2013

Reducing parameter estimation bias for data with missing values using simulation extrapolation

Yu-Yi Hsu
Iowa State University

Reducing parameter estimation bias for data with missing values using simulation extrapolation

by

Yu-Yi Hsu

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:

Alicia L. Carriquiry, Co-major Professor

Yongming Qu, Co-major Professor

Mark S. Kaiser

Jae Kwang Kim

Kenneth Koehler

Dan Nordman

Iowa State University

Ames, Iowa

2013

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with the writing of this thesis. I would like to thank Dr. Alicia L. Carriquiry and Dr. Yongming Qu for their guidance, patience and support throughout this research and the writing of this thesis. I would also like to thank my committee members for their efforts and contributions to this work.

ABSTRACT

Missing data is a common problem in data analysis, and has been studied extensively. We propose using simulation extrapolation(SIMEX), a general simulation-based approach to adjust the bias in the estimator due to missing values assuming the model for missingness is known. The SIMEX approach was originally proposed for measurement error models. The SIMEX method includes simulation steps that use information from the missing mechanism and an extrapolation step to adjust the bias. While EM and multiple imputation methods rely on the correct assumptions on the conditional distribution of missing data given observed data, the proposed SIMEX method assumes the correct model for the missingness. Therefore, SIMEX is more robust on an incorrect specification of the probability model for the unobserved data. We discuss the properties of the SIMEX estimator and compare this method with existing methods using simulation. The advantages and limitations of our approach are also discussed.

CHAPTER 1. OVERVIEW

Incomplete data is a common problem in data analysis and has been studied extensively. It can occur in cross-sectional studies where data are collected at a given time and some measurements on a subject are unavailable, or in longitudinal studies where subjects are measured repeatedly over the study period. For example, a subject in a clinical trial may drop out the study early for a variety of reasons including lack of treatment efficacy, increased safety concerns, family relocation and many others (Daniels and Hogan, 2008; Molenberghs and Kenward, 2007; Robbins and White, 2011). Simply ignoring the incomplete records and analyzing the observed measurements as if they comprised the complete dataset may result in biased parameter estimates and incorrect interpretation of the study results. When the probability of data missing depends on observed covariates we can improve the analyses and statistical inference by setting up the appropriate conditional model for the missing mechanism. Many strategies to include the additional information in analyses have been proposed and their theoretical properties have been discussed.

In this section, we review the categories of missing-data mechanisms and strategies for analyzing incomplete data. We then review the simulation extrapolation (SIMEX) method for measurement error models. We propose an approach to extend SIMEX to analyze incomplete data.

1.1 Missing Data Analysis

1.1.1 Categories of missing-data mechanism

Assume that our goal is to make inferences about the unknown parameter(s) θ in the response model $f_Y(y|x;\theta)$ where y is the dependent variable(s), and x are independent variables. Suppose that some of the responses are missing. In the presence of missing data, we use superscripts (o) and (m) to denote the observed and missing portions of the vector of responses respectively. For example, $Y^{(o)}$ and $Y^{(m)}$ denote the observed and missing portions of full dataset Y which can be described by the response model. Let the missing indicator R be a dummy variable which has value one if the corresponding measurement is observed and has value zero otherwise.

The joint distribution of (Y, R) can be factorized as

$$f_{Y,R}(y, r|x, z; \eta, \theta) = f_Y(y|x;\theta) f_{R|Y}(r|y, x, z; \eta, \theta), \tag{1.1}$$

where Z are additional explanatory variables and η are parameters for the missing-data mechanism given Y. Equation (1.1) is called the selection modeling approach (Rubin, 2004). The selection model proposes that the missing-data mechanism is a process to select observed data from the full dataset. We focus on the selection modeling approach in this paper since the SIMEX method is built on equation (1.1).

The observed $Y^{(o)}$ follows the distribution

$$\int_{Y^{(m)}} f_{Y|R}(y^{(o)}, y^{(m)}|r, x, z; \eta, \theta) dy^{(m)}$$

instead of the original distribution $f_Y(y|x;\theta)$. Therefore, the simple naïve method which estimates the parameter θ assuming that $Y^{(o)}$ are samples from $f_Y(y|x;\theta)$ may produce biased estimators. In the presence of missing data, the full likelihood which includes the response model and the missing-data mechanism should be used; however, the maximization of this complex likelihood can be a challenge. Currently, there are no readily available software packages to implement this full likelihood approach for general cases.

To simplify the analysis, we make assumptions about the relationship between the response model and the missing data mechanism. The following categorization proposed by Rubin (1976) is commonly used to identify situations in which we can safely simplify the analysis by ignoring incomplete records or ignoring the missing data mechanism.

Let X be independent variables in the response model. The missing data mechanism is called "covariate-dependent" dropout (Little, 1995), when the missing-data mechanism depends not only on Y but also on X in the response model. With respect to the relationship between missing indicators (R) and response variables (Y) conditionally on explanatory variables (X), missing-data mechanisms are categorized into three categories:

1. Missing completely at random (MCAR) occurs when missingness depends on neither observed ($Y^{(o)}$) nor unobserved dependent variables ($Y^{(m)}$).

2. Missing at random (MAR) occurs when missingness only depends on observed dependent variables ($Y^{(o)}$).

3. Missing not at random (MNAR) occurs when missingness depends on unobserved dependent variables ($Y^{(m)}$).

The categories of missing-data mechanisms are important for selecting appropriate methods to analyze the data. To find evidence of not MCAR, Little (1988) proposes a single test statistic for multivariate data. Park and Davis (1993) add interaction terms of missing pattern indicators and model parameters then test those interaction effects for repeated categorical measurements of the response variable. Park and Lee (1997) add missing pattern indicators as explanatory variables and test their main effects for longitudinal data using the GEE method.

The missing-data mechanism is called *ignorable* for likelihood inference provided that the following two conditions hold: i) the missing mechanism is MAR, which means that

$$f(r|y, x; \eta) = f(r|y^{(o)}, x; \eta)$$

for all x and η, and ii) the spaces of parameters for the model and the missing mechanism are distinct (Rubin, 1976; Little and Rubin, 2002). The likelihood based methods, which ignore the missing data mechanism, require ignorability, an assumption that cannot be tested. When the missing mechanism is ignorable, likelihood-based or Bayesian inferences about θ can be made from $f_{Y^{(0)}} = \int_{y^{(m)}} f_Y(y^{(o)}, y^{(m)})dy^{(m)}$ without involving R and η (Rubin, 1976). Little and Rubin (2002) provide details and examples of maximum likelihood inferences for monotone or general missing-patterns under the assumption of ignorability. Dempster et al. (1977) reviews theories and applications of the EM algorithm for likelihood based inferences from incomplete data under the ignorability assumption.

When the missing data mechanism is MNAR, adding additional information (assumptions, structures or covariates) may simplify the analysis or reintegrate part of the lost information caused by missingness. For example,

- Covariate-dependent MAR missingness (MAR-cov) occurs when the missing data mechanism $f(r|y, x)$ depends on $(y^{(o)}, y^{(m)})$ given x, but only depends on $y^{(o)}$ when we can condition on (x, z). MAR-cov also called auxiliary variables MAR (A-MAR), and the additional covariates Z are known as auxiliary variables (Daniels and Hogan, 2008) .

- Random effect dependent missingness occurs when the missing-data mechanism depends on the random effects of a mixed effects model. The joint distribution of Y and R can be factorized as

$$f(y, r|x, z, b; \theta, \eta) = f(y|x, z, b; \theta, \eta)f(r|y^{(o)}, x, z, b; \theta, \eta),$$

  where the "share parameter" b is an unobservable random effect in the response model (Little, 1995) .

- The pattern-mixture model stratifies missing data according to the pattern of the missing values and treats the missing pattern as part of the response model (Little,

1993). Little (1993) uses pattern-mixture models on multivariate incomplete data. Little (1995) and Hedeker and Gibbons (1997) involve missing-pattern indicators in the random effects model for longitudinal data. Qu and Lipkovich (2009) extend the Hedeker and Gibbons (1997) approach to the inverse probability weighted estimator by including both multiple imputation procedures and missing-pattern indicators (MIMP) for calculating a propensity score for propensity-score based estimation in the presence of missing covaiates. They showed through simulations that MIMP has similar performance as multiple imputations for the case of MAR and has better performance than multiple imputations for the case of MNAR.

- Daniels and Hogan (2008) describe a model in which the response model is conditional on R instead of the pattern of R as a "mixture model". Compared to the pattern-mixture model that assumes similarity between unobserved and observed data within each pattern (each stratum), the mixture model assumes that the distribution of unobserved data is an extrapolation distribution which is identifiable under the MAR assumption.

1.1.2 Strategies for analyzing missing data

Little and Rubin (2002) summarize four categories of methods for analyzing incomplete data: i) procedures based on complete-cases (CC) or available-cases (AC) ii) weighting procedures, iii) imputation-based procedures, and iv) model-based procedures. These four categories are not mutually exclusive.

The SIMEX approach discussed in this paper is similar to the weighting procedures. They both model the missing-data mechanism separately from the response model and both require good estimates of parameters for the missing-data mechanism. On the other hand, imputation-based procedures impute each one of the missing cells with a single value or multiple values. The multiple imputation (MI) technique replaces each missing value with two or more imputed values generated from the distribution of the missing

data (Rubin, 2004). The MI method has both the flexibility of specifying separate models for the response and missing mechanism and the convenience of fitting response models without involving the missing mechanism. The response model can be parametric or non-parametric. For example, one can implement MI on the Kaplan-Meier estimator in the case of right censored data with missing censoring indicators (Subramanian, 2009). The imputation procedure may be difficult when the dimension of missing variables increases, or the variable type is not normal.

The choice of strategy for analyzing data with missing observations is made according to the knowledge about the response model and the missing-data mechanism. We compare the SIMEX procedure that we propose with the weighting and imputation procedures in the simulation studies.

1.2 The SIMEX Method for Measurement Error Problems

The simulation extrapolation (SIMEX) method proposed by Cook and Stefanski (1994) is a simulation-based method to do inference in non-linear models where covariates are observed with measurement error. The method is noteworthy for its ease of implementation and general applicability. The SIMEX method includes two basic steps. The first reveals the effect of increasing the magnitude of the measurement error on an estimator using simulation. The bias of the estimator tends to increase as measurement error increases, and the association can be estimated in the first step. Second, one finds the value of the reduced bias estimate by extrapolation using the trend found in the previous step. The idea is to use the estimated function to estimate the true parameters when the measurement error vanishes. The notation in this section is different from notation used in other sections.

Suppose we are interested in the model

$$Y \sim f(y; U, V, \theta),$$

where Y denotes the response variable and (U, V) denote the explanatory variables measured without error. The true value of U is unobservable. Assume that

$$X_j = U_j + \sigma Z_j, j = 1, \ldots, n$$

where X_j are observable measurements, U_j are fixed and unknown, and Z_j are independent standard normal random variables with known variance one. Since U_j is unobservable, the consistent estimator, $\hat{\theta}_{\text{true}} = T(\{Y_j, U_j, V_j\}_1^n)$, is unavailable. Therefore, $\hat{\theta}_{\text{naïve}} = T(\{Y_j, X_j, V_j\}_1^n)$ can be used for estimation but it may be biased due to the unobservable measurement error. The object is to find a SIMEX estimator $\hat{\theta}_{\text{SIMEX}}$ which is closer to the true estimator than the naïve estimator.

Cook and Stefanski (1994) defined a function

$$\begin{aligned} \theta(\lambda) &\equiv E_{\{Z_{b,j}\}_{j=1}^n} T(\{Y_j, X_j + \sqrt{\lambda}\sigma Z_{b,j}, V_j\}_1^n) \\ &= E_{\{Z_{b,j}\}_{j=1}^n} T(\{Y_j, U_j + \sigma Z_j + \sqrt{\lambda}\sigma Z_{b,j}, V_j\}_1^n) \end{aligned} \tag{1.2}$$

as a function of $\lambda \in [-1, \infty)$ with new standard normal distributed random vector $\{Z_{b,j}\}_{j=1}^n$, which are also called pseudoerrors. The values of the function $\theta(\lambda)$ for $\lambda \geq 0$ is

$$\hat{\theta}(\lambda) \equiv \begin{cases} T(\{Y_j, X_j, V_j\}_1^n), & \lambda = 0 \\ \frac{1}{B}\sum_{b=1}^B T(\{Y_j, X_j + \sqrt{\lambda}\sigma Z_{b,j}, V_j\}_1^n), & \lambda > 0 \end{cases}$$

where $\{\{Z_{b,j}\}_{j=1}^n\}_{b=1}^B$ are independent standard normal random variables which are also independent of the data (Y, X, V). To extrapolate to the case of no measurement error, we assume a parametric form for $\theta(\lambda)$, say $\theta(\lambda, c)$. The value of c can be estimated by regressing $\{\lambda_k\}$ on $\{\hat{\theta}(\lambda_k)\}$ using the parametric funtion $\theta(\lambda, c)$ for $k = 1, \ldots, K$ where $\lambda_k > 0$.

The SIMEX estimator is defined as $\hat{\theta}_{\text{SIMEX}} \equiv \hat{\theta}(-1, \hat{c})$. If the closed form of the smooth function $\theta(\lambda, c)$ is unknown, the exact extrapolation function is approximated by a second order polynomial.

1.2.1 The expectation of $\hat{\theta}_{SIMEX}$

This section is a short summary of proofs from Stefanski and Cook (1995). See (Carroll and Stefanski, 1997) and (Carroll et al., 2006) for more assumptions and asymptotic theoretic properties of the SIMEX estimator.

The expectation of $\theta(\lambda)$ given (Y_j, U_j, V_j) is

$$\begin{aligned}\mathbb{E}_{\{Z_j\}_{j=1}^n}(\theta(\lambda)) &= \mathbb{E}_{\{Z_j\}_{j=1}^n}\mathbb{E}_{\{Z_{b,j}\}_{j=1}^n} T(\{Y_j, X_j + \sqrt{\lambda}\sigma Z_{b,j}, V_j\}_{j=1}^n) \\ &= \mathbb{E}_{\{Z_j\}_{j=1}^n}\mathbb{E}_{\{Z_{b,j}\}_{j=1}^n} T(\{Y_j, U_j + \sigma Z_j + \sqrt{\lambda}\sigma Z_{b,j}, V_j\}_{j=1}^n),\end{aligned}$$

Stefanski and Cook (1995) consider the special case with sample of size one ($n = 1$) or the case where $\bar{X}$ alone is a sufficient statistic for estimation. The expectation of $\theta(\theta; c)$ given (Y_j, U_j, V_j) becomes

$$\begin{aligned}\mathbb{E}_Z(\theta(\lambda; c)) &= \mathbb{E}_Z\mathbb{E}_{Z_b} T(\{Y\}, X + \sqrt{\lambda}\sigma Z_b, \{V\}) \\ &= \mathbb{E}_Z\mathbb{E}_{Z_b} T(\{Y\}, U + \sigma Z + \sqrt{\lambda}\sigma Z_b, \{V\})\end{aligned} \tag{1.3}$$

where $\lambda \geq -1$ and T is a function of complex random variables. The Taylor expansion of $\theta(\lambda)$ at $\lambda = 0$ is

$$\theta(\lambda) = \mathbb{E}_{Z_b} T(\{Y\}, U, \{V\}) + \sum_{k=1}^{\infty} \frac{T^{(k)}(\{Y\}, U, \{V\})}{k!}(\sigma Z + \sqrt{\lambda}\sigma Z_b)^k.$$

The expectation of the Taylor expansion of T when $\lambda \to -1$ is

$$\begin{aligned}\lim_{\lambda\to-1} \mathbb{E}_Z(\theta(\lambda)) &= \lim_{\lambda\to-1} \mathbb{E}_Z\mathbb{E}_{Z_b}\left(T(Y, U + \sigma Z + \sqrt{\lambda}\sigma Z_b, V)\right) \\ &= \lim_{\lambda\to-1} \mathbb{E}_{Z,Z_b}\left(T(Y, U, V) + \sum_{k=1}^{\infty} \frac{T^{(k)}(Y, U, V)}{k!}(\sigma Z + \sqrt{\lambda}\sigma Z_b)^k\right).\end{aligned}$$

Stefanski and Cook (1995) define that a function is "sufficiently smooth" if

$$\begin{aligned}\lim_{\lambda\to-1} \mathbb{E}_Z(\theta(\lambda)) &= T(\{Y\}, U, \{V\}) + \sum_{k=1}^{\infty} \frac{T^{(k)}(Y, U, V)}{k!}\mathbb{E}_{Z,Z_b} \lim_{\lambda\to-1}(\sigma Z + \sqrt{\lambda}\sigma Z_b)^k, \\ &= T(\{Y\}, U, \{V\}).\end{aligned} \tag{1.4}$$

where $\mathbb{E}_{Z,Z_b}(\sigma Z + i\sigma Z_b)^k = 0$ for $k = 1, 2, \ldots$ and independent standard normal distributed (Z, Z_b).

The extrapolation function $\hat{\theta}(\lambda, c) = \theta(\lambda, \hat{c})$ is estimated by minimizing

$$\sum_{k=1}^{K} \Big(\hat{\theta}(\lambda_k) - \theta(\lambda_k, c)\Big)^2$$

over $c \in \mathbb{R}^{dim(c)}$. If the functional form of $\theta(\lambda)$ is known and used as the extrapolation function, then $\theta(\lambda) = \theta(\lambda, c)$ for $c \geq -1$, If the true extrapolating function is known, and the functional forms of $\theta\lambda$ is linear, then the estimator $\hat{c}$ is unbiased. If the true extrapolating function is known, and the functional forms of $\theta(\lambda)$ is nonlinear, the estimator $\hat{c}$ is consistent under some conditions (Wu, 1981).

If the true extrapolating function is known, and the parameter c can be estimated unbiasedly, then

$$\begin{aligned} \mathbb{E}_{Y,Z}(\hat{\theta}_{SIMEX}) &= \mathbb{E}_{Y,Z} E_{Z_b}(\theta(-1, \hat{c})) \\ &= \mathbb{E}_{Y,Z}(\theta(-1, c)) \\ &= \mathbb{E}_{Y,Z}(\lim_{\lambda \to -1} \hat{\theta}(\lambda)). \end{aligned} \tag{1.5}$$

Additionally, if $T(Y, U, V)$ is an unbiased estimator of θ, $\theta(\lambda)$ is "sufficiently smooth", and σ is known, then by Equation (1.4) and Equation (1.5),

$$\begin{aligned} \mathbb{E}_{Y,X}(\hat{\theta}_{SIMEX}) &= \mathbb{E}_Y \mathbb{E}_Z \mathbb{E}_{Z_b}(T(Y, U + \sigma Z + i\sigma Z_b, V)) \\ &= \mathbb{E}_Y(T(Y, U, V)) \\ &= \theta, \end{aligned} \tag{1.6}$$

which means that the SIMEX estimator is unbiased.

1.2.2 The variance of $\hat{\theta}_{SIMEX}$

The SIMEX estimator is a combination of several correlated estimators with different λ's, so the exact variance of the SIMEX estimator is hard to derive. One practical but

time consuming way to estimate the variance is to employ a bootstrap procedure (Cook and Stefanski, 1994). Stefanski and Cook (1995) propose another estimator of variance introduced below.

