\left(\underline{\beta}\right) = \frac{2}{N}\sum_{i=1}^{N} e_1^2(y_{i1})\left[A_{1\underline{\beta}}(y_{i1}) + B_{1\underline{\beta}}(y_{i1}) + C_1(y_{i1})\right]\left[\frac{\partial}{\partial \underline{\beta}} A_{1\underline{\beta}}(y_{i1}) + \frac{\partial}{\partial \underline{\beta}} B_{1\underline{\beta}}(y_{i1})\right].$$

Then

$$\begin{aligned}\frac{\partial^2}{\partial \underline{\beta}^2} M_N\left(\underline{\beta}\right) =& \frac{2}{N}\sum_{i=1}^{N}\left\{ e_1^2(y_{i1})\left[\frac{\partial}{\partial \underline{\beta}} A_{1\underline{\beta}}(y_{i1}) + \frac{\partial}{\partial \underline{\beta}} B_{1\underline{\beta}}(y_{i1})\right]\left[\frac{\partial}{\partial \underline{\beta}} A_{1\underline{\beta}}(y_{i1}) + \frac{\partial}{\partial \underline{\beta}} B_{1\underline{\beta}}(y_{i1})\right]^T \right. \\ & \left. + e_1^2(y_{i1})\left[A_{1\underline{\beta}}(y_{i1}) + B_{1\underline{\beta}}(y_{i1}) + C_1(y_{i1})\right]\left[\frac{\partial^2}{\partial \underline{\beta}^2} A_{1\underline{\beta}}(y_{i1}) + \frac{\partial^2}{\partial \underline{\beta}^2} B_{1\underline{\beta}}(y_{i1})\right]\right\} \\ \frac{\partial^2}{\partial \underline{\beta}^2} M_N\left(\underline{\beta}^0\right) \approx& \frac{2}{N}\sum_{i=1}^{N} e_1^2(y_{i1})\frac{\partial}{\partial \underline{\beta}} A_{1\underline{\beta}^0}(y_{i1})\frac{\partial}{\partial \underline{\beta}} A_{1\underline{\beta}^0}(y_{i1})^T.\end{aligned}$$

The approximation is due to the fact that $B_{1\underline{\beta}}(y_{i1})$, $C_1(y_{i1})$ and $\frac{\partial}{\partial \underline{\beta}} B_{1\underline{\beta}}(y_{i1})$ are $o_p(1)$. Then the probability limit of the second derivative is

$$E\left[\frac{\partial^2}{\partial \underline{\beta}^2} M_N\left(\underline{\beta}^0\right)\right] \approx 2E\left[e_1^2(Y_1)\,\underline{g}(Y_1)\,\underline{g}(Y_1)^T\right]. \qquad \blacksquare$$

Proof of Lemma 4.9. For $\varphi_n^{(1)}\left(\underline{\beta}^0\right)$, we have

$$\varphi_n^{(1)}\left(\underline{\beta}^0\right) = \frac{1}{nN}\sum_{i=1}^{n}\sum_{j=1}^{N}\left[2e_2^2(y_{i2})\,\underline{k}(y_{i2})\,T_2\left(y_{j1}, y_{j2}, y_{i2}, w_j\right)\right].$$

Then

$$\begin{aligned} h_1^{(1)}\left(\underline{X}_j\right) =& E\left[2e_2^2\left(Y_{i2}\right)\underline{k}\left(Y_{i2}\right)T_2\left(Y_{j1},Y_{j2},Y_{i2},W_j\right)\mid \underline{X}_j\right] \\ =& \int 2e_2^2\left(y_{i2}\right)\underline{k}\left(y_{i2}\right)\int \frac{W_j K_{h_1}\left(y_1-Y_{j1}\right)K_{h_2}\left(y_{i2}-Y_{j2}\right)}{1/\left(1+\exp\left(-\beta_0^0-\beta_1^0 y_1-\beta_2^0 y_{i2}\right)\right)}dy_1 f_2\left(y_{i2}\right)dy_{i2} \\ & - E\left[2e_2^2\left(Y_{i2}\right)\underline{k}\left(Y_{i2}\right)f_2(Y_{i2})\right]. \end{aligned}$$

