AN IMPUTATION METHOD UNDER A PSEUDOLIKELIHOOD METHOD FOR ANALYSIS OF MULTIVARIATE MISSING DATA

by

Yu–Mi Kwon

MS, Rutgers University, 2006

BA, University of Arizona, 2004

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2010

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Yu–Mi Kwon

It was defended on

August 10th 2010

and approved by

Gong Tang, PhD, Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Abdus Wahed, PhD, Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Dianxu Ren, MD, Assistant Professor, Department of Health and Community Systems, School of Nursing, University of Pittsburgh

Ada Youk, PhD, Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Dissertation Director: Gong Tang, PhD, Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

AN IMPUTATION METHOD UNDER A PSEUDOLIKELIHOOD METHOD FOR ANALYSIS OF MULTIVARIATE MISSING DATA

Yu–Mi Kwon, PhD

University of Pittsburgh, 2010

Missing data are prevalent in many public health studies for various reasons. For example, some subjects do not answer certain questions in a survey, or some subjects drop out of a longitudinal study prematurely. It is important to develop statistical methodologies to appropriately address missing data in order to reach valid conclusions. For regression analysis on data with missing values in the response variable, when data are not missing at random, usually the missing-data mechanism needs to be modeled. When the missingness only depends on the response variable, a pseudolikelihood method that avoids modeling the nonignorable missing-data mechanism was developed in the past. A corresponding mean imputation method was used to impute the missing responses under this pseudolikelihood method. In this dissertation, we consider the inference on the moments of the response variable for missing data analyzed by this pseudolikelihood method. At first, we compared three methods: the delta method, the bootstrap method and a re-sampling method, for estimating the variance of the corresponding pseudolikelihood estimate in simulation studies. Second, we modified that mean imputation method and developed a corresponding stochastic imputation method. Multiple imputations were subsequently used to obtain estimates of the moments and the corresponding variance estimates. We compared the performance of these two imputation methods in simulation studies and illustrated them through analysis of the data from a Schizophrenia clinical trial. Compared to the mean imputation method, the stochastic imputation method leads to less and negligible bias.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

Missing data are very common in many biomedical studies for various reasons. For example, some subjects do not answer certain questions in a survey, or some subjects drop out of a longitudinal study prematurely. These missing data are often troublesome because most standard statistical methods require complete data. There are several methods for analysis of missing data. The simplest method is the complete–case analysis (CC). The CC discards all cases which have any missing value, and perform analysis on cases where all variables are present. This method is simple, and has a comparability of univariate statistics. But, it is inefficient because some data are discarded, and often leads to biased estimates. Therefore, it is not recommended in general (Little& Rubin, 2002).

The likelihood–based method is the most common method to analyze missing data by specifying the missing–data mechanism in the likelihood function in addition to a model for the hypothetical complete data. Missing data indicator is used to denote whether a value is observed or not: The missing data indicators are defined as 1 if the corresponding value is observed, and it is defined as 0 if the corresponding value is missing. The likelihood–based method fully specifies the joint distribution of the variables of interest and the missing data indicators. There are two different model frameworks according to how to factor the joint distribution of the variables of interest and the missing data indicators. One is selection models, and the other is pattern–mixture models (Little & Rubin, 2002). Selection models factor the joint distribution into the product of the distribution of the hypothetically complete data and the conditional distribution of the missing data indicator given the hypothetically complete data. Pattern–mixture models stratify the hypothetically complete data according to the missing patterns, and model the distribution of hypothetically complete data within each stratum. When one concerns inference about characteristics of the entire

population, selection models have more natural interpretation, and are more popular. In this dissertation, we focus on selection models.

Selection models generally require specifying the missing–data mechanism, and make inferences based on the full likelihood function. The missing–data mechanism is ignorable if missing data are missing at random (MAR), and parameters of interest and parameters for the missing–data mechanism are distinct (Little & Rubin, 2002). If missing data are ignorable, then the likelihood–based inferences for parameters of interest from full likelihood function is the same as likelihood–based inferences for the parameters of interest from the ignorable likelihood function that is solely based on observed values. However, the observed data do not provide evidence whether the missing–data mechanism is MAR or not, let alone the functional form of the missing–data mechanism. If the missing–data mechanism is mis–specified, the maximum likelihood method often leads to biased estimates and wrongful conclusion. Tang et al. (2003) proposed a pseudolikelihood method to estimate regression paramters for a class of Not–MAR mechanisms and avoid specifying the missing–data mechanisms. They proved that the pseudolikelihood (PL) estimates of regression parameters follow asymptotically normal distribution. However, the covariance matrix of regression parameters by the PL method is very complicated. Suppose that one concerns estimating the variance of the PL estimate of the data. Under this premise, several standard methods can be considered. The standard methods such as the Delta method may not be the most convenient for this case. The Delta method computes asymptotic variance estimates of that function using the estimated covariance matrix of regression parameter estimates and the marginal distribution of covariates. The PL estimate of a general function, say the first moment of the response variable, is a function of the regression parameter estimates and the empirical distribution of the covariates. Therefore, the implementation of the Delta method is very computationally intensive in general.

The Bootstrap method is relatively simpler. The Bootstrap generates random samples with replacement, and estimate regression parameters by the PL method for each bootstrap sample. The function of interest is computed with those PL regression estimates, and the sample variance becomes the variance estimate of the function of interest. However, the Bootstrap requires lots of computing time for regression parameter estimates from the

bootstrap samples because the PL regression parameter estimates have to be numerically searched for each bootstrap sample.

As an alternative method, a direct resampling method is newly developed in the dissertation besides the Delta method and the Bootstrap. The direct resampling method derives samples of the PL regression parameter estimates from the asymptotic normal distribution of the PL regression parameter estimates that are obtained from the original dataset, and samples of covariates are generated by sampling with replacement. Then predictive values of the response variables are drawn from normal distributions whose means and variances are calculated from the covariates and the PL regression parameter estimates. The function of interest is computed with these predictive values in multiple times, and corresponding sample variance is used as the variance estimate of the corresponding PL estimate. The direct resampling method requires less computing time than the Bootstrap because one directly draws parameter estimates from their asymptotic normal distributions.

Beside the direct resampling method, imputation methods may be more useful in that one can avoid complicated computation. The imputed dataset is treated like complete data with appropriate imputation methods, and the variance of the PL estimate of that function can be simply estimated via multiple imputation. Imputation methods replaces missing values with predictive values, and have two generic approaches to generate predictive values. If the predictive values are generated from a formal statistical model, the imputation methods are called explicit model, and if the predictive values are generated from an algorithm instead of an explicit model, the imputation methods are called implicit model. We focus on explicit imputation methods in the dissertation, and the explicit imputation methods include regression imputation, and stochastic imputation. These methods are reviewed in detail in chapter 2. Imputation methods are simple, and the imputed datasets are treated as complete datasets, so most standard statistical analysis can be applied to these imputed datasets. However, imputation methods create predictive values based on observed data, MAR assumption is required for the missing data. Hence, many imputation methods may yield severe bias in estimates when the data are not MAR. Tang devised an imputation method for the PL method (2002) with a mean imputation approach. This imputation method is designed to fill missing values with predictive values that are drawn from the

estimated conditional distribution of the missing values given the observed values for complete cases, and he used Natharaya–Watson (NW) regression estimator (Nadaraya, 1964) to derive means from the complete cases. We modify this mean imputation method with a piece–wise linear regression instead of the NW estimator, and newly introduce a corresponding stochastic imputation method in the dissertation. These imputation methods take into account the population mean in the predictive distribution. Therefore, one can expect less bias in the imputation methods. In the dissertation, we assume that the distribution of the missing data mechansim depends only on response variables, and parameters of interest are estimated by the pseudolikelihood method by Tang et al. (2003). As mentioned earlier, we suppose that we concern estimating the variance of the PL estimate of any function of the data, more specifically, moments of the response variable, and observe the performances of two imputation methods – the mean imputation and the stochastic imputation– in the dissertation. This dissertation consists of several chapters. After the introduction, we review the missing–data mechanisms and three missing data analysis methods in chapter 2. The likelihood–based method, generalized estimating equation and imputation method are discussed in chapter 2. In chapter 3, we introduce the pseudolikelihood method for bivariate and multivariate monotone missing data under the assumption of that the distribution of the missing–data mechanism only consists of dependent variables. In chapter 4, we study the standard methods to estimate the variance of a function of interest under the pseudolikelihood, and examine the advantages and the disadvantages in practice. In chapter 5, we propose two imputation methods: One is a modified mean imputation method based on a piece–wise linear regression from Tang (2002) and the other is the corresponding stochastic imputation method. We conduct the standard methods and two imputation methods to estimate the first moment and the second moment of PANSS data for schizophrenia patients, and compare the results among the different methods in chapter 6. In the final chapter, we summarize the related issues about two imputation methods.

2.0 STANDARD METHODS FOR ANALYSIS OF MISSING DATA

The complete–case analysis is the simplest method for analysis of missing data. Because the complete–case analysis only use the cases where all the variables are presenet, one can apply standard statistical analysis. But this method is inefficient because some data are discarded, and it can also cause a severe bias in estimates when the data are not MCAR. Another method for missing data is available analysis. This method includes all cases where the variable of interest is present, but the sample base changes from variable to variable according to the missing pattern, so available analysis yields a comparability problem across the variables (Little & Rubin, 2002).

The most common method for analysis of missing data is the likelihood–based method. When data are complete, the likelihood–based method estimates parameters of interest based on the likelihood functions where the likelihood function is a function of paramters of interest that is proportional to probabilty density function of the data. If data are incomplete, the likelihood–based method is based on specific modeling assumptions about the missing–data mechanism to estimate parmeters of interest. The missing–data mechanism is ignorable if data are MAR and parameters of interest and the parameter about the missing–data mechanism are distinct. Although data are incomplete, if the missing–data mechanism is ignorable, the likelihood method is relatively simple to use because one does not have to specify the missing–data mechanism in the likelihood functions. If missing data are ignorable, one can estimate parameters of interest by ignorable likelihood because full likelihood functions are proportional to observed likelihood functions where ignorable likelihood is the likelihood of parameters of interest based on observed data ignoring the missing–data mechanism (Little& Rubin, 2002). However, if missing data are not–missing at random (NMAR), likelihood based method requires the full specification of the joint probability of variables of interest

and missing data indicators where missing data indicators are defined as 1 if corresponding variables of interest are present, otherwise they are 0. In fact, it is almost impossible to correctly specify the missing–data mechanism in likelihood functions, and mis–specification of the missing–data mechanism often leads to biased estimates. However, one can obtain consistent estimates without specifying the missing–data mechanism in likelihood–based method under certain assumptions, and this method and assumptions are reviewed in chapter 3.

Instead of likelihood–based method, generalized estimating equation (GEE) can be applied to missing data (Liang & Zeger, 1986). While full likelihood functions need to specify the joint probability structure of the observations, GEE does not have to be associated with likelihood functions. GEE only requires the mean and variance functions, and computes consistent estimates of parameters of interest by treating correlation structure as a nuisance parameter. However, GEE has a limitation to apply to an missing dataset because GEE assumes that the missing–data mechanism is MCAR when data are incomplete. If data are not MCAR, this method does not yield consistent estimates with missing data.

Besides, imputation methods are also frequently used for missing data analysis. Imputations are techniques that replace missing values to reasonable predictive values according to formal statistical models or underlying models. Once missing values are imputed, the imputed dataset is treated as the complete dataset. If one creates more than one complete data by imputation, then imputation methods are called multiple imputation. Single imputation undermines the variability within the employed predictive model, so usually multiple imputation are necessary in order to account for the variability of data, and we more focused on multiple imputation methods in the dissertation. Imputation methods are easy to conduct, and one can directly employ standard statistical procedures on the imputed dataset. We review the concept of the missing–data mechanisms that lead to missing data in the first section, and study three common methods for missing data analysis are studied in the second section.

2.1 MISSING–DATA MECHANISM

"The missing–data mechanisms are crucial for missing data analysis because the properties of missing data methods depend very strongly on the nature of the dependencies in these missing–data mechanisms" (Little & Rubin, 2002). According to the theory of Rubin (1976), the concept of the missing–data mechanism begins at the definition of the missing data indicators. We denote the missing data indicator as R, and R is defined as 1 if the value is observed, otherwise it is defined as 0. This missing data indicator R is treated as a random variable. Based on the defintion of the missing data indicator, the missing–data mechanisms are statistically formalized by Rubin (1976) according to the relationship between the hypothetical complete data and the missing data indicators, and they explain how variables of interest are related with underlying values in missing data.

We denote an independent variable and a dependent variable as X and Y where X is fully observed and Y is partially observed. Y is expressed as $\{Y_{obs}, Y_{mis}\}$ where Y_{obs} denotes observed part of Y and Y_{mis} denotes missing part of Y. The missing–data mechanism is characterized by the conditional distribution of R given $[X, Y]$ where α and ψ denote parameters of interest and the parameter of the missing–data mechanism. The missing–data mechanism is categorized into three : Missing Completely At Random (MCAR), Missing At Random (MAR) and not–missing at random (NMAR) (Rubin & Little, 2002).

1. MCAR, with the assumption:

$$pr[R|y, x; \psi] = pr[R; \psi]$$

   Missing completely at random (MCAR) is the strongest assumption on missing data. If the data are MCAR, it implies that there is no relationship between variables of interest and the missing data indicator at all. The assumption of MCAR does not imply that the missing pattern itself is random, but rather that the missingness does not depend on the data values (Little & Rubin, 2002).

2. MAR, with the assumption:

$$pr[R|y, x; \psi] = pr[R|y_{obs}, x; \psi]$$

An assumption of missing at random (MAR) is less restrictive than MCAR. MAR implies that the missingness is not dependent on the missing values after conditioning on the observed values. So, if the missing data are missing at random, one does not have to specify full likelihood functions under the specification of the missing–data mechanism. Because full likelihoods are proportional to the observed likelihoods, the missing–data mechanism can be ignored in likelihood–based method. When the data are incomplete, most statistical packages such as SAS assume that data are missing at random.

3. NMAR, with the assumption:

$$pr[R|y, x; \psi] = pr[R|y_{obs}, y_{mis}, x; \psi]$$

Not–missing at random (NMAR) includes all missing–data mechanisms that do not belong to either MCAR or MAR. NMAR implies that the missingness depends on the missing values even after conditioning on the observed values, and is the condition that makes the missing data analysis complicated. When missing data are NMAR, the missing–data mechanism should be specified in the likelihood functions to yield consistent estimates, and many imputation methods yield severeo biases for estimation.

The distribution of observed data is obtained by integrating Y_{mis} out of the joint density of $[Y, M]$ as follows:

$$f(Y_{obs}, M|\alpha, \psi) = \int f(Y_{obs}, Y_{mis}|\alpha) f(M|Y_{obs}, Y_{mis}, \psi) dY_{mis} \tag{2.1}$$

If data are MAR, and parameters of interest and the parameter of the missing–data mechanism are distinct, the missing data are usually called ignorable. If the missing–data mechanism is ignorable, the missing–data mechanism does not depend on Y_{mis}.

$$f(M|Y_{obs}, Y_{mis}, \psi) = f(M|Y_{obs}, \psi) \quad \text{for all Y} \tag{2.2}$$

Therefore, (2.1) is summarized as follows:

$$f(Y_{obs}, M|\alpha, \psi) = f(Y_{obs}|\alpha) f(M|Y_{obs}, \psi) \tag{2.3}$$

This ignorable condition does not require the specification of the missing–data mechanism in the likelihood–based approach according to (2.3). So one performs the likelihood–based method by ignoring the missing–data mechanism for a valid inference under ignorable condition. In addition, most imputation methods also assume this condition.