Stefanski and Cook (1995) first define variance and covariance for complex random variables. Let $W = \mu + \sigma Z_1 + i\sigma Z_2$, where Z_1 and Z_2 are independent standard normal random variables. The moments of W are

$$\mathbb{E}(W^m) = \sum_{k=0}^{m} \binom{m}{k} \mu^k \mathbb{E}(\sigma Z_1 + i\sigma Z_2)^{m-k} = \mu^m, m = 1, 2, \ldots$$

Let W_1 and W_2 denote two complex-valued random variables. Stefanski and Cook (1995) define $\mathrm{Var}^{(c)}(W) \equiv \mathbb{E}(W^2) - (\mathbb{E}(W))^2$ and $\mathrm{Cov}^{(c)}(W_1, W_2) \equiv \mathbb{E}(W_1 W_2) - \mathbb{E}(W_1)\mathbb{E}(W_2)$. By definition, $\mathrm{Var}^{(c)}(i\sigma Z_b) = -\sigma^2$ and

$$\begin{aligned}
\mathrm{Var}^{(c)}(W) &= \mathrm{Var}^{(c)}(\mu + \sigma Z_1 + i\sigma Z_2) \\
&= \mathbb{E}\left((\mu + \sigma Z_1 + i\sigma Z_2)^2\right) - (\mathbb{E}(\mu + \sigma Z_1 + i\sigma Z_2))^2 \\
&= \mathbb{E}\left((\mu + \sigma Z_1)^2\right) - \mathbb{E}(\sigma Z_2^2) - (\mathbb{E}(\mu + \sigma Z_1))^2 + (\mathbb{E}(\sigma Z_2))^2 \\
&\quad +2i\mathbb{E}\left((\mu + \sigma Z_1)(\sigma Z_2)\right) - 2i\mathbb{E}(\mu + \sigma Z_1)\mathbb{E}(\sigma Z_2) \\
&= \mathrm{Var}(\mu + \sigma Z_1) - \mathrm{Var}(\sigma Z_2) \\
&= 0.
\end{aligned}$$

Assume that $B = \infty$. Since $\hat{\theta}(\lambda) = \mathbb{E}_{Z_b}\left(\hat{\theta}_b(\lambda)|X\right)$, the covariance of $\hat{\theta}_b(\lambda)$ and $\hat{\theta}(\lambda)$ is

$$\begin{aligned}
\mathrm{Cov}^{(c)}\left(\hat{\theta}_b(\lambda), \hat{\theta}(\lambda)\right) &= \mathbb{E}_{Z,Z_b}\left(\hat{\theta}_b(\lambda)\hat{\theta}(\lambda)\right) - \mathbb{E}_{Z,Z_b}\left(\hat{\theta}_b(\lambda)\right)\mathbb{E}_{Z,Z_b}\left(\hat{\theta}(\lambda)\right) \\
&= \mathbb{E}_Z\left(\hat{\theta}(\lambda)\mathbb{E}_{Z_b}\left(\hat{\theta}_b(\lambda)|X\right)\right) \\
&\quad -\mathbb{E}_Z\left(\mathbb{E}_{Z_b}\left(\hat{\theta}_b(\lambda)|X\right)\right)\mathbb{E}_Z\left(\hat{\theta}(\lambda)\right) \\
&= \mathbb{E}_Z\left(\hat{\theta}(\lambda)^2\right) - \left(\mathbb{E}_Z\left(\hat{\theta}(\lambda)\right)\right)^2 \\
&= \mathrm{Var}^{(c)}(\hat{\theta}(\lambda)). \qquad (1.7)
\end{aligned}$$

Let $\Delta_b(\lambda) = \hat{\theta}_b(\lambda) - \hat{\theta}(\lambda) = T\left(Y, U + \sigma Z + \sqrt{\lambda}\sigma Z_b, V\right) - \mathbb{E}_{Z_b}\left(\hat{\theta}(\lambda)|X\right)$. By Equation (1.7), the variance of $\Delta_b\lambda$ is

$$\begin{aligned}\mathrm{Var}^{(c)}(\Delta_b(\lambda)) &= \mathrm{Var}^{(c)}(\hat{\theta}_b(\lambda)) + \mathrm{Var}^{(c)}(\hat{\theta}(\lambda)) - 2\mathrm{Cov}^{(c)}(\hat{\theta}_b(\lambda), \hat{\theta}(\lambda)) \\ &= \mathrm{Var}^{(c)}(\hat{\theta}_b(\lambda)) - \mathrm{Var}^{(c)}(\hat{\theta}(\lambda)).\end{aligned}$$

The variance of the SIMEX estimator is

$$\begin{aligned}\mathrm{Var}(\hat{\theta}_{SIMEX}) &= \mathrm{Var}^{(c)}(\lim_{\lambda\to-1}\hat{\theta}(\lambda)) \\ &= \lim_{\lambda\to-1}\mathrm{Var}^{(c)}(\hat{\theta}(\lambda)) \\ &= \lim_{\lambda\to-1}\left(\mathrm{Var}^{(c)}(\hat{\theta}_b(\lambda)) - \mathrm{Var}^{(c)}(\Delta_b(\lambda))\right). \qquad (1.8)\end{aligned}$$

The first term, $\lim_{\lambda\to-1}\mathrm{Var}^{(c)}(\hat{\theta}_b(\lambda))$, can be estimated by a SIMEX estimator of variance of $T(Y, U, V)$. The second term, $\lim_{\lambda\to-1}\mathrm{Var}^{(c)}(\Delta_b(\lambda))$, can be estimated by the sample variance of $\Delta_b(\lambda)$, and then extrapolating to $\lambda = -1$.

In the special cases when $n = 1$ described in Equation (1.3), the variance of the SIMEX estimator is

$$\mathrm{Var}(\hat{\theta}_{SIMEX}) = -\lim_{\lambda\to-1}\mathrm{Var}^{(c)}(\Delta_b(\lambda)) \qquad (1.9)$$

since

$$\begin{aligned}\lim_{\lambda\to-1}\mathrm{Var}^{(c)}(\hat{\theta}_b(\lambda)) &= \lim_{\lambda\to-1}\left(\mathbb{E}(\hat{\theta}_b(\lambda)^2) - \mathbb{E}(\hat{\theta}_b(\lambda)^2)\right) \\ &= \lim_{\lambda\to-1}\big(\mathbb{E}(T^2(\{Y\}, X + \lambda\sigma Z_b, \{V\})) \\ &\quad -\mathbb{E}(T(\{Y\}, X + \lambda\sigma Z_b, \{V\}))^2)\big) \\ &= 0\end{aligned}$$

by the "sufficiently smooth" condition defined in Equation (1.4).

Stefanski and Cook (1995) show that the variance estimator in Equation (1.8) is unbiased when both the variance of the measurement error, σ^2, and the exact extrapolation function are known. The variance estimator is also relatively less computationally

intensive compared to the bootstrap estimator but it does not take the variation from the estimated σ^2 into consideration. The simulation study of Stefanski and Cook (1995) shows that the estimated variance of the SIMEX estimator in Equation (1.9) underestimates the variability of the SIMEX estimator when an estimator of σ^2 and a linear extrapolation function are used.

1.2.3 An example - estimate e^{μ}

Stefanski and Cook (1995) use the following example to demonstrate the SIMEX method. The objective is to estimate $\theta = f(\mu) = e^{\mu}$ from X. The observations $\{X\}$ are equal to μ plus measurement errors which are normally distributed with mean zero and variance σ^2 where σ is known. Two estimators for θ are presented: the naïve estimator which is the maximum likelihood estimator without any adjustment and the SIMEX estimator obtained using the true extrapolation function.

The naïve estimate of θ is $\hat{\theta}_{\text{naïve}} = e^{\bar{X}}$, because the maximum likelihood of μ is $\bar{X}$. Since $\bar{X}$ is assumed to be normally distributed, the naïve esitmator is lognormally distributed with mean $e^{\mu+\frac{\sigma^2}{2n}}$ and variance $e^{2\mu+\frac{\sigma}{\sqrt{n}}}(e^{\frac{\sigma}{\sqrt{n}}}-1)$.

To find the true extrapolation function, consider the observable measurement $\bar{X} = U + Z$ as the true unobservable measurement U plus a measurement error $Z \sim N(0, \frac{\sigma^2}{n})$. Let $\hat{\theta}(0) = \hat{\theta}_{\text{naïve}}$. The new estimator based on measurements with additional measurement errors is

$$\hat{\theta}(\lambda^*) = \frac{1}{B}\sum_{b=1}^{B} e^{\bar{X}+\sqrt{\lambda^*}\frac{\sigma}{\sqrt{n}}Z_b} = \frac{1}{B}\sum_{b=1}^{B} e^{\mu+\bar{Z}+\sqrt{\lambda^*}\frac{\sigma}{\sqrt{n}}Z_b} \to e^{\bar{X}+\frac{\lambda^*\sigma^2}{2n}}, a.s.$$

as $B \to \infty$ by the strong law of large number for one $\lambda^* > 0$. In this example, we can find the true expectation of $\hat{\theta}(\lambda)$, $\mathbb{E}_{Z_b}(\hat{\theta}(\lambda)|\bar{X}) = e^{\bar{X}+\frac{\lambda\sigma^2}{2n}}$ for $\lambda > 0$. Therefore, $\log(\mathbb{E}_{Z_b}(\hat{\theta}(\lambda)|\bar{X})) = \bar{X} + \frac{\lambda\sigma^2}{2n}$. Assume that we know that the relationship between $\log \mathbb{E}(\hat{\theta}(\lambda)|\bar{X})$ and λ is linear. The straight line that passes through $(0, \log(\hat{\theta}(0)))$ and

$(\lambda^*, \log(\hat{\theta}(\lambda^*)))$ is

$$\log(\hat{\theta}(\lambda)) = \log(\hat{\theta}(0)) + \frac{1}{\lambda^*}\left(\log(\hat{\theta}(\lambda^*)) - \log(\hat{\theta}(0))\right)\lambda = \log\left(\hat{\theta}(0)^{\frac{\lambda^*-\lambda}{\lambda^*}}\hat{\theta}(\lambda^*)^{\frac{\lambda}{\lambda^*}}\right)$$

for $\lambda > 0$. By extrapolating $\log(\hat{\theta}(\lambda))$ to $\lambda = -1$, the SIMEX estimator is

$$\hat{\theta}_{SIMEX} = e^{\log(\hat{\theta}(-1))} = \frac{\hat{\theta}(0)^{\frac{\lambda^*+1}{\lambda^*}}}{\hat{\theta}(\lambda^*)^{\frac{1}{\lambda^*}}} \xrightarrow{d} \frac{e^{\bar{X}\frac{\lambda^*+1}{\lambda^*}}}{e^{\left(\bar{X}+\frac{\lambda^*\sigma^2}{2n}\right)\left(\frac{1}{\lambda^*}\right)}} = e^{\bar{X}-\frac{\sigma^2}{2n}}$$

as $B \to \infty$ by the continuous mapping theorem (Athreya and Lahiri, 2006). $\hat{\theta}_{SIMEX}$ is lognormally distributed and with mean $\mathbb{E}_{\bar{X}}(\hat{\theta}_{\text{SIMEX}}) = e^{\mu}$ and variance $\text{Var}_{\bar{X}}(\hat{\theta}_{SIMEX}) = e^{2\mu}(e^{\frac{\sigma^2}{n}} - 1)$. When $B \to \infty$, σ is known and the true extrapolation function is known, and the SIMEX estimator is an unbiased estimator. Assume that we have sample size $n = 4$ and sample values $x = (-.20544,\ .33879,\ 1.39088,\ -1.02414)$. Figure 1.1 shows $\log(\hat{\theta}(\lambda))$ versus λ for $-1 < \lambda < 1$.

The variance of $\hat{\theta}_{\text{SIMEX}}$ can also be estimated by another extrapolation step as shown in Equation (1.9). The sample variance of $\hat{\theta}_b(\lambda)$ converges to

$$v(\bar{X}, \lambda) \equiv \text{Var}_{Z_b}(\hat{\theta}_b(\lambda; c)|\bar{X}) = \text{Var}_{Z_b}(e^{\bar{X}+\sqrt{\lambda}\sigma\bar{Z}_b}|\bar{X}) = e^{2\bar{X}+\frac{\sigma^2}{n}(\lambda)}(e^{\frac{\sigma^2}{n}(\lambda)} - 1)$$

as $B \to \infty$. Assuming that the true extrapolation function $v(\bar{X}, \lambda)$ is known, the SIMEX estimator of $\text{Var}(\hat{\theta}_{\text{SIMEX}})$ is

$$-v(\bar{X}, -1) = e^{2\bar{X}-\frac{\sigma^2}{n}}(1 - e^{-\frac{\sigma^2}{n}}).$$

1.2.4 Extensions and applications

Cook and Stefanski (1994) and Stefanski and Cook (1995) have shown that the SIMEX method works well for a measurement error model with normal additive error on one explanatory variable. The SIMEX method has been used to reduce biases in other models with additive measurement error. (Stefanski and Bay, 1996) use the SIMEX method to reduce "much of the bias" on estimators of a finite population cumulative distribution function (cdf) which is a nonlinear function of observations when observations are measured with normally distributed error.

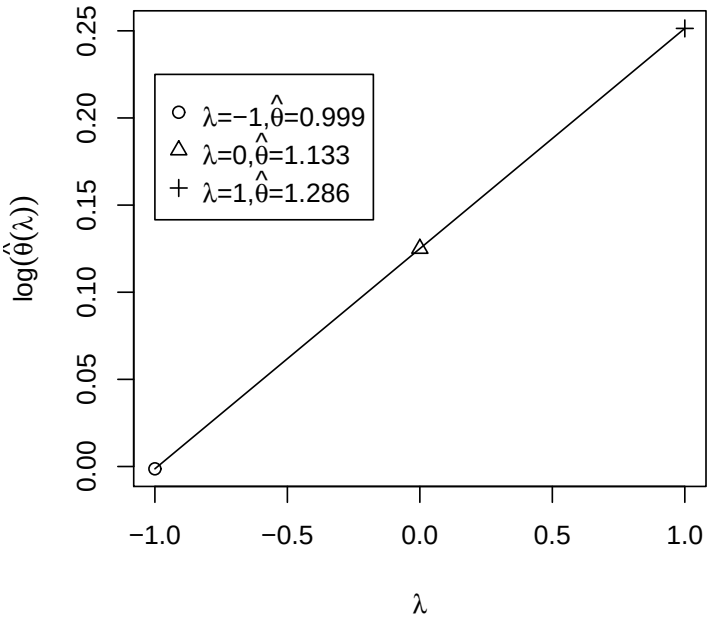

Figure 1.1 An example of using the SIMEX method with true extrapolation function, $\log(\mathbb{E}(\hat{\theta}(\lambda))) = a + b\lambda$ where a and b are two real numbered coefficients.

The SIMEX method has also been applied to generalized linear mixed models(GLMMs) in the presence of normal additive measurement error on explanatory variables. Wang et al. (1998) propose generalized linear mixed measurement error models(GLMMeMs) to describe a GLMM model with normal additive measurement error on one normally distributed covariate. They show that when the measurement error model is combined into the GLMM model, the observed data still follow a GLMM, but the structure of fixed effects is changed and variance structure becomes more complex. They apply the SIMEX method to estimate regression coefficients and variance components and make inferences. Lin and Carroll (1999) follow Wang et al. (1998) and apply the SIMEX method on score tests to test if the measurements within one specific cluster are correlated and if all variance components across clusters are zero in a GLMMeM.

Another application of SIMEX is in the context of accelerated failure time (AFT) models with measurement error on covariates (He et al., 2007). An R package has been designed for performing SIMEX for AFT models with measurement error on covariates (Xiong et al., 2010).

The SIMEX method also has been extended to the case of misclassification (MC-SIMEX) in discrete covariates or responses in the longitudinal regression model (Küchenhoff et al., 2006; Lederer and Küchenhoff, 2006) and the clustered survival model (Slate and Bandyopadhyay, 2009). The asymptotic variance for MC-SIMEX has been developed by Küchenhoff et al. (2007). An R package has been designed for performing SIMEX and MC-SIMEX (Lederer and Kchenhoff, 2009). The SIMEX method also has been extended to nonparametric regression in the presence of covariate measurement error (Staudenmayer and Ruppert, 2004),

The benefits of the SIMEX method include its flexibility and the ability of visually presenting the biases. However, approximation between the extrapolation function and the expectation of biased estimators is not promising in general. One difficulty of using the SIMEX method is to find a good extrapolation function for $\hat{\theta}(\lambda)$. The extrapolation

function is important and critical for the extrapolation steps, but typically the exact parametric extrapolation function is unknown. Therefore, a quadratic function is generally used to approximate the extrapolation function. The simulated example showed that the estimators that are based on the quadratic function perform well in terms of bias reduction under some smoothness assumptions (Stefanski and Cook, 1995). Cook and Stefanski (1994) demonstrate the use of differences between simulated results and values of extrapolation functions to make choices between candidate extrapolation functions. We will use a plot to demonstrate the idea in our simulation study.

1.3 The SIMEX method for missing data

We propose an approach to reduce the bias of parameter estimates in data with missing values using the SIMEX idea. This approach can be applied to parametric, nonparametric or semi-parametric models. The basic idea is to assume that the effect of missingness on the parameter estimator is a function of the parameter of the missing-data mechanism. We first estimate the parameter of the missing-data model, and then increase the proportion of missing observations gradually. The model of missing data is not limited, it can be parametric, nonparametric, or a combination of several models when there are different missing mechanisms, as long as we can estimate the probability of observing for each record. We build a function that describes the relationship between the bias in the data model parameters and the missing rate and use the function to extrapolate to the situation where there are no missing observations.

Most strategies for dealing with missing data require assumptions about the missing data. In the context of the pattern-mixture model, it is assumed that within each pattern, the missingness is ignorable. In the context of the selection model, the imputation methods rely on assumptions about the missing portion of the data; and the weight based methods make assumptions about the structure of the missing data mechanism

combined with the response model. One advantage of the SIMEX method for missing data is that the SIMEX method does not require that we make assumption on the distribution of the unobserved data. We must, however, assumes that the effect of the missing data mechanism can be consistently estimated. Since we assume that the effect of the missing data mechanism can be reliably estimated, the SIMEX method works well when the missing data model is correctly specified.

One other advantage of SIMEX is that it directly utilizes existing estimation methods and is easy to program. The SIMEX method avoids the complexity inherent in modeling the missing mechanism jointly with the data, and at the same time incorporates information provided by the missing mechanism using standard estimating procedures that are available for fully observed data. Since the missing mechanism is modeled separately, we can keep the response model simple and the type of response model is not limited by the type of the model of missing data mechanism.

In Chapter 2, we formulate the expectation of naïve estimators and describe our approach based on SIMEX to reduce the bias of the naïve estimator in the presence of missing values. An algorithm is also proposed, and issues such as variance estimation are discussed. In Chapter 3, we demonstrate the effectiveness of SIMEX for reducing bias in an example via simulation studies. In Chapter 4, we use a longitudinal study to demonstrate the use of SIMEX. In Chapter 5, we discuss the advantages and the limitations of SIMEX in the missing data context, and describe some future research directions.

CHAPTER 2. METHOD

2.1 Notation and Model

Assume that the full dataset contains n records, and let $i = 1, 2, \ldots, n$ index the n records. For simplicity, a single index i is used, but the records may be clustered or correlated. The vector (Y_i, X_i, R_i, Z_i) denotes the measurements in the ith record, where Y_i denotes the response variable (which could be a multidimensional vector), X_i denotes the explanatory variables for predicting Y_i, R_i denotes the indicator for missingness of Y_i, and Z_i denotes the additional explanatory variables for predicting R_i. The missing indicator $R_i = 1$ if the corresponding variable Y_i is observed; otherwise $R_i = 0$. The letters without subscript i denote the whole set from $i = 1$ to n. For example, $Y = \{Y_i : i = 1, \ldots, n\}$ and $X = \{X_i : i = 1, \ldots, n\}$.

The following sections develop notation for the response model and the missing data mechanism, the simulation model which contains the consistently estimated probability of observing for each record, the SIMEX algorithm, and large sample and finite sample properties of the SIMEX estimator.

2.1.1 The response model

Assume that the conditional distribution of the response variable Y_i is

$$Y_i|x_i \quad \sim \quad f(y|x_i;\theta)$$

where θ is a vector of parameters for the distribution of Y_i.

The main objective is to make inference about θ in the response model from samples drawn from the distribution $f(Y|x;\theta)$. Assume that we have a consistent estimator for θ, say $\hat{\theta} = T(Y)$. When missing data occurr, we need to make inference based on $Y^{(o)}$ instead of Y.

Assume that the mass function of missing indicator $R_i|(y_i, x_i, z_i)$ is

$$f(r|y_i, x_i, z_i; \eta) = p_i^r (1 - p_i)^{(1-r)}, \tag{2.1}$$

where z denotes the covariates associated with the missing mechanisms, η denotes the parameters for the distribution of $R_i|Y_i$, and $p_i \equiv P(R_i = 1|y, x, z; \eta) \in (0, 1)$. Given $R = r$, the observed sample $Y^{(o)}$ has the following distribution (Rubin, 1976)

$$Y^{(o)}|r \sim \int_{\Omega(Y^{(m)})} \frac{f(y|x;\theta) f(r|y, x, z; \eta)}{f(r|x, z; \theta, \eta)} dy^{(m)}. \tag{2.2}$$

The correct inference about θ should be made from either the conditional distribution of observed data given r, $f(y^{(o)}|r, x, z; \theta, \eta)$, or the joint distribution $f(y^{(o)}, r|x, z; \theta, \eta)$. Either way, it is more complicated than to make inference from $Y = (Y^{(o)}, Y^{(m)}) \sim f(y|x, \theta)$.

2.1.2 The missing-data mechanism

Recall that the mass function of missing indicator $R_i|(y_i, x_i, z_i)$ is defined in Equation 2.1. The probability of observing the ith record, p_i, is a function of (y, x, z, η). Define $\mathcal{P} = \{p_i : i = 1, \ldots, n\}$ as the collection of probabilities. Given (y, x, z, η), the distribution of R is $f(r|y, x, z; \eta) = f(r|\mathcal{P})$. Assume that $\{R_i|\mathcal{P} : i = 1, \ldots, n\}$ are independent given $\mathcal{P}$, the joint distribution of $R = (R_1, \ldots, R_n)'$ can be expressed as $f(r|y, x, z; \eta) = f(r|\mathcal{P}) = \Pi_{i=1}^n p_i^{r_i} (1 - p_i)^{1-r_i}$. In this chapter, we assume that all the values of $p_i = P(R_i = 1|Y^{(o)}, x, z; \theta, \eta)$ are either fixed and known or can be consistently estimated.

If we ignore the missingness and use $Y^{(o)}$ as Y, the estimator is called the naïve estimator and it is denoted as $\hat{\theta}_{\text{naïve}} = T(Y^{(o)}) = T(Y|R)$. Since the distribution of

observed $Y^{(o)}$ is no longer $f(Y|x;\theta)$, the distribution of naïve estimator $T(Y|R)$ may be different from the estimator $T(Y)$.

2.1.3 The simulation model

To estimate the expectation of naïve estimator with increasing missing rate, we generate new missing indicators given observed missing indicators R in the simulation step from the simulation model. In this section, we discuss the simulation model and the marginal distribution of new missing indicators generated in the simulation step.