Let $u_1=\frac{y_1-Y_{j1}}{h_1}$, $y_1=Y_{j1}+h_1u_1$, $dy_1=h_1du_1$ and $u_2=\frac{y_{i2}-Y_{j2}}{h_2}$, $y_{i2}=Y_{j2}+h_2u_2$, $dy_{i2}=h_2du_2$. With change of variable and the Taylor expansion

$$\begin{aligned} h_1^{(1)}\left(\underline{X}_j\right) \approx& \int 2e_2^2\left(Y_{j2}\right)\underline{k}\left(Y_{j2}\right)\int \frac{W_j K\left(u_1\right)K\left(u_2\right)}{1/\left(1+\exp\left(-\beta_0^0-\beta_1^0 Y_{j1}-\beta_2^0 Y_{j2}\right)\right)}du_1 f_2\left(Y_{j2}\right)du_2 \\ & - E\left[2e_2^2\left(Y_{i2}\right)\underline{k}\left(Y_{i2}\right)f_2(Y_{i2})\right] \\ =& 2e_2^2\left(Y_{j2}\right)\underline{k}\left(Y_{j2}\right)W_j\left(1+\exp\left(-\beta_0^0-\beta_1^0 Y_{j1}-\beta_2^0 Y_{j2}\right)\right)f_2\left(Y_{j2}\right) \\ & - 2E\left[e_2^2\left(Y_{i2}\right)\underline{k}\left(Y_{i2}\right)f_2(Y_{i2})\right]. \end{aligned}$$

Then $\Sigma_2^{(1)}=Var\left[h_1^{(1)}\left(\underline{X}\right)\right]=E\left[h_1^{(1)}\left(\underline{X}\right)h_1^{(1)}\left(\underline{X}\right)^T\right]=\left\{E\left\{h_{ij}^{(1)}\left(\underline{X}\right)\right\}\right\}_{i,j=1}^{3}$, where $h_{ij}^{(1)}\left(\underline{X}\right)$ is the ij^{th} element in the matrix $h_1^{(1)}\left(\underline{X}\right)h_1^{(1)}\left(\underline{X}\right)^T$ such that

$$\begin{aligned} h_{ij}^{(1)}\left(\underline{X}\right) =& 4e_2^4\left(Y_2\right)k_i\left(Y_2\right)k_j\left(Y_2\right)W^2\left(1+\exp\left(-\beta_0^0-\beta_1^0 Y_1-\beta_2^0 Y_2\right)\right)^2 f_2^2\left(Y_2\right) \\ & - 4e_2^2\left(Y_2\right)k_i\left(Y_2\right)W\left(1+\exp\left(-\beta_0^0-\beta_1^0 Y_1-\beta_2^0 Y_2\right)\right)f_2\left(Y_2\right)E\left[e_2^2\left(Y_2\right)k_j\left(Y_2\right)f_2(Y_2)\right] \\ & - 4e_2^2\left(Y_2\right)k_j\left(Y_2\right)W\left(1+\exp\left(-\beta_0^0-\beta_1^0 Y_1-\beta_2^0 Y_2\right)\right)f_2\left(Y_2\right)E\left[e_2^2\left(Y_2\right)k_i\left(Y_2\right)f_2(Y_2)\right] \\ & + 4E\left[e_2^2\left(Y_2\right)k_i\left(Y_2\right)f_2(Y_2)\right]E\left[e_2^2\left(Y_2\right)k_j\left(Y_2\right)f_2(Y_2)\right]. \end{aligned}$$

Then we have for each element of the matrix $h_1^{(1)}(\underline{X})\, h_1^{(1)}(\underline{X})^T$,