However, if the data are not–missing at random , one has to specify the missing–data mechanism in full likelihood functions, and the mis–specified missing–data mechanism results in a severe bias for estimation problems. If the data are not–missing at random, imputation methods are also troublesome. Because most imputation methods assume MAR, those imputation methods bring about severe biases for not MAR data. However, if the distribution of the missing–data mechanism is a function of dependenet variables, one can obtain consistent estimates without specifying the missing–data mechanism in likelihood functions by the pseudolikelihood method (Tang et al. 2003).

2.2 METHODS FOR ANALYSIS OF MISSING DATA

2.2.1 Likelihood–based Methods

The likelihood–based method is the most common method for analysis of missing data. The likelihood–based method specifies the joint distribution of the missing data indicator and variables of interest with the assumption about the missing–data mechanism when the data are incomplete, and estimates parameters of interest by maximizing likelihood functions.

The likelihood–based method has two model frameworks to express the joint distribution of the missing data indicators and variables of interest when the data are incomplete. Where R and Y denote the missing data indicators and variables of interest, the selection models express the joint distribution of $[Y, R]$ as the product of $[Y]$ and $[R|Y]$ (Heckman, 1976), and the pattern–mixture models express $[Y, R]$ with $[R]$ and $[Y|R]$ after stratifying the missing data according to the missing data patterns. The expressions of these two models are exchangeable: The selection models can be written as the pattern–mixture models, and the pattern–mixture models can be written as the selection models.

- Selection Models

  If one concerns parameters estimates of entire population, the selection model framework is more natural expression of the joint distribution of $[Y, R]$ in likelihood–based method. In the selection model framework, the joint distribution of $[Y, R]$ is factored as the product of distribution of $[Y]$ and the conditional distribution of $[R|y]$ as follows where α and ψ denote parameters of interest and the parameter of the missing–data mechanism.

$$pr(X, Y, R; \alpha, \psi) = pr[X, Y|\alpha] \cdot pr[R|X, Y, \psi]$$

  The selection models focus on the inferences of the population parameters, α, while the pattern–mixture models focus on the properties of the missing data patterns. When the missing–data mechanism is ignorable such that where the missing–data mechanism is MAR and α and ψ are distinct, the conditional distribution of $[R|y, \psi]$ is ignored to estimate parameters of interest α in the selection models.

- Pattern– Mixture Models

  The pattern–mixture models are an alternative model framework to express the joint distribution of $[Y, R]$ in the likelihood–based method (Glynn, Laird and Rubin, 1986). Unlike the selection model, the pattern–mixture models stratify the missing data by the missing data patterns, and express the joint distribution of variables of interest and the missing data indicator as follows where δ and γ denote a parameter of interest and the parameter of the missing–data mechanism in a given stratum.

$$pr[X, Y, R|\delta, \gamma] = pr[X, Y|R, \delta] \cdot pr[R|\gamma]$$

  When the missing dataset consists of multiple sub–populations across the missing data patterns, the pattern–mixture models may be more useful when one is interested in observing the properties of sub–populations within each stratum. However, the inference of the population paramter is drawn by the mixture form of the distributions of the sub–populations due to the characteristics of the expression, so the identifiability problem often occurs under the pattern–mixture model framework.

We use the selection model framework to find regression parameter estimates in the likelihood based method among the above two models in the dissertation. If the data are complete, maximum likelihood method (ML) is the most efficient likelihood method. Maximum likelihood method is to estimate parameters of interest that maximize the likelihood functions or the log–likelihood functions about the parameters of interest. If the data are incomplete, and satisfy the ignorable condition, the conditional distribution of the missing data indicator R given Y is not associated with parameters of interest in the likelihood functions. Therefore the maximum likelihood method can be used to estimate parameters of interest without any difficulty.

For analysis of missing data, the maximum likelihood method may not always be easy to use especially when the missing–data mechanism is NMAR because the selection models require the full specification of the missing–data mechanism. Mis–specification of the missing–data mechanism often leads to biased estimates.

2.2.2 Generalized Estimating Equations

Generalized estimating equation model (GEE) is another method to analyze missing data. The GEE is widely used for longitudinal analysis (Liang & Zeger, 1986) because the joint distribution of the repeated responses does not have to be fully specified, so it is easy to be applied to data with repeated measurements. The likelihood–based method requires the full specification of a joint probability structure, but the objective function of GEE is only associated with the mean and variance function as follows:

$$\frac{\partial \mu}{\partial \alpha} V^{-1}(Y - \mu) = 0$$

Where $\mu = \mu(\alpha)$ and α denote the mean function and the parameters of interest, GEE only specifies the mean and the variance, the shape of the distribution remains free. So, it is especially useful in analysis of non–gaussian data. In addition, a correlation structure is treated like a nuisance parameter in generalized estimating equation (GEE), only mean and variance are used to estimate a parameter of interest. If data are complete, the solution of

the above equation is known to provide asymptotically consistent estimates of α under mild regularity conditions (Liang & Zeger, 1986). However, this method is sometimes problematic in analysis of missing data. Because this method is based on the observed data to estimate parameters of interest, MCAR should be assumed for missing data. If the missing–data mechanism is not MCAR, simple generalized estimating equations do not yield consistent estimates. One can use weighted GEE (Robins,Rotnitsky and Zhao, 1995) using a auxiliary variable z_i that predict whether or not y_i is completed as follows:

$$\frac{\partial \mu}{\partial \alpha} w(\hat{\eta}) V^{-1}(Y - \mu) = 0$$

Where $w(\hat{\eta})$ is the inverse of an estimate of the probability of being a complete case obtained by a logistic regression of R_i on x_i and z_i, and η is the parameter of the logistic regression by maximum likelihood, $w(\hat{\eta})$ allows the missingness to depend on the auxiliary variables as well as the covariates, so weighted GEE is known to correct the bias of unweighted GEE that attribute to the dependency of the missing–data mechanism on z_i (Robins,Rotnitsky and Zhao, 1995). However, the dissertation is focused on NMAR missing data, the generalized estimating equation is not considered in the dissertation.

2.2.3 Multiple Imputation

Imputation methods are direct and simple for missing data analysis. Because the imputed datasets are treated as the complete data, most standard statistical analysis can be employed to these imputed datasets. Imputation methods fill missing values with predictive values, and there are two generic methods to generate these predictive values. The first method is to draw predictive values from formalized statistical models, and the second method is to draw predictive values from underlying models. The first method is refered to explicit modeling method and the second method is refered to implicit modeling method. The dissertation is focused on explicit model based imputation method, and mean imputation, regression imputation, and stochastic imputation are included in explicit based modeling method. Those imputation methods can be summarized as follows:

- Mean Imputation

  Missing values are substituted by means from the responding units in the sample in the mean imputation. Means can be formed within cells or classes, and mean imputation leads to estimates similar to those found by weighting provided the sampling weights are constant within weighting classes (Little & Rubin, 2002).

- Regression imputation

  Missing values are imputed by predictive values from a regression of the missing item on items observed for the unit in regression imputation methods. The regression equation to draw the predictive values is usually calculated from the units with both observed and missing variables are present together. Mean imputation method is regarded as a special case of regression imputation method if the predictor variables are dummy indicator variables for the units (Little & Rubin, 2002).

- Stochastic regression imputation

  Stochastic regression imputation method fills missing values with predictive values that are computed with values predicted by regression imputation method plus a residual. If normal linear regression is considered, one can assume the residual following normal distribution where the expectation is zero and the variance is the residual variance in regression. Because stochastic regression imputation method reflects the sampling uncertainty in the predicted value, it is more preferred to regression imputation (Little & Rubin, 2002).

Imputation methods are characterized according to how to draw predictive values for missing values, but it can be also categorized as single imputation and multiple imputation. Single imputation fills missing value once, and creates one complete data, and analyze the imputed dataset. Multiple imputation creates D complete datasets by separate and independent D imputations. Generally, multiple imputation is D repetition from the posterior predictive distribution of Y_{mis} for the considered model, and each repetition corresponds to an independent drawing of parameters and missing values.

The mulitiply–imputed dataset is analyzed using the same complete data method. Let $\hat{\alpha}_i$ and W_i for $i = 1, \cdots, D$ be the estimates of interest and their associated variances of $\hat{\alpha}$ from D imputed datasets. The combined estimate of $\bar{\alpha}$ is calculated by (2.4).

$$\bar{\alpha} = \frac{1}{D} \sum_{i=1}^{D} \hat{\alpha}_i \tag{2.4}$$

Averaging over D imputed datasets increases the efficiency of estimates over a single imputed dataset (Little & Rubin, 2002). The variability by multiple imputations consists of two components: One is the 'Within imputation variance' component of (2.5), and the other is the 'Between imputation' component as (2.6). But the variability by single imputation is only expressed with the 'Within imputation variance'component, and the variance estimate by the single imputation may not be valid when the data are not–missing at random (NMAR) because the variances between the complete cases and the missingcases are not generally same.

$$\bar{W}_D = \frac{1}{D} \sum_{i=1}^{D} W_i \tag{2.5}$$

$$B_D = \frac{1}{D-1} \sum_{i=1}^{D} (\hat{\alpha}_i - \bar{\alpha}_D)^2 \tag{2.6}$$

According to (2.5) and (2.6), total variability of $\bar{\alpha}_D$ is computed by adding together in multiple imputation as (2.7).

$$T_D = \bar{W}_D + \frac{D+1}{D} \cdot B_D \tag{2.7}$$

Multiple imputation helps reducing the imputation bias (Little & Rubin, 1983) and it performs favorably to produce the unbiased estimate in comparison with single imputation [Graham & Schafer, (1999), Schafer & Graham, (2002)]. We conduct both single imputation and multiple imputation in a simulation study, and compare their performances with regards to 95% coverage rates. Also, we use (2.7) to compute the variance estimates of the first moment and the second moment for multiple imputation.

3.0 A PSEUDOLIKELIHOOD METHOD FOR ANALYSIS OF MULTIVARIATE MONOTONE MISSING DATA

The likelihood–based method, which is the most common method for missing data analysis, needs to fully specify the joint distribution of the missing data indicator and variables of interest under a assumed missing–data mechanism. The likelihood functions can be expressed as two model frameworks depending how to express the joint distribution of $[Y, R]$. One is the selection model framework (Heckman, 1976) and the other is the pattern–mixture model framework (Glynn, Laird & Rubin, 1986). The selection models express the joint distribution of $[Y, R]$ as the product of a marginal distribution of the dependent variables $[Y]$ and a conditional distribution of the missing data indicators given the dependent variables $[R|y]$ while the pattern–mixture models express $[Y, R]$ as a conditional distribution of the dependent variables given the missing data indicators $[Y|R]$ after stratifying the missing data according to the missing data patterns. As mentioned in the previous chapter, the pattern–mixture models compute the population parameters of interest as the combination of the parameters which are driven from the conditional distributions of all strata, so the selection model is more natural to interpret the population parameters in the likelihood–based method. Between these two model frameworks, we use the selection model framework in the likelihood–based method in the dissertation.

If missing data are missing at random (MAR), the selection model does not require the specification of the missing–data mechanism in the likelihood functions. Under those assumptions, the parameters of interest are not associated with the parameter about the missing–data mechanism, so maximum likelihood method can be used to estimate the parameters of interest by the ignorable likelihood function. However, the selection models need a full specification of the missing–data mechanism in the likelihood functions when the data

are not–missing at random, and the mis–specification of the missing–data mechanism leads to inconsistent estimates. Hence, maximum likelihood method is not easy to conduct with not–missing at random missing data, but one can conduct the pseudolikelihood method to estimate paramters of interest instead of maximum likelihood method with not–missing at random missing data under certain assumptions. The first assumption is that distributions of dependent variables follow known parametric functions, and the second assumption is that the missingness only depends on the underlying values of a dependent variable. Under these two assumptions, one can estimate parameters of interest without specifying the missing–data mechanism by the pseudolikelihood method (Tang et al. 2003). They used the pseudolikelihood method to compute regression parameters estimates, and compute the asymptotic distribution of these regression parameter estimates. The covariance matrix of these regression parameter estimates are also suggested, and the estimate of the covariance matrix can be refered to Appendix A.

We study the pseudolikelihood method to estimate regression parameter estimates with not–missing at random data which have a monotone missing pattern in this chapter where the monotone missing pattern is the pattern that all outcomes are missing since the previous outcome is missing. This monotone pattern frequently occurs due to drop–outs in a longitudinal study design. In the first section, we study the pseudolikelihood method with bivariate monotone missing data, and extend this pseudolikelihood method to the multivariate monotone missing data in the second section. For these not–missing at random missing data, we assume that the missing–data mechanisms depend upon the response variables, and the conditional distributions of the response variables are known parametric functions.

3.1 A PSEUDO LIKELIHOOD METHOD FOR BIVARIATE MONOTONE MISSING DATA

Consider a bivariate monotone missing dataset of $\{x_i, y_i\}$ for $i = 1, \cdots, n$ such that the covariate X is fully observed and the response variable Y is partially but monotonely missing. According to the monotone missing pattern, y_i are observed for $i = 1, \cdots, m$, but y_i are

missing for $i = m+1, \cdots, n$ where $n > m > 0$. The missing data indicator R_i is defined as 1 if corresponding y_i is observed, and R_i is defined as 0 if y_i is missing. We concern estimating the parameters of interest from the conditional distribution of Y given X, $[Y|X;\alpha]$, where α is the vector of parameters of interest. The missingness of Y is assumed to depend upon a function of Y as (3.1) where $\omega(\cdot)$ is an arbitrary function.

$$P[R = 1|X, Y] = \omega(Y;\psi) \tag{3.1}$$

Under the assumption of (3.1), the full likelihood function based on the selection model can be expressed as follows:

$$\begin{aligned}
L(\alpha,\eta,\psi;X,Y,R) &= \prod_{i=1}^{m} p(x_i,y_i,R_i|\alpha,\eta,\psi) \prod_{i=m+1}^{n} \int p(x_i,y,R_i|\alpha,\eta,\psi)dy \\
&= \prod_{i=1}^{m} p(x_i,y_i|\alpha,\eta)\omega(y|\psi) \prod_{i=m+1}^{n} \int p(x_i,y|\alpha,\eta)(1-\omega(y|\psi))dy \\
&= \prod_{i=1}^{m} p(x_i|\eta)p(y_i|x_i,\alpha)\omega(y|\psi) \prod_{i=m+1}^{n} p(x_i|\eta) \int p(y|x_i,\alpha)(1-\omega(y|\psi))dy \\
&= \prod_{i=1}^{n} p(x_i|\eta) \prod_{i=1}^{m} p(y_i|x_i,\alpha)\omega(y|\psi) \prod_{i=m+1}^{n} \int p(y|x_i,\alpha)(1-\omega(y|\psi))dy \\
&= \prod_{i=1}^{n} p(x_i|\eta) \prod_{i=1}^{m} \frac{p(y_i|x_i,\alpha)}{\int p(y_i|x,\alpha)p(x|\eta)dx} \\
&\times \prod_{i=1}^{m} \omega(y_i|\psi) \int p(y_i|x_i,\alpha)p(x_i|\eta)dx \prod_{m+1}^{n} \int p(y|x_i,\alpha)(1-\omega(y|\psi))dy
\end{aligned}$$

From (3.1), the complete cases are a random sample of the conditional distribution of X given Y, Tang et al. (2003) considered the following conditional likelihood to make inference on α.

$$L(\alpha;\eta) \propto \prod_{i=1}^{m} \frac{p(y_i|x_i,\alpha)}{\int p(y_i|x,\alpha)p(x|\eta)dx} \tag{3.2}$$

Where α is the parameters of interest and η is the nuisance parameter from the marginal distribution of X.