First, consider a random process $R_i^{(u)}$ for the ith record for $i = 1, \ldots, n$ and $u >= 0$ where u is a nonnegative real number that controls the probability of $R_i^{(u)} = 1$ for all i. Assume the random process $R^{(u)}$ has independent and stationary increments. Given u, $R_i^{(u)}$ is a binary random variable with mass function

$$P(R_i^{(u)} = r|\mathcal{P}) = \left(\pi_i^{(u)}\right)^r \left(1 - (\pi_i^{(u)})^{(1-r)}\right), \tag{2.3}$$

where $\pi_i^{(u)} \equiv P(R_i^{(u)} = 1|\mathcal{P}) = 1 - P(R_i^{(u)} = 1|\mathcal{P}) = p_i^u$ for $u \geq 0$. For example, $\pi_i^{(0)} = 1$ and $\pi_i^{(1)} = p_i = P(R = 1|Y)$ for all $i = 1, \ldots, n$ given $\mathcal{P}$. The random vector $R^{(u)} = \{R_1^{(u)}, R_2^{(u)}, \ldots, R_n^{(u)}\}$ has values in $\Omega\left(R^{(u)}\right) = \bigotimes_{i=1}^n \{0, 1\}$. For each $r^{(u)} \in \Omega(R^{(u)})$, define the corresponding observed and missing sets of indexes

$$\begin{aligned} \mathcal{I}^o(R^{(u)}) &= \{i : i = 1, \ldots, n, r_i^{(u)} = 1\}, \\ \mathcal{I}^m(R^{(u)}) &= \{i : i = 1, \ldots, n, r_i^{(u)} = 0\}. \end{aligned} \tag{2.4}$$

Let $u > 1$ be a real number. Let $R_i^{(u-1)}$ be a random variables with mass function described in (2.3). The marginal probability $P(R_i^{(u-1)} = 1|\mathcal{P}) = p_i^{u-1}$. The random variable $R_i^{(u)} \equiv R_i^{(u-1)} \times R_i^{(1)} = max(R_i^{(u-1)}, R_i^{(1)})$ has value one if both $R_i^{(u-1)} = 1$ and $R_i^{(1)} = 1$ and value zero otherwise. Since we assume the random process has independent and stationary increments, the probability of $R_i^{(u)} = 1$ given $\mathcal{P}$ is

$$P(R_i^{(u)} = 1|\mathcal{P}) = P(R_i^{(u-1)} = 1|\mathcal{P})P(R_i^{(1)} = 1|\mathcal{P}) = p_i^{u-1+1} = p_i^u = \pi_i^{(u)}.$$

Then, consider the situation that one specific $\{R_i^{(1)}, i = 1, \ldots, n\}$ is observed at $u = 1$. When the value of the random process $R_i^{(u)}$ is observed at a fixed $u = 1$, we can find the conditional distribution of the random process at any other point. The conditional probability of $R_i^{(u)} = 1$ given $R_i^{(1)}$ is

$$\pi_i^{(u|1)} \equiv P(R_i^{(u)} = 1|R_i^{(1)}, \mathcal{P}) = \begin{cases} p_i^u & \text{if } R_i^{(1)} = 1, \\ 0 & \text{if } R_i^{(1)} = 0. \end{cases} \tag{2.5}$$

Assume that the distribution of $R = \{R_i\}_{i=1}^n$ can be factorized as $\Pi_{i=1}^n p_i^{r_i}(1-p_i)^{1-r_i}$ given $\mathcal{P}$. The joint distribution of $r^{(u)}$ is

$$f(r^{(u)}|\mathcal{P}) = \prod_{i=1}^{n} (p_i^u)^{r_i^{(u)}}(1-p_i^u)^{1-r_i^{(u)}}.$$

Similarly, the conditional distribution of $R^{(u|1)}$ given $R^{(1)}$ is

$$f(r^{(u|1)}|r^{(1)}, \mathcal{P}) = \prod_{i=1}^{n} (\pi_i^{(u|1)})^{r_i^{(u|1)}}(1-\pi_i^{(u|1)})^{1-r_i^{(u|1)}}. \tag{2.6}$$

We will use the conditional distribution described above to generate new missing indicators in the simulation step. For each set of new missing indicators $R^{(u|1)}$ generated from (2.6), $\hat{\theta}^{(u)} = T(Y, R^{(u|1)})$ is the naïve estimator calculated from $Y^{(o,u|1)} = \{Y_i; i = 1, \ldots, n, R_i^{(u|1)} = 1\}$.

In summarizing, we need the following assumptions:

- The sample size is moderately large such that there are enough samples left to estimate the expectation of naïve estimators with more missing values.

- The joint distribution of $P(R = r|\mathcal{P})$ can be factorized as

$$\Pi_{i=1}^n P(R_i = 1|\mathcal{P})^{r_i}(1 - P(R_i = 1|\mathcal{P}))^{1-r_i}. \tag{2.7}$$

- $p_i \equiv P(R_i = 1|y, x, z; \eta) \in (0, 1)$ for $i = 1, \ldots, n$. can be consistently estimated which means that the missing data are MAR given model explanatory variables x and auxiliary variables z.

- The random process $R^{(u)}$ has independent and stationary increments.

The assumption in (2.7) corresponds to arbitrary missing patterns with independent $\{R_i\}$ given $\mathcal{P}$ or a monotone missing pattern. The missing-data mechanism may be assigned to each record (eg. probability of missing of the jth visit of the ith patient) or to each variable of each record (eg. probability of missing of the answer of the kth question of jth visit of the ith patient).

2.2 The SIMEX Algorithm

We start from an estimator of θ, $\hat{\theta} = T(Y)$, which is consistent when data are fully observed or the missingness is MCAR. The consistent estimator $T(Y)$ has expectation $\mathbb{E}(T(Y)) \to \theta$ as $n \to \infty$. For fixed $u \geq 0$, define $\hat{\theta}^{(u)}_{\text{naïve}} = T(Y, R^{(u|1)}) = T(Y^{(o,u|1)})$ as the simulated naïve estimator involving only $Y^{(o,u|1)}$ selected by $R^{(u|1)}$.

We invent an single additional parameter $u \geq 0$ and generate additional missing indicators $R^{(u|1)} = \{R_i^{(u|1)}\}$ from $\{P(R_i^{(u|1)} = 1 | R_i^{(1)} = 1) = p_i^{u-1}, i = 1, \ldots, n\}$. We connect the conditional expectation of $T(Y, R^{(u|1)})$ given (Y, R) with an augmented value of u and analyze the trend of expectation of T as a function of u. The purpose of inventing u is to describe values of conditional expectations of simulated naïve estimators, $T(Y, R^{(u|1)})$, along a single parameter u. The conditional expectations form a smooth function of u for $u \geq 1$.

Let y and r be specific realizations of Y and R respectively. Let $r^{(u|1)}$ be realization of $R^{(u|1)}$ which has distribution described in (2.5). Assume that the probability of observing $p_i = P(R_i = 1 | Y^{(o)}, X, Z)$ are consistently estimated for each i. The algorithm for finding the SIMEX estimator based on a Kth order polynomial with coefficients $c = (c_0, \ldots, c_K)$ is summarized below and in Figure **??**.

- **Simulation step:**

1. Let $u_0 = 1$ and $u_k = u_{k-1} + (u_{max} - 1)/K^*$ for $k = 1, \ldots, K^*$ where $K^* \geq K + 1$ is the number of conditional expectations that will be estimated in the simulation step and $u_{max} > 1$ is the maximum value of u used in the simulation step. The value of u_{max} depends on the order K and is limited by the sample size.

2. Let $R^{(u_0)} = R$ and $\hat{m}(u_0|Y, R) = T(Y|R) = T(Y^{(o)})$, which is the naïve estimator calculated from observed data set $Y^{(o)}$.

3. Let B be the number of iterations in the simulation step. For each $b \in \{1, \ldots, B\}$, new missing indicators are generated at each u_k, $k = 1, \ldots, K^*$, by following steps:

   (a) Generate $R_{b,i}^{(u_k - u_{k-1})} \sim Ber(p_i^{u_k - u_{k-1}})$.

   (b) Let $R_{b,i}^{(u_k|u_{k-1})} = R_{b,i}^{(u_{k-1})} \times R_{b,i}^{(u_k - u_{k-1})}$.

   (c) Let $\hat{\theta}_b^{u_k} = T(Y, \{R_{b,i}^{(u_k|u_{k-1})}\})$,

4. For $k = 1, \ldots, K^*$, $\hat{m}(u_k|y, r) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^{u_k}$.

- **Extrapolation step:**

1. Estimate coefficients $c = (c_0, \ldots, c_K)$ in the extrapolation function, $M(u; c) = \sum_{k=0}^{K} c_k u^k$, by using the least squares method on points $\{(u_k, \hat{m}(u_k|Y, R)), k = 0, \ldots, K^*\}$.

2. The SIMEX estimator of θ is defined as $\hat{\theta}_{SIMEX} = M(0; \hat{c}) = \hat{c}_0$.

In summary, we estimate $\hat{m}(u_k|y, r^{(u|1)})$ for $k = 1, 2, \ldots, K^*$ by simulation from the conditional distribution. In the simulation step, we generate new missing indicators for $u \in \{u_1, \ldots, u_{K^*}\}$ where $1 < u_1 < \ldots < u_{K^*}$. The first step is to generate independent increments $R^{(u_k|u_{k-1})}$ for $k = 1, \ldots, K^*$ from the conditional distribution given in Equation (2.5). Then, let $R_i^{(u|1)} \equiv R_i^{(u_{k-1}|1)} \times R_i^{(u_k|u_{k-1})}$ which means $R_i^{(u|1)} = 1$ if both $R_i^{(u_{k-1}|1)} = 1$ and $R_i^{(u_k|u_{k-1})} = 1$. If the missing pattern is assumed monotone, the value

of $R^{(u|1)}$ is then adjusted according to the missing pattern assumption. The simulated naïve estimator $T(y|R^{(u|1)})$ is calculated via only $\{y_i; i \in \mathcal{I}^{(o,u|1)}\}$.

In the extrapolation step, we find a Kth order polynomial $M(u;\hat{c})$ that minimizes

$$\sum_{k=0}^{K^*} \left(\hat{m}(u_k|y,r) - M(u_k;c)\right)^2$$

over $c \in \mathbb{R}^{K+1}$. The polynomial $M(u;\hat{c})$ approximates $m(u|y,r^{(u|1)})$ for $1 \leq u \leq u_{K^*}$. In the extrapolation step, define $\hat{\theta}_{SIMEX} = M(0;\hat{c})$.

Figure 2.1 shows an example of marginal expectation of $T(Y,R^{(u)})$, and a fourth order polynomial $N^E(u;c)$ that approximates for $1 \leq u \leq 2$. When a full random sample (y) is completely observed, the red cross denotes the expectation of the consistent estimator $T(y)$. When a partial random sample $(y^{(o,1)})$ is observed, the red diamond($\diamond$) denotes the biased naïve estimator $T(y^{(o,1)})$. For $0 \leq u \leq 1$, the *random path* from $T(y)$ to $T(Y^{(o,u)})$, between the red cross and the red diamond in Figure 2.1, is not observable. For $1 < u \leq 3$, the *conditional expectation* of naïve estimator with increased missing probability given $(y, r^{(1)})$, the red dashed line in Figure 2.1, is estimated by the simulation steps and under a smoothness assumption. A fourth order polynomial, $M(u;\hat{c})$, approximates $\hat{m}(u|y,r)$ within $1 \leq u \leq 2$. The SIMEX estimator $\hat{\theta}_{SIMEX} = M(0;\hat{c})$ is labeled by a inverse triangle($\bigtriangledown$). The SIMEX estimator($\bigtriangledown$) is not meant to capture the value of consistent estimator $T(y)(\times)$. Instead, the marginal expectation of the SIMEX estimator, which is the plus sign(+), is meant to approximate the expectation of $T(Y)$ which goes to true parameter θ when n goes to infinity.

The large sample and finite sample properties are discussed in the following sections. The SIMEX method relies on the functional form for the relationship between the index u and the expectation of naïve estimators. We show that the extrapolation is reasonable especially for missing data analysis.

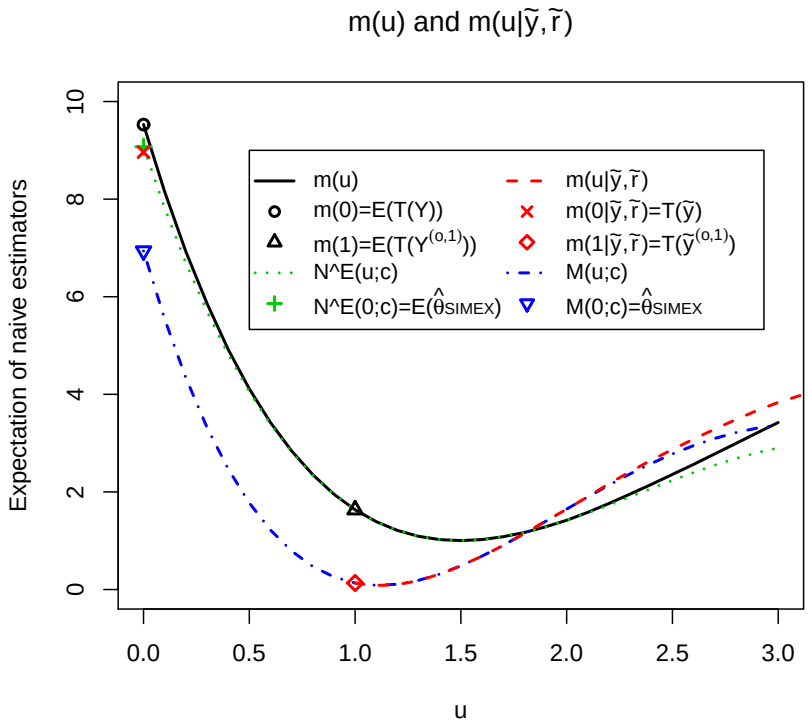

Figure 2.1 An example of marginal expectation function $m(u)$ of $T(Y, R^{(u)})$ on $0 \leq u \leq 3$ and conditional expectation function $m(u|y, r)$ given a fixed sample (y, r) on $1 \leq uleq3$. The fourth order polynomial that approximates $m(u)$ for $1 \leq u \leq 2$ is close to $m(u)$ for $0 < u < 1$ and $2 < u < 3$. The fourth order polynomial that approximates $m(u|y, r)$ for $1 \leq u \leq 2$ is close to $m(u|y, r)$ for $2 < u < 3$.

2.2.1 The choice of coefficients (K, K^*, u_{K^*})

A Kth order polynomial can be defined by $K + 1$ points. Therefore, the minimum value of K^* is K. If $K^* = K$, Runge's phenomenon describes the problem of higher interpolation error near the edges of the interval when interpolating a smooth bounded function by a high order polynomial. The problem gets worse for extrapolation. Dahlquist (1974) demonstrate this phenomenon by an example and suggests using the method of least squares for a lower order polynomial with number of estimated conditional means $K^* > \left(\frac{K}{2}\right)^2$ (e.g. $K = 6$ and $K^* > 9$) such that the polynomial fit is not ill-conditioned.

The value of u_{K^*} depends on the observed sample size and variance of naïve estimator

$T(Y|R^{(u|1)})$ given (Y, R). The value of u_{K^*} should not be too large. One reason is that for any finite sample, the probability of having any sample remaining is nearly zero when u is large, since $p_i^{(u-1)} \to 0$ as $u \to \infty$ for $0 < p_i < 1$. Another reason is that the variance is lager when u_{K^*} is larger. Therefore, a very large number of simulation iteration B may be needed for having a stable estimator of $m(u_{K^*}|Y, R)$ when u_{K^*}. Given finite B, if the plot of $m(u_k|Y, R)$ vs B does not seem to converge at $k = K^*$, the value of K^* should be reduced to have a stable approximation polynomial.

Ideally, the order K of the extrapolation polynomial can be any positive integer and $M(u; c)$ with higher order K should approximate $m(u|Y, R)$ better on $1 \leq u \leq u_{K^*}$. But limited by calculation time and calculation precision, higher order polynomial may be unstable given finite simulation iteration B and finite K^*.

The residual plot, $M(u_k; \hat{c}) - \hat{m}(u_k|\mathcal{P})$, provides information about the suitability of a extrapolation function and provides clues to choose between several extrapolation functions. The smaller residual on $1 \leq u \leq u_{K^*}$ indicates that the polynomial approximates the function $m(u|Y, R)$ well. Some extra points, like $\{m(u_k|Y, R), k = K^* + 1, \ldots\}$, are suggest to be estimated in the simulation step for the diagnose purpose. The smaller residual on $u_{K^*} \leq u \leq u_{K^*} + 1$ indicates better approximation beyond u_{K^*}. Although, good approximation for $u > u_{K^*}$ can not promise good approximation for $u = 0$, but it would be a sign of good overall approximation.

Here are some general suggestion for selecting (K, K^*, u_{K^*}):

- Starting:

  A simple approach is to set $K = 2$ or 3, $u_{K^*} = 2$ and $K^* > 5$.

- Diagnosing:

  A better tuned approach is to increase simulation iterations B and make sure the estimated $m(u|Y, R)$ for $1 < u < 3$ are stable. Draw the residual plot of polynomials with order 1 to 5 and select the order K polynomial with residuals closest to zero

on $1 < u < 3$.

- Improving:

  A even finer tuned approach is to increase the value of B, K, K^* and u_{K^*} one at each time and repeat the process when possible, and try to improve the extrapolation polynomial as good as possible.

2.3 Finite Sample Properties of $\hat{\theta}_{SIMEX}$

This section contains discussions of finite sample properties of $\hat{\theta}_{SIMEX}$. The discussion starts from finding the conditional expectation of the naïve estimator, $T(Y^{(o,u|1)})$. Then, the marginal expectation of those naïve estimators yields properties of the marginal distribution of the SIMEX method. The relationships between functions discussed in this section, including relationships between the conditional and marginal mean functions, the extrapolation function, and the mean of extrapolation function discussed in this section, are summarized in Figure 2.2.

Let y and $r = r^{(1)}$ be specific realizations of Y and $R^{(1)}$ respectively, and let $y^{(o,1)} = \{y_i; i \in \mathcal{I}^o(r)\}$ be the observed data. Let $T(y^{(o,1)}) = T(y, r)$ be the naïve estimator calculated from $y^{(o,1)}$. The distribution of the new naïve estimator, $T(y, R^{(u|1)})$ which is calculated from from $Y^{(o,u|1)}$, depends on observed y and r. Define a function $m(u|y, r)$ as

$$m(u|y,r) \equiv \begin{cases} T(y^{(o,1)}) & u = 1 \\ \mathbb{E}_{R^{(u|1)}|r^{(1)},\mathcal{P}}(T(y, r^{(u|1)})) & u \geq 1 \end{cases} \tag{2.8}$$

The function $m(u|y, r)$ describes the conditional mean of the naïve estimator $T(y, R^{(u|1)})$ given $r^{(1)}$ for $u \geq 1$. By increasing the value of u, the expectation of missing rate is increased and the conditional mean $m(u|y, r)$ is changed according to u.

Since the closed form of $m(u|y, r)$ is usually unknown, in the extrapolation step, we find a polynomial function $M(u; c)$ to approximate $m(u|y, r)$ for $u > 1$. One way to

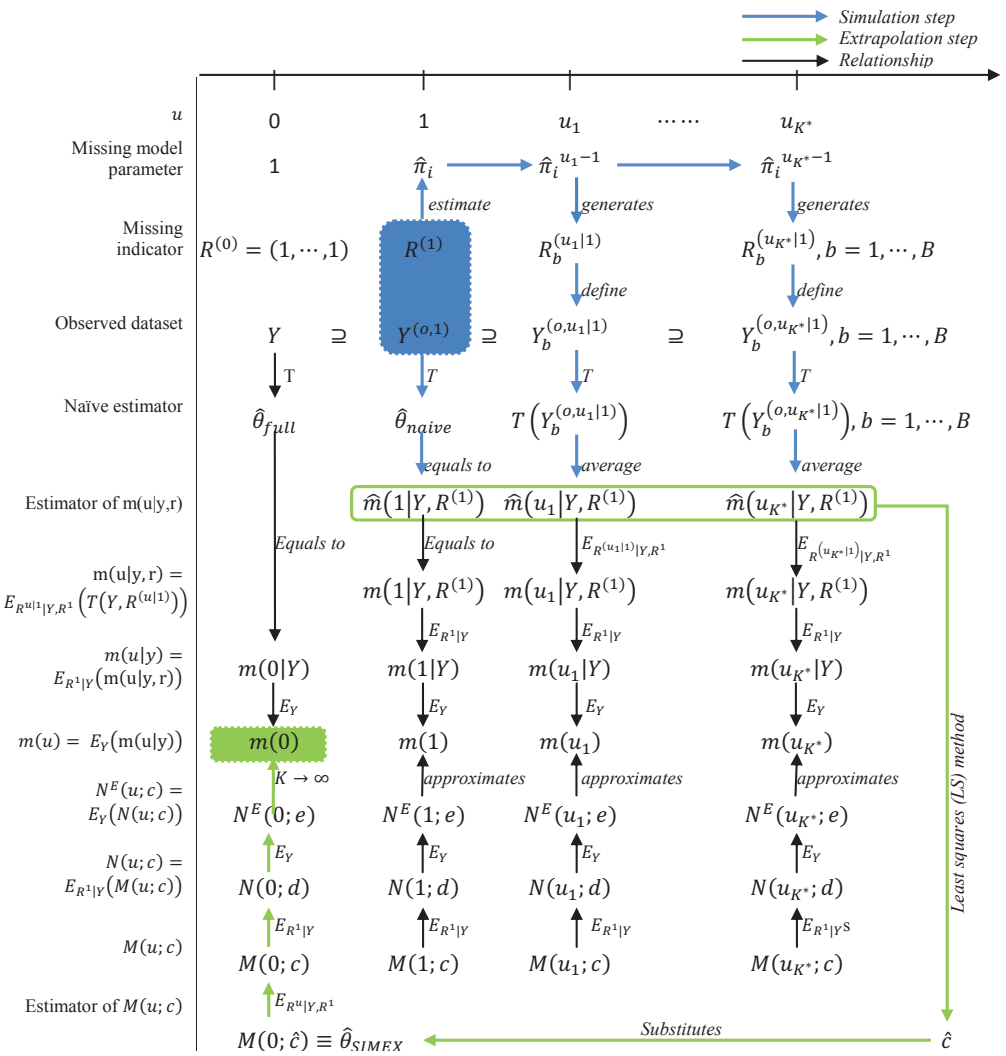

Figure 2.2 Flowchart of the SIMEX method applied to parameter estimation for incomplete data.

approximate $m(u|y,r)$ around $u = 1^+$ is by the Kth order Taylor polynomial expanded at $u = 1^+$. Assume that the exact form of the first K derivatives of $m(u|y,r)$ are known for $u > 1$. The Kth order Taylor polynomial of $m(u|y,r)$ at $u = 1^+$ is

$$M(u;c) = \sum_{k=0}^{K} c_k (u-1)^k, u \geq 1 \tag{2.9}$$

where $c_k = \frac{1}{k!} \lim_{x \to 1^+} \frac{d^k}{dx^k} m(x|y,r)$.