$$\begin{aligned}
E\left(h_{ij}^{(1)}(\underline{X})\right) =& 4E\left[e_2^4(Y_2)\, k_i(Y_2)\, k_j(Y_2)\left(1+\exp\left(-\beta_0^0-\beta_1^0 Y_1-\beta_2^0 Y_2\right)\right) f_2^2(Y_2)\right] \\
&- 4E\left[e_2^2(Y_2)\, k_i(Y_2)\, f_2(Y_2)\right] E\left[e_2^2(Y_2)\, k_j(Y_2)\, f_2(Y_2)\right] \\
&- 4E\left[e_2^2(Y_2)\, k_j(Y_2)\, f_2(Y_2)\right] E\left[e_2^2(Y_2)\, k_i(Y_2)\, f_2(Y_2)\right] \\
&+ 4E\left[e_2^2(Y_2)\, k_i(Y_2)\, f_2(Y_2)\right] E\left[e_2^2(Y_2)\, k_j(Y_2)\, f_2(Y_2)\right] \\
=& 4E\left[e_2^4(Y_2)\, k_i(Y_2)\, k_j(Y_2)\left(1+\exp\left(-\beta_0^0-\beta_1^0 Y_1-\beta_2^0 Y_2\right)\right) f_2^2(Y_2)\right] \\
&- 4E\left[e_2^2(Y_2)\, k_i(Y_2)\, f_2(Y_2)\right] E\left[e_2^2(Y_2)\, k_j(Y_2)\, f_2(Y_2)\right].
\end{aligned}$$

Define

$$U_N^* = \frac{1}{N}\sum_{j=1}^{N} h_1^{(1)}(\underline{X}_j).$$

By central limit theorem

$$\sqrt{N}U_N^* \xrightarrow{d} N\left(\underline{0}, \Sigma_2^{(1)}\right),$$

We want to show that $\varphi_n^{(1)}\left(\underline{\beta}^0\right) \xrightarrow{p} U_N^*$, this is equivalent to show that

$$Var\left[\varphi_n^{(1)}\left(\underline{\beta}^0\right) - U_N^*\right] \longrightarrow 0 \qquad as \qquad N \longrightarrow \infty.$$

First we have

$$
\begin{aligned}
& Var\left[\varphi_n^{(1)}\left(\underline{\beta}^0\right)\right] \\
=& \frac{1}{n^2N^2}\sum_{i=1}^{n}\sum_{j=1}^{N} Var\left[2e_2^2(Y_{i2})\underline{k}(Y_{i2})T_2(Y_{j1},Y_{j2},Y_{i2},W_j)\right] \\
& + \frac{1}{n^2N^2}\sum_{i=1}^{n}\sum_{j=1}^{N}\sum_{j\neq j'} Cov\left[2e_2^2(Y_{i2})\underline{k}(Y_{i2})T_2(Y_{j1},Y_{j2},Y_{i2},W_j),\right. \\
& \left. 2e_2^2(Y_{i2})\underline{k}(Y_{i2})T_2(Y_{j'1},Y_{j'2},Y_{i2},W_{j'})\right] \\
& + \frac{1}{n^2N^2}\sum_{i=1}^{n}\sum_{j=1}^{N}\sum_{i\neq i'} Cov\left[2e_2^2(Y_{i2})\underline{k}(Y_{i2})T_2(Y_{j1},Y_{j2},Y_{i2},W_j),\right. \\
& \left. 2e_2^2(Y_{i'2})\underline{k}(Y_{i'2})T_2(Y_{j1},Y_{j2},Y_{i'2},W_j)\right] \\
\approx& \frac{1}{nN} Var\left[2e_2^2(Y_{i2})\underline{k}(Y_{i2})T_2(Y_{j1},Y_{j2},Y_{i2},W_j)\right] + \frac{n(n-1)}{n^2N}\Sigma_1 \\
\approx& \frac{1}{N}\Sigma_2^{(1)}.
\end{aligned}
\tag{A.6}
$$

And

$$
Var(U_N^*) = \frac{1}{N}\Sigma_2^{(1)}. \tag{A.7}
$$

And

$$
\begin{aligned}
& Cov\left[\varphi_n^{(1)}\left(\underline{\beta}^0\right), U_N^*\right] \\
=& Cov\left[\frac{1}{nN}\sum_{i=1}^{n}\sum_{j=1}^{N}\left[2e_2^2(Y_{i2})\underline{k}(Y_{i2})T_2(Y_{j1},Y_{j2},Y_{i2},W_j)\right], \frac{1}{N}\sum_{j'=1}^{N} h_1^{(1)}\left(\underline{X}_{j'}\right)\right] \\
=& \frac{1}{nN^2}\sum_{i=1}^{n}\sum_{j=1}^{N}\sum_{j=j'} Cov\left[2e_2^2(Y_{i2})\underline{k}(Y_{i2})T_2(Y_{j1},Y_{j2},Y_{i2},W_j), h_1^{(1)}\left(\underline{X}_{j'}\right)\right] \\
=& \frac{1}{N}\Sigma_2^{(1)}.
\end{aligned}
\tag{A.8}
$$