This term of (3.2) is completely factored from the distribution of the missingness and the nuisance parameter of η. Because one only concerns the estimation of the parameters of interest, α, a natural approach is to substitute the nuisance parameter η in (3.2) by a consistent estimate $\hat{\eta}$. This leads to a pseudolikelihood function in (3.3).

$$L_2(\alpha; \hat{\eta}) = \prod_{i=1}^{m} \frac{p(y_i|x_i, \alpha)}{\int p(y_i|x_i, \alpha) p(x_i; \hat{\eta}) dx} \tag{3.3}$$

If the parameteric form of $p(X; \eta)$ is unknown, one can use empirical distribution instead of $p(X; \eta)$. The distribution function of $[X]$ is denoted $F(x)$ and the empirical distribution of $[X]$ is denoted $F_n(X)$. In the following context, we consider the pseudolikelihood method that substitutes F(x) by its empirical estimate $F_n(X)$, and corresponding pseudolikelihood function becomes (3.4).

$$L_2(\alpha; F_n(x)) \quad \propto \quad \prod_{i=1}^{m} \frac{p(y_i|x_i, \alpha)}{\frac{1}{n} \sum_{j=1}^{n} p(y_i|x_j, \alpha)} \tag{3.4}$$

According to the above expressions, parameter estimates of interest $\hat{\alpha}$ are defined as follows:

$$\hat{\alpha} \quad = \quad \arg\max_{\alpha} \sum_{i=1}^{m} [\log p(y_i|x_i; \alpha) - \log\{\frac{1}{n} \sum_{j=1}^{n} p(y_i|x_j, \alpha)\}]$$

By the above definitions, the score functions are obtained by taking the first derivatives about α from the log pseudolikelihood function, and the regression parameter estimates of $\hat{\alpha}$ are computed by setting these score functions to zeros. However, these score functions do not help to estimate the parameters of interest due to a complicated form of the denominator in the pseudolikelihood function, so the regression parameters of interest should be numerically obtained. Asymptotically, these regression parameter estimates $\hat{\alpha}$ follow a multivariate normal distribution as follows:

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \xrightarrow{d} N(0, \Sigma), \quad \text{as} \quad N \to \infty$$

The estimate of the covariance matrix, $\hat{\Sigma}$ was derived by Tang et al. (2003), and is referred to Appendix A. in detail. This estimate of variance of the PL estimates is used to estimate the first moment and the second moment later in a simulation study in the dissertation.

3.2 THE EXTENSION TO ANALYSIS OF MULTIVARIATE MONOTONE MISSING DATA

The pseudolikelihood method for multivariate missing data is conducted by extending the pseudolikelihood method for bivariate not–missing at random missing data described in the previous section to a multivariate dataset. Generally, the pseudolikelihood method factors the joint conditional distribution of k–variate data as k different pseudolikelihood functions where each function has a distinct parameter of interest, and estimates this distinct parameter of interest from the corresponding pseudolikelihood function. This procedure can be illustrated with k-variate data as follows.

Suppose a multivariate dataset of $\{x_i, y_i\}$ for $i = 1, \cdots, n$ such that a covariate X is fully observed and a dependent variable Y is partially but monotonely missing where Y is a k–dimension vector as $Y = \{Y_1, \cdots, Y_k\}$. Then, the missing data indicator of R is defined as a k–dimension vectore as $R = \{R_1, \cdots, R_k\}$ where R_i corresponds to Y_i. Namely, R_i is defined as 1 if Y_i is observed, and R_i is defined as 0 if Y_i is not observed. Because the missing pattern is monotone, the dataset has $k-1$ missing patterns from k–variate dependent variable, and $R_i = j$ indicates that subject i belongs to a j missing pattern such that $\{y_{i,1}, \cdots, y_{i,j}\}$ are observed and $\{y_{i,j+1}, \cdots, y_{i,k}\}$ are missing. The missing–data mechanism of the j missing pattern is specified as (3.5) where j is between 1 and $k-1$ and $\omega_j(\cdot)$ indicates a arbitrary function of Y_j.

$$P[R = j|x, y_1, \cdots, y_k, R \geq j] = \omega_j(Y_j) \tag{3.5}$$

The joint conditional distribution of these multivariate data of $\{Y_1, \cdots, Y_k|x; \alpha\}$ is expressed as follows where parameters of $\alpha = \{\alpha_1, \cdots, \alpha_k\}$ are distinct parameters of interest as (3.6). We assume that the each factorized distribution on (3.6) follows a different known parametric distribution, and this known parametric distribution is denoted as $g_j(\cdot)$ for $j = 1, \cdots, k$ and is plugged in the pseudolikelihood function instead of the conditional distribution on (3.6).

$$
\begin{aligned}
p[Y_1, \cdots, Y_k | x; \alpha] &= p[Y_1|x; \alpha_1] p[Y_2|y_1, x; \alpha_2] \cdots p[Y_k|y_1, \cdots, y_{k-1}; \alpha_k] \\
&= p[Y_1|x; \alpha_1] p[Y_2|Y_1, \cdots, Y_k, x, R \geq 2; \alpha_2] \\
&\cdots \\
&\quad p[Y_k|Y_1, \cdots, Y_k, x, R \geq k; \alpha_k] \qquad (3.6)
\end{aligned}
$$

The pseudolikelihood function of these multivariate data are driven from these factorized conditional distributions of the multivariate data, $\{Y_1, \cdots, Y_k|x; \alpha\}$, on (3.6). Analogous to the pseudolikelihood function from a bivariate missing data on (3.3), the pseudolikelihood function of k–variate data can be set up with k pseudolikelihood functions for $j = 1, \cdots, k$ as follows because the parameters of interest, $\alpha = \{\alpha_1, \cdots, \alpha_k\}$, are distinct.

$$
\begin{aligned}
L_1(\alpha_1) &= \prod_{i=1}^{n} g_1(Y_1|x; \alpha_1) \\
L_2(\alpha_2) &= \prod_{R \geq 2} \frac{g_2(Y_2|Y_1, x; \alpha_2)}{\int g_2(Y_2|Y_1, x; \alpha_2) dF_n(x, y_1)} \\
&\cdots \\
&\cdots \\
L_k(\alpha_k) &= \prod_{R \geq k} \frac{g_k(Y_k|y_1, \cdots, y_{k-1}; \alpha_k)}{\int g_k(Y_2|y_1, \cdots, y_{k-1}; \alpha_k) dF_n(x, y_1, \cdots, y_{k-1})}
\end{aligned}
$$

According to (3.7), one has k pseudolikelihood functions. Because each function is only associated with one parameter of interest out of α, each associated α_i is computed with the pseudolikelihood function of $L_i(\alpha_i)$. Same as the pseudolikelihood method with a bivariate missing data, one can obtain k log pseudolikelihood functions by taking logarithm to these k pseudolikelihood functions where the log pseudolikelihood function about the parameter of interest, α_i, is denoted as $l_i(\alpha_i)$. Hence, the regression parameter estimates $\hat{\alpha} = \{\hat{\alpha}_1, \cdots, \hat{\alpha}_k\}$ are obtained by maximizing corresponding log pseudolikelihood functions as (3.7).

$$
\hat{\alpha}_k = \arg\max_{\alpha_k} l_k(\alpha_k) \quad \text{where} \quad k = 1, \cdots, K \qquad (3.7)
$$

Both these pseudolikelihood functions and log pseudolikelihood functions have complicated forms of the denominators like the case of the bivariate missing data, so one can not expect the regression parameter estimates as closed–forms by setting the score functions to zeros where the score functions are defined as log pseudolikelihood functions that are taken the first derivative about the corresponding parameters of interest. Therefore, the regression parameter estimate $\hat{\alpha}_k$ should be numerically computed, and these regression parameter estimates follow aysmptotic normal distribution also.

4.0 THREE METHODS FOR VARIANCE ESTIMATION UNDER THE PSEUDOLIKELIHOOD METHOD

Consider regression analysis of a dataset where all covariates are observed and the response variable is partially observed. If the missingness only depends on the response itself, consistent estimates and the corresponding asymptotic variance matrix of the regression parameter estimates can be obtained by the pseudolikelihood method (Tang et al. 2003). However, because the empirical process of the covariates are involved in the pseudolikelihood method, the regression parameter estimates and the covariance matrix of these regression parameter estimates are computationally very intensive.

Suppose that one concerns estimating the variance of the PL estimates of a function of missing data Y. We denote a function of the missing data as $h(X, Y)$, and denote the function of interest as ϕ, which is the expectation of a function from missing data, $E[h(X, Y)]$. $E[h(X, Y)]$ can be expressed as the function of the distribution of $[X]$ and the regression parameters of $[Y|X]$. Then the function of interest ϕ is formally expressed as follows:

$$\begin{aligned} \phi &= E[h(X,Y)] \\ &= E[E[h(X,Y)|x]] \\ &= E[\int_{-\infty}^{\infty} h(X,y) \cdot g(y|X;\alpha)dy] \end{aligned} \tag{4.1}$$

$g(\cdot)$ indicates a known parametric function, and α is the unknown regression parameters from the conditional distribution of $[Y|X]$ on (4.1). From (4.1), the estimate of the function of interest, $\hat{\phi}$ can be obtained by replacing unknown regression parameters of α to the estimates of the regression parameters, $\hat{\alpha}$ as follows where $\hat{\alpha}$ is computed by the pseudolikelihood method.

$$\hat{\phi} \quad = \quad \frac{1}{n} \cdot \sum_{i=1}^{n} [\int_{-\infty}^{\infty} h(x_i, y) \cdot g(y|x_i; \hat{\alpha}) dy] \tag{4.2}$$

According to the formalization of (4.1) and (4.2), we can compute the variance estimate of the function of interest with missing data. One of the most widely used standard methods for variance estimation is the Delta method with a missing dataset. The Delta method is an analytical method to estimate a variance of a function of interest using the first order approximation of the Taylor series expansion. The Delta method uses the extended covariance matrix estimates between the regression parameter estimates and covariates, and derives the variance estimate of the function of interest from this extended covariance matrix estimate and the first derivatives of the function of interest about all related variables. The theory of the Delta method is mathematically solid, and if the covariance matrix and the first derivatives about the regression parameters are easy to compute, the Delta method is a good way to obtain the variance of the estimates. However, the covariance matrix of the regression parameter estimates has a complicated form under the pseudolikelihood method, so its computation is not simple. In addition, this function of interest is associated with empirical distribution of covariates, so the application of the Delta method to the function of the PL estimates is computationally very intensive in practice.

Another method to estimate the variance of the function of interest is the Bootstrap. The Bootstrap is a resampling technique that generates random samples of missing data with replacement. One estimates the regression parameters per each bootstrap sample, and computes the functions of interest with the regression parameter estimates. Then, one can compute the sample variance among the functions of interest, and this sample variance becomes the variance estimate of the function of interest in the Bootstrap method. This method is known to provide a consistent variance estimate especially as the number of sample size or the number of bootstrap samples increases (Efron, 1979). However, if the regression parameters need to be estimated by the pseudolikelihood method, the Bootstrap needs a procedure to numerically estimate the regression parameters from each bootstrap sample, which takes a lot of the computation time in practice.

Besides these standard methods, a direct resampling method is newly developed in the dissertation. Because the Delta method and the Bootstrap have their own difficulties in practice, the direct resampling method attempts to address some of these difficulties. This method is designed to draw multiple samples of parameter estimates from the asymptotic normal distribution of the regression parameter estimates, so one can expect to save the computation time which is required to estimate the regression parameters by the pseudolikelihood method in the Bootstrap.

We review these three methods, and study their properties as well as their procedures in this chapter. These three methods are employed to estimate the variances of the PL estimates of the first moment and the second moment of Y in a simulation study, and their performances are compared with regard to averages of 95% confidence interval widths and 95% coverage rates. In addition, we examine the advantages and the disadvantages of these three methods in practice.

4.1 DESCRIPTION OF THE THREE METHODS

4.1.1 The Delta Method

The Delta method is a widely used method to estimate the variance of a function of interest based on the first order approximation of the Taylor series expansion, and can be applied to both univariate and multivariate data. The Delta method derives an approximate probability of distribution function for a function of the asymptotic normal estimator from limiting the variance of that estimator, and provides an analytical solution about the variance of the function of interest.

Denote the first set of parameters and the corresponding estimates as α and T. Assume that T follows an asymptotic normal distribution as follows:

$$\sqrt{n}(T-\alpha) \rightarrow \quad N(0,\Sigma) \tag{4.3}$$

For $h(\alpha)$, a smooth function of α, the natural estimate is h(T). Then, by the Delta method, one can compute the asymptotic variance of $h(T)$ as follows:

$$h(T) - h(\alpha) = (T - \alpha)(\frac{\partial h(\alpha)}{\partial \alpha}) + o(1) \tag{4.4}$$

$$\sqrt{n}(h(T) - h(\alpha)) \rightarrow N(0, (\frac{\partial h(\alpha)}{\partial \alpha})^T \hat{\Sigma} (\frac{\partial h(\alpha)}{\partial \alpha})) \tag{4.5}$$

If the distribution of the missing–data mechanism is (3.1), one can obtain the regression parameter estimates and corresponding covariance matrix estimate of these regression parameter estimates by the pseudolikelihood method. However, the covariance matrix estimate of the regression parameter estimates, $\hat{\Sigma}$, has a complicated form, which requires a computationally intensive procedure. In addition, if the function of interest is a function of the PL estimates, the extended covariance matrix for the Delta method is more complicated where it consists of the empirical distribution of covariates and the regression parameter estimates. Therefore, the Delta method is a computationally intensive practice. We conduct the Delta method to estimate the first moment and the second moment in a simulation study after estimating the regression parameters by the pseudolikelihood method, and examine the advantages and the disadvantages in practice.

4.1.2 The Bootstrap Method

The Bootstrap, which was first introduced by Efron (1979), generates random samples with replacement from an independent and identically distributed dataset. Unlike the Delta method, this method is a computer–intensive resampling method, and is known to perform better and consistent relatively in comparison to other non–parametric techniques (Efron, 1981). The Bootstrap is simple and straightforward to derive estimates of variances and confidence intervals although a function of interest is composed of a complex form of the parameters.

Let ϕ denote a function of interest which is composed of the regression parameter estimates $\hat{\alpha}$, and one is interested in estimating the variance of ϕ. If $\hat{\phi}$ indicates an estimate of ϕ from the original dataset, $\hat{\phi}_{boot}$ indicates a Bootstrap estimate of ϕ. Accordingly, $V(\hat{\phi})$

denotes the variance of the estimate of the function, and V_{boot} denotes the Bootstrap variance estimate of the estimate of the function. When one conduct the Bootstrap to estimate the variance of ϕ, it is known that the Bootstrap estimate of $\hat{\phi}_{boot}$ is less biased than the estimate of the function of interest with the original dataset, $\hat{\phi}$, (Little& Rubin, 2002), and the Bootstrap variance estimate, V_{boot} is known as a consistent estimate especially as the sample size of the original dataset, N, or the number of the repetition, B, tends to infinity [Efron, 1979, Little& Rubin, 2002].