Another way to approximate $m(u|y,r)$ for $u \geq 1$ is to use least squares method to

find a Kth order polynomial based on K^*+1 points where $K^* \geq K$. Let

$$M(u;c) = \sum_{k=0}^{K} C_k u^k, u \geq 1. \tag{2.10}$$

Let $1 = u_0 < u_1 < \cdots < u_K^*$ where $K^* \geq K$. The coefficients c can be estimated by minimizing

$$\sum_{k=1}^{K^*} (\hat{m}(u_k|y,r) - M(u_k;c))^2.$$

If $K < K^*$, the value of $M(u;\hat{c})$ may not be $T(y^{(o)})$ when $u = 1$ If $K = K^*$, the vector of estimated coefficients $\hat{c}$ is the solution of the following $K+1$ equations,

$$\sum_{k=0}^{K} C_k u_j^k = \hat{m}(u_k|y,r). \tag{2.11}$$

and $M(u;\hat{c})$ has value $T(y^{(o)})$ when $u = 1$.

There difficulties of using Taylor polynomials as the extrapolation function: (1) When the sample size is large, the derivatives may take more calculation time. (2) The derivatives are complicated when $K > 2$. (3) We have ignored the probability of having $R^{(u|1)}$ which makes $T(y, R^{(u|1)})$ does not exist in the derivatives listed in Appendix A. Therefore, finding the polynomial from K^*+1 points is usually easier than finding the Taylor polynomial from derivatives.

2.3.1 The function $M(u;c)$ converges to $m(u|Y,R)$ as $K \to \infty$

Recall that $R^{(1)}$ denotes the observed missing indicators and $R^{(u|1)}$ denotes the new missing indicators generated from $f_{R^{(u|1)}}(r)$ conditional on $R^{(1)}$ with larger probability of missing controlled by increased $u > 1$. The following theorem shows that the function $m(u|y,r)$ is a continuous smooth analytic function for $u \in \{1, u_{K^*}\}$. Therefore, the extrapolation function $M(u;c)$ which is either the Taylor polynomial expanded at $u = 1$ or the least squares polynomial, converges to $m(u|y,r)$ as order K goes to infinity for every $u \in \{1, u_{K^*}\}$.

Theorem 2.3.1. *The function $m(u|Y,R)$ defined in (2.8) is analytic on $(1, u_{K^*}+1)$.*

Proof. For $u \in (1, u_{K^*}+1)$, the function $m(u|Y,R)$ defined in Equation 2.8 is a linear combination of $\{f(r^{(u|1)}|R^{(1)},\mathcal{P}), r^{(u|1)} \in \Omega(R^{(u|1)})\}$. For each $r \in \Omega(R^{(u|1)})$, let $n(r) \equiv \sum_i R_i^{(1)} - \sum_i r_i$ which is the number of indexes in the set $\mathcal{I}^m(r) \cap \mathcal{I}^o(R^{(1)})$. Let $\mathcal{S}(r) = \{s_j = (s_{ji}), s_{ji} \in \{0,1\}, i \in \mathcal{I}^m(r) \cap \mathcal{I}^o(R^{(1)})\}$. $S(r)$ is a vector with length $n(r)$. Each element of $S(r)$ has value either 0 or 1. and let $J(r) = 2^{n(r)}$ be the size of $\mathcal{S}(r)$. For each $r \in \Omega(R^{(u|1)})$,

$$\begin{aligned} f(r|R^{(1)},\mathcal{P}) &= \Pi_{i\in\mathcal{I}^o(R^{(1)})} p_i^{(u-1)r_i}(1-p_i^{u-1})^{1-r_i} \\ &= \left(\Pi_{i\in\mathcal{I}^o(r)} p_i^{u-1}\right)\left(\Pi_{i\in\mathcal{I}^m(r)\cap\mathcal{I}^o(R^{(1)})}(1-p_i^{u-1})\right) \\ &= a_0(r,u|1)\left(\sum_{j=1}^{J(r)} a_j(r,u|1)\right), \end{aligned}$$

where $a_0(r,u|1) = \Pi_{i\in\mathcal{I}^o(r)} p_i^{u-1}$, and $a_j(r,u|1) = \Pi_{i\in\mathcal{I}^m(r)\cap\mathcal{I}^o(R^{(1)})}(-p_i)^{s_{ji}} \in (-1,0)\cup(0,1)$. Let $g(u) = |a_j|^{u-1} \in (0,1)$. The function $g(u)$ has the kth order derivative $(log(|a_j|))^k |a_j|^{u-1}$ for any positive integer k. For every $u \in (1, u_{K^*}+1)$,

$$\begin{aligned} \left|\frac{d^k}{du^k} a_j\right| &= (-log(|a_j|))^k |a_j|^{u-1} \\ &\leq (-log(|a_j|))^k \\ &\leq (\max_{j=1}^{J} -log(|a_j|))^k \\ &= \left(-log\left(\Pi_{i\in\mathcal{I}^o(R^{(1)})} p_i^{u-1}\right)\right)^k, \end{aligned}$$

since the smallest value of $|a_j|$ is $\Pi_{i\in\mathcal{I}^o(R^{(u|1)})} p_i^{u-1} \Pi_{i\in\mathcal{I}^m(R^{(u|1)})\cap\mathcal{I}^o(R^{(1)})}(p_i^{u-1})$. Therefore, a_j^{u-1} is analytic on $(1, u_{K^*})$. The functions $f(r|R^{(1)},\mathcal{P})$ and $m(u|Y,R)$, which are linear combinations of a_j^{u-1}, are analytic on $(1, u_{K^*}+1)$. □

We only focus on the finite open set $(1, u_{K^*}+1)$ since the extrapolation function approximates $m(u|y,r)$ in that set.

When $K \to \infty$, $M(u;c) = \sum_{k=1}^{\infty} c_k u^k \to m(u|Y,R)$ for any $u \in (1, u_{K^*}+1)$. Further, $M(u;c)$ and $m(u|Y,R)$ are continuous on $[1, u_{K^*}+1]$. Therefore, $lim_{u\to 1^+} M(u;c) = M(1;c) = m(1|Y,R) = lim_{u\to 1+} m(u|Y,R)$.

Note that we are not interested in the function $M(u;c)$ for $0 \leq u < 1$, but we are interested in the expectation of $M(u;c)$ with respect to $f_{R^{(1)}|Y}$ or $f_{R^{(1)},Y}$. Since, for $0 \leq u < 1$, the expectation of $M(u;c)$ approximates the expectation of $m(u|Y,R)$. In later sections, we show that the marginal expectation of the naïve estimator is a smooth and analytic function of u for $0 \leq u \leq u_{K^*}$. Therefore, the marginal expectation of $M(u;c)$ converges to the marginal expectation of naïve estimators for $0 \leq u \leq u_{K^*}$. Which means the marginal expectation of the SIMEX estimator converges to the marginal expectation $E_Y(\hat{\theta})$ when $K \to \infty$.

2.3.2 The conditional expectations $m(u|Y)$ and $N(u;d)$

We first define functions that describe the conditional expectation of the naïve estimator given Y,

$$\begin{aligned} m(u|Y) &\equiv E_{R^{(u)}|Y}\left(T(Y,R^{(u)})\right) \\ &= \sum_{r^{(u)}\in\Omega(R^{(u)})} T(Y,r^{(u)})f(r^{(u)}|\mathcal{P}), \end{aligned} \tag{2.12}$$

for $u \geq 0$ where $f(r^{(u)}|\mathcal{P})$ is defined in (2.3) and $\mathcal{P} = (Y,X,\theta,\eta)$ is fixed. The function $m(u|Y)$ is a smooth continuous function of u for $u \geq 0$. The first two derivatives of $m(u|Y)$ exist for every $u > 0$ and are described in Appendix A.3. The function $m(u|Y)$ is the conditional expectation of $m(u|Y,R)$ for $u \geq 1$,

$$\begin{aligned} m(u|Y) &= E_{R|Y}\left(E_{R^{(u)}|Y,R}\left(T(Y,R^{(u|1)})\right)\right) \\ &= E_{R|Y}\left(m(u|Y,R)\right), \end{aligned}$$

for $u \geq 1$.

Additionally, define the polynomial $N(u;d)$ as the conditional expectation of the

extrapolation function $M(u;c)$ given Y,

$$\begin{aligned} N(u;d) &\equiv E_{R|Y}\left(M(u;c)\right) \\ &= \sum_{k=0}^{K} E_{R|Y}(c_k)u^k \\ &= \sum_{k=0}^{K} d_k u^k, \end{aligned} \tag{2.13}$$

where $d_k = E_{R|Y}(c_k)$ for $u \geq 0$. Each of the elements of coefficient $c = (c_0, c_1, \ldots, c_K)$ in (2.10) is a linear combination of $\{m(u_k|Y,R), k = 0, \ldots, K^*\}$ for $k = 0, 1, \ldots, K$. Let

$$c_k = \sum_{j=0}^{K^*} v_{k,j} m(u_k|Y,R)$$

where $v_{k,j}$ is a function of $\{u_j^s; j = 0, 1, \ldots, K^*; s = 0, 1 \ldots, K\}$. Each of the elements of coefficient $d = (d_0, d_1, \ldots, d_K)$ in (2.13) is a linear combinations of $\{m(u_k|Y), k = 0, \ldots, K^*\}$ for $k = 0, 1, \ldots, K$, since

$$\begin{aligned} d_k &= \sum_{j=0}^{K^*} v_{k,j} E_{R|Y}(m(u_k|Y,R)) \\ &= \sum_{j=0}^{K^*} v_{k,j} m(u_k|Y). \end{aligned}$$

The function $N(u;d)$ is actually the polynomial that minimizes the sum of squared differences between $m(u_k|Y)$ and $N(u_k;d)$ for $k = 0, \ldots, K^*$. Which means the coefficient d minimizes

$$\sum_{k=0}^{K^*} \left(m(u_k|Y) - N(u_k;d)\right)^2.$$

Theorem 2.3.2. *The function $m(u|Y)$ defined in (2.12) is analytic on $(0, u_{K^*} + 1)$.*

Proof. For $u \in (0, u_{K^*} + 1)$, the function $m(u|Y)$ defined in Equation 2.12 is a linear combinations of $\{f(r^{(u)}|\mathcal{P}), r^{(u)} \in \Omega(R^{(u)})\}$. For each $r \in \Omega(R^{(u)})$, let $n(r) \equiv n - \sum_i r_i$ be the number of indexes in the set $\mathcal{I}^m(r)$. Let $\mathcal{S}(r) = \{s_j = (s_{ji}), s_{ji} \in \{0,1\}, i \in \mathcal{I}^m(r)\}$ which is a length $n(r)$ vector with elements 0 or 1 and let $J(r) = 2^{n(r)}$ which is the size

of $\mathcal{S}(r)$. For each $r \in \Omega(R^{(u|1)})$,

$$\begin{aligned} f(r|\mathcal{P}) &= \Pi_{i=1,\ldots,n}(p_i^{(u)})^{r_i}(1-p_i^u)^{1-r_i} \\ &= \left(\Pi_{i\in\mathcal{I}^o(r)}p_i^{u-1}\right)\left(\Pi_{i\in\mathcal{I}^m(r)}(1-p_i^{u-1})\right) \\ &= a_0(r,u)\left(\sum_{j=1}^{J(r)} a_j(r,u)\right), \end{aligned}$$

where $a_0(r,u) = \Pi_{i\in\mathcal{I}^o(r)}p_i^u$, and $a_j(r,u) = \Pi_{i\in\mathcal{I}^m(r)}(-p_i)^{s_{ji}} \in (-1,0)\cup(0,1)$. Similar to the proof of Theorem 2.3.1, a_j^u is analytic on $(0, u_{K^*})$. The functions $f(r|\mathcal{P})$ and $m(u|Y)$, which are linear combination of a_j^{u-1}, are analytic on $(0, u_{K^*}+1)$. □

By Theorem 2.3.2, the function $m(u|y)$ is analytic for $u \in (0, u_{K^*}+1)$, and the least squares polynomial $N(u;d)$ converges to $m(u|y)$. Additionally, $N(u;d)$ and $m(u|y)$ are continuous on $[0, u_{K^*}+1)$. The conditional expectation of the SIMEX estimator given Y, $N(0;d) = E_{R|Y}(\hat{\theta}_{SIMEX})$, converges to $m(0|y)$ when order $K \to \infty$.

2.3.3 The mean square error and the variance estimator of $\hat{\theta}_{SIMEX}$

Let $\mu_T = E(T(Y))$ be the expectation of the consistent estimator $T = \hat{\theta}$ when the full dataset Y is observed. Let $\mu_S = E(\hat{\theta}_{SIMEX}) = N^E(0;e)$ be the expectation of the SIMEX estimator T when only $Y^{(o)}$ is observed. The MSE of $\hat{\theta}_{SIMEX}$ is

$$\begin{aligned} MSE(\hat{\theta}_{SIMEX}) &= E((\hat{\theta}-\mu_T)^2) \\ &= E((\hat{\theta}_{SIMEX}-\mu_{SIMEX})^2) + (\mu_{SIMEX}-\mu_T)^2 \qquad (2.14) \\ &= Var(\hat{\theta}_{SIMEX}) + Bias(\hat{\theta}_{SIMEX})^2 \end{aligned}$$

The second term in Equation (2.14) will converge to zero as $B \to \infty$, $K \to \infty$ and samples size $n \to \infty$. If $K < \infty$, the second term is the square of unobservable bias. We would like the extrapolation function $M(u;c)$ to be a higher order polynomial for lower bias, but the order will be limited by the sample size and calculation time.

The first term in Equation (2.14) is the variance of the SIMEX estimator. The randomness of the SIMEX estimator came from the randomness of

$$\left(Y, R, \{R_b^{(u_k|1)}\}_{b=1,\ldots,B,k=1,\ldots,K^*}\right).$$

Let $\mathcal{R} \equiv \{R_b^{(u_k|1)}; b = 1,\ldots,B; k = 1,\ldots,u_{K^*}\}$. Let $\mu_{Y,R,\mathcal{R}} \equiv E_{Y,R,\mathcal{R}}(\hat{\theta}_{SIMEX})$ which converges to $\mu_Y = E_Y(\hat{\theta})$ as order of approximation function $K \to \infty$. Let $\mu_{\mathcal{R}|Y,R} \equiv E_{\mathcal{R}|Y,R}(\hat{\theta}_{SIMEX})$. The variance of the SIMEX estimator is

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}_{SIMEX}) &= \mathbb{E}_{Y,R,\mathcal{R}}\left(\left(\hat{\theta}_{SIMEX} - \mu_{Y,R,\mathcal{R}}\right)^2\right) \\
&= \mathbb{E}_{Y,R}\mathbb{E}_{\mathcal{R}|Y,R}\left(\left(\hat{\theta}_{SIMEX} - \mu_{\mathcal{R}|Y,R} + \mu_{\mathcal{R}|Y,R} - \mu_{Y,R,\mathcal{R}}\right)^2\right) \\
&= \mathbb{E}_{Y,R}\mathbb{E}_{\mathcal{R}|Y,R}\left(\left(\hat{\theta}_{SIMEX} - \mu_{\mathcal{R}|Y,R}\right)^2\right) + \mathbb{E}_{Y,R}\mathbb{E}_{\mathcal{R}|Y,R}\left(\left(\mu_{\mathcal{R}|Y,R} - \mu_{Y,R,\mathcal{R}}\right)^2\right) \\
&\quad +2\mathbb{E}_{Y,R}\left(\left(\mu_{\mathcal{R}|Y,R} - \mu_{Y,R,\mathcal{R}}\right)\mathbb{E}_{\mathcal{R}|Y,R}\left(\hat{\theta}_{SIMEX} - \mu_{\mathcal{R}|Y,R}\right)\right) \\
&= E_{Y,R}\underbrace{\mathbb{E}_{\mathcal{R}|Y,R}\left(\left(\hat{\theta}_{SIMEX} - \mu_{\mathcal{R}|Y,R}\right)^2\right)}_{P1} + \underbrace{\mathbb{E}_{Y,R}\left(\left(\mu_{\mathcal{R}|Y,R} - \mu_{Y,R,\mathcal{R}}\right)^2\right)}_{P2} \qquad (2.15)
\end{aligned}$$

The portion $(P1)$ is the variation from the simulation step. We find the intercept of the extrapolation function for each simulation iteration $b = 1,\ldots,B$. We can estimate $(P1)$ by calculating the variance of those intercepts then divided the variance by B. The portion $(P1)$ is the variation from distribution of $Y^{(o,1)}$, and it can be arbitrary small (but limited by computation precision) by increasing the value of B.

Let $\mu_{\mathcal{R},R|Y} \equiv E_{\mathcal{R},R|Y}(\hat{\theta}_{SIMEX}) = \hat{\theta}$. When $K \to \infty$,

$$\begin{aligned}
(P2) &= \mathbb{E}_{Y,R}\left(\left(\mu_{\mathcal{R}|Y,R} - \mu_Y\right)^2\right) \\
&= \mathbb{E}_{Y,R}\left(\left(\mu_{\mathcal{R}|Y,R} - \mu_{\mathcal{R},R|Y} + \mu_{\mathcal{R},R|Y} - \mu_Y\right)^2\right) \\
&= \mathbb{E}_Y\mathbb{E}_{R|Y}\left(\left(\mu_{\mathcal{R}|Y,R} - \mu_{\mathcal{R},R|Y}\right)^2\right) + \mathbb{E}_Y\left(\left(\mu_{\mathcal{R},R|Y} - \mu_Y\right)^2\right) \\
&= \mathbb{E}_Y\underbrace{\mathbb{E}_{R|Y}\left(\left(\mu_{\mathcal{R}|Y,R} - \hat{\theta}\right)^2\right)}_{P2a} + \underbrace{\mathbb{E}_Y\left(\left(\hat{\theta} - \mu_Y\right)^2\right)}_{P2b} \qquad (2.16)
\end{aligned}$$

The portion $(P2b)$ is the variation of the estimator $\hat{\theta}$ from the full dataset. We can estimate $(P2b)$ by using the SIMEX method on variance estimators. The three portions $(P1)$, $(P2a)$ and $(P2b)$ are variations from f_Y, $f_{R|Y}$ and $f_{\mathcal{R}|R,Y}$.

One way to estimate $(P2a)$ is by the bootstrap method. Consider Y as a fixed population and $Y^{(o,1)}$ as a sample selected by $f_{R|Y}$. Then, S new sets of sample are selected from $Y^{(o,1)}$ with replacement and $\hat{\theta}_{SIMEX,s}$ are calculated for each new set of sample for $s = 1, \ldots, S$. Then, $(P2a)$ is estimated by the variance of $\hat{\theta}_{SIMEX,s}$. The bootstrap method will increase the already long calculating time.

Another way is to approximate $(P2a)$ by a Taylor approximation.

$$\begin{aligned}
(P2a) &= \mathbb{E}_{R|Y}\left((M(0;c) - m(0|y))^2\right) \\
&= \mathbb{E}_{R|Y}\left((M(1;c) - m(1|y))^2\right) \\
&\quad +2\mathbb{E}_{R|Y}\left((M(1;c) - m(1|y))\left(\frac{d}{du}M(1;c) - \frac{d}{du}m(1|y)\right) + res\right) \\
&= \mathbb{E}_{R|Y}\left((M(1;c) - m(1|y))^2\right) \\
&\quad +2\mathbb{E}_{R|Y}\left(M(1;c)\frac{d}{du}M(1;c) - m(1|y)\frac{d}{du}M(1;c)\right. \\
&\quad \left. -M(1;c)\frac{d}{du}m(1|y) + m(1|y)\frac{d}{du}m(1|y) + res\right) \\
&= \mathbb{E}_{R|Y}\left((m(1|y,r) - m(1|y))^2\right) \\
&\quad +2\mathbb{E}_{R|Y}\left(M(1;c)\frac{d}{du}M(1;c)\right) - 2m(1|y)\frac{d}{du}m(1|y) + \mathbb{E}_{R|Y}(res) \\
&= var_{R|Y}\left(M(1;c)\right) + 2cov\left(M(1;c), \frac{d}{du}M(1;c)\right) + \mathbb{E}_{R|Y}(res) \\
&\approx \underbrace{var_{R|Y}\left(M(1;c)\right)}_{P2a1} + 2\underbrace{cov\left(M(1;c), \frac{d}{du}M(1;c)\right)}_{P2a2}. \qquad (2.17)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}_{SIMEX}) &\approx E_{Y,R}\underbrace{\mathbb{E}_{\mathcal{R}|Y,R}\left(\left(\hat{\theta}_{SIMEX} - \mu_{\mathcal{R}|Y,R}\right)^2\right)}_{P1} + \underbrace{\mathbb{E}_Y\left(\left(\hat{\theta} - \mu_Y\right)^2\right)}_{P2b} \\
&\quad + \underbrace{var_{R|Y}\left(m(1|y,r)\right)}_{P2a1} + 2\underbrace{cov\left(m(1|y,r), \frac{d}{du}m(1|y,r)\right)}_{P2a2} \qquad (2.18)
\end{aligned}$$

where $(P2a1)$ and $(P2a2)$ are calculated from bootstrap samples from $Y^{(o,1)}$ The resampling is taking on each subject when the missing pattern is monotone and missing indicators R are correlated within each subject. The derivative $\frac{d}{du}M(1;c) = \frac{d}{du}m(1+|y,r)$ can be calculated from equations i in Appendix A.1 When the sample size is large, it takes a longer time to calculate the derivative. The derivative can also be approximated by the slope of the linear function that passes through $(1, T(Y^{(o,1)(j)})$ and $(1+\delta, m(1|Y^{(o,1+\delta)(j)})$ for small δ and for the jth bootstrap sample $Y^{(o,1)(j)}$.