With Eq (A.6), (A.7) and (A.8)

$$\begin{aligned} Var\left[\varphi_n^{(1)}\left(\underline{\beta}^0\right) - U_N^*\right] =& Var\left[\varphi_n^{(1)}\left(\underline{\beta}^0\right)\right] + Var\left[U_N^*\right] - 2Cov\left[\varphi_n^{(1)}\left(\underline{\beta}^0\right), U_N^*\right] \\ \approx& 0. \end{aligned}$$

Thus, $\varphi_n^{(1)}\left(\underline{\beta}^0\right) \xrightarrow{p} U_N^*$ and one has

$$\sqrt{N}\varphi_n^{(1)}\left(\underline{\beta}^0\right) \sim N\left(\underline{0}, \Sigma_2^{(1)}\right). \qquad \blacksquare$$

Proof of Lemma 4.10. We have

$$\begin{aligned} E\left[h^{(2)}\left(Y_{i2}, Y_{l2}\right)\right] =& E\left[E\left[e_2^2\left(Y_{i2}\right)\underline{k}\left(Y_{i2}\right)\left(f_2(Y_{i2}) - K_{h_2}\left(Y_{i2} - Y_{l2}\right)\right) \mid Y_{i2}\right]\right] \\ &+ E\left[E\left[e_2^2\left(Y_{l2}\right)\underline{k}\left(Y_{l2}\right)\left(f_2(Y_{l2}) - K_{h_2}\left(Y_{l2} - Y_{i2}\right)\right) \mid Y_{l2}\right]\right] \\ =& E\left[e_2^2\left(Y_{i2}\right)\underline{k}\left(Y_{i2}\right)\left[f_2(Y_{i2}) - \int K_{h_2}\left(Y_{i2} - y_{l2}\right) f_2(y_{l2}) dy_{l2}\right]\right] \\ &+ E\left[e_2^2\left(Y_{l2}\right)\underline{k}\left(Y_{l2}\right)\left[f_2(Y_{l2}) - \int K_{h_2}\left(Y_{l2} - y_{i2}\right) f_2(y_{i2}) dy_{i2}\right]\right] \\ \approx& E\left[e_2^2\left(Y_{i2}\right)\underline{k}\left(Y_{i2}\right)\left[f_2(Y_{i2}) - f_2(Y_{i2})\right]\right] \\ &+ E\left[e_2^2\left(Y_{l2}\right)\underline{k}\left(Y_{l2}\right)\left[f_2(Y_{l2}) - f_2(Y_{l2})\right]\right] \\ =& \underline{0}. \end{aligned}$$

Define

$$\begin{aligned} h_1^{(2)}(Y_{i2}) &= E\left(h^{(2)}(Y_{i2},Y_{l2}) \mid Y_{i2}\right) \\ &= E\left[e_2^2(Y_{l2})\underline{k}(Y_{l2})(f_2(Y_{l2}) - K_{h_2}(Y_{l2}-Y_{i2})) \mid Y_{i2}\right] \\ &= E\left[e_2^2(Y_{l2})\underline{k}(Y_{l2})f_2(Y_{l2})\right] - \int e_2^2(y_{l2})\underline{k}(y_{l2})K_{h_2}(y_{l2}-Y_{i2})f_2(y_{l2})\,dy_{l2} \\ &\approx E\left[e_2^2(Y_{l2})\underline{k}(Y_{l2})f_2(Y_{l2})\right] - e_2^2(Y_{i2})\underline{k}(Y_{i2})f_2(Y_{i2}). \end{aligned}$$