Suppose that the regression parameters, α, are estimated by the pseudolikelihood method under the assumption of that the distribution of the missing–data mechanism of this dataset depends only on the response variables. If one conducts the Bootstrap to compute the variance of ϕ with this dataset, one has to compute the regression parameter estimates by the pseudolikelihood method per each bootstrap sample after generating the random samples with replacement from the original dataset. With these regression parameter estimates obtained from the bootstrap samples, the Bootstrap estimate of the function of interest is computed from each bootstrap samples, and the Bootstrap variance estimate is computed. Therefore, the computation procedure is multiplied as much as the number of the bootstrap samples in this case, and the computation time increases a lot for implementing the Bootstrap with not–missing at random missing data.

The general procedure of the Bootstrap with missing data is summarized to estimate the variance of a function of interest as follows where regression parameter estimates and a function of interest are denoted as $\hat{\alpha}$ and ϕ. The sample size of the original dataset is N, and the number of the repetition of the Bootstrap is B. The original dataset is denoted as $\mathcal{D}$, and the Bootstrapping datasets are denoted as $\mathcal{D}^{(b)}$ for $b = 1, \cdots, B$.

- Step 1. Generate a sample $\mathcal{D}^{(b)}$ with replacement from the original missing dataset of $\mathcal{D}$.
- Step 2. Estimate the regression parameters, $\hat{\alpha}^{(b)}$ from the bootstrap sample of $\mathcal{D}^{(b)}$.
- Step 3. Estimate the function of interest, $\hat{\phi}^{(b)}$ based on $\mathcal{D}^{(b)}$.
- Step 4. Repeat Step 1.– Step 3. for $b = 1, \cdots, B$.
- Step 5. Compute the bootstrap estimate of $\hat{\phi}_{boot} = \frac{1}{B}\sum_{b=1}^{B}\hat{\phi}^{(b)}$.
- Step 6. Compute the bootstrap variance estimate $\hat{V}_{boot} = \frac{1}{B-1}\sum_{i=1}^{B}(\hat{\phi}^{(b)} - \hat{\phi}_{boot})^2$

4.1.3 A Direct Resampling Method

A direct resampling method is newly introduced here to estimate a variance of a function of interest in the dissertation. The direct resampling method depends on a repeated sampling technique like the Bootstrap, but one can expect less computation time than the Bootstrap when data are incomplete. This direct resampling method is designed to draw the samples from asymptotic normal distribution of regression parameter estimates and the samples of covariates by the Bootstrap. Then, predictive values for missing values are generated from normal distribution where the parameters are made up of the samples of the regression parameter estimates and the covariates, and these predictive values are used to estimate the function of interest and corresponding sample variance.

Let α denote regression parameters and ϕ is the function of interest where ϕ is a function of the parameter α. The consistent estimate of the parameter, $\hat{\alpha}$, is obtained from missing data, and $\hat{\alpha}$ asymptotically follows normal distribution such that the expectation is α, and the variance is Σ. Then the general procedure of the direct resampling method is summarized to estimate the variance of the function of interest as follows:

- Step 1. Estimate consistent estimates of $\hat{\alpha}$ from the original dataset.
- Step 2. Obtain the asymptotic distribution of $\hat{\alpha}$.
- Step 3. Randomly draw a sample of $\alpha^{(b)}$ from $N(\hat{\alpha}, \hat{\Sigma})$.
- Step 4. Generate a bootstrap sample from covariates X.
- Step 5. Randomly draw $\{y_1^{(b)}, \cdots, y_n^{(b)}\}$ from asymptotic normal distributions whose means and variances are composed of $\alpha^{(b)}$ and the bootstrap samples of the covariates.
- Step 6. Estimate $\hat{\phi}^{(b)}$ using $\{y_1^{(b)}, \cdots, y_n^{(b)}\}$.
- Step 7. Repeat Step 1.– Step 6. for $b = 1, 2, \cdots, B$.
- Step 8. Compute the resamling estimate for the function of interest, $\hat{\phi}_r = \frac{1}{B}\sum_{b=1}^{B} \hat{\phi}^{(b)}$.
- Step 9. Compute the variance estimate by $V_r = \frac{1}{B-1}\sum_{b=1}^{B} (\hat{\phi}^{(b)} - \hat{\phi}_r)^2$.

Based on the above procedure, $\hat{\phi}_r$ and V_r are the direct resampling estimates and corresponding variance estimate of the function of interest. There are some common procedures with the Bootstrap, but the direct resampling method does not require estimating parameters

of α per each repetition time because it directly draws $\hat{\alpha}$ from the asymptotic distribution obtained from the original dataset while the Bootstrap computes $\hat{\alpha}$ from each bootstrap sample. Therefore, once the asymptotic distribution of the parameter estimates from the original dataset is well defined, one can obtain the variance estimate of the function of interest faster than the Bootstrap.

4.2 A SIMULATION STUDY FOR THE THREE METHODS

4.2.1 Simulation Procedure

The simulation is conducted with a bivariate dataset such that the covariate X is fully observed and the response variable Y is partially but monotonely observed. The bivariate datasets have four different sample sizes of 100, 300, 500 and 1000, and 1000 bivariate missing datasets are generated per each different sample size. The missing data are created from the complete bivariate datasets after specifying the designed missing–data mechanism, and the specific procedure for the bivariate missing data is summarized as follows:

- Step 1. Generate randomly the covariate X according to standard normal distribution N(0,1).
- Step 2. Generate the response variable Y based on the conditional distribution of $[Y|X]$ from $N(\beta_0 + \beta_1 \cdot x, \sigma^2)$ where $\alpha = \{\beta_0, \beta_1, \sigma^2\} = (1, 1, 1)$.
- Step 3. Specify the cases whose response variables are missing according to the following mechanism (4.6).

$$P[R = 0|x, y] = \Phi(\psi_0 + \psi_1 \cdot y) \tag{4.6}$$

  Where $(\psi_0, \psi_1) = (-1, 1)$ and $\Phi(\cdot)$ refers to the cumulative distribution function(C.D.F.) of standard normal distribution.
- Step 4. The missing datasets are created by erasing Y values of the specified cases from the datasets.

As a result of the above procedure, about 50% of Y's are missing on average, and these missing datasets are not–missing at random. In the simulation, we consider estimating the variances of two functions of interest. We denote the functions of interest as ϕ_1 and ϕ_2, and ϕ_1 and ϕ_2 are specified the first moment and the second moment of missing data Y. At first, we compute the regression parameter estimates of $\hat{\alpha} = \{\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2\}$ and the covariance matrix estimates of these parameter estimates, $\hat{\Sigma}$, by the pseudolikelihood method to compute the variance estimates of two specified moments. The first moment and the second moment are expressed as (4.7) and (4.8).

$$\phi_1 \quad = \quad \beta_0 + \beta_1 \cdot \mu_x \tag{4.7}$$

$$\phi_2 = \beta_1^2 \cdot \mu_2 + \sigma^2 + \beta_0^2 + 2\beta_0 \cdot \beta_1 \cdot \mu_x \tag{4.8}$$

Where $\mu_x = E[X]$ and $\mu_2 = E[X^2]$.

The estimates of the functions of interest are denoted as $\hat{\phi}_1$ and $\hat{\phi}_2$, and they are obtained by replacing the regression parameters to the regression parameter estimates as follows:

$$\hat{\phi}_1 = \hat{\beta}_0 + \hat{\beta}_1 \cdot \hat{\mu}_x \tag{4.9}$$

$$\hat{\phi}_2 = \hat{\beta}_1^2 \cdot \hat{\mu}_2 + \hat{\sigma}^2 + \hat{\beta}_0^2 + 2\hat{\beta}_0 \cdot \hat{\beta}_1 \cdot \hat{\mu}_x \tag{4.10}$$

Where $\hat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$, and n is the sample size.

The variances of these estimates are obtained by three methods : One is the Delta method, another is the Bootstrap, and the other is the direct resampling method. According to (4.7) and (4.8), ϕ_1 is a function of $\{\mu_x, \alpha\}$, and ϕ_2 is a function of $\{\mu_x, \mu_2, \alpha\}$. The Delta method computes the variances of $\hat{\phi}_1$ and $\hat{\phi}_2$ by taking the first derivatives about those components of the functions of interest with the extended covariance matrix estimates. The Delta method directly compute the variance estimates, so it does not provide estimates of the functions of interest unlike the Bootstrap and the direct resampling method. The specific procedures of the Delta method for the first moment and the second moment of Y are referred to Appendix A–Appendix B in detail. Unlike the Delta method, the Bootstrap and the direct resampling method are computer–intensive techniques, and have multiple regression parameter estimates, $\hat{\alpha}^{(b)}$ for $b = 1, \cdots, B$ in different ways where B is the number of repetition. These two methods compute multiple functions of interest from multiple regression parameter estimates. The Bootstrap and the direct resampling method compute the variances of $\hat{\phi}_1$ and $\hat{\phi}_2$ as sample variances among the multiple functions of interest. The Bootstrap and the direct resampling method are conducted with repetition of two different numbers 50 and 128 in the simulation.

The specific procedure for $\hat{\phi}_1$ by the Bootstrap method is summarized as follows:

1. Generate a random sample of $\mathcal{D}^{(b)}$ from the original sample $\mathcal{D}$ with replacement.
2. Estimate the regression parameters of $[\hat{\beta}_0^{(b)}, \hat{\beta}_1^{(b)}, \hat{\sigma}^{2(b)}]$ using the pseudolikelihood method which is described on (3.3) and $\hat{\mu}_x^{(b)}$ on the bootstrap sample of $\mathcal{D}^{(b)}$.
3. Compute $\hat{\phi}^{(b)} = \beta_0{}^{(b)} + \beta_1^{(b)} \cdot \mu_x^{(b)}$ on the bootstrap sample of $\mathcal{D}^{(b)}$ where $\mu_x^{(b)} = \frac{1}{N}\sum_{i=1}^{N} x_i^{(b)}$.
4. Repeat 1.–3. for $b = 1, 2, \cdots, B$
5. Compute the bootstrap estimate of $\hat{\phi}_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B} \hat{\phi}^{(b)}$ from $\{\hat{\phi}^{(1)}, \hat{\phi}^{(2)}, \cdots, \hat{\phi}^{(B)}\}$.
6. Compute the estimate of the variance,$\hat{V}_{\text{boot}} = \frac{1}{B-1}\sum_{b=1}^{B} (\hat{\phi}^{(b)} - \hat{\phi}_{\text{boot}})^2$ from $\{\hat{\phi}^{(1)}, \hat{\phi}^{(2)}, \cdots, \hat{\phi}^{(B)}\}$.

The specific procedure for $\hat{\phi}_1$ by the direct resampling method is summarized as follows:

1. Derive the regression parameter estimates, $\hat{\boldsymbol{\alpha}} = [\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2]$ of $[Y|x]$ by the pseudolikelihood method, and estimates the variance of $\hat{\boldsymbol{\alpha}}$ using (A.8).
2. Randomly draw $\hat{\boldsymbol{\alpha}}^{(b)} = [\hat{\beta}_0^{(b)}, \hat{\beta}_1^{(b)}, \hat{\sigma}^{2(b)}]$ from $N(\hat{\boldsymbol{\alpha}}, \hat{Var}(\hat{\boldsymbol{\alpha}}))$ with restriction of $(\hat{\sigma}^2)^b > 0$.
3. Generate random samples of $\{x_1^{(b)}, \cdots, x_n^{(b)}\}$ with replacement from X.
4. Randomly draw $y_i^{(b)}$ from $N(\beta_0{}^{(b)} + \beta_1{}^{(b)} \cdot x_i{}^{(b)}, \sigma^{2(b)})$.
5. Estimate the function of interest, $\hat{\phi}^{(b)} = \frac{1}{N} \cdot \sum_{i=1}^{N} y_i^{(b)}$ from (4).
6. Repeat 1.–5. for $b = 1, 2, \cdots, B$.
7. Compute the mean of the estimate of parameter of interest, $\hat{\phi}_d = \frac{1}{B} \cdot \sum_{b=1}^{B} \hat{\phi^{(b)}}$.
8. Compute the sample variance, $\hat{Var}(\hat{\phi}) = \frac{1}{B-1} \cdot \sum_{b=1}^{B} (\hat{\phi^{(b)}} - \hat{\phi}_d)^2$ from $\{\hat{\phi}^{(1)}, \hat{\phi}^{(2)}, \cdots, \hat{\phi}^{(B)}\}$.

When the function of interest is the second moment of missing data Y, the $\hat{\phi}^{(b)}$ is based on (4.8) for $b = 1, \cdots, B$ on the above procedures of the Bootstrap and the direct resampling method. After we obtain the variance estimates of the first moment and the second moment by three different methods, we compare their performances with 1000 bivariate datasets with regard to averages of 95% confidence interval widths and 95% coverage rates. Averages of biases are also computed.

4.2.2 Simulation Results

We estimate the variances of the first moment and the second moment with 1000 bivariate datasets of sizes 100, 300, 500 and 1000 by three different methods in the simulation. One method is the Delta, another is the Bootstrap and the other is the direct resampling method. According to the above procedures of three methods, we obtained averages of biases, averages of 95% confidence interval widths and 95% coverage rates.

Table 2 and Table 3 show averages of biases, averages of 95% confidence interval widths and 95% coverage rates for the first moment and the second moment which are obtained with 1000 bivariate missing datasets by the different sample sizes. True values of the first moment and the second moment are 1 and 3, and the averages of 95% confidence interval widths are multiplied by 1000 in the tables. Averages of biases are computed by the average differences of the true values from the estimates with 1000 datasets, and those values are negligible for both estimates of the first moment and the second moment regardless of the sample sizes on Table 2 and 3. Therefore, we concern more about averages of 95% confidnece interval widths and 95% coverage rates than averages of biases. According to Table 2 and 3, we can see that the averages of 95% confidence interval widths become similar among these three methods as the sample size increases. Also, 95% coverage rates of the first moment and the second moment are distributed around 95% and none go below 90%, so these three methods show stable performances with regard to the coverage rates. In addition, there are no distinct repetition effect between 50 and 128 in both the Bootstrap and the direct resampling method in our simulation, and this result is consistent with Efron's (1979).