2.3.4 Sensitivity of assumptions on missing data mechanism

The SIMEX method explores the functional relationship between bias and the power of probability of not missing. Therefore, it is necessary to have good estimates of probability of not missing. We may try several different missing models to see the effect on the SIMEX estimator (Baker et al., 2003). One common choice for the binary missing indicator is a generalized linear model with binary response and logit link to a linear function of the covariates. We change the structure of the linear predictor to explore the assumption on the missing data mechanism. An option under the MCAR assumption is the logistic model with only an intercept in the linear predictor.

When the expectation of an estimator depends on the sample size n, which is very likely, the corresponding mean function shows upward or downward trends because of the sample size reduction. That can yield confusion between effects of different sample sizes and effects of different linear predictors on each value of $u > 1$. We can adjust the total sample size. The adjustment of sample size only emphasizes the effect of different assumptions for $u > 1$, but it won't change the value of the SIMEX estimator under substitute linear predictors of R. The adjustment is only needed if we are interested in comparing the trends of $m(u|y,r)$ for $u > 1$.

2.3.5 When the remaining sample size $\sum_i r_i^{(u)}$ is too small for estimation

Consider the random set $(Y, R^{(u)}) \sim f(y|x;\theta)f(r^{(u)}|\mathcal{P})$ where $\mathcal{P} = \{p_i : i = 1, \ldots, n\}$. Let $Y^{(o,u)} = \{Y_i : i = 1, \ldots, n, R_i^{(u)} = 1\} = \{Y_i : i \in I^{(o,u)}\}$. The probability of all data being missing is $P(Y^{(o,u)} = \emptyset) = P(\sum_{i=1}^n R_i^{(u)} = 0) = \prod_{i=1}^n (1 - p_i^u) > 0$. When all data are missing, there is no information for making inference and this can cause some problems during computation. Let n_T be the minimum sample size needed for calculating T. We are indeed sampling $R^{(u)}$ from the conditional distribution, $f(r^{(u)}|\mathcal{P}, \sum_{i=1}^n r_i^{(u)} > n_T) = \frac{f(r^{(u)}, \sum_{i=1}^n r_i^{(u)} > n_T|\mathcal{P})}{f(\sum_{i=1}^n r_i^{(u)} > n_T|\mathcal{P})}$, instead of $f(r^{(u)}|\mathcal{P})$. The last row in Table A.1 is an example of no sample remaining for estimation. When $u < 3$, n is moderately large and the p_i are not very close to zero, $P(\sum_{i=1}^n r_i^{(u)} > n_T) \approx 1$. The problem of insufficient sample remaining only affects the Taylor extrapolation function since we find the derivatives without considering the probability of $P(\sum_{i=1}^n r_i^{(u)} < n_T)$. During the simulation step, the random vector $\{R^{(u_k|1)}\}$ is obtained given $\sum_{i=1}^n r_i^{(u)} > n_T$, so we don't need to worry about this problem. If for each variable, there are moderate size of missing but jointly small percentage of records are complete, combine imputation in each simulation step is suggested.

2.3.6 Transformation of u

For easier explanation, we may plot the extrapolation function $M(u)$ against $g(u)$ which is any strictly monotone function of u, or we can build another extrapolation function based on $g(u)$. The extrapolation function $M(u)$ performs better when the function $m(u|\tilde{y}, \tilde{r})$ is flat for $0 \leq u \leq 2$.

For example, we are interested in the averaged missing proportion. Let

$$g(u) \equiv 1 - \tilde{\bar{p}}^{(u)} = 1 - \frac{1}{n}\sum_{i=1}^n \tilde{p}_i^u = 1 - \frac{1}{n}E(I^{(o,u)}) = \frac{1}{n}E(I^{(m,u)})$$

which has $\frac{d}{du}g(u) = -\frac{1}{n}\sum_{i=1}^n \tilde{p}_i^u log(p_i)$ and $\frac{d^2}{du^2}g(u) = -\frac{1}{n}\sum_{i=1}^n \tilde{p}_i^u (log(p_i))^2$. We can draw $(g(u), M(u))$ instead of $(u, M(u))$. Further, the extrapolation function based on

$M(g(u))$ is

$$\hat{M}(g(0)) = \hat{M}(g(1)) - \lim_{g(u)\to g(1^+)} \frac{d}{dg(u)} M(g(u)) + \frac{1}{2} \lim_{g(u)\to g(1^+)} \frac{d^2}{dg(u)^2} M(g(u))$$

where

$$\frac{d}{dg(u)} M(g(u)) = \frac{d}{du} M(u) \frac{1}{\frac{d}{du} g(u)}$$

and

$$\frac{d^2}{dg(u)^2} M(g(u)) = \frac{1}{\left(\frac{d}{du} g(u)\right)^2} \frac{d^2}{du^2} M(u) - \frac{1}{\left(\frac{d}{du} g(u)\right)^3} \frac{d}{du} M(u) \frac{d^2}{du^2} g(u).$$

We can draw $(g(u), M(g(u)))$ and compare with $(u, M(u))$. Comparing to $g(u) = u$, this function $g(u) = 1 - \bar{\tilde{p}}^{(u)}$ is steeper at $u \leq 1$ which is less favored as an extrapolation function. In later simulation, we have plots for the comparison and the resulting extrapolation estimator from $M(g(u))$ has more bias than estimators from $M(u)$.

2.3.7 Combine with imputation based methods

When missing values occur on more than one variable, the methods using only complete records may lose too much efficiency. In the simulation step, the imputation method can fill in missing data before each model fitting. When the imputation model is not correctly specified, as long as the missing-data mechanism is correctly specified, the extrapolation function reflects both of effects of missingness and inaccurate imputation model. The expectations of naïve estimators still produce a smooth extrapolation function. Some further discussion is provided in the simulation study.

2.4 The Large Sample Properties of $\hat{\theta}_{SIMEX}$

Let Y_n denote a size n data. The function $m(u|Y_n)$ defined in (2.12) is a smooth and analytic functions for $0 < u < u_{K^*}$. Previously, we have shown that when the paremeters

η are known, the conditional expectation of SIMEX estimator

$$\begin{aligned} E_{R|Y_n}(\lim_{K\to\infty}\hat{\theta}_{SIMEX}) &= E_{R|Y_n}(\lim_{K\to\infty}M(0;c)) \\ &= \lim_{K\to\infty}E_{R|Y}(M(0;c)) \\ &= \lim_{K\to\infty}N(0;d) \\ &= m(0|Y_n) \\ &= T(Y_n). \end{aligned}$$

Therefore, for any fixed sample size n, the expectation of $\hat{\theta}_{SIMEX}$ is exactly the expectation of $T(Y_n)$,

$$E_{(R,Y_n)}(\lim_{K\to\infty}\hat{\theta}_{SIMEX}) = E_{Y_n}(T(Y_n)).$$

For simiplicity, define

$$\begin{aligned} \hat{\theta}_s &\equiv \lim_{K\to\infty}M(0;c) \\ &= \lim_{K\to\infty}\lim_{B\to\infty}M(0;\hat{c}). \end{aligned}$$

The expectation $E_{R|Y}(\hat{\theta}_S) = T(Y)$ for any fixed sample size n. In following sections, we discuss the limit of a SIMEX estimator when n goes to ∞.

2.4.1 When $T(Y_n)$ is a consistent estimator of θ

Assume that $T(Y_n)$ is a consistent estimator. For every $\epsilon > 0$ and for every $\theta \in \Theta$,

$$\lim_{n\to\infty}P_\theta\left(|T(Y_n)-\theta|<\epsilon\right) = 1.$$

Therefore, for every $\epsilon > 0$ and for every $\theta \in \Theta$,

$$\lim_{n\to\infty}P_\theta\left(\left|E_{R|Y_n}\left(\hat{\theta}_S\right)-\theta\right|<\epsilon\right) = 1. \tag{2.19}$$

The expectation $E_{R|Y}(\hat{\theta}_S)$ converges to θ in probability for every $\theta \in \Theta$.

But the SIMEX estimator $\hat{\theta}_S$ itself may not converge to θ. Consider a partition of sample space, $\Omega(Y_n, R_n) = \cup_{j=1}^{4} \Omega_j(Y_n, R_n)$ where

$$\begin{aligned}
\Omega_{1,n} &= \{(Y_n, R_n); \left| E_{R_n|Y_n}(\hat{\theta}_S) - \theta \right| \geq \epsilon, \left| \hat{\theta}_S - \theta \right| < \epsilon\}, \\
\Omega_{2,n} &= \{(Y_n, R_n); \left| E_{R_n|Y_n}(\hat{\theta}_S) - \theta \right| \geq \epsilon, \left| \hat{\theta}_S - \theta \right| \geq \epsilon\}, \\
\Omega_{3,n} &= \{(Y_n, R_n); \left| E_{R_n|Y_n}(\hat{\theta}_S) - \theta \right| < \epsilon, \left| \hat{\theta}_S - \theta \right| < \epsilon\}, \text{and} \\
\Omega_{4,n} &= \{(Y_n, R_n); \left| E_{R_n|Y_n}(\hat{\theta}_S) - \theta \right| < \epsilon, \left| \hat{\theta}_S - \theta \right| \geq \epsilon\}.
\end{aligned}$$

To discuss the limit of $\hat{\theta}_S$, consider the probability

$$P_{\theta,\eta}\left(\left| \hat{\theta}_S - \theta \right| \geq \epsilon \right) = P_{\theta,\eta}\left((Y_n, R_n) \in \Omega_{2,n} \right) + P_{\theta,\eta}\left((Y_n, R_n) \in \Omega_{4,n} \right). \quad (2.20)$$

where η denotes the parameters for the distribution of R_i given Y_i. By Equation 2.19, for every $\epsilon > 0$, for every $\theta \in \Theta$ and for a fixed η,

$$\lim_{n\to\infty} P_{\theta,\eta}\left((Y_n, R_n) \in \Omega_{2,n} \right) = 0.$$

That means the first term in Equation 2.20 converges to 0 as $n \to \infty$.

The second term in Equation 2.20,

$$\begin{aligned}
P_{\theta,\eta}\left((Y_n, R_n) \in \Omega_{4,n} \right) &= P_{\theta,\eta}\left(\left| E_{R_n|Y_n}(\hat{\theta}_S) - \theta \right| < \epsilon, \left| \hat{\theta}_S - \theta \right| \geq \epsilon \right) \\
&= P_\theta\left(\left| E_{R_n|Y_n}(\hat{\theta}_S) - \theta \right| < \epsilon \right) \\
&\quad \times P_\eta\left(\left(\left| \hat{\theta}_S - \theta \right| \geq \epsilon \right) \middle| \left(\left| E_{R_n|Y_n}(\hat{\theta}_S) - \theta \right| < \epsilon \right) \right) \\
&= P_\theta\left((Y_n, R_n) \in \Omega_{3,n} \cup \Omega_{4,n} \right) \\
&\quad \times P_\eta\left((Y_n, R_n) \in \Omega_{4,n} \middle| (Y_n, R_n) \in \Omega_{3,n} \cup \Omega_{4,n} \right)
\end{aligned}$$

By Equation 2.19, the limit probability

$$lim_{n\to\infty} P_\theta\left((Y_n, R_n) \in \Omega_{3,n} \cup \Omega_{4,n} \right) = 1.$$

The probability

$$\begin{aligned}
& P_\eta\left((Y_n, R_n) \in \Omega_{4,n} \middle| (Y_n, R_n) \in \Omega_{3,n} \cup \Omega_{4,n} \right) \\
< \; & P_\eta\left(\left(\left| \hat{\theta}_S - E_{R_n|Y_n}(\hat{\theta}_S) \right| \geq \epsilon - \left| E_{R_n|Y_n}(\hat{\theta}_S) - \theta \right| \right) \middle| (Y_n, R_n) \in \Omega_{3,n} \cup \Omega_{4,n} \right)
\end{aligned}$$

Therefore, if

$$\lim_{n\to\infty} P_\eta\left((Y_n, R_n) \in \Omega_{4,n} \,|(Y_n, R_n) \in \Omega_{3,n} \cup \Omega_{4,n}\right) = 0 \tag{2.21}$$

or

$$\lim_{n\to\infty} P_\eta\left(\left(\left|\hat{\theta}_S - E_{R_n|Y_n}(\hat{\theta}_S)\right| \geq \epsilon - \left|E_{R_n|Y_n}(\hat{\theta}_S) - \theta\right|\right) |(Y_n, R_n) \in \Omega_{3,n} \cup \Omega_{4,n}\right) = 0 \tag{2.22}$$

then

$$\lim_{n\to\infty} P_{\theta,\eta}\left(\left|\hat{\theta}_S - \theta\right| > \epsilon\right) = 0.$$

The condition in Equation 2.21 is true if $\lim_{n\to\infty} P(|m(u_k|Y_n, R_n) - m(u_k|Y_n)| < \epsilon^*|Y_n) = 0$ for every Y_n, for every η and for every u_k,$k = 0, 1, \ldots, K^*$. The condition in Equation 2.22 is true if $\lim_{n\to\infty} E_{R|Y}(m(u_k|Y_n, R_n)) = m(u|Y_n)$ and $\lim_{n\to\infty} var_{R|Y}(m(u_k|Y_n, R_n)) = 0$ for every Y_n, for every η and for every u_k,$k = 0, 1, \ldots, K^*$.

2.4.2 When $T(Y_n) \to \theta$ almost surely for every $\theta \in \Theta$

Assume that $T(Y_n) \to \theta$ almost surely for every $\theta \in \Theta$. For every $\epsilon > 0$ and for every $\theta \in \Theta$,

$$P_\theta\left(\lim_{n\to\infty} |T(Y_n) - \theta| < \epsilon\right) = 1.$$

Therefore, for every $\epsilon > 0$ and for every $\theta \in \Theta$,

$$P_\theta\left(\lim_{n\to\infty} \left|E_{R|Y_n}\left(\hat{\theta}_S\right) - \theta\right| < \epsilon\right) \;=\; 1. \tag{2.23}$$

The expectation $E_{R|Y}(\hat{\theta}_S)$ converges to θ almost surely for every $\theta \in \Theta$.

CHAPTER 3. SIMULATION

Here we use simulation to demonstrate the use of the SIMEX estimator. The non-parametric McNemar's test statistic tests marginal homogeneity on two correlated binary outcomes. The structure of the test statistic is simple and the closed forms of marginal and conditional means exist when data are full or partially observed. We compare the approximated extrapolation function and the marginal mean, and to illustrate other aspects of the SIMEX method under the MAR assumption. We also show that when we have partial knowledge about the distribution of the missing values, and partial knowledge about the missing-data mechanism, we can make use of all the information without building a new massive model by generating data from MNAR and use auxiliary variables to recover part of the bias.

In section 3.1, we describe the model for simulations. In section 3.2 and 3.3, we show that the marginal and conditional means of the naïve estimator are smooth functions of u for $u \geq 0$. These mean functions can be approximated well by some polynomials on $1 \leq u \leq u_{K^*}$ for some $u_{K^*} > 1$. These polynomials are expectations of extrapolation functions used in the SIMEX method and the expectation of extrapolated values at $u = 0$, which is the expectation of SIMEX estimators, can be very close to the true marginal expectation when we carefully choose a suitable polynomial. In section 3.4, we demonstrate the simulation and extrapolation steps of the SIMEX method by given one randomly selected dataset $(\tilde{y}^{(o,1)}, \tilde{r}^{(o)})$. The extrapolation function is estimated for $1 \leq u \leq u_{K^*}$ by simulation and extrapolated to both $0 \leq u < 1$ and $u_{K^*} < u \leq u_{K^*} + 1$. In section 3.5, we simulate the distribution of the SIMEX estimators and compare it

with the distribution of the consistent estimator from fully observed data in section 3.2 at $u = 0$. In section 3.6, the effects of misspecifying the model for the missing data mechanism are examined under two simulation model.

3.1 McNemar's Chi-squared Test for Paired Binary Data

Here we use an example of McNemar's test (Agresti, 1990) to demonstrate the use of the SIMEX method. McNemar's test is used for testing marginal homogeneity of two correlated binary outcomes. One application is testing whether a binary measurement taken before and after a treatment have the same marginal distribution. Another application is to test if a genetic marker and a quantitative trait locus tend to be inherited together. When measurements of some subjects in the designed sampling frame are absent, the naïve estimator which uses only complete pairs of measurements is appropriate when the missing is completely at random. Here we assume that there is evidence that missingness could be described by one of the following three situations: depends on an auxiliary variable, depends on the observed response variable (which assumes that only one of each pair could be missing), or depends on the unobserved response variable. We compare the mean functions by simulation and discuss some details of the simulation and extrapolation procedures.

3.1.1 Models

Consider two correlated binary random variables (Y_{1i}, Y_{2i}) with $p_1 = P(Y_{1i} = 1) = (1+\exp(1))^{-1} = 0.2689$, $p_2 = P(Y_{2i} = 1) = (1+\exp(0.5))^{-1} = 0.3775$ and $cor(Y_{1i}, Y_{2i}) = 0.4$ for $i = 1, \ldots, n = 200$. Since $p_{22} - p_1 p_2 = 0.4\sqrt{p_1(1-p_1)p_2(1-p_2)}$, the probabilities of discordance are

$$p_{21} = P(Y_1 = 1, Y_2 = 0) = p_1 - p_1 p_2 - 0.4\sqrt{p_1(1-p_1)p_2(1-p_2)} = 0.0814$$

	Second visit (Y_2)		
First visit (Y_1)	0	1	Total
0	n_{11}	n_{12}	n_{1+}
1	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Table 3.1 Summary table of a binary reponse variable measured before (Y_1) and after (Y_2) a treatment.

and

$$p_{12} = P(Y_1 = 0, Y_2 = 1) = p_2 - p_1 p_2 - 0.4\sqrt{p_1(1-p_1)p_2(1-p_2)} = 0.1900.$$

The probabilities of concordance are

$$p_{22} = P(Y_1 = 1, Y_2 = 1) = p_1 p_2 + 0.4\sqrt{p_1(1-p_1)p_2(1-p_2)} = 0.1875$$

and

$$p_{11} = P(Y_1 = 0, Y_2 = 0) = 1 - 0.0387 - 0.2338 - 0.1438 = 0.5410.$$

The data are summarized in Table 3.1.1. The data have marginal homogeneity if $E(n_{1+} - n_{+1}) = E(n_{2+} - n_{+2}) = 0$ or equivalently $E(n_{12}) = E(n_{21})$. The McNemar's statistic is

$$T(Y_1, Y_2) = \begin{cases} \frac{(n_{12}-n_{21})^2}{n_{12}+n_{21}}, & \text{if } n_{12} + n_{21} > 0 \\ 0, & \text{if } n_{12} + n_{21} = 0 \end{cases}.$$

The McNemar's statistic has asymptotic distribution $\chi^2(1)$ under the null hypothesis of marginal homogeneity. Let $n_d = n_{12} + n_{21}$ be the total number of discordant pairs. The count $n_{12}|(n_d, n_d > 0) \sim Bin(n_d, p_{12|d})$ where $p_{12|d} = \frac{p_{12}}{p_{12}+p_{21}} = 0.3$. Given $n = 200$,

$E(n_d) = P(n_d > 0)E(n_d|n_d > 0) = 54.2895$ and the statistic T has expectation

$$\begin{aligned}
E_{Y_1,Y_2}(T) &= E_{n_{12},n_{21}}(T) \\
&= P(n_d > 0)E_{n_{12},n_{21}|n_d>0}(T) + 0 \\
&= P(n_d > 0)E_{n_d|n_d>0}\left(E_{n_{12}|n_d,n_d>0}(T)\right) \\
&= P(n_d > 0)E_{n_d|n_d>0}\left(\frac{4}{n_d}E_{n_{12}|n_d,n_d>0}(n_{12}^2) - 4E_{n_{12}|n_d,n_d>0}(n_{12}) + n_d\right) \\
&= P(n_d > 0)E_{n_d|n_d>0}\left(\frac{4}{n_d}n_d p_{12|d}(1-p_{12|d}) + n_d^2 p_{12|d}^2 - 4n_d p_{12|d} + n_d\right) \\
&= 4p_{12|d} - 4p_{12|d}^2 + 4p_{12|d}^2 E(n_d) - 4p_{12|d}E(n_d) + E(n_d) \qquad (3.1) \\
&= 9.5295.
\end{aligned}$$

Under the null hypothesis $H_0 : n_{1+} = n_{+1}$, the probability that a chi-square random variable with one degree of freedom exceeds 9.5295 is 0.002.

Assume now that some subjects do not come for the second visit. Therefore, some Y_{2i} are unobserved. Since the McNemar statistic counts subjects with both Y_1 and Y_2 observed, the incomplete pairs are removed for calculation. Li et al. (2002) suggest using a conditional logistic regression method to estimate and test. Here, we focus on the McNemar statistic and adjust its bias by the SIMEX method.