Define $\Sigma_2^{(2)} = Var\left[h_1^{(2)}(Y_2)\right] = E\left[h_1^{(2)}(Y_2)h_1^{(2)}(Y_2)^T\right] = \left\{E\left[h_{ij}^{(2)}(Y_2)\right]\right\}_{i,j=1}^{3}$, where $h_{ij}^{(2)}(y_2)$ is the ij^{th} element in the matrix $h_1^{(2)}(y_2)h_1^{(2)}(y_2)^T$ such that

$$\begin{aligned} h_{ij}^{(2)}(y_2) =& \left[E\left[e_2^2(Y_2)k_i(Y_2)f_2(Y_2)\right] - e_2^2(y_2)k_i(y_2)f_2(y_2)\right] \\ &\times \left[E\left[e_2^2(Y_2)k_j(Y_2)f_2(Y_2)\right] - e_2^2(y_2)k_j(y_2)f_2(y_2)\right] \\ =& E\left[e_2^2(Y_2)k_i(Y_2)f_2(Y_2)\right]E\left[e_2^2(Y_2)k_j(Y_2)f_2(Y_2)\right] \\ &- E\left[e_2^2(Y_2)k_i(Y_2)f_2(Y_2)\right]e_2^2(y_2)k_j(y_2)f_2(y_2) \\ &- e_2^2(y_2)k_i(y_2)f_2(y_2)E\left[e_2^2(Y_2)k_j(Y_2)f_2(Y_2)\right] \\ &+ e_2^4(y_2)k_i(y_2)k_j(y_2)f_2^2(y_2). \end{aligned}$$

Then we have for each element of the matrix $h_1^{(2)}(y_2)h_1^{(2)}(y_2)^T$,

$$\begin{aligned} E\left(h_{ij}^{(2)}(Y_2)\right) =& E\left[e_2^4(Y_2)k_i(Y_2)k_j(Y_2)f_2^2(Y_2)\right] \\ &- E\left[e_2^2(Y_2)k_i(Y_2)f_2(Y_2)\right]E\left[e_2^2(Y_2)k_j(Y_2)f_2(Y_2)\right]. \end{aligned}$$

By the property of a V-statistics, one has

$$\sqrt{n}\varphi_n^{(2)}\left(\underline{\beta}^0\right) \sim N\left(\underline{0}, 4\Sigma_2^{(2)}\right). \qquad \blacksquare$$

Proof of Lemma 4.11. Recall

$$\begin{aligned} M_n\left(\underline{\beta}\right) &= \frac{1}{n}\sum_{i=1}^{n}\left[e_2\left(y_{i2}\right)\left[\widetilde{f}_2(y_{i2}\mid\underline{\beta}) - \widehat{f}_2(y_{i2})\right]\right]^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[e_2\left(y_{i2}\right)\left[A_{2\underline{\beta}}(y_{i2}) + B_{2\underline{\beta}}(y_{i2}) + C_2(y_{i2})\right]\right]^2. \end{aligned}$$

And

$$\begin{aligned} \frac{\partial}{\partial\underline{\beta}}M_n\left(\underline{\beta}\right) =& \frac{2}{n}\sum_{i=1}^{n} e_2^2\left(y_{i2}\right)\left[A_{2\underline{\beta}}(y_{i2}) + B_{2\underline{\beta}}(y_{i2}) + C_2(y_{i2})\right] \\ & \left[\frac{\partial}{\partial\underline{\beta}}A_{2\underline{\beta}}(y_{i2}) + \frac{\partial}{\partial\underline{\beta}}B_{2\underline{\beta}}(y_{i2})\right]. \end{aligned}$$