However, these values on Table 2 and 3 are overestimated values. Because the regression parameter estimates are numerically obtained, and the estimate of the covariance matrix of the regression parameter estimates is not expressed as a closed form, one can encounter some difficulties to estimate the variances of the estimates of two moments in practice. If the regression parameter estimates are too different from the true values in a dataset, then the dataset is considered having a convergence problem, and the dataset and the regression parameter estimates are excluded from the computation. Also, if a covariance matrix does not satisfy the positive definite condition, the dataset are also excluded from the computa-

tion. The Delta method is an analytical way to compute variances based on the Taylor series expansion. This method is theoretically solid and it does not require repeated calculation. However, when the missing–data mechanism depends on a function of response variables and the regression parameters are estimated by the pseudo–likelihood, the Delta method is not the best way to use for the variance estimation. Because a covariance matrix by the pseudolikelihood method is not a closed form, the Delta method is computationally intensive to derive the variance estimates of the moments from the covariance matrix of the regression parameter estimates. Compared to the Delta method, one does not have to estimate the complicated covariance matrix of the regression parameter estimates in the Bootstrap. The Bootstrap randomly generates multiple samples with replacement, and estimates regression parameters from the generated samples, and computes the moments and their variances of these moments. Because the variances of these moments are estimated with the sample variances from the samples, the variance estimation is straightforward in the Bootstrap. However, the regression parameters should be estimated with each bootstrap sample, so the Bootstrap needs a large amount of computation time to search the regression parameter estimates per each sample by the pseudolikelihood method when the missing–data mechanism depends on a function of response variables. For example, it takes about 25 hours to compute 1000 datasets of size 300 with 50 bootstrap samples per each dataset on Table 1. The direct resampling method attempts to reduce a large amount of the computation time by a repeated calculation. This method is based on the resampling technique, but it generates regression parameter estimates directly from the asymptotic normal distribution unlike the Bootstrap. Although this method needs a repeated computation, but its computation time is much less than the Bootstrap because the regression parameters are directly drawn from a normal distribution. Also, the direct resampling method does not require computing the covariance matrix between the PL estimates $\hat{\alpha}$ and $\hat{\mu}_x$ or the covariance between $\hat{\alpha}$ and $(\hat{\mu}_x, \hat{\mu}_2)$ unlike the Delta method. However, it has a similar difficulty to the Delta method in that one has to compute the asymptotic covariance matrix of the regression parameter estimates by the pseudolikelihood method.

Table 1 shows the computation time of the three methods in our simulation, and the computation time is measured with CPU time of the three methods with 1000 datasets of

Table 1: Computing Time Comparison

Computing Time	Delta Method	Bootstrap	Resampling
Hour : Minute :Second	2 : 01 : 48	25 : 23 : 20	2 : 01 : 31

size 300. Regarding this computation time, these three methods have advantages and disadvantages to apply to not–missing at random missing data under pseudolikelihood in practice.

Table 2: The Variance Estimate For The First Moment Under Three Methods

N	B	Bias	Delta Method		Bootstrap		Resampling	
			Width	Cvg	Width	Cvg	Width	Cvg
100	50	-0.0030	1246	972	982	940	1163	966
	128				981	938	1156	966
300	50	-0.0020	607	966	546	945	566	950
	128				551	946	561	949
500	50	0.0026	462	959	451	949	457	951
	128				452	950	454	950
1000	50	0.0006	309	961	308	961	309	961
	128				308	961	307	960

Table 3: The Variance Estimate For The Second Moment Under Three Methods

N	B	Bias	Delta Method		Bootstrap		Resampling	
			Width	Cvg	Width	Cvg	Width	Cvg
100	50	0.0277	7339	971	6988	952	6611	956
	128				7649	963	6776	959
300	50	0.0018	2786	956	2685	952	2690	954
	128				2688	952	2701	954
500	50	0.0102	2112	961	2101	956	2079	951
	128				2107	956	2086	952
1000	50	0.0125	1163	957	1154	953	1121	943
	128				1156	953	1116	941

5.0 IMPUTATION METHODS UNDER THE PSEUDOLIKELIHOOD METHOD FOR BIVARIATE MISSING DATA

5.1 INTRODUCTION

Most statistical analysis methods require the complete data, but missing data frequently occur in many areas of research with various reasons such as non–response in survey data, or drop–outs in clinical trial data. The methods to analyze these missing data are relatively few, and imputation methods are one of them.

Consider a bivariate dataset $\{X, Y\}$ where an independent variable of X is fully observed and a response variable of Y is partially observed. The missing data are not–missing at random (NMAR), and we assume that the distribution of the missing–data mechanism is only composed of a dependent variable and the conditional distribution of $[Y|X;\alpha]$ follows a parametric distribution. Under these assumptions, one can estimate regression parameters of α by the pseudolikelihood method without specifying the missing–data mechanism with this dataset. However, the pseudolikelihood method by Tang (2003) has some difficulties to perform in practice. Suppose that one is interested in estimating the variance of the PL estimates of any function of missing data Y. Let $h(X, Y)$ denote an arbitrary function of the missing data of $\{X, Y\}$, and $E[h(X, Y)]$ be the expectation of this function. $E[h(X, Y)]$ is the function of interest ϕ, and $\hat{\phi}$ is computed by the pseudolikelihood method. This specific problem of interest is to estimate the variance of $\hat{\phi}$. One can consider the standard methods to estimate the variance of ϕ such as the Delta method and the Bootstrap for this problem. However, these standard methods have some difficulties to be employed in practice when the missing–data mechanism depends on a function of response variables and the regression parameter estimates of $\hat{\alpha}$ are obtained by the pseudolikelihood method.

As mentioned in chapter 4, the covariance matrix of regression parameter estiamtes by the pseudolikelihood method is very complicated. The Delta method derives the variance estimate of this function from the covariance matrix estimate of the regression parameter estimates and the first derivatives about regression parameters, α, so the computation by the Delta method is computationally very intensive. The Bootstrap has also difficulty in implementation with missing data. Because regression parameter estimates have to be numerically searched per each bootstrap sample, it takes a lot of time in implementation with missing data when the missing–data mechanism depends on the response variables.

Besides these standard methods, the imputation method can be used for the variance estimation. The imputation method is simple. One can draw predictive values from a formal statistical model, and replace missing values with them. Once imputation is completed, one can apply standard statistical analysis with this imputed data. Therefore, the imputation method can be more advantageous for the variance estimation of the PL estimates. But many imputation methods assume missing at random (MAR), so these imputation methods bring about severe biases in estimation with not–missing at random missing data. Tang devised a mean imputation method for NMAR multivariate normal data (2002). The missing–data mechansim of missing data is assumed to be only expressed with functions of dependent variables, and the conditional distribution of $[Y|X;\alpha]$ is assumed to be known parameteric distribution with unknown paramters of α. He estimated predictive values that are drawn from the estimated conditional distribution of the missing values given the observed values for complete cases. He computed parameter estimates of entire population $\hat{\alpha}$ by the pseudolikelihood method, and used Natharaya–Watson regression estimator to derive means from the complete cases. But his approach does not show satisfactory peformance with regard to coverage rates. In this dissertation, we propose a mean imputation method by replacing the NW estimator to piece–wise linear regression estimator from his approach, and newly introduce stochastic imputation method. These imputation methods take into account the population mean in the predictive distribution, so one can prevent a severe bias with NMAR missing data where the estimated condition mean of entire population is predicted with regression parameter estimates obtained by the pseudolikelihood method in the imputation methods. These methods are studied in the following section in detail.

5.2 TWO IMPUTATION METHODS UNDER THE PSEUDO LIKELIHOOD METHOD

5.2.1 A Mean Imputation Method

Suppose a bivariate missing dataset such that a covariate of X is fully observed and a response variable of Y is partially observed. The missing data indicator is denoted as R, and R is defined as 1 if Y is observed, and R is defined as 0 if Y is missing. Based on the definition of the missing data indicator of R, the conditional distribution of entire population of $[Y|x]$ is expressed as (5.1) where $R = 0$ is a group of missingcases and $R = 1$ is a group of complete cases.

$$pr[Y|x] = \sum_{k=0}^{1} pr[Y|x, R = k] \cdot pr[R = k|x] \tag{5.1}$$

From (5.1), $pr[Y|x, R = 0]$ is expressed as follows where $pr[R = 0|x] = 1 - pr[R = 1|x]$.

$$pr[Y|x, R = 0] = \frac{pr[Y|x] - pr[R = 1|x] \cdot pr[Y|x, R = 1]}{(1 - pr[R = 1|x])} \tag{5.2}$$

$E[Y|x, R = 0]$ is derived from above (5.2) as (5.3).

$$\begin{aligned} E[Y|x, R = 0] &= \frac{\int y \cdot pr[Y|x]dy - p[R = 1|x] \int y \cdot pr[Y|x, R = 1]dy}{(1 - pr[R = 1|x])} \\ &= \frac{E[Y|x] - pr[R = 1|x] \cdot E[Y|x, R = 1]}{(1 - pr[R = 1|x])} \end{aligned} \tag{5.3}$$

Using the representation of (5.3), Tang (2002) introduced a mean imputation method for multivariate normal missing data under the assumption of (3.1) on the missing–data mechanism. According to the above representation of (5.3), one needs to specify the estimates of three components of $E[Y|x]$,$pr[R = 1|x]$ and $E[Y|x, R = 1]$. In Tang (2002), $\hat{E}[Y|x]$ were derived from regression parameter estimates of $\hat{\alpha}$ by the pseudolikelihood method. $\hat{pr}[R = 1|x]$ and $\hat{E}[Y|x, R = 1]$ were derived from the kernel regression estimators as (5.4) and (5.5).

$$\hat{pr}[R = 1|x_i] = \frac{\sum_{i=1}^{N} K(\frac{x - x_i}{\hat{h}}) \cdot I(R_i = 1)}{\sum_{i=1}^{N} K(\frac{x - x_i}{\hat{h}})} \tag{5.4}$$

$$\hat{E}[Y|x_i, R=1] = \frac{\sum_{i=1}^{N} K(\frac{x-x_i}{\hat{h}}) \cdot y_i}{\sum_{i=1}^{N} K(\frac{x-x_i}{\hat{h}})} \tag{5.5}$$

In this dissertation, we use a PWL estimator to estimate the conditional mean from the complete cases instead of the kernel regression estimator of (5.5). The piece–wise linear regression (PWL) segments the range of X with the minimum number of breakpoints to consider the changes of the slopes where the breakpoints are defined as the thresholds where the slopes change (Quandt.R, 1958). The piece–wise linear (PWL) regression assumes that (1) a linear model is continuous within a segment and (2) it enforces the continuity at breakpoints (Sprent, 1961). Because PWL is a non–parametric method, its consistent coefficents are determined iteratively as the arguments that minimize the sum of square errors. As the number of breakpoints increases, the flexibility of the model increases, but it needs more computation procedures because the number of coefficients, which we have to estimate, increases together.

We choose the PWL model for $\hat{E}[Y|X, R=1]$ using approximated F–test under the fixed significant level $\alpha = 0.05$ by a forwording algorithm which increases the number of breakpoints by one. Because the sum of square errors decreases as the number of breakpoints increases, this algorithm stops at the model with the minimum possible number of breakpoints under the fixed α. The general procedure for the forwarding algorithm of the piece–wise linear (PWL) model selection can be summarized to compute the estimates of $E[Y|x, R=1]$ on (5.3) as follows :

- Step 1. Begin with no breakpoint, and conduct a linear regression using the whole range of X where a and b are the PWL coefficients. After calculating $\hat{a}$, $\hat{b}$ which minimize the sum of square errors according to the following model, and test approximated F–test under $\alpha_0 = 0.05$.

  $$Y = a + b \cdot X$$

  If p–value from the approximated F–test is significant, one may add breakpoint one more and compare the change of F-statistics value. If F-statistics does not increase, one derives $E[Y|x, R=1]$ using $\hat{a} + \hat{b} \cdot X$.

- Step 2. If the previous model is not significant, we increase the number of breakpoints by one, and set a new model as follows:

$$\begin{aligned} Y &= a + b \cdot X & X \leq x_0 \\ Y &= a + x_0 \cdot (b - b_1) + b_1 \cdot X & X > x_0 \end{aligned}$$

where x_0 is the breakpoint, and it is determinted together with the coefficients of $a, b, b1$ that minimize the sum of square errors by the above model. After numerically estimating the coefficient estimates with the breakpoint, we test the model fitness according to approximated F–test under fixed significant level again. If p–value is significant, but F-statistics is not increased comparing with the previous model, one chooses the previous model and derives $\hat{E}[Y|x, R = 1]$ from the previous model. However, p–value is significant and F-statistics is increased comparing with the previous model, one adds one breakpoint more, and compares F-statistics with current model to make a decision to derive $\hat{E}[Y|x, R = 1]$. But if p–value is not significant, one discards the current model, and increase the number of breakpoints by one and set a new model again.

We can repeat this procedure until a PWL model is significant under the fixed significant level to estimate $\hat{E}[Y|x, R = 1]$. Once one determine the PWL model according to the above algorithm, the estimates of the conditional means $hatE[Y|x, R = 1]$ are computed, and these values are plugged in (5.3) for imputation. $\hat{pr}[R = 1|x]$ is computed using a kernel estimator of (5.4), and $\hat{E}[Y|X]$ is computed using the PL regression estimates. Therefore, predictive values of $\hat{y_i}$ at given x_i by the proposed mean imputation method is derived from (5.6).

$$\hat{y_i} = \frac{\hat{E}[Y|x_i; \alpha = \hat{\alpha}] - \hat{pr}[R = 1|x_i] \cdot \hat{E}[Y|x_i, R = 1]}{(1 - \hat{pr}[R = 1|x_i])} \tag{5.6}$$

5.2.2 A Stochastic Imputation Method

Stochastic imputation methods draw predictive values by adding residuals to the predictive values that are drawn by mean imputation method on (5.6) for missing values in not–mssing at random (NMAR) missing data. Namely, $pr[R=1|x]$ and $E[Y|x,R=1]$ have the same estimators specified in mean imputation method: $\hat{pr}[R=1|x]$ is drawn from a kernel regression estimator of (5.4), and $\hat{E}[Y|x,R=1]$ is drawn from piece–wise linear regression procedure explained in the above section. In addition, population means of $E[Y|x;\alpha]$ are derived from regression parameter estimates of $\hat{\alpha}$ which are computed by the pseudolikelihood method.

Besides these three components on (5.3), one needs to specify a residual term in stochastic imputation method. We assume that residuals follow normal distribution where expectation of the residuals is zero, and variance of the residuals is $Var[Y|x,R=0]$. When missing data are not–missing at random (NMAR), $Var[Y|x,R=0]$ and $Var[Y|x,R=1]$ usually depend on the values of X. Naturally, the conditional variance of $Var[Y|x,R=0]$ can be obtained using similar technique to the previous section. The conditional variance of the population is simply expressed as follows:

$$\begin{aligned} Var[Y|x] &= E[Y^2|x]-E[Y|x]^2 \\ &= (pr[R=1|x]\cdot E[Y^2|x,R=1]+pr[R=0|x]\cdot E[Y^2|x,R=0]) \\ &\quad -(pr[R=1|x]\cdot E[Y|x,R=1]+pr[R=0|x]\cdot E[Y|x,R=0])^2 \\ &= pr[R=1|x]\cdot Var[Y|x,R=1]+pr[R=0|x]\cdot Var[Y|x,R=0] \\ &\quad +pr[R=1|x]\cdot pr[R=0|x]\cdot (E[Y|x,R=1]-E[Y|x,R=0])^2 \end{aligned} \tag{5.7}$$

From (5.7), we can rearrange the above expression about $Var[Y|x,R=0]$ using the relationship between the conditional variance of the population,$Var[Y|x]$ and the variance of $Var[Y|x,R=1]$ as follows:

$$\begin{aligned} Var[Y|x,R=0] &= \frac{Var[Y|x]-pr[R=1|x]\cdot Var[Y|x,R=1]}{(1-pr[R=1|x])} \\ &\quad -\frac{pr[R=1|x]\cdot(1-pr[R=1|x])\cdot(E[Y|x,R=1]-E[Y|x,R=0])^2}{(1-pr[R=1|x])} \end{aligned} \tag{5.8}$$

Therefore, (5.8) is the residual variance for stochastic imputation, and an estimate of the residual is randomly drawn from $N(0, \hat{\text{Var}}[Y|x_i, R=0])$. After randomly drawing a residual estimate from $N(0, \hat{\text{Var}}[Y|x_i, R=0])$, a predictive value of $\hat{y_i}$ at given x_i is drawn from (5.9) for stochastic imputation where $\hat{\epsilon_i}$ denotes an estimate of the residual as follows:

$$\hat{y_i} = \frac{\hat{E}[Y|x_i;\hat{\alpha}] - \hat{pr}[R=1|x_i] \cdot \hat{E}[Y|x_i, R=1]}{(1 - \hat{pr}[R=1|x_i])} + \hat{\epsilon_i} \tag{5.9}$$

5.3 SIMULATION STUDY FOR THE TWO IMPUTATION METHODS

5.3.1 Simulation Procedure

We conduct mean imputation and stochastic imputation, and observe their performances with 1000 bivariate missing datasets. The bivariate missing datasets are generated from bivariate normal distribution, and the missing–data mechanisms of missing datasets are specified as (3.1). We consider four different sample sizes as 100, 300, 500 and 1000 for this simulation, and an independent variable X is fully observed, and a dependent variable of Y is partially observed. The specific procedure for this bivariate missing dataset is summarized as follows:

- Step 1. Generate an independent variable of X randomly according to a parametric distribution function of N(0,1).
- Step 2. Generate Y based on the conditional distribution of $[Y|x]$ which is specified to be $N(\beta_0 + \beta_1 \cdot x, \sigma^2)$ where $\alpha = \{\beta_0, \beta_1, \sigma^2\}$ are set to be $(1, 1, 1)$.
- Step 3. The missing mechanisms of the missing datasets are specified as follows:

  $$P[R|X, Y] = \Phi(\psi_0 + \psi_1 \cdot y)$$

  Where $(\psi_0, \psi_1) = (-1, 1)$, specify the cases whose response variables have missing values according to the above missing–data mechanism.
- Step 4. The missing datasets are created by erasing Y values of the specified cases from the datasets.