We generate missing data by first selecting a model for the missing indicator R_i for Y_{2i}:

$$\pi_i = P(R_i = 1|Y_{1i}) = (1 + \exp(-2Y_{1i}))^{-1}.$$

This generates about forty percent missing second visit observations. The estimator $\hat{\pi}_i$ of π_i is estimated by fitting a logistic model with correctly specified mean structure. The McNemar statistic uses only complete case in which both (Y_{1i}, Y_{2i}) are observed. When we use only the observed portion of data defined by R, the estimator $T\left(\{(Y_{1i}, Y_{2i})\}_{i\in\{i;R_i=1\}}\right)$ is called the naïve estimator. Figure 3.1.1 shows the marginal mean and 0.05 and 0.95 quantiles of the naïve estimator. The test statistic T has expectation 9.52, standard deviation 5.52 and p-value 0.002 when we use the full dataset. The

test statistic T has expectation 1.63, standard deviation 2.06 and p-value 0.2 when only pairs with $R_i^{(1)} = 1$ are observed. The inference made from only the observed portion of dataset without any adjustment is subject to a large bias.

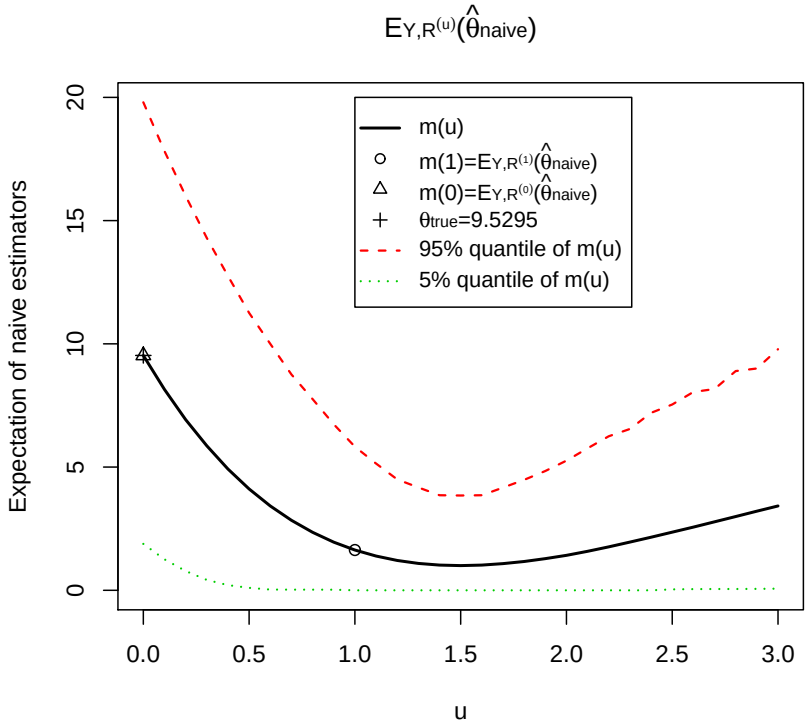

Figure 3.1 The expectation and standard deviations of naïve estimators estimated from 40,000 datasets.

The mean of the McNemar's test is proportion to sample size n. The function $m(u)$ is defined as the marginal expectation of naïve estimators. Here we take a look at the exact form of $m(u)$. The probability of missing indicators $R = 1$ is

$$\begin{aligned} \pi(Y_1) &= P(R = 1|Y_1) \\ &= \begin{cases} \frac{1}{1+\exp(-1)} = 0.7311, & \text{if } Y_1 = 0 \\ \frac{1}{1+\exp(-2)} = 0.8808, & \text{if } Y_1 = 1 \end{cases}. \end{aligned}$$

The probability of simulated missing indicators $R^{(u)} = 1$ is

$$\begin{aligned} \pi(Y_1)^{(u)} &= P(R^{(u)} = 1|Y_1) \\ &= \begin{cases} \left(\frac{1}{1+\exp(-1)}\right)^u, & \text{if } Y_1 = 0 \\ \left(\frac{1}{1+\exp(-2)}\right)^u, & \text{if } Y_1 = 1 \end{cases}. \end{aligned} \tag{3.2}$$

Given missing indicators $R^{(u)} = 1$, the distribution of $(Y_1, Y_2)|R^{(u)} = 1$ is

$$\begin{aligned} p_{11|R^{(u)}=1} &= P(Y_1 = 0, Y_2 = 0|R^{(u)} = 1) \\ &= \frac{P(Y_1 = 0, Y_2 = 0, R^{(u)} = 1)}{P(R^{(u)} = 1)} \\ &= \frac{P(R^{(u)} = 1|Y_1 = 0, Y_2 = 0)P(Y_1 = 0, Y_2 = 0)}{P(R^{(u)} = 1)} \\ &= \frac{P(R^{(u)} = 1|Y_1 = 0)p_{11}}{p_1 P(R^{(u)} = 1|Y_1 = 1) + (1 - p_1)P(R^{(u)} = 1|Y_1 = 0)} \\ &= \frac{\pi^{(u)}(Y_1 = 0)p_{11}}{p_1\pi^{(u)}(Y_1 = 1) + (1 - p_1)\pi^{(u)}(Y_1 = 0)}, \\ p_{12|R^{(u)}=1} &= P(Y_1 = 0, Y_2 = 1|R^{(u)} = 1) \\ &= \frac{\pi^{(u)}(Y_1 = 0)p_{12}}{p_1\pi^{(u)}(Y_1 = 1) + (1 - p_1)\pi^{(u)}(Y_1 = 0)}, \\ p_{21|R^{(u)}=1} &= P(Y_1 = 1, Y_2 = 0|R^{(u)} = 1) \\ &= \frac{\pi^{(u)}(Y_1 = 1)p_{21}}{p_1\pi^{(u)}(Y_1 = 1) + (1 - p_1)\pi^{(u)}(Y_1 = 0)}, \\ p_{22|R^{(u)}=1} &= P(Y_1 = 1, Y_2 = 1|R^{(u)} = 1) \\ &= \frac{\pi^{(u)}(Y_1 = 1)p_{22}}{p_1\pi^{(u)}(Y_1 = 1) + (1 - p_1)\pi^{(u)}(Y_1 = 0)}. \end{aligned}$$

The naïve estimator $T(Y^{(o,u)})$ has expectation

$$\begin{aligned}
m(u) &= E_{Y_1,Y_2|R^{(u)}=1}(T) \\
&= P\left(n_d > 0\right) E_{Y_1,Y_2|R^{(u)}=1,n_d>0}(T) + 0 \\
&= P\left(n_d > 0\right) E_{n_d|R^{(u)}=1,n_d>0}\left(E_{n_{12}|n_d,R^{(u)}=1,n_d>0}(T)\right) \\
&= P\left(n_d > 0\right) E_{n_d|R^{(u)}=1,n_d>0}\left(E_{n_{12}|n_d,R^{(u)}=1,n_d>0}\left(\frac{4}{n_d}n_{12}^2 - 4n_{12} + n_d\right)\right) \\
&= P\left(n_d > 0\right) E_{n_d|R^{(u)}=1,n_d>0}\Big(4p_{12|d,R^{(u)}=1}(1 - p_{12|d,R^{(u)}=1}) + n_d^2 p_{12|d,R^{(u)}=1}^2 \\
&\quad -4n_d p_{12|d,R^{(u)}=1} + n_d\Big) \\
&= 4p_{12|d,R^{(u)}=1} - 4p_{12|d,R^{(u)}=1}^2 \\
&\quad +P\left(n_d > 0\right) E(n_d|R^{(u)} = 1, n_d > 0)(4p_{12|d,R^{(u)}=1}^2 - 4p_{12|d,R^{(u)}=1} + 1) \\
&= 4p_{12|d,R^{(u)}=1} - 4p_{12|d,R^{(u)}=1}^2 + E(n_d|R^{(u)} = 1)(2p_{12|d,R^{(u)}=1} - 1)^2, \qquad (3.3)
\end{aligned}$$

where

$$\begin{aligned}
p_{12|d,R^{(u)}=1} &= \frac{p_{12|R^{(u)}=1}}{p_{12|R^{(u)}=1} + p_{21|R^{(u)}=1}} \\
&= \frac{\pi^{(u)}(Y_1 = 0)p_{12}}{\pi^{(u)}(Y_1 = 0)p_{12} + \pi^{(u)}(Y_1 = 1)p_{21}} \\
&= \frac{1}{1 + \frac{\pi^{(u)}(Y_1=1)p_{21}}{\pi^{(u)}(Y_1=0)p_{12}}} \\
&= \frac{1}{1 + \frac{p_{21}}{p_{12}}\left(\frac{\pi(Y_1=1)}{\pi(Y_1=0)}\right)^u} \\
&= \frac{1}{1 + \frac{p_{21}}{p_{12}}\left(\frac{1+\exp(-1)}{1+\exp(-2)}\right)^u},
\end{aligned}$$

and

$$\begin{aligned} E(n_d|R^{(u)}=1) &= P(n_d>0)E(n_d|R^{(u)}=1,n_d>0) \\ &= n(p_{21|R^{(u)}=1}+p_{21|R^{(u)}=1}) \\ &= n\frac{\pi^{(u)}(Y_1=0)p_{12}+\pi^{(u)}(Y_1=1)p_{21}}{p_1\pi^{(u)}(Y_1=1)+(1-p_1)\pi^{(u)}(Y_1=0)} \\ &= n\frac{p_{12}+p_{21}\left(\frac{1+\exp^{-1}}{1+\exp^{-2}}\right)^u}{p_1\left(\frac{1+\exp^{-1}}{1+\exp^{-2}}\right)^u+(1-p_1)} \\ &= n\frac{p_{12}+p_{21}\left(\frac{1+\exp^{-1}}{1+\exp^{-2}}\right)^u}{p_1\left(\frac{1+\exp^{-1}}{1+\exp^{-2}}\right)^u+(1-p_1)}. \end{aligned}$$

The closed form expression for $m(u)$ in Equation 3.3 is a smooth function of u but it would be hard to use this as an extrapolation function family. Instead, we approximate the mean function by fitting a curve using the least squares.

3.2 The Expectation of Naïve Estimators ($m(u)$) And The Expectation of Extrapolation Functions ($N^E(u;d) = E_{Y,R}(M(u;c))$)

We have shown that the marginal means of naïve estimators, $m(u)$, is a smooth function in Chapter 2. We find Kth order polynomial extrapolation functions, $M(u;c)$, which have marginal means $N^E(u;c) = E_{Y,R}(M(u;c))$, to approximate $m(u|y,r^{(1)})$ within $1 \le u \le u_{K^*}$ in the extrapolation step. In this section we take a look at $m(u)$, and discuss the closeness between $m(u)$ and the means of several extrapolation functions for $0 \le u \le 4$, $K = 1,2,\ldots,5$ and $u_{K^*} = 2$ or 3.

We generate $\{(Y_{i1}^{(j)}, Y_{i2}^{(j)}); i = 1,\ldots,n, j = 1,\ldots,J\}$ from the joint distribution of correlated binary random variables described in Section 3.1.1. For each $i = 1,\ldots,n$, $j = 1,\ldots,J$ and $k = 1,\ldots,K^*$, we generate missing indicators $R_i^{(u_k,j)}$ for Y_{i2} from $P(R_i^{(u_k)} = 1|Y_{1i}) = \pi_i^{u_k}$ which yields higher probability of missing when u_k increases.

The function of $m(u)$ is then estimated by

$$\hat{m}(u) = \frac{1}{J}\sum_{j=1}^{J} T\left(\{(Y_{1i}^{(j)}, Y_{2i}^{(j)})\}_{i\in\{i;R_i^{(u,j)}=1\}}\right)$$

for $u \in \{u_k; k = 0, 1, \ldots, K^*\}$. Figure 3.2 shows the marginal expectation of naïve estimators, $m(u)$, as a black solid line. As we indicated before, the probability $P(R^{(u)} = r)$ for each $r \in \{0,1\}^n$ is a smooth function of u. The function $m(u)$ is known to be a smooth continuous function but it may not be monotone and may not be unimodal. We can find that the bias of the naïve estimator is $m(1) - E(T) = 1.6554 - 9.5295 = -7.8741$.

For each $K = 1, \ldots, 6$, we use least squares estimation to find the closest Kth order polynomial, $N^E(u; c)$. Since $\hat{c}$ is a linear combination of $\{\hat{m}(u_k); k = 1, \ldots, K^*\}$ with coefficients that depend only on $\{u_k; k = 1, \ldots, K^*\}$, the polynomial $N^E(u; c)$ that approximates $m(u)$ is the expectation of the polynomial $M(u; c)$ that approximates $m(u|y, r^{(1)})$. Figure 3.2 shows polynomials with order from one to six. The function $m(u)$ is approximated well by polynomials with order greater than one on $1 \leq u \leq 2$. For a fixed K^*, the maximum polynomial order is $K = K^*$, which yields a polynomial passing through all $(u_k, m(u_k|\tilde{y}, \tilde{r}))$ for $k = 0, 1, \ldots, K^*$. For a fixed u_{K^*}, the polynomial order K can be any integer between one and K^*. The first order polynomial in Figure 3.2 yields the largest bias amount all six polynomials. The 6th order polynomial in Figure 3.2 is far away from $m(u)$ for $u > 2.5$ and also away from $m(u)$ for $u < 0.5$ and yields the second largest bias.

Figure 3.3 shows residuals computed as the difference between the expectation of extrapolation functions, $N^E(u; c)$, and the expectation of the mean function, $m(u)$. The flat residuals for $1 \leq u \leq 2$ indicate good approximation of $m(u)$.

We can improve the fitting of higher order polynomials by increasing the value of u_{K^*} for $K = 5, 6$. or decrease the value of u_{K^*} for $K = 1, 2$. Figure 3.4 shows polynomials with order from one to six with $u_{K^*} = 3$. The first and second order polynomials are significantly worse than the first and second order polynomials in Figure 3.2.

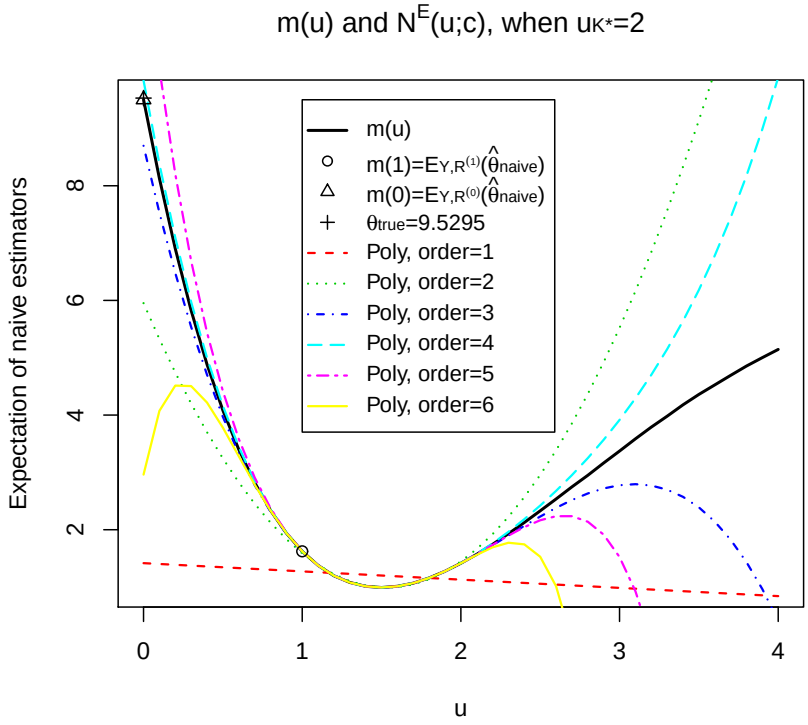

Figure 3.2 The marginal expectations of naïve estimators, $m(u)$, and expectations of extrapolation functions, $N^E(u;d)$, estimated from 40,000 datasets with $u_{K^*} = 2$ and $K = 1, \ldots, 6$.

Choosing the values of K and u_{K^*} is important for better bias reduction. Table 3.2 shows percentage of biases for different K, K^* and u_{K^*}. The percentage bias of the naïve estimator $(100 \times (m(1) - 9.5295)/9.5295)$ is -82.97. For each row in Table 3.2, there is at least one u_{K^*} such that the absolute percentage of bias is smaller than 82.97. For each of the last four rows in Table 3.2, there is at least one u_{K^*} such that the absolute percentage of bias is reduced from 82.97 to less than 6.

The relatively larger values in the first column of Table 3.2 suggest that the range $1 \leq u \leq 1.5$ is too long for polynomials of order one and two and too short for polynomials of order five and six. For smaller order polynomials, u_{K^*} should be close to one such that the polynomial can approximate the first or second order Taylor polynomial of $m(u)$ expanded at $u = 1$. For higher order polynomials, u_{K^*} should be close to three or even

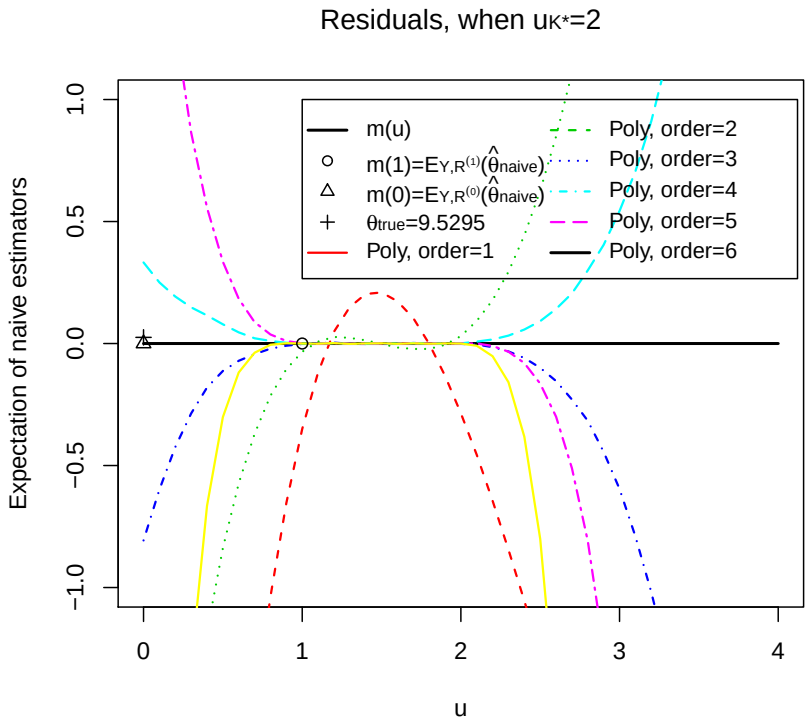

Figure 3.3 The residual of extrapolation functions, $N^E(u;\hat{c})) - \hat{m}(u)$ from 40,000 datasets.

larger when possible.

There are disadvantages for choosing u_{K^*} that is too large. The value of u_{K^*} is limited by sample size. When u_{K^*} gets very large, most data are missing. We will need extreme large simulation iterations to get one dataset with enough data left for estimation. For fixed sample size n, we can only estimated $m(u)$ in a limited range of u.

We also tried to increase the value of K^*. Comparing the third and fourth columns, when we increase K^* from 10 to 20, the bias is not further reduced.

Since we can only estimate $m(u|y, r^{(1)})$ for $u \geq 1$ in real data, the closeness between the extrapolation functions and $m(u)$ for $u_{K^*} \leq u \leq u_{K^*} + 1$ is a sign of a better approximation for $0 \leq u \leq 1$ although it is not a guarantee. Table 3.3 and Table 3.4 show the difference between $N^E(u;c)$ and $m(u)$ at $u = 0$ and $u_{K^*}+1$, which have distance one from both end of $1 \leq u \leq u_{K^*}$. The order within each column in Table 3.4 does not

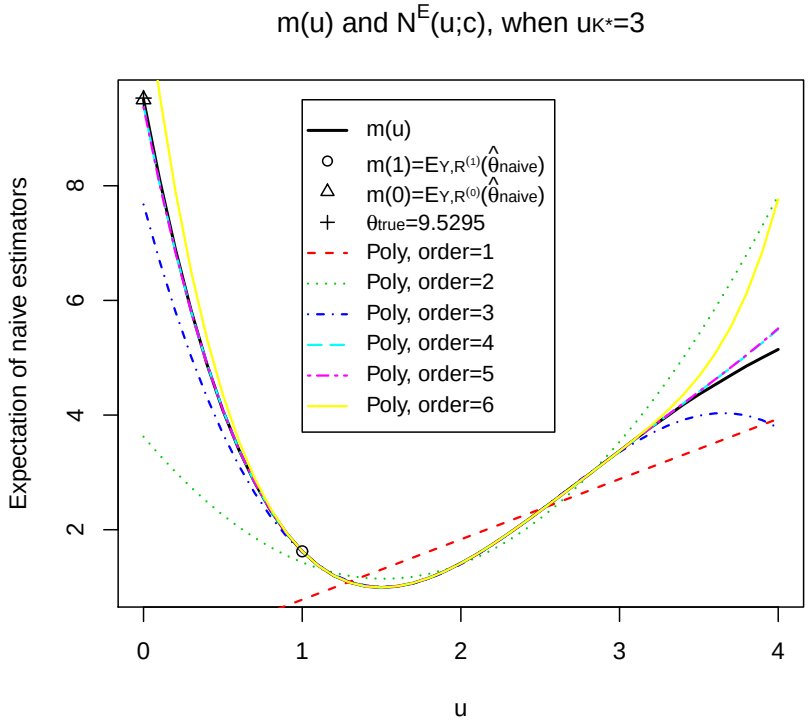

Figure 3.4 The marginal expectations of naïve estimators, $m(u)$, and expectations of extrapolation functions, $N^E(u;d)$, estimated from 40,000 datasets with $u_{K^*} = 3$ and $K = 1, \ldots, 6$.

exactly match the order in Table 3.3 but it does suggest which are the better and worse choices within each column.

To find a set of (K, K^*, u_{K^*}) for the simulation step, we can predetermine the desired order K then choose the value of u_{K^*} and K^* from the corresponding row in Table 3.4. As shown in Table 3.2, it is important to select a value of u_{K^*} according to the order K. When $K^* > 3K$, the increase of K^* has less impact on the expectation and standard deviation of the SIMEX estimator.