Then

$$\begin{aligned} \frac{\partial^2}{\partial\underline{\beta}^2}M_n\left(\underline{\beta}\right) =& \frac{2}{n}\sum_{i=1}^{n}\left\{e_2^2\left(y_{i2}\right)\left[\frac{\partial}{\partial\underline{\beta}}A_{2\underline{\beta}}(y_{i2}) + \frac{\partial}{\partial\underline{\beta}}B_{2\underline{\beta}}(y_{i2})\right]\left[\frac{\partial}{\partial\underline{\beta}}A_{2\underline{\beta}}(y_{i2}) + \frac{\partial}{\partial\underline{\beta}}B_{2\underline{\beta}}(y_{i2})\right]^T\right. \\ & \left. + e_2^2\left(y_{i2}\right)\left[A_{2\underline{\beta}}(y_{i2}) + B_{2\underline{\beta}}(y_{i2}) + C_2(y_{i2})\right]\left[\frac{\partial^2}{\partial\underline{\beta}^2}A_{2\underline{\beta}}(y_{i2}) + \frac{\partial^2}{\partial\underline{\beta}^2}B_{2\underline{\beta}}(y_{i2})\right]\right\}, \\ \frac{\partial^2}{\partial\underline{\beta}^2}M_n\left(\underline{\beta}^0\right) \approx& \frac{2}{n}\sum_{i=1}^{N} e_2^2\left(y_{i2}\right)\frac{\partial}{\partial\underline{\beta}}A_{2\underline{\beta}^0}(y_{i2})\frac{\partial}{\partial\underline{\beta}}A_{2\underline{\beta}^0}(y_{i2})^T. \end{aligned}$$

The approximation is due to the fact that $B_{2\underline{\beta}}(y_{i2})$, $C_2(y_{i2})$ and $\frac{\partial}{\partial\underline{\beta}}B_{2\underline{\beta}}(y_{i2})$ are $o_p(1)$. The leading term of each summation is $e_2^2\left(y_{i2}\right)\frac{\partial}{\partial\underline{\beta}}A_{2\underline{\beta}^0}(y_{i2})\frac{\partial}{\partial\underline{\beta}}A_{2\underline{\beta}^0}(y_{i2})^T$. Then the

probablity limit of the second derivative is

$$E\left[\frac{\partial^2}{\partial \underline{\beta}^2} M_n\left(\underline{\beta}^0\right)\right] \approx 2E\left[e_2^2(Y_2)\,\underline{k}(Y_2)\,\underline{k}(Y_2)^T\right]. \qquad \blacksquare$$

Proof of Theorem 4.3. Now we have three multivariate normal distributed vectors, namely $\varphi_N\left(\underline{\beta}^0\right)$, $\varphi_n^{(1)}\left(\underline{\beta}^0\right)$ and $\varphi_n^{(2)}\left(\underline{\beta}^0\right)$. The sum of these vectors is again a multivariate normal distributed vector. By the relationship between V-statistics and U-statistics and the proof of asymptotic properties of an U-statistics, we can rewrite these three random vectors as

$$\varphi_N\left(\underline{\beta}^0\right) \approx \widetilde{\varphi}_N\left(\underline{\beta}^0\right) = \frac{2}{N}\sum_{i=1}^{N} h_1\left(\underline{X}_i\right) \sim N\left(\underline{0}, \frac{4}{N}\Sigma_1\right),$$

$$\varphi_n^{(1)}\left(\underline{\beta}^0\right) \approx \widetilde{\varphi}_n^{(1)}\left(\underline{\beta}^0\right) = \frac{1}{N}\sum_{j=1}^{N} h_1^{(1)}\left(\underline{X}_j\right) \sim N\left(\underline{0}, \frac{1}{N}\Sigma_2^{(1)}\right),$$

$$\varphi_n^{(2)}\left(\underline{\beta}^0\right) \approx \widetilde{\varphi}_n^{(2)}\left(\underline{\beta}^0\right) = \frac{2}{n}\sum_{l=1}^{n} h_1^{(2)}\left(Y_{l2}\right) \sim N\left(\underline{0}, \frac{4}{n}\Sigma_2^{(2)}\right),$$

where $\underline{X}_i$ and $\underline{X}_j$ represent the sample from the panel and Y_{l2} the sample from refreshment, therefore they are independent. As a result

$$Cov\left[h_1\left(\underline{X}_i\right), h_1^{(2)}\left(Y_{l2}\right)\right] = 0 \text{ and } Cov\left[h_1^{(1)}\left(\underline{X}_j\right), h_1^{(2)}\left(Y_{l2}\right)\right] = 0.$$