With 1000 incomplete datasets, mean imputation and stochastic imputation are performed. Regression parameter estimates $\hat{\alpha} = \{\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2\}$ are computed from the original dataset by the pseudolikelihood method, and the predictive values for mean and stochastic imputations are derived from (5.6) and (5.9) after fixing regression parameter estimates at $\hat{\alpha}$ on (5.6) and (5.9). The detail procedure for these two imputation methods can be summarized as follows:

- Step 1. Compute regression parameter estimates of $\hat{\alpha} = \{\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2\}$ by the pseudolikelihood method from the original data of $[Y|x;\alpha]$.
- Step 2. Derive $\hat{E}[Y|x;\alpha]$ with $\hat{\alpha} = \{\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2\}$, and plug $\hat{E}[Y|x;\alpha]$ in (5.6) and (5.9).
- Step 3. Reproduce a bivariate missing dataset which is the same to the original bivariate missing dataset.
- Step 4. Draw predictive values for missing values in a reproduced bivariate missing dataset from (5.6) and (5.9) and replace the missing values with the predictive values.
- Step 5. Repeat Step 3.–Step 5. for 20 times.
- Step 6. Estimate the variance of the function of interest with 20 complete datasets.

In this simulation, we consider estimating the variances of the first moment and the second moment of missing data Y by mean imputation and stochastic imputation. According to the above procedure, the mean imputation method is the same to single imputation, but the stochastic imputation is mulitple imputation because we have 10 imputed datasets after the imputation. The variance estimates of the first moment and the second moment are computed from these 10 imputed datasets per each original dataset. After implementing these procedures with 1000 missing datasets, we compared their performances of two imputation methods with regard to the imputation biases, averages of 95% confidence interval widths and 95% coverage rates.

5.3.2 Simulation Results

We conduct mean and stochastic imputation methods under the specified algorithm. After imputation, we compute the variance estimates of the first moment and the second moment with the imputed datasets according to (2.7), and examine their performances with regard to imputation biases, averages of 95% confidence interval widths and 95% coverage rates. Under the specified algorithm, we fix the regression parameter estimates of the population to the estimates which are obtained from the original dataset, so the mean imputation is the same to single imputation, but the stochastic imputation is multiple imputation. Table 4 shows the imputation biases, averages of 95% confidence interval widths and 95% coverage rates of the first moment, and Table 5 shows those results of the second moment

Total variance consists of two components in multiple imputation: One is the within imputation variance component, and the other is the between imputation variance component. But the variance of single imputation is only composed of the within imputation variance component. The variance estimate by single mean imputation may not be valid with NMAR missing data because the 'Within imputation variance' component of the complete cases is not generally the same to that of the population. According to Table 4, the imputation biases of the first moment are negligible, but 95% coverage rates based on the variance estimates by the mean imputation are below 90%. Also, the coverage rates of the second moment are poor for the mean imputation, and the imputation estimates of the second moment are negative–biased on Table 5. However, the stochastic imputation results in nominal coverage rates with negligible imputation biases for both the first moment and the second moment. In comparison with Table 2 and 3, averages of 95% confidence interval widths are smaller because these imputations fix the regression parameters $\hat{\alpha}$.

Fig. 1 displays biases and 95% coverage rates obtained by mean and stochastic imputation methods under the specified algorithm. Black lines and blue lines represent the mean imputation method and the stochastic imputation method on the plots. Two upper plots show biases at each different sample sizes for the first moment and the second moment, and two lower plots show 95% coverage rates at each different sample sizes for the first moment and the second moment. The absolute values are used for bias in the plots. In terms of

biases, the second moments by the mean imputation show much higher values than the stochastic imputation. Accordingly, the mean imputation shows poor coverage rates while the stochastic imputation shows the nominal coverage rates.

Table 4: The Variance Estimate of $\hat{\phi}_1$ Under Two Imputation Methods (1)

N	Mean Imputation			Stochastic Imputation		
	Bias	C.I.Width	Cvg.rate	Bias	C.I.Width	Cvg.rate
100	-0.0046	513	914	-0.0060	610	949
300	-0.0020	293	898	-0.0013	351	944
500	0.0026	226	891	0.0024	271	933
1000	0.0009	161	872	0.0013	192	921

Table 5: The Variance Estimate of $\hat{\phi}_2$ Under Two Imputation Methods (2)

N	Mean Imputation			Stochastic Imputation		
	Bias	C.I.Width	Cvg.rate	Bias	C.I.Width	Cvg.rate
100	-0.2898	1099	675	0.0498	1840	933
300	-0.3237	621	480	0.0253	1072	953
500	-0.3232	481	351	0.0232	832	936
1000	-0.3258	338	353	0.0311	592	937

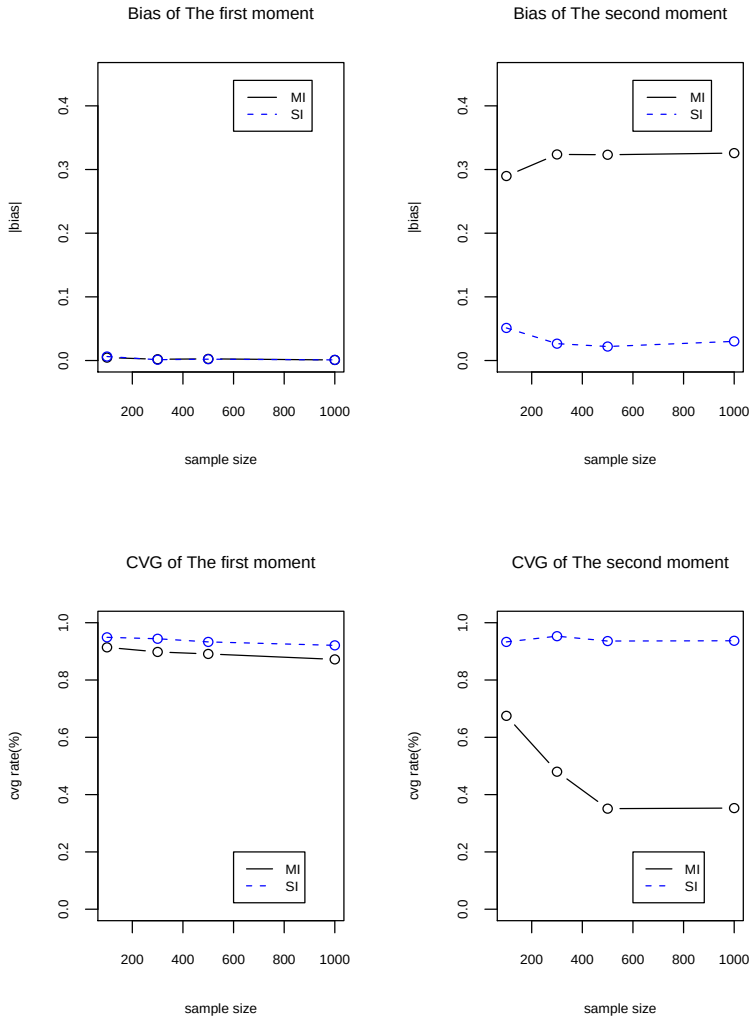

Figure 1: Comparison about biases & 95% coverage rates

6.0 APPLICATION TO A SCHIZOPHRENIA TRIAL

We consider the positive and negative syndrome scale data (PANSS) from the Schizophrenia trial which was introduced in Diggle et al. (2002) as an example to illustrate the imputation methods as well as the standard methods for the variance estimation with missing data. The positive and negative syndrome scale or PANSS is a medical scale used for measuring symptom severity of patients with schizophrenia, where higher PANSS scores indicate more severe symptoms. This data set is collected from a Phase III clinical trial data to compare the different drug regimes of the treatment of schizophrenia by Diggle et al. (2002). The clinical study has a longitudinal study design with five time–points of 1, 2, 4, 6 and 8 weeks besides the baseline and the selection procedure, and the PANSS scores are measured at each designed time–points. The total sample size of the trial is 523 subjects selected between ages of 18 and 65. All subjects are randomly assigned to three different drug regimes of placebo, haloperidol and risperidone. Haloperidol is common drug for schizophorenia patients at present, but risperidone is newly developed drug for the patients. Among these drug regimes, the risperidone group has been treated with four different dosages of 2mg, 6mg, 10mg, and 16mg. The risperidone group has the most subjects and shows faster decrease of PANSS scores by the time than other groups.

We selected the PANSS scores of the risperidone group at the baseline and four weeks for the illustration. From this selected dataset, three cases which have missing values at baseline are excluded, and 345 cases total are selected for our analysis. The PANSS scores at baseline are all observed, and those at four weeks are partially missing. Table 6 shows the frequency of the risperidone group at four weeks where 'Complete' and 'Drop–out' indicate observed cases and missing cases.

Table 6: The Frequency of The Risperidone Group at 4 weeks

	2mg	4mg	10mg	16mg	Total
Complete	49	56	39	55	199
Drop–out	36	31	48	31	146
Total	87	86	87	86	345

146 patients dropout at four week out of 345 patients, so about 42.3% of our dataset have missing values according to Table 6, and the PANSS scores range from 40 to 147 in the dataset. The mean PANSS score of the complete cases is 92.23 at baseline, and that of the missingcases is 92.76 at baseline. The subjects with the higher PANSS scores tend to drop out more at four weeks. With this tendency, we assume that the missing–data mechanism of the selected dataset follows (3.1) and the distribution of the PANSS scores follows a parametric density function.

Let Y_4 denote the PANSS scores of the risperidone group at four weeks. The conditional distribution of $[Y_4|y_1]$ are assumed to follow normal distribution as (6.1) where $[\beta_0, \beta_1, \sigma^2]$ are unknown parameters.

$$p[Y_4|y_1] \quad \sim \quad N(\beta_0 + \beta_1 \cdot y_1, \sigma^2) \tag{6.1}$$

The parameter estimates of $[\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2]$ are numerically obtained by the pseudolikelihood method described in (3.7) without specifying the missing–data mechanism, and the covariance matrix estimate of these parameter estimates is also computed. Once we obtain both the parameter estimates and the covariance estimate of these paramter estimates, we can compute the variances of the functions of interest ϕ where $\phi = E[Y_4]$. We employ two imputation methods as well as three standard methods to estimate the variances of the first moment of Y_4 with our dataset. Two imputation methods are mean imputation and stochastic imputation. Three standard methods are the Delta, the Bootstrap and the direct resampling method.

The same algorithm in chapter 5 is applied to the stochastic imputation with this dataset. This algorithm does not consider the variability of the regression parameter estimates from the conditional distribution of $[Y_4|y_1]$. The population paramter estimates of $[\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2]$ are fixed after obtaining them from the original dataset. Because the mean imputation according to the specified algorithm in chapter 5 does not provide a valid inference about the variance estimates of the functions of interest, we do not perform the mean imputation under this algorithm. Instead, we generated 20 bootstrap samples, and estimates regression parameter estimates of the population with these bootstrap samples according to (6.1). With these twenty different population parameters, we conducted mean imputation. After the mean and the stochastic imputations, we estimated the first moment of Y_4 and corresponding standard error with 20 multiply–imputed datasets.

Besides these two imputation methods, we conduct three standard methods with this dataset, and compute the estimates of the first moment and their corresponding standard error estimates. The Delta method is conducted with the covariance matrix estimate of these regression parameter estimates and empirical distribution of Y_1 which are obtained from the original dataset. The estimates of the first moment in the row of the Delta method are computed with regression parameter estimates of the original dataset on Table 7. The detail procedure for the Delta method is analogous to Appendix A–Appendix B.

The Bootstrap estimates and the Direct resampling estimates are computed with 100 random samples per each. The Bootstrap generates 100 random samples with replacement, and computes regression parameter estimates with 100 bootstrap samples. We compute the Bootstrap estimates of the first moment with these regression parameter estimates, and calculate the sample variance of the first moment with these bootstrap samples. The direct resampling method randomly draws 100 regression parameter estimates from the asymptotic normal distribution of the regression paramter estimates that are computed with the original dataset. We resample the predictive values of Y_4 from normal distributions having paramters that are composed of randomly drawn regression parameter estimates, and compute the resampling estimates of the first moment. The variance estimate of the first moment is computed with sample variances.

Table 7: The Performance Comparison Among Five Different Methods

Method	The first moment (ϕ)	
	Estimate	Standard Error
Delta method	76.69	1.4639
Bootstrap	76.82	1.4098
Resampling Method	76.59	1.5653
Mean Imputation	76.48	1.5142
Stochastic Imputation	76.52	1.2875

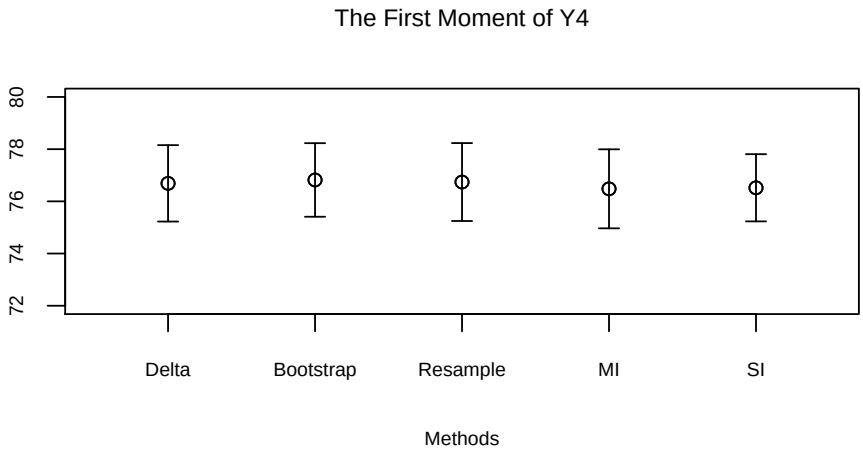

Figure 2: The Performance Comparison Among Five Different Methods

Table 7 shows the estimates of the first moment of Y_4 and the corresponding standard–errors by five different methods. The estimates in the row of the Delta method are directly computed with the parameter estimates of $[\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2]$ with the original dataset on Table 7, but those estimates in other rows are computed with multiple sample sets generated by the four different methods: the estimates in the row of the Bootstrap are computed from the randomly generated bootstrap samples, and those in the direct resampling method are computed from the predictive values which are drawn from normal distributions. In addition, the estimates in the rows of the imputations are deriven from the multiply–imputed datasets on Table 7. We can examine the performances of the five different methods according to Table 7. At fisrt, the estimates of the first moment of Y_4 are closely distributed among the five methods. All figures are distributed between 76.48 and 76.82, so those methods show similar results. The corresponding standard errors of the first moment range from 1.2875 to 1.5653. The stochastic imputation method with the specified algorithm in chapter 5 shows the smallest standard error and the resampling method shows the largest standard error. But, these estimated standard errors are very similar each other.