We can also predetermine the desired value of u_{K^*} then decide the order K by comparing values within the corresponding column in Table 3.4 and inspect residual plots like that in Figure 3.3. When we increase the order K, the standard deviation of the estimate $N^E(0;c)$ increases. Table 3.5 shows the standard deviation of the estimator of

Table 3.2 Table of bias percentages $(100 \times (N^E(0;\hat{c}) - 9.5295)/9.5295)$ of Kth order polynomial made from points $\{(u_k, m(u_k)); k = 1, \ldots, K^*\}$ by least squares method with orders $K = 1, \ldots, 6$, $u_{K^*} = 1.5, 2$ or 3 and $K^* = 10$ or 20.

	(K^*, u_{K^*})			
Order	$(10, 1.5)$	$(10, 2)$	$(10, 3)$	$(20, 3)$
1	-71.09	-85.02	-103.05	-104.14
2	-26.37	-37.58	-61.53	-62.53
3	0.40	-9.38	-19.21	-19.49
4	-6.05	-6.98	-3.88	-3.93
5	31.44	47.02	-5.41	-7.82
6	3339.86	159.73	3.08	3.18

Table 3.3 Table of $N^E(u;c) - m(u)$ at $u = 0$, which are expected biases of SIMEX estimators.

	(K^*, u_{K^*})			
Order	$(10, 1.5)$	$(10, 2)$	$(10, 3)$	$(20, 3)$
1	-6.77	-8.09	-9.81	-9.92
2	-2.51	-3.57	-5.86	-5.95
3	$0.05^{(1)}$	$-0.89^{(2)}$	-1.82	-1.85
4	$-0.57^{(2)}$	$-0.66^{(1)}$	$-0.36^{(2)}$	$-0.37^{(2)}$
5	3.00	4.49	-0.51	-0.74
6	318.28	15.23	$0.30^{(1)}$	$0.31^{(1)}$

(1) The lowest bias of each column.
(2) The second lowest bias of each column.

$N^E(0;c)$ when we estimate $N^E(0;c)$ by one set of $(Y, R^{(1)}, \ldots, R^{(u_{K^*})})$. The standard deviation of estimator of $N^E(0;c)$ increases dramatically as the order K increases considering that the standard deviation of the test statistic $\hat{\theta}$ is only 5.52 when data are fully observed.

In a real situation, the SIMEX estimator finds the intercept of the polynomial based on one set of $(Y, R^{(1)})$ and estimates $m(u_k|Y, R^{(1)})$, $k = 1, \ldots, K^*$ by simulation. That is different from the randomness displayed in Table 3.5. The standard deviation of the SIMEX estimator is difficult to derive analytically. We later propose a bootstrap method to compute the standard deviation the SIMEX estimator.

Table 3.4 Table of $N^E(u;c) - m(u)$ at $u = u_{K^*} + 1$.

	(K^*, u_{K^*})			
Order	$(10, 1.5)$	$(10, 2)$	$(10, 3)$	$(20, 3)$
1	-2.69	-2.43	-1.19	-1.16
2	1.57	2.10	2.77	2.80
3	$-0.98^{(1)}$	-0.59	-1.27	-1.30
4	-1.59	$-0.36^{(1)}$	$0.19^{(1)}$	$0.18^{(1)}$
5	$-5.16^{(2)}$	$-5.51^{(2)}$	0.34	0.56
6	$310.11^{(2)}$	$5.23^{(2)}$	1.15	1.60

(1) The lowest difference of each column. These match cells either (1) or (2) in Table 3.3.
(2) The higher values of differences at $u = u_{K^*} + 1$ suggest worse approximation in Table 3.3.

Table 3.5 Table of standard deviation of estimator of $N^E(0;c)$ (w.r.t distribution of $(Y, R^{(1)}, R^{(u_k|u_{k-1})}; k = 1, \ldots, u_{K^*}, u_0 = 1)$) from single set of $(Y, R^{(1)}, \ldots, R^{(u_{K^*})})$.

	(K^*, u_{K^*})			
Order	$(10, 1.5)$	$(10, 2)$	$(10, 3)$	$(20, 3)$
1	4.54	4.57	4.55	4.63
2	9.25	9.30	9.29	9.47
3	38.35	38.80	38.67	39.34
4	279.41	269.57	271.97	278.54
5	2278.79	2268.47	2287.78	2305.27
6	21191.55	21217.34	20821.00	21588.28

3.3 Estimate $m(u|\tilde{y})$ When One Full Dataset $\tilde{y}$ Is Observed

Assume that one particular realization of Y, $\tilde{y}$, is observed. We find mean functions $m(u|\tilde{y})$ and show that the function is smooth and continuous everywhere for $u \geq 0$. The function $m(u|\tilde{y})$ itself is not estimable when data are not fully observed. The purpose of exploring $m(u|\tilde{y})$ is that it is the conditional expectation of $m(u|Y, R^{(1)})$ for $0 \leq u \leq 1$. Although $m(u|Y, R^{(1)})$ itself is not a smooth function at $u = 1$, $m(u|\tilde{y})$ is a smooth function at $u \geq 0$. The smooth extrapolation function, $M(u;c)$, has expectations that is smooth and approximates $m(u|\tilde{y})$ at $0 \leq u \leq 1$. The flowchart in Figure 2.2 shows their relationships.

Let $\tilde{y}$ be a particular realization of $Y \sim f_Y(y|\theta)$. The statistic $T(\tilde{y}) = \frac{(n_{12}-n_{21})^2}{n_{12}+n_{21}}$ is a

consistent estimator of the expectation of the asymptotic distribution of T,

$$\frac{n(p_{1+}-p_{+1})^2}{p_{1+}(1-p_{1+})+p_{+1}(1-p_{+1})-2(p_{11}p_{22}-p_{12}p_{22})}+1$$

where $p_{1+}=p_{11}+p_{12}$ and $p_{+1}=p_{11}+p_{21}$. We generate missing indicators $R^{(u)}$ from Equation 3.2. We find the function $m(u|\tilde{y})$ to describe the expectation of $T(\tilde{y}|R^{(u)})$.

Table 3.6 shows the mean and standard deviation of the estimated $N(0;c)$. The standard deviation is for single simulated $(R^{(1)},\ldots,R^{(u_{K^*})})$ given $\tilde{y}$. Therefore, we can reduce the standard deviation by increasing the number of datasets in the simulation when possible. Reducing the percentage of bias is more difficult since that cannot be achieved by simply increasing the simulation size. Table 3.6 shows that the bias percentages can be reduced by choosing smaller u_{K^*} for $K=1,2,3$ and larger value of u_{K^*} for $K=5,6$ but we pay the price of a higher standard deviation.

Figure 3.3 shows the polynomials with u_{K^*} chosen from Table 3.6. When we reduced u_{K^*} to 1.1, the first order polynomial adjusts the bias in the correct direction but still is conservative. The higher order polynomials approximate $m(u|\tilde{y})$ better, but we need more iterations in the simulation to reduce the standard deviation of the SIMEX estimator.

3.4 Estimate $m(u|\tilde{y},\tilde{r}^{(1)})$ when only the incomplete dataset $(\tilde{y}^{(o,1)},\tilde{r}^{(1)})$ is observed

In this section, we assume only $\tilde{y}^{(o,1)}$ and the missing indicators $\tilde{r}$ are observed. We first estimate the parameters of the missing-data mechanism. Then, We find the mean function, $m(u|\tilde{y},\tilde{r}^{(1)})$ and a approximation function, $M(u;c)$. Finally, we compute the SIMEX estimator, $M(0;c)$ by extrapolation. The relationships between $m(u), m(u|\tilde{y})$ and $m(u|\tilde{y}^{(o)},\tilde{r})$ are explained in Figure 2.2.

Figure 3.4 shows traces of SIMEX estimators against simulation size B. For polynomials with order one to four, 10,000 iterations is enough to achieve stable estimators, but

Table 3.6 Table of mean and standard deviation of estimators of $N(0;c)$ from one set of Y. The last column is percentage of biases conditional on $\tilde{y}$.

K	K^*	u_{K^*}	**mean**	**sd**	$100\frac{N(0;c)-T(\tilde{y})}{T(\tilde{y})}$
1	6	1.3	2.70	3.30	-69.84
	4	$1.2^{\downarrow}$	3.05	$3.65^{\uparrow}$	$-65.98^{\downarrow}$
	2	$1.1^{\downarrow}$	3.41	$4.51^{\uparrow}$	$-61.92^{\downarrow}$
2	6	1.3	6.93	15.86	-22.66
	4	$1.2^{\downarrow}$	7.15	$27.02^{\uparrow}$	$-20.25^{\downarrow}$
	2	1.1	6.56	81.02	-26.82
3	10	2.0	8.20	16.83	-8.46
	10	$1.5^{\downarrow}$	8.54	$52.73^{\uparrow}$	$-4.74^{\downarrow}$
4	10	3.0	8.88	22.54	-0.93
	13	3.6	8.56	16.48	-4.51
5	10	3.0	8.54	68.57	-4.77
	13	$3.6^{\uparrow}$	9.31	$41.41^{\downarrow}$	$3.91^{\downarrow}$
6	10	3.0	10.00	239.68	11.56
	13	$3.6^{\uparrow}$	8.13	$115.81^{\downarrow}$	$-9.25^{\downarrow}$

we need more than 40,000 iterations to have stable estimators for polynomial with order five or six. As discussed previously, we need higher u_{K^*} and higher iteration numbers for higher order polynomials. Higher order polynomials require much more calculation time. Limited by calculation precision, higher order polynomials may not yield a better approximation.

Given $\tilde{y}$ and $\tilde{r}^{(1)}$, the naïve estimator is $T(\tilde{y}^{(o,1)}) = 0.71$. Starting from 0.71, we estimate $m(u_k|\tilde{y},\tilde{r})$ for $1,\ldots,K^{(max)}$, then try several values of (K,u_{K^*}), where $K^* \leq K^{(max)}$, to find a better approximation function $M(u;c)$. Table 3.7 shows the SIMEX estimators. The standard deviations in parentheses show the variation of the SIMEX estimators from the simulation step, and are inversely proportional to the square root of simulation number B.

We suggest choosing (K,u_{K^*}) such that the polynomial $M(u;c)$ approximates the estimated $m(u|\tilde{y},\tilde{r})$ well within $1 \leq u \leq u_{K^*}+1$. A simple but not always best criterion is the differences of $m(u|\tilde{y},\tilde{r})$ and $M(u;c)$ at $u = u_{K^*}+1$, although the order of differences at $u = u_{K^*}+1$ cannot predict the order of differences at $u = 0$. The differences shown

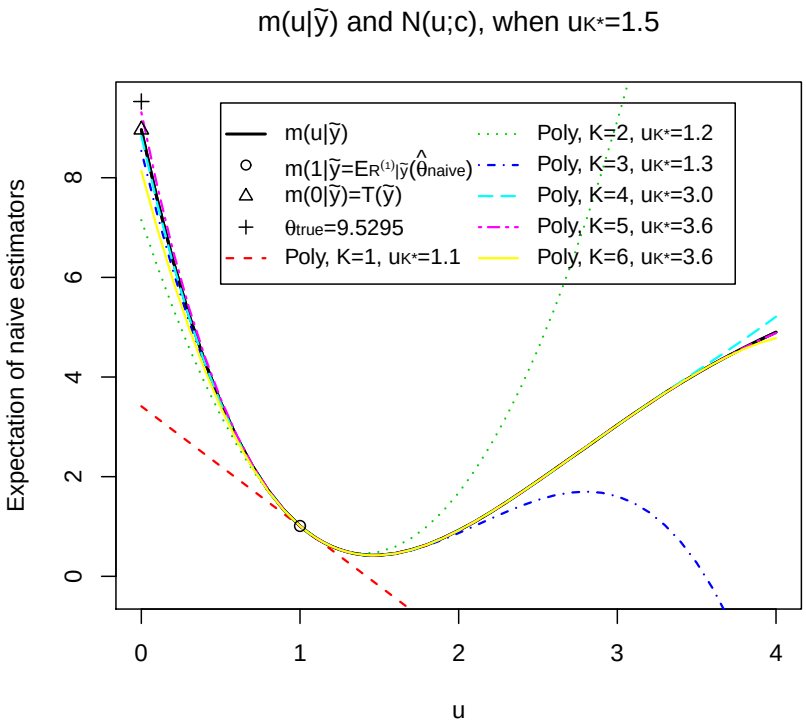

Figure 3.5 The $\hat{m}(u|\tilde{y})$, estimated from 40,000 datasets, and several approximation functions.

in Table 3.8 suggests that we choose $u_{K^*} = 1.5$ for $K = 2, 3$, $u_{K^*} = 2$ for $K = 4$ and $u_{K^*} = 3$ for $K = 1, 5, 6$. Another clue can be obtained from the plot of $m(u|\tilde{y}, \tilde{r})$ and $M(u; c)$ against u, which shows more details about the approximation. Figure 3.4 shows the first order polynomial with $u_{K^*} = 1.5, 2$, or 3. Given $K = 1$, the polynomial with $u_{K^*} = 1.5$ yields better approximation comparing to $u_{K^*} = 2, 3$. We can further increase the value of u_{K^*} for $K = 5, 6$ or decrease the value of u_{K^*} for $K = 1, 2, 3$. But in either case, the variance of SIMEX estimators will increase, and extra simulation iterations are required to have reliable SIMEX estimators.

Figure 3.4 shows residuals of all six polynomials with u_{K^*} chosen from Table 3.8 and Figure 3.4. The first and second order polynomials still have non-constant residuals in $1 \leq u \leq 1.5$. The fifth and sixth order polynomials also have non-constant residual patterns and additionally, require relatively longer simulations as shown in Figure 3.4.

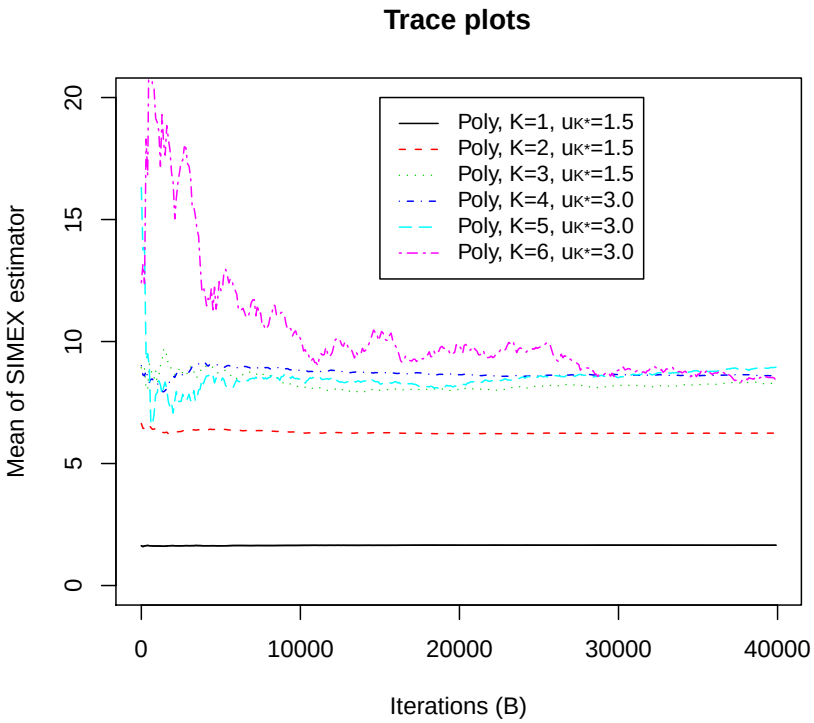

Figure 3.6 Trace of SIMEX estimators.

The third and fourth order polynomial have residuals that are approximately equal to 0 in $1 \leq u \leq 1.5$ or 2. The polynomials $M(u;c)$ with $K = 3$ or 4 seem to better approximate $m(u|\tilde{y},\tilde{r})$. They also have differences across columns in Table 3.8.

The SIMEX estimators in Table 3.7 range widely from -127 to 12. It is essential to find a good approximation for $m(u|\tilde{y},\tilde{r})$ for a good SIMEX estimator. A good approximating polynomial $M(u;c)$ should at least have a flat residual plot for $1 \leq u \leq u_{K^*}$ and a small difference between $M(u;c)$ and $m(u|\tilde{y},\tilde{r})$ at $u = u_{K^*}$.

We find the variance of the SIMEX estimator using a bootstrap procedure. To do so, draw 10,000 samples from $\{(\tilde{y}_{1i}, \tilde{y}_{2i}, \tilde{r}_i^{(1)}); i = 1, \ldots, 200\}$ with replacement and repeat the process 1,000 times. The standard deviation of the SIMEX estimators from the 1,000 datasets with $B = 5,000$ or $10,000$ are shown in Table 3.9. The bootstrap estimates are larger than the estimated marginal standard deviations in Table 3.11. The standard

Table 3.7 Table of means and standard deviations (in parentheses; w.r.t the distribution of $\{R_b^{(u_k|1)}; k = 1, \ldots, K^*, b = 1, \ldots, B\}$) of $M(0; c)$ with several sets of (B, K, u_{K^*}) given $(\tilde{y}, \tilde{r}^{(1)})$.

		(K^*, u_{K^*})		
Order	B	$(10, 1.5)$	$(10, 2)$	$(10, 3)$
1	10,000	1.64(0.01)	0.20(0.01)	-1.62(0.02)
2	10,000	6.29(0.06)	5.08(0.03)	2.57(0.03)
3	10,000	8.17(0.38)	8.10(0.13)	6.92(0.07)
4	10,000	7.89(2.84)	8.46(0.55)	8.82(0.21)
5	40,000	5.16(23.50)	8.96(2.62)	8.95(0.63)
6	40,000	-127.74(216.90)	12.91(14.07)	8.40(2.18)

Table 3.8 Table of differences of $m(u|\tilde{y}, \tilde{r})$ and $M(u; c)$ at $u = u_{K^*} + 1$.

	(K^*, u_{K^*})		
Order	$(10, 1.5)$	$(10, 2)$	$(10, 3)$
1	-2.88	-2.58	-1.25$^{(1)}$
2	1.71$^{(1)}$	2.27	2.97
3	-0.31$^{(1)}$	-0.61	-1.31
4	1.54	0.22$^{(1)}$	0.39
5	6.50	0.06	0.04$^{(1)}$
6	-126.40	4.01	-0.51$^{(1)}$

(1) Lowest absolute differences of each row.

deviation will not serve as a criterion for choosing K or u_{K^*} but can provide clues of insufficient value of u_{K^*} or B.

3.5 Estimate Marginal Distribution of $\hat{\theta}_{SIMEX}$ by Simulation With True Parameters

In Section 3.4 we show the process of finding the SIMEX estimator from one single set of $(\tilde{y}, \tilde{r})$. In this section we discuss the marginal distribution of the SIMEX estimator.

We estimate the marginal mean and standard deviation of the SIMEX estimator by generating 5,000 sets of (Y, R). Then, we calculate the SIMEX estimators with $B = 5,000$ and $u_{K^*} = 1.5, 2, 3$ for each dataset. The marginal percentage of bias of SIMEX estimators are listed in Table 3.10. The iteration number $B = 5,000$ is large

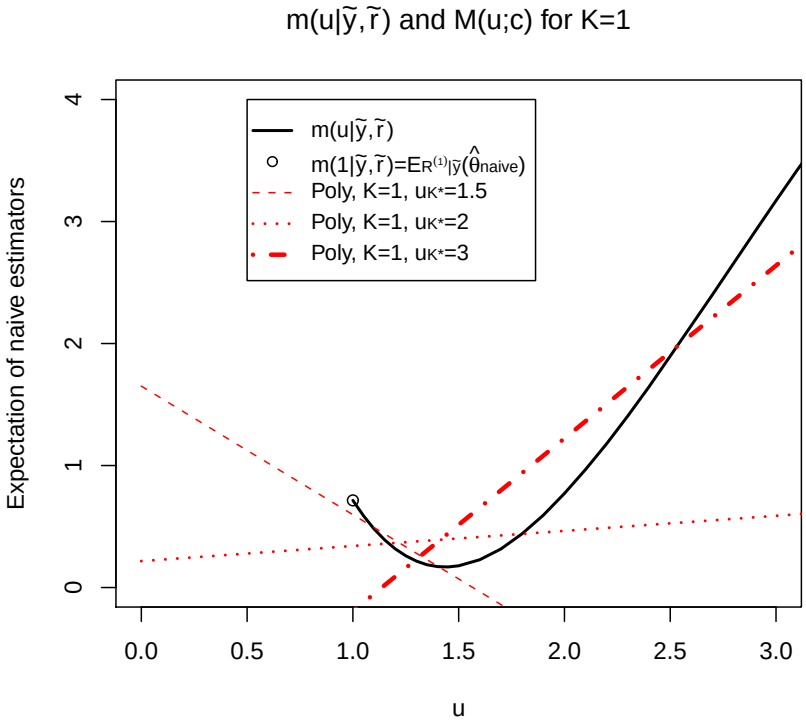

Figure 3.7 The first order polynomial, $M(u;c)$ and $m(u|\tilde{y}^{(o)},\tilde{r})$, estimated from 40,000 datasets.

enough for finding the SIMEX estimator. The bias of the SIMEX estimator is decided by the values of K, the order of the polynomial, and u_{K^*}. When order $K \geq 4$, we can find u_{K^*} such that the absolute bias smaller than 3 percent.