The covariance contribution is between $\varphi_N\left(\underline{\beta}^0\right)$ and $\varphi_n^{(1)}\left(\underline{\beta}^0\right)$. For $i \neq j$, we have $Cov\left[h_1\left(\underline{X}_i\right), h_1^{(1)}\left(\underline{X}_j\right)\right] = 0$. And for $i = j$, we have

$$\Sigma_{cov} = Cov\left[h_1\left(\underline{X}\right), h_1^{(1)}\left(\underline{X}\right)\right] = E\left[h_1\left(\underline{X}\right) h_1^{(1)}\left(\underline{X}\right)^T\right] = \left\{E\left[h_{ij}^{cov}\right]\right\}_{i,j=1}^{3},$$

where h_{ij}^{cov} is the ij^{th} element of matrix $h_1(x)\,h_1^{(1)}(x)^T$ and

$$\begin{aligned}
h_{ij}^{cov} =& \left[e_1^2(y_1)\,g_i(y_1)\,wf_1(y_1)\left(1+\exp\left(-\beta_0^0-\beta_1^0y_1-\beta_2^0y_2\right)\right)-e_1^2(y_1)\,g_i(y_1)\,f_1(y_1)\right]\\
&\times\left[2e_2^2(y_2)\,k_j(y_2)\,w\left(1+\exp\left(-\beta_0^0-\beta_1^0y_1-\beta_2^0y_2\right)\right)f_2(y_2)\right.\\
&\left.-2E\left[e_2^2(Y_2)\,k_j(Y_2)\,f_2(Y_2)\right]\right]\\
=&2e_1^2(y_1)\,e_2^2(y_2)\,g_i(y_1)\,k_j(y_2)\,w^2\left(1+\exp\left(-\beta_0^0-\beta_1^0y_1-\beta_2^0y_2\right)\right)^2f_1(y_1)\,f_2(y_2)\\
&-2e_1^2(y_1)\,g_i(y_1)\,wf_1(y_1)\left(1+\exp\left(-\beta_0^0-\beta_1^0y_1-\beta_2^0y_2\right)\right)E\left[e_2^2(Y_2)\,k_j(Y_2)\,f_2(Y_2)\right]\\
&-2e_1^2(y_1)\,e_2^2(y_2)\,g_i(y_1)\,k_j(y_2)\,w\left(1+\exp\left(-\beta_0^0-\beta_1^0y_1-\beta_2^0y_2\right)\right)f_1(y_1)\,f_2(y_2)\\
&+2e_1^2(y_1)\,g_i(y_1)\,f_1(y_1)\,E\left[e_2^2(Y_2)\,k_j(Y_2)\,f_2(Y_2)\right].
\end{aligned}$$

Then

$$\begin{aligned}
E\left[h_{ij}^{cov}\right] =&2E\left[e_1^2(Y_1)\,e_2^2(Y_2)\,g_i(Y_1)\,k_j(Y_2)\left(1+\exp\left(-\beta_0^0-\beta_1^0Y_1-\beta_2^0Y_2\right)\right)f_1(Y_1)\,f_2(Y_2)\right]\\
&-2E\left[e_1^2(Y_1)\,e_2^2(Y_2)\,g_i(Y_1)\,k_j(Y_2)\,f_1(Y_1)\,f_2(Y_2)\right]\\
=&2E\left[e_1^2(Y_1)\,e_2^2(Y_2)\,g_i(Y_1)\,k_j(Y_2)\,f_1(Y_1)\,f_2(Y_2)\exp\left(-\beta_0^0-\beta_1^0Y_1-\beta_2^0Y_2\right)\right].
\end{aligned}$$

Let $N=rn$, r is the ratio between N and n. Then we have

$$\sqrt{N}\left[\varphi_N\left(\underline{\beta}^0\right)+\varphi_n^{(1)}\left(\underline{\beta}^0\right)+\varphi_n^{(2)}\left(\underline{\beta}^0\right)\right]\sim N\left(\underline{0},4\Sigma_1+\Sigma_2^{(1)}+4r\Sigma_2^{(2)}+4\Sigma_{cov}\right).$$