Figure 2 is a comparison plot based on Table 7 among these methods where 'o' indicates the location of the estimate and '|' indicates the magnitude of the standard error from the location of the estimate. The plot displays the comparison result about the estimates of the first moment and the lengths of corresponding standard errors among the methods. We can see the location of the first moment estimates are almost parallel in Figure 2, and the estimates and corresponding standard errors by the imputation methods show the equivalent patterns to the other methods with regard to the locations of the estimates and the lengths of the standard errors. However, the stochastic imputation method results in the smallest standard error. In addition, we would like to know whether the risperidone group shows clinical improvement for schizophrenia patients at four weeks based on PANSS data or not according to the estimates of the first moment of Y_4 in Table 7. The clinical improvement can be defined based on (6.2), and if (6.2) is over 20%, then we can conclude that the risperidone group is clinically improved at four weeks based on PANSS data.

$$\frac{E[Y_4] - E[Y_1]}{E[Y_1]} \times 100(\%) \tag{6.2}$$

The estimate of the first moment at baseline and four weeks are 92.34 and 76.69. The declination rate at four weeks is computed as about 16.94% according to (6.2). Therefore, there is no clear clinical proof for the improvement of the risperidone group based on the declination rate at four weeks with PANSS data even though the PANSS score seems to be decreased at four weeks.

7.0 DISCUSSION

When the missing–data mechanism only depends on the response variables, one can compute regression parameter estimates by the pseudolikelihood method without specifying the missing–data mechanism (Tang et al, 2003). However, the covariance estimates of the regression parameter estimates are computationally intensive. If one concerns estimating a variance of the PL estimates, then standard methods like the Delta method and the Bootstrap may not be the best methods. For this case, imputation methods may be more advantageous because they do not require complicated form of covariance estimates by the pseudolikelihood method. However, many imputation methods assume MAR missing–data mechanism. In the dissertation, we introduce a mean imputation method which modifies a component from Tang's approach(2002) and newly develop a stochastic imputation method. We conducted a simulation with 1000 bivariate datasets, and applied to a real dataset of PANSS data for Schizophrenia patients (Diggle et al., 2002) to examine the performances of the imputation methods. Our imputation methods showed equivalent results to other standard methods with regard to averages of 95% confidence interval widths and 95% coverage rates. Specifically, the stochastic imputation method performed showed the smallest 95% confidence interval widths for the first moment. However, the mean imputation method showed severe bias about other than the first moment, and a residual estimate may not be valid for the stochastic imputation when the conditional variance estimate of the complete cases, $\hat{Var}[Y|x, R = 1]$, is greater than that of the population, $\hat{Var}[Y|x]$. In addition, when the complete cases are sparse or sample sizes are too small, it may encounter a convergence problem to implement these imputation methods.

APPENDIX A

THE DELTA METHOD FOR ESTIMATING THE VARIANCE OF THE PL ESTIMATE OF E[Y]

Consider a bivariate missing data of $[X, Y]$ where X is fully observed and Y is partially observed. The missing–data mechanism is defined as (A.1).

$$pr[R|x, y] = \omega(y) \tag{A.1}$$

Under the assumption of (A.1), the regression parameter of $[Y|x]$ can be obtained by pseudolikelihood method without specifying the missing–data mechanism. Suppose that one is interested in estimating the variance of marginal mean of missing data Y. According to (4.2), the marginal mean of Y, ϕ, is expressed by setting $h(X, Y) = y$ as follows.

$$\begin{aligned} \phi &= E[y] = E[E[y|X]] \\ &= E[\beta_0 + \beta_1 \cdot x] \\ &= \beta_0 + \beta_1 \cdot \mu_x \end{aligned} \tag{A.2}$$

Where $\mu_x = E[x]$ and $\alpha = [\beta_0, \beta_1, \sigma^2]$ is regression parameter of $[Y|x]$ in (A.2), the estimate of ϕ is expressed as (A.3).

$$\hat{\phi} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \hat{\mu_x} \tag{A.3}$$

In (A.3), $\hat{\mu}_x = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ and $\hat{\alpha} = [\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2]$ is the regression parameter estimates by pseudolikelihood method. If both a vector of regression parameter estimates from $[Y|x]$ and

the estimate of the marginal mean of X can be pack into the vector of $\hat{\theta} = [\hat{\mu}_x, \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2]$, then the variance estimate of $\hat{\phi}$, which is a function of $\hat{\theta}$, is estimated using (A.4).

$$\hat{\text{Var}}(\hat{\phi}) \approx \frac{1}{n} D(\hat{\theta}) \cdot \text{Cov}(\sqrt{n} \cdot (\hat{\theta} - \theta)) \cdot D(\hat{\theta})^T \tag{A.4}$$

On (A.4), $D(\hat{\theta})$ is the vector of the first derivative of $\phi = \phi(\theta)$ about θ at $\theta = \hat{\theta}$. The first derivative of $D(\hat{\theta})$ is computed as (A.5), and the covariance matrix of $\hat{\theta}$ is denoted as (A.6).

$$D(\hat{\theta}) = [\frac{\partial \phi}{\partial \mu_x}, \frac{\partial \phi}{\partial \beta_0}, \frac{\partial \phi}{\partial \beta_1}, \frac{\partial \phi}{\partial \sigma^2}] = [\hat{\beta}_1, 1, \hat{\mu}_x, 0] \tag{A.5}$$

$$\text{Cov}(\sqrt{n} \cdot (\hat{\theta} - \theta)) = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix} \tag{A.6}$$

On (A.6), Σ_{11} is the variance of the estimate of μ_x, Σ_{22} is the variance of the regression parameter estimates, $\hat{\boldsymbol{\alpha}} = [\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2]$, and Σ_{12} is the covariance between these two. Tang et al.(2003) obtained the asymptotic variance of the regression parameter estimates of $\hat{\boldsymbol{\alpha}}$ by pseudolikelihood method, and Σ_{22} in (A.6) is this asymptotic variance of $\hat{\boldsymbol{\alpha}}$. Each component

of (A.6) can be expressed as follows:

$$\Sigma_{22} = \mathrm{E}(-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P^x))^{-1}[\mathrm{E}\{l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P^x) l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P^x)^T\} - \Sigma_1 - \Sigma_1^T + \Sigma_2]\mathrm{E}(-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P^x))^{-1} \quad \text{(A.7)}$$

Where

$$l(\boldsymbol{\alpha}) = \sum_{i=1}^{m}[\log g(y_i|x_i;\alpha) - \log\sum_{j=1}^{n} g(y_i|x_j;\alpha) + \log n]\text{Where}$$

$$\Sigma_1 = -\mathrm{P}^{\tilde{Z}}(l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P^x;\tilde{z})P^z[g(y|\tilde{X},\boldsymbol{\alpha}_0)\frac{I(R=1)P^x g_{\boldsymbol{\alpha}}(y|x,\boldsymbol{\alpha}_0)}{\{P^x(g(y|x,\boldsymbol{\alpha}_0))\}^2}$$

$$- \; g_{\boldsymbol{\alpha}}(y|\tilde{X},\boldsymbol{\alpha}_0)\frac{I(R=1)}{P^x g(y|x,\boldsymbol{\alpha}_0)}]$$

$$\Sigma_2 = I_1 - I_2 - I_2^T + I_3$$

Where

$$I_1 = P^{\tilde{X}}(P^z[g(y|\tilde{x},\boldsymbol{\alpha}_0)\frac{I(R=1)P^x g_{\boldsymbol{\alpha}}(y|x,\boldsymbol{\alpha}_0)}{\{P^x g(y|x,\boldsymbol{\alpha}_0)\}^2}]P^z[g(y|\tilde{x},\boldsymbol{\alpha}_0)\frac{I(R=1)P^x g_{\boldsymbol{\alpha}}(y|x,\boldsymbol{\alpha}_0)}{\{P^x g(y|x,\boldsymbol{\alpha}_0)\}^2}]^T)$$

$$- \; P^{\tilde{X}}(P^z[g(y|\tilde{x},\boldsymbol{\alpha}_0)\frac{I(R=1)P^x g_{\boldsymbol{\alpha}}(y|x,\boldsymbol{\alpha}_0)}{\{P^x g(y|x,\boldsymbol{\alpha}_0)\}^2}])P^{\tilde{X}}(P^z[g(y|\tilde{x},\boldsymbol{\alpha}_0)\frac{I(R=1)P^x g_{\boldsymbol{\alpha}}(y|x,\boldsymbol{\alpha}_0)}{\{P^x g(y|x,\boldsymbol{\alpha}_0)\}^2}])^T$$

$$I_2 = P^{\tilde{X}}(P^z[g(y|\tilde{x},\boldsymbol{\alpha}_0)\frac{I(R=1)P^x g_{\boldsymbol{\alpha}}(y|x,\boldsymbol{\alpha}_0)}{\{P^x g(y|x,\boldsymbol{\alpha}_0)\}^2}]P^Z[g(y|\tilde{x},\boldsymbol{\alpha}_0)\frac{I(R=1)}{P^x g(y|x,\boldsymbol{\alpha}_0)}]^T)$$

$$- \; P^{\tilde{X}}(P^z[g(y|\tilde{x},\boldsymbol{\alpha}_0)\frac{I(R=1)P^x g_{\boldsymbol{\alpha}}(y|x,\boldsymbol{\alpha}_0)}{\{P^x g(y|x,\boldsymbol{\alpha}_0)\}^2}])P^{\tilde{X}}(P^Z[g(y|\tilde{x},\boldsymbol{\alpha}_0)\frac{I(R=1)}{P^x g(y|x,\boldsymbol{\alpha}_0)}])^T$$

$$I_3 = P^{\tilde{X}}(P^Z[g(y|\tilde{x},\boldsymbol{\alpha}_0)\frac{I(R=1)}{P^x g(y|x,\boldsymbol{\alpha}_0)}]P^Z[g(y|\tilde{x},\boldsymbol{\alpha}_0)\frac{I(R=1)}{P^x g(y|x,\boldsymbol{\alpha}_0)}]^T)$$

$$- \; P^{\tilde{X}}(P^Z[g(y|\tilde{x},\boldsymbol{\alpha}_0)\frac{I(R=1)}{P^x g(y|x,\boldsymbol{\alpha}_0)}])P^{\tilde{X}}(P^Z[g(y|\tilde{x},\boldsymbol{\alpha}_0)\frac{I(R=1)}{P^x g(y|x,\boldsymbol{\alpha}_0)}])^T$$

More details about the asymptotic covariance matrix of the regression parameter estimates, $\hat{\boldsymbol{\alpha}} = [\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2]$, is referred to Tang et al.(2003).

$$\Sigma_{11} = \mathrm{Var}[\sqrt{n}\cdot(\hat{\mu_x} - \mu_x)] = \sigma_x^2 \quad \text{(A.8)}$$

$$\Sigma_{12} = Cov(\sqrt{n}\cdot(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0), \sqrt{n}\cdot(\hat{\mu_x} - \mu_x))$$

$$= Cov(E(-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}})^{-1}\cdot[\sqrt{n}\cdot P_n^z l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P^x) + \sqrt{n}\cdot P_n^z(l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P_n^x) - l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P^x))]$$

$$, \sqrt{n}\cdot P_n^z(x - \mu_x)) + o_{p(1)}$$

$$= I_3 + I_4 + o_p(1) \quad \text{(A.9)}$$

where

$$
\begin{aligned}
I_3 &= E(-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}})^{-1} \cdot \mathrm{Cov}(\sqrt{n} \cdot P_n^z l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P^x), \sqrt{n} \cdot P_n^z(x - \mu_x)) \\
&= E(-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}})^{-1} \cdot \mathrm{E}(l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P^x) \cdot (x - \mu_x)) \\
&= E(-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}})^{-1} \cdot \mathrm{E}(l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P^x) \cdot x) \\
I_4 &= E(-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}})^{-1} \cdot \mathrm{Cov}(\sqrt{n} \cdot P_n^z(l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}, P_n^x) - l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P^x)), \sqrt{n} \cdot P_n^z(x - \mu_x)) \\
&\approx E(-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}})^{-1} \cdot \mathrm{Cov}(\sqrt{n} \cdot (P_n^{\tilde{x}} - P^{\tilde{x}}) P^z [\frac{g(Y|\tilde{x}) \cdot I(R=1) \cdot P^x g_{\boldsymbol{\alpha}_0}(Y|x, \boldsymbol{\alpha}_0)}{(P^x g(Y|x, \boldsymbol{\alpha}_0))^2} \\
&\quad - \frac{g_{\boldsymbol{\alpha}_0}(y|\tilde{x}, \boldsymbol{\alpha}_0) \cdot I(R=1)}{P^x g(y|x, \boldsymbol{\alpha}_0)}], \sqrt{n} \cdot (P_n^{\tilde{x}} - P^{\tilde{x}}) \cdot \tilde{x}) \\
&= E(-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}})^{-1} \cdot \mathrm{Cov}(P^z [\frac{g(Y|\tilde{x}, \boldsymbol{\alpha}_{00}) \cdot I(R=1) \cdot P^x g_{\boldsymbol{\alpha}_0}(Y|x, \boldsymbol{\alpha}_0)}{(P^x g(Y|x, \boldsymbol{\alpha}_0))^2} - \frac{g_{\boldsymbol{\alpha}_0}(y|\tilde{x}, \boldsymbol{\alpha}_0) \cdot I(R=1)}{P^x g(y|x, \boldsymbol{\alpha}_0)}], \tilde{x}) \\
&= E(-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}})^{-1} \cdot [\mathrm{E}(P^z [\frac{g(Y|\tilde{x}, \boldsymbol{\alpha}_{00}) \cdot I(R=1) \cdot P^x g_{\boldsymbol{\alpha}_0}(Y|x, \boldsymbol{\alpha}_0)}{(P^x g(Y|x, \boldsymbol{\alpha}_0))^2} - \frac{g_{\boldsymbol{\alpha}_0}(y|\tilde{x}, \boldsymbol{\alpha}_0) \cdot I(R=1)}{P^x g(y|x, \boldsymbol{\alpha}_0)}], \tilde{x}) \\
&\quad - \mathrm{E}(P^z [\frac{g(Y|\tilde{x}, \boldsymbol{\alpha}_{00}) \cdot I(R=1) \cdot P^x g_{\boldsymbol{\alpha}_0}(Y|x, \boldsymbol{\alpha}_0)}{(P^x g(Y|x, \boldsymbol{\alpha}_0))^2} - \frac{g_{\boldsymbol{\alpha}_0}(y|\tilde{x}, \boldsymbol{\alpha}_0) \cdot I(R=1)}{P^x g(y|x, \boldsymbol{\alpha}_0)}]) \cdot \mathrm{E}(\tilde{x})] \quad \text{(A.10)}
\end{aligned}
$$

Where $Z = [X, Y, R]$, and P_n^z denots empirical process of (X,RY,R) on (A.8) and (A.9)

APPENDIX B

THE DELTA METHOD FOR ESTIMATING THE VARIANCE OF THE PL ESTIMATE OF E$[Y^2]$

Consider a bivariate missing data of $[X, Y]$ where X is fully observed and Y is partially observed. Under the assumption of (A.1) about the missing–data mechanism, the regression parameter of $[Y|x]$ can be obtained by pseudolikelihood method without specifying the missing–data mechanism. Suppose that one is interested in estimating the variance of the second moment of missing data Y. According to (4.2), the second moment of Y, ϕ, is expressed by setting $h(X, Y) = Y^2$ as follows.

$$\phi \quad = \quad \beta_1^2 \cdot \mu_2 + \sigma^2 + \beta_0^2 + 2\beta_0 \cdot \beta_1 \cdot \mu_x \tag{B.1}$$

Where $\mu_x = E(x)$ and $\mu_2 = E(x^2)$, the estimate of ϕ is:

$$\hat{\phi} = \hat{\beta}_1^2 \cdot \hat{\mu}_2 + \hat{\sigma}^2 + \hat{\beta}_0^2 + 2 \cdot \hat{\beta}_0 \cdot \hat{\beta}_1 \cdot \hat{\mu}_x$$

Where $\hat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\hat{\mu_2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$ on (B.1), and $\hat{\boldsymbol{\alpha}} = [\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2]$ are pseudolikelihood estimates.

$\text{Var}(\hat{\phi})$ is estimated by delta method as follows where $\hat{\theta} = \{\hat{\mu_x}, \hat{\mu}_2, \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2\}$.

$$\hat{\text{Var}}(\sqrt{n} \cdot \hat{\phi}) \approx D(\hat{\theta}) \cdot \text{Cov}(\sqrt{n} \cdot (\hat{\theta} - \theta)) \cdot D(\hat{\theta})^T$$

$$D(\hat{\theta}) = [2 \cdot \hat{\beta}_0 \cdot \hat{\beta}_1, \hat{\beta}_1^2, 2 \cdot (\hat{\beta}_0 + \hat{\beta}_1 \cdot \hat{\mu}_x), 2 \cdot (\hat{\beta}_1 \cdot \hat{\mu}_2 + \hat{\beta}_0 \cdot \hat{\mu}_x), 1] \tag{B.2}$$

Covariance matrix of $\hat{\alpha}$ is derived as below.

$$\mathrm{Cov}(\sqrt{n}\cdot(\hat{\theta}-\theta)) = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix} \tag{B.3}$$

Σ_{22} is same as (A.8), and Σ_{11} is asymptotic variance–covariance matrix of $[\sqrt{n}(\hat{\mu}_x - \mu_x), \sqrt{n}(\hat{\mu}_2 - \mu_2)]$ as (B.4).

$$\Sigma_{11} = \begin{bmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{bmatrix} = \begin{bmatrix} \mu_2 - \mu_x^2 & \mu_3 - \mu_x \cdot \mu_2 \\ \mu_3 - \mu_x \cdot \mu_2 & \mu_4 - \mu_2^2 \end{bmatrix} \tag{B.4}$$

Where $\mu_4 = E(x^4)$, $\mu_3 = E(x^3)$, $\hat{\mu}_4 = \frac{1}{n}\sum_{i=1}^n x_i^4$ and $\hat{\mu}_3 = \frac{1}{n}\sum_{i=1}^n x_i^3$.
Σ_{12} is the covariance matrix between $\hat{\boldsymbol{\alpha}}$ and $[\hat{\mu}_x, \hat{\mu}_2]^T$, and it has two elements of $[\tau_1, \tau_2]$. Let τ_1 denote the covariance between $\hat{\boldsymbol{\alpha}}$ and $\hat{\mu}_x$, and τ_2 denote the covariance between $\hat{\boldsymbol{\alpha}}$ and $\hat{\mu}_2$. τ_1 and τ_2 are computed as follows:

$$\begin{aligned} \tau_1 &= Cov(\sqrt{n}\cdot(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0), \sqrt{n}\cdot(\hat{\mu_x} - \mu_x)) \\ &= \mathrm{Cov}(E(-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}})^{-1}\cdot[\sqrt{n}\cdot P_n^z l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P^x) + \sqrt{n}\cdot P_n^z(l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P_n^x) - l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P^x))] \\ &\quad , \sqrt{n}\cdot P_n^z(x - \mu_x)) + o_{p(1)} \\ &= I_3 + I_4 + o_p(1) \end{aligned} \tag{B.5}$$

where

$$
\begin{aligned}
I_3 &= E[-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}}]^{-1} \cdot \mathrm{Cov}[\sqrt{n} \cdot P_n^z l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P^x), \sqrt{n} \cdot P_n^z (x - \mu_x)] \\
&= E[-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}}]^{-1} \cdot \mathrm{E}[l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P^x) \cdot (x - \mu_x)] \\
&= E(-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}})^{-1} \cdot \mathrm{E}[l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P^x) \cdot x] \\
I_4 &= E[-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}}]^{-1} \cdot \mathrm{Cov}[\sqrt{n} \cdot P_n^z (l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}, P_n^x) - l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P^x)), \sqrt{n} \cdot P_n^z (x - \mu_x)] \\
&\approx E[-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}}]^{-1} \cdot \mathrm{Cov}[\sqrt{n} \cdot (P_n^{\tilde{x}} - P^{\tilde{x}}) P^z [\frac{g(Y|\tilde{x}) \cdot I(R=1) \cdot P^x g_{\boldsymbol{\alpha}}(Y|x, \boldsymbol{\alpha})}{(P^x g(Y|x, \boldsymbol{\alpha}))^2} \\
&\quad - \frac{g_{\boldsymbol{\alpha}}(y|\tilde{x}, \boldsymbol{\alpha}) \cdot I(R=1)}{P^x g(y|x, \boldsymbol{\alpha})}], \sqrt{n} \cdot (P_n^{\tilde{x}} - P^{\tilde{x}}) \cdot \tilde{x}]_{\alpha=\alpha_0} \\
&= E(-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}})^{-1} \cdot [\mathrm{E}(P^z [\frac{g(Y|\tilde{x}, \boldsymbol{\alpha}_{00}) \cdot I(R=1) \cdot P^x g_{\boldsymbol{\alpha}_0}(Y|x, \boldsymbol{\alpha}_0)}{(P^x g(Y|x, \boldsymbol{\alpha}_0))^2} - \frac{g_{\boldsymbol{\alpha}_0}(y|\tilde{x}, \boldsymbol{\alpha}_0) \cdot I(R=1)}{P^x g(y|x, \boldsymbol{\alpha}_0)}] \cdot \tilde{x}) \\
&\quad - \mathrm{E}(P^z [\frac{g(Y|\tilde{x}, \boldsymbol{\alpha}_{00}) \cdot I(R=1) \cdot P^x g_{\boldsymbol{\alpha}_0}(Y|x, \boldsymbol{\alpha}_0)}{(P^x g(Y|x, \boldsymbol{\alpha}_0))^2} - \frac{g_{\boldsymbol{\alpha}_0}(y|\tilde{x}, \boldsymbol{\alpha}_0) \cdot I(R=1)}{P^x g(y|x, \boldsymbol{\alpha}_0)}]) \cdot \mathrm{E}(\tilde{x})]
\end{aligned}
$$

$$
\begin{aligned}
\tau_2 &= \mathrm{Cov}[\sqrt{n} \cdot (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0), \sqrt{n} \cdot (\hat{\mu}_2 - \mu_2)] \\
&= Cov(E(-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}})^{-1} \cdot [\sqrt{n} \cdot P_n^z l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P^x) + \sqrt{n} \cdot P_n^z (l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P_n^x) - l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P^x))] \\
&\quad , \sqrt{n} \cdot P_n^z (x^2 - \mu_2)) + o_{p(1)} \\
&= I_5 + I_6 + o_p(1) \qquad \text{(B.6)}
\end{aligned}
$$

where

$$
\begin{aligned}
I_5 &= E(-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}})^{-1} \cdot \mathrm{Cov}(\sqrt{n} \cdot P_n^z l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P^x), \sqrt{n} \cdot P_n^z (x^2 - \mu_2)) \\
&= E(-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}})^{-1} \cdot \mathrm{E}(l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P^x) \cdot (x^2 - \mu_2)) \\
&= E(-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}})^{-1} \cdot \mathrm{E}(l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P^x) \cdot x^2) \\
I_6 &= E(-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}})^{-1} \cdot \mathrm{Cov}(\sqrt{n} \cdot P_n^z (l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}, P_n^x) - l_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_0, P^x)), \sqrt{n} \cdot P_n^z (x^2 - \mu_2)) \\
&\approx E(-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}})^{-1} \cdot \mathrm{Cov}(\sqrt{n} \cdot (P_n^{\tilde{x}} - P^{\tilde{x}}) P^z [\frac{g(Y|\tilde{x}) \cdot I(R=1) \cdot P^x g_{\boldsymbol{\alpha}_0}(Y|x, \boldsymbol{\alpha}_0)}{(P^x g(Y|x, \boldsymbol{\alpha}_0))^2} \\
&\quad - \frac{g_{\boldsymbol{\alpha}_0}(y|\tilde{x}, \boldsymbol{\alpha}_0) \cdot I(R=1)}{P^x g(y|x, \boldsymbol{\alpha}_0)}], \sqrt{n} \cdot (P_n^{\tilde{x}} - P^{\tilde{x}}) \cdot \tilde{x^2}) \\
&= E(-l_{\boldsymbol{\alpha}\boldsymbol{\alpha}})^{-1} \cdot [\mathrm{E}(P^z [\frac{g(Y|\tilde{x}, \boldsymbol{\alpha}_{00}) \cdot I(R=1) \cdot P^x g_{\boldsymbol{\alpha}_0}(Y|x, \boldsymbol{\alpha}_0)}{(P^x g(Y|x, \boldsymbol{\alpha}_0))^2} - \frac{g_{\boldsymbol{\alpha}_0}(y|\tilde{x}, \boldsymbol{\alpha}_0) \cdot I(R=1)}{P^x g(y|x, \boldsymbol{\alpha}_0)}] \cdot \tilde{x^2}) \\
&\quad - \mathrm{E}(P^z [\frac{g(Y|\tilde{x}, \boldsymbol{\alpha}_{00}) \cdot I(R=1) \cdot P^x g_{\boldsymbol{\alpha}_0}(Y|x, \boldsymbol{\alpha}_0)}{(P^x g(Y|x, \boldsymbol{\alpha}_0))^2} - \frac{g_{\boldsymbol{\alpha}_0}(y|\tilde{x}, \boldsymbol{\alpha}_0) \cdot I(R=1)}{P^x g(y|x, \boldsymbol{\alpha}_0)}]) \cdot \mathrm{E}(\tilde{x^2})]
\end{aligned}
$$

BIBLIOGRAPHY

Altman, N.& Macgibbon B. (1998). "Consistent bandwidth selection for kernel binary regression". Journal of Statistical Planning and Inference 70, 121–137.

Chen, K. (2001) "Parametric Models for Response-Biased Sampling", Journal of Royal Statistical Society, Ser. B, 63, 4:775–789.

Dempster A.P. Laird N.M. Rubin D.B. (1977). "Maximum Likelihood from incomplete data via the EM algorithm" Journal of Royal Statistics Ser. B.39 1–38.

Diggle P., Heagerty P., Liang K.Y. & Zeger S.C.(2002). "Analysis of longitudinal data", New York, Oxford University Press.

Efron, B. (1979) "Bootstrap Methods: Another Look at the Jackknife". The Annals of Statistics 7 (1) 1–26

Efron, B. (1981) "Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods". Biometrika 68. 589–599

Efron, B.(1994) "Missing Data, Imputation, and the Bootstrap, Theory and Methods", Journal of the American Statistical Association, vol.89, 426:463–475.

Glynn, R.J. & Laird, N.M. & RUBIN D.B.(1986) "Selection modeling versus mixture modeling with nonignorable nonresponse", in Drawing Inferences from Self–Selected Samples (ed. H. Wainer), New York, 115–142.

Gong, G. & Samaniego, F. J.(1981) "Pseudo Maximum Likelihood Estimation: Theory and Application", The Annals of Statistics,9,4:861–869

Graham, J.W., Schafer, J. L. (1999). "On the performance of multiple imputation for multivariate data with small sample size". Statistical Strategies for Small Sample Research.

Lawless, J.F. & Kalbfleisch, J.D. & Wild, C.J.(1999) "Semiparametric methods for response –selective and missing data problems in regression", Journal of Royal Statistics, Ser. B, 61, 413–438.

Little K.& Zeger S.(1986) "Longitudinal analysis using generalized linear models" Biometrika 73:13–22.

Little, R.J.A.& Rubin D.B. (2002). "Statistical analysis with missing data". Willy series 2nd. New York.

Little, R.J.A.(1995) "Modelling the Drop-Out Mechanism in Repeated-Measure Studies", Journal of the American Statistical Association,90, 431:1112–1121.

Nadaraya, E.A. (1964). "On estimating regression". Theory Prob. Applic. 9, 141–142.

Neyman J. & Scott E.L.(1948) "Consistent estimates based on partially consistent observations". Econometrika, 16, 1–32.

Quandt, R. E. (1958). "'The estimation of the parameter of a linear regression system obeying two separate regimes". Journal of American Statistical Association, 53, 873–880.

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

Rubin, D.B. (1976). "Inference and Missing data". Biometrika. 63, 581–592.

Rubin, D.B. (1978). "Multiple imputations in sample surveys:a phenomenological Bayesian approach to nonresponse", Proc. Section on Surv. Res. Meth.,Amer.Statist.Assoc. pp.20–34.

Rubin, D.B. (1987) "Multiple Imputation for Nonresponse in Surveys", J. Wiley & Sons. New York

Schuster, E.F. (1972). "Joint asymptotic distribution of the estimated regression function at a finite number of distinct points". Ann. Math. Statist. 43, 84–88.

Schafer, J.L. Graham, J. W. (2002). "Missing data: Our view of the state of the art". Psychological Methods, 7 (2), 147–177.

Sheather, S.J. Jones M.C. (1991). "A reliable data data-based bandwidth selection method for kernel density estimation". Journal of Royal Statistics Society Ser.B. 53, 683–690.

Sprent, P. (1961). "Some hypothesis concerning two-phase regression lines." Biometrics, 17, 634–645

Tang, G. (2002). "An Imputation method for data with non–ignorable nonresponse [Presentation section: Abstract No. 301532]", Joint Statistical Meetings, 2002.

Tang, G. Little R.J.A. RAGHUNATHAN T.E. (2003). "Analysis of multivariate missing data with nonignorable nonresponse". Biometrika 90, 747.764.

Watson, G.S. (1964). "Smooth regression analysis". Sankhya A. 26, 359–372.

Yates, F.(1933) "An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias", Journal of the American Statistical Association, vol.57, 348–368.