Define $\Sigma=4\Sigma_1+\Sigma_2^{(1)}+4r\Sigma_2^{(2)}+4\Sigma_{cov}$ and $V=E\left[\frac{\partial^2}{\partial\underline{\beta}^2}M_N\left(\underline{\beta}^0\right)\right]+E\left[\frac{\partial^2}{\partial\underline{\beta}^2}M_n\left(\underline{\beta}^0\right)\right]$. By Theorem 5.21 Van der Vaart (2000) we have the asymptotic porperty for $\widehat{\underline{\beta}}$ as follow

$$\sqrt{N}\left(\widehat{\underline{\beta}}-\widehat{\underline{\beta}}_0\right)\sim N\left(\underline{0},\left(V^{-1}\right)\Sigma\left(V^{-1}\right)^T\right). \qquad \blacksquare$$

Lemma A.1. *Relationship between U-statistics and V-statistics. Let $h(x,y)$ be a symmetric function and let $\theta = E\left[h\left(X_i,X_j\right)\right]$, one has*

$$\begin{aligned}\sqrt{n}\left(V_n-\theta\right) &\sim \sqrt{n}\left(U_n-\theta\right)\\ &\sim N\left(0,4\zeta_1\right),\end{aligned}$$

given $\zeta_1 > 0$ and $E\left[h^2\left(X_i,X_j\right)\right] < \infty$, where $\zeta_1 = Var\left[h_1\left(X_1\right)\right] = Var\left[E\left(h\left(X_1,X_2\right) \mid X_1\right)\right]$.

Proof of Lemma A.1. Let $\widetilde{h}(x,y) = h(x,y) - \theta$. Let U_n be an U-statistics and V_n be the corresponding V-statistics such as

$$U_n = \frac{1}{\binom{n}{2}}\sum_c h\left(x_i,x_j\right),$$

$$V_n = \frac{1}{n^2}\sum_i\sum_j h\left(x_i,x_j\right).$$

Define the corresponding centered U_n and V_n as

$$\widetilde{U}_n = \frac{1}{\binom{n}{2}}\sum_c \widetilde{h}\left(x_i,x_j\right) = U_n - \theta,$$

$$\widetilde{V}_n = \frac{1}{n^2}\sum_i\sum_j \widetilde{h}\left(x_i,x_j\right) = V_n - \theta.$$

From Lemma 5.7.3 Serfling (2009) we have the relationship

$$n^2\left(U_n - V_n\right) = \left(n^2 - n_{(2)}\right)\left(U_n - W_n\right),$$

where $n_{(2)} = n(n-1)$ and $W_n = \frac{1}{n}\sum_i h\left(x_i, x_j\right)$. Then it leads to the centered version

$$\begin{aligned}
n^2\left(\widetilde{U}_n - \widetilde{V}_n\right) &= \left(n^2 - n_{(2)}\right)\left(\widetilde{U}_n - (W_n - \theta)\right), \\
n^2\widetilde{U}_n - n^2\widetilde{V}_n &= n\widetilde{U}_n - n\widetilde{W}_n, \\
n\left(\widetilde{V}_n - \widetilde{U}_n\right) &= \widetilde{W}_n - \widetilde{U}_n, \\
\sqrt{n}\left(\widetilde{V}_n - \widetilde{U}_n\right) &= \frac{1}{\sqrt{n}}\left(\widetilde{W}_n - \widetilde{U}_n\right), \\
\sqrt{n}\widetilde{V}_n &= \sqrt{n}\widetilde{U}_n + \frac{1}{\sqrt{n}}\left(\widetilde{W}_n - \widetilde{U}_n\right).
\end{aligned}$$

Since $\frac{1}{\sqrt{n}} \overset{n\longrightarrow\infty}{\longrightarrow} 0$, $\widetilde{U}_n \overset{wp1}{\longrightarrow} 0$ and $\widetilde{W}_n \overset{wp1}{\longrightarrow} \sum_{k=1}^{\infty}\lambda_k$, by Theorem A 5.5.1 Serfling (2009) we have

$$\begin{aligned}
\sqrt{n}\left(V_n - \theta\right) &\sim \sqrt{n}\left(U_n - \theta\right) \\
&\sim N\left(0, 4\zeta_1\right),
\end{aligned}$$

given $\zeta_1 > 0$ and $E\left[h^2\left(X_i, X_j\right)\right] < \infty$, where $\zeta_1 = Var\left[h_1\left(X_1\right)\right] = Var\left[E\left(h\left(X_1, X_2\right) \mid X_1\right)\right]$.

■