The marginal Monte Carlo standard deviations of SIMEX estimators are listed in Table 3.11. These marginal standard deviations are all smaller than the bootstrap estimates in Table 3.9. The bootstrap estimates tend to overestimate the standard deviations. The standard deviation of $T(Y)$ (w.r.t density of Y) is 5.52 when Y is fully observed. The standard deviation of $T(Y^{(o,1)})$ (w.r.t density of $(Y, R^{(1)})$) is 2.06 when only $Y^{(o,1)}$ is observed. The SIMEX estimator is a linear combination of $T(Y^{(o,1)})$ and estimated $m(u|Y, R^{(1)})$ which are nonlinear functions of $T^{(o,1)}$. The total variation of $\hat{\theta}_{SIMEX}$ is composed of the variation of $T(Y^{(o,1)})$ and the variation from simulation step shown in Table 3.7, but the actual functional relationship is usually unknown.

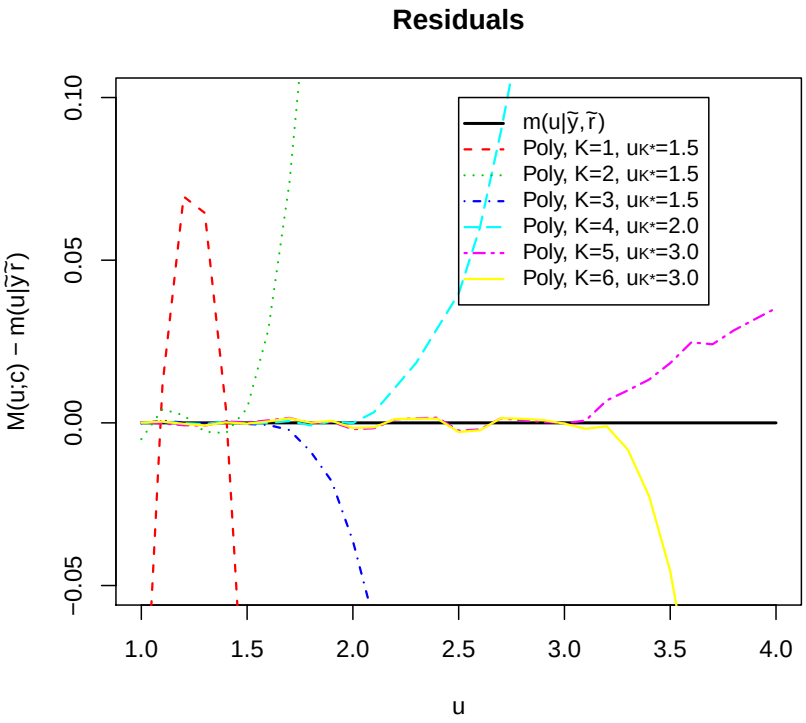

Figure 3.8 The residual $M(u;\hat{c}) - \hat{m}(u|\tilde{y}^{(o)},\tilde{r})$, estimated from 40,000 datasets.

Figure 3.5 shows the mean and the fifth and ninety fifth percentiles of $m(u|Y.R)$ and $M(u;c)$ with $K = 4$ and $u_{K^*} = 2$. If we choose u_{K^*} carefully, the mean of the extrapolation function (blue) built from $m(u^{(k)}|\tilde{y},\tilde{r})$ (black) is very close to the marginal mean $m(u)$ (yellow) for $0 \leq u \leq 2$. The extrapolation function may be fall outside of range of $T(Y^{(u)})$. The marginal probability of getting negative SIMEX estimator is about four percent. The probabilities of rejecting the null hypothesis of marginal homogeneity under an 0.05 significance level are listed in Table 3.12. The probabilities of rejecting the false hypothesis are all higher than 75 percent when $K > 2$ and with u_{K^*} choosen from Table 3.8

The first and second order polynomials are usually suggested for an extrapolation in general, since they are relatively conservative and stable. But the in the SIMEX procedure the special structure of $m(u|\tilde{y},\tilde{r})$, which is a sum of $\Pi_i T(z) p^{uz_i}(1-p^u)^{1-z_i}$,

Table 3.9 Table of bootstrap estimates of standard deviation of $M(0;c)$.

		(K^*, u_{K^*})		
Order	B	$(10, 1.5)$	$(10, 2)$	$(10, 3)$
1	5,000	6.95	6.36	5.23
	10,000	6.95	6.36	5.24
2	5,000	8.79	8.65	8.27
	10,000	8.78	8.65	8.27
3	5,000	9.24	9.08	9.12
	10,000	9.13	9.07	9.12
4	5,000	13.35	9.22	9.13
	10,000	11.80	9.09	9.12
5	5,000	86.62	12.69	9.12
	10,000	60.33	10.79	9.08
6	5,000	$764.22^{(1)}$	46.10	10.19
	10,000	$526.70^{(1)}$	34.01	11.34

(1) The standard deviation is significantly reduced by increased values of B but it is still too big to make reliable inferences.

can be approximated well by high order polynomials. Therefore, when the residual plots shows curvier results for $K = 1, 2$, utilizing higher order polynomials for extrapolation can reduce both the bias and standard deviation of the SIMEX estimator.

Figure 3.5 shows the mean and fifth and ninety fifth percentiles of $m(u|Y.R)$ and $M(u;c)$ with $K = 2$ and $u_{K^*} = 1.5$. The mean of the extrapolation function has a significant gap from $m(u)$ for $0 \leq u \leq 1$. The marginal probability of getting negative SIMEX estimator is about thirteen percent.

When $u = 10$, the probability of observing a subject with $p(R = 1) = 0.7$ is $p^{(u)} = 0.03$ and it is harder to have enough sample to calculate $T(Y^{(o,u|1)})$ when u gets larger. Although we define the domain of $m(u|\tilde{y}, \tilde{r})$ as $u \geq 0$, we can only estimate $m(u|\tilde{y}, \tilde{r})$ in a limited range of u given finite calculation time. The order K is limited, because the simulation number increases significantly when K or u_{K^*} increases.

Table 3.10 Table of marginal percentage of bias $100(M(0;c) - 9.5295)/9.5295$ (w.r.t density of (Y, R)) when $B = 10,000$.

		(K^*, u_{K^*})		
Order	B	$(10, 1.5)$	$(10, 2)$	$(10, 3)$
1	5,000	-74.14$^{(1)}$	-87.34	-103.10
	10,000	-74.70$^{(1)}$	-87.99	-103.87
2	5,000	-30.49$^{(1)}$	-42.82	-66.88
	10,000	-30.80$^{(1)}$	-43.14	-67.40
3	5,000	-8.02$^{(1)}$	-12.56	-25.27
	10,000	-8.37$^{(1)}$	-12.82	-25.42
4	5,000	-2.72$^{(1)}$	-3.43	-6.53
	10,000	-2.69$^{(1)}$	-3.62	-6.67
5	5,000	-9.55	-1.54$^{(1)}$	-2.20
	10,000	-3.51	-3.26	-2.75$^{(1)}$
6	5,000	-43.22	-2.93	-1.52$^{(1)}$
	10,000	-55.18	-1.12$^{(1)}$	-1.86

(1)The bias of SIMEX estimators largely depend on the values of K and u_{K^*}.

3.6 When The Missing Model Is Incorrectly Specified

The SIMEX method adjusts for bias by estimating the effect of the missing-data mechanism. We need to fit the missing model and estimate the probability of missing for each record as accurately as we can. We have already shown that the percentage of bias can be reduced to less than three percent, when we estimate the probability of observing from a correctly specified structure of a linear predictor in the generalized linear model for missing indicators R. In this section, we try several different linear predictors to examine the effects on SIMEX estimators.

We generate missing indicators from a MAR model, or one of these two MNAR models,

$$(G1 : MAR) \qquad \pi_i = P(R_i = 1|Y_{1i}) = (1 + \exp(-2Y_{1i}))^{-1},$$

$$(G2 : MNAR_1) \qquad \pi_i = P(R_i = 1|Y_{1i}) = (1 + \exp(-2Y_{2i}))^{-1},$$

$$(G3 : MNAR_2) \qquad \pi_i = P(R_i = 1|Y_{1i}) = (1 + \exp(2 - 2Y_{2i}))^{-1}.$$

Table 3.11 Table of marginal standard deviation of $M(0;c)$ (w.r.t density of $(Y, R^{(1)})$).

		(K^*, u_{K^*})		
Order	B	$(10, 1.5)$	$(10, 2)$	$(10, 3)$
1	5,000	4.92	4.49	3.60
	10,000	4.84	4.40	3.51
2	5,000	6.27	6.21	5.98
	10,000	6.19	6.12	5.88
3	5,000	6.49	6.46	6.55
	10,000	6.38	6.40	6.47
4	5,000	$10.17^{(1)}$	6.57	6.48
	10,000	$8.47^{(1)}$	6.39	6.42
5	5,000	$65.91^{(1)}$	9.32	6.50
	10,000	$46.97^{(1)}$	7.91	6.40
6	5,000	$622.94^{(1)}$	$37.51^{(1)}$	7.59
	10,000	$428.68^{(1)}$	$26.95^{(1)}$	6.90

(1) higher order polynomial with relatively small u_{K^*} yields very large standard deviations. It suggests that we should choose either increase the value of B or u_{K^*}.

Table 3.12 Table of probability of rejecting the null hypothesis of marginal homogeneity under 0.05 significance level.

		(K^*, u_{K^*})		
Order	B	$(10, 1.5)$	$(10, 2)$	$(10, 3)$
1	10,000	0.3230	0.2232	0.1150
2	10,000	0.6082	0.5328	0.3964
3	10,000	0.7498	0.7182	0.6368
4	10,000	0.7576	0.7838	0.7590
5	10,000	0.5410	0.7688	0.7814
6	10,000	0.5104	0.5914	0.7840

Additionally, we generate two auxiliary dichotomize variables Z_{1i} and Z_{2i} with probability

$$P(Z_i = 1|Y_{2i}) \quad = \quad \begin{cases} \frac{p_z - \rho\sqrt{p_2(1-p_2)p_z(1-p_z)} - p_2 p_z}{1-p_2} = 0.7311, \text{ if } Y_2 = 0 \\ \frac{\rho\sqrt{p_2(1-p_2)p_z(1-p_z)} + p_2 p_z}{p_2} = 0.8808, \text{ if } Y_2 = 1 \end{cases}$$

where $p_w = 0.z$ and $\rho = 0.6$ for Z_1 and $\rho = 0.9$ for Z_2.

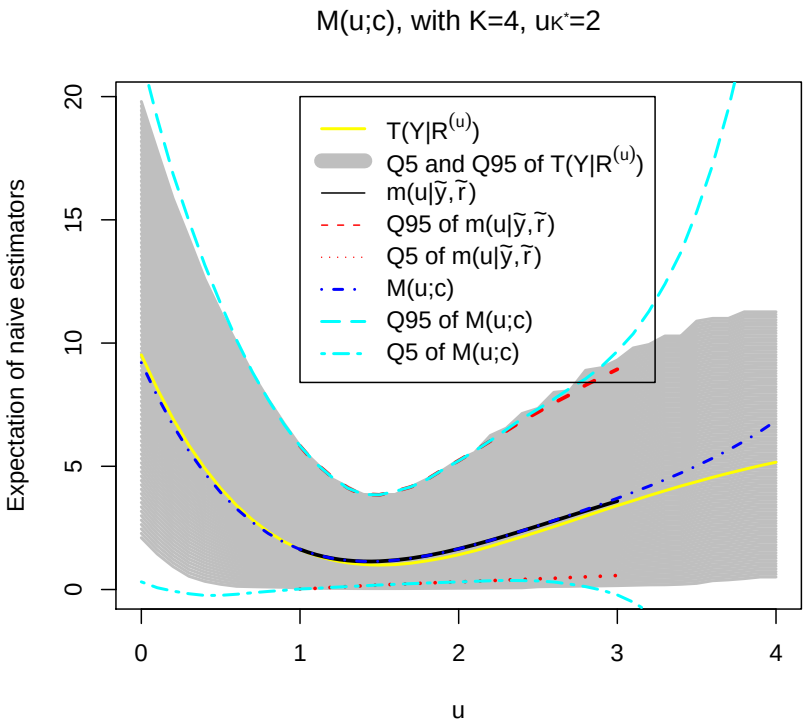

Figure 3.9 The mean and fifth and ninety-fifth percentiles (w.r.t density of (Y, R)) of $m(u|Y, R)$ with $B = 5,000$ and $M(u; \hat{c})$ with $K = 4$ and $u_{K^*} = 2$. These are estimated from 5,000 iterations. The gray area is between the 0.05 and 0.95 percentiles of $T(Y^{(o,u)})$ estimated from 40,000 simulation iterations, Note that the test statistic is nonnegative.

Then, we estimate $\pi_i = P(R_i = 1|Y)$ using four linear predictors:

$$\begin{aligned} (M1 : MCAR) \quad & \beta_0, \\ (M2 : MCAR - cov) \quad & \beta_0 + \beta_1 Z_1, \\ (M3 : MAR - cov) \quad & \beta_0 + \beta_1 Y_1 + \beta_2 Z_1, \\ (M4 : MCAR - cov) \quad & \beta_0 + \beta_1 Z_2. \end{aligned}$$

Table 3.13 shows the bias percentages using either the $(M1 : MCAR)$ or the $(M2 : MCAR - cov)$ working model when the true missing generating model is $(G1 : MAR)$. Both working models are wrong. In the first column, the SIMEX estimators have percentages of biases smaller than 82.79 even when the working model is just MCAR, That

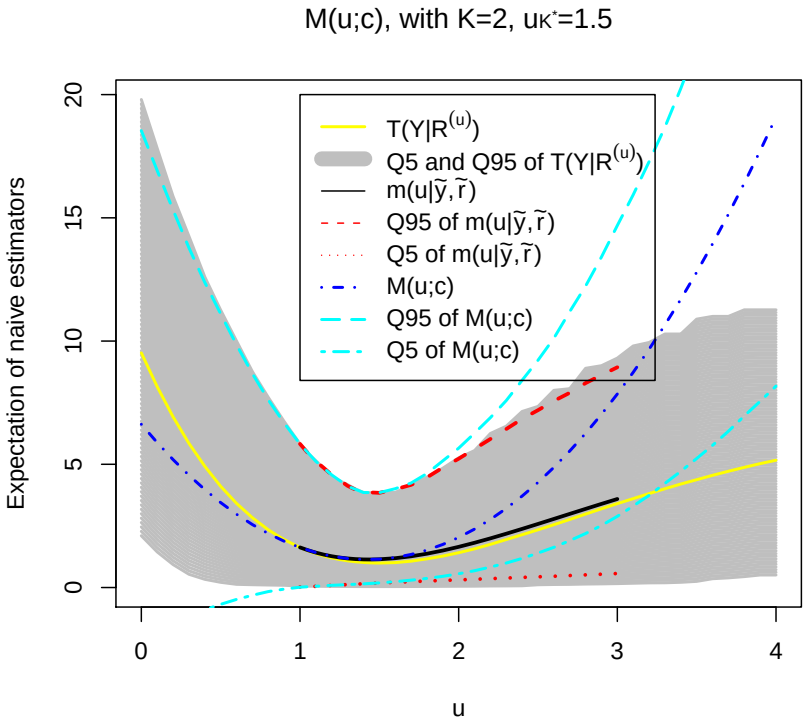

Figure 3.10 The mean and 5 and 95 percentiles (w.r.t density of (Y, R)) of $m(u|Y, R)$ with $B = 5,000$ and $M(u; \hat{c})$ with $K = 2$ and $u_{K^*} = 1.5$. These are estimated from 5,000 iterations. The gray area is between 0.05 and 0.95 percentiles of $T(Y^{(o,u)})$ estimated from 40,000 simulation iterations,

is because that the expectation of $T(Y)$ is a function of total size n as shown in Equation 3.1 and the trend of $m(u)$ adjust part of the bias from smaller sample size. However, in the second column, the probability of missing is assumed to depend on an auxiliary variable Z_1 when the true probability depends on Y_1. The SIMEX estimators have larger bias because the missing model contains an unnecessary explanatory variable Z_1. The process of fitting the missing model shows that $(M2)$ is a poor linear predictor. The averaged pvalue of testing if the coefficient of Y_1 is zero in model $(M2)$ is 0.3298. When the missing data mechanism can not be appropriately estimated, the SIMEX estimator can cause biases that are higher then bias of the naïve estimator from observed data set. Figure 3.11 shows $m(u)$ estimated under $(M1)$ or $(M2)$ assumption and the expectation

of SIMEX estimator $N^E(0;c)$ with $K = 4$ and $u_{K^*} = 2$.

Table 3.13 Table of bias percentages of the SIMEX estimator $(100(E_{Y,R,\{R_b^{(u|1)}\}}(\hat{\theta}_{SIMEX}) - 9.5295)/9.5295)$ when missing is MAR(G1). The simulation number is 1,000. We fix $K = 10$ and $B = 10,000$. The percentage of bias of naïve estimator $T(Y^{(o,1)})$ is -82.89.

		(K^*, u_{K^*})	
Order	u_{K^*}	**M1**	**M2**
1	1.5	-79.95	-83.91
2	1.5	-78.89	-84.75
3	1.5	-78.14	-85.11
4	2.0	-79.15	-85.40
5	3.0	-78.80	-85.22
6	3.0	-80.30	-86.03

Table 3.14 shows the percentage of biases using $(M1)$, $(M2)$ or $(M3)$ working model when the true missing generating model is $(G2 : MNAR_1)$. Figure 3.12 shows $m(u)$ estimated under $(M1)$, $(M2)$ or $(M3)$ assumption and the expectation of SIMEX estimator $N^E(0;c)$ with $K = 4$ and $u_{K^*} = 2$. In the first column, $N^E(0;c)$ with $K = 1,\ldots,6$ estimated under $(M1 : MCAR)$ assumption. The SIMEX procedure adjusts only the effect of smaller sample size. The second column shows expectations of SIMEX estimator under $(M2 : MCAR - cov)$ assumption. With the additional information carried by Z_1, which is correlated to the incomplete true predictor Y_2, the biases in the second column are further reduced to about three fourth of the naïve estimator. The third column shows expectations of SIMEX estimator under $(M3 : MAR - cov)$ assumption. The biases raise again when we add Y_1 into the missing model. During the process of fitting the missing model, Y_1 usually has an insignificant pvalue in the model comparison anova table. The averaged pvalue of testing if the coefficient of Y_1 is zero in model $(M3)$ is 0.2448.

Table 3.15 shows the percentage of biases using $(M1)$, $(M2)$, $(M3)$ or $(M4)$ working model when the true missing generating model is $(G3 : MNAR_2)$ which is another MNAR model. The first three columns in Table 3.15 shows that none of $(M1 : MCAR)$, $(M2 : MCAR - cov)$ and $(M3 : MAR - cov)$ can help to reduce the bias. Figure 3.13

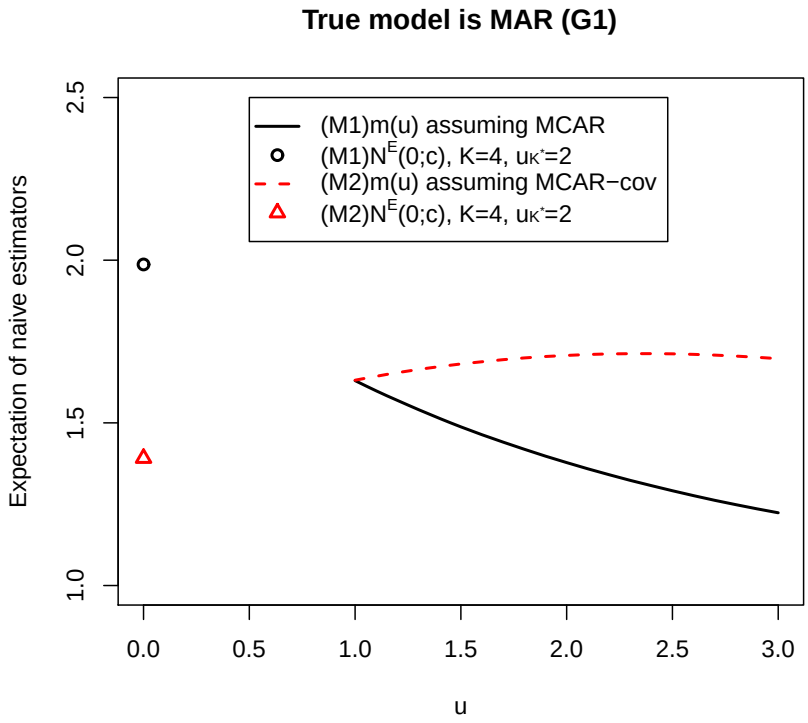

Figure 3.11 The mean function $m(u|Y,R)$ with $B = 10,000$ and $M(0;\hat{c})$ with $K = 4$ and $u_{K^*} = 2$. These are estimated from 1,000 iterations. The missing indicators are generated from MAR model $(G1 : MAR)$. The true parameter value is 9.5295.

shows $m(u)$ estimated under $(M1)$, $(M2)$, $(M3)$ or $(M4)$ assumption and the expectation of SIMEX estimator $N^E(0;c)$ with $K = 4$ and $u_{K^*} = 2$. The decreased expectation of naïve estimator yielded from $(M1 : MCAR)$ shows the effect of decresed sample size as shown in all of Figure 3.11, Figure 3.12 and Figure 3.13. As shown in Figure 3.11 and Figure 3.12, the SIMEX estimators under $(M1 : MCAR)$ assumption have smaller bias than the standard GEE estimator has. The SIMEX estimator under $(M1 : MCAR)$ assumption has higher bias, since the value of the naïve estimator based on observed data set $T(\tilde{y}^{(o,1)}) > 9.5295$,

In the fourth column of Table 3.15, an auxiliary variable Z_2, which is highly correlated to the true but incomplete Y_2, is added into the missing model. With the information

Table 3.14 Table of percentage bias of SIMEX estimator when missing is MNAR(G2). The simulation number is 1,000. We fix $K = 10$ and $B = 10,000$. The percentage of bias of naïve estimator $T(Y^{(o,1)})$ is -82.79.

		(K^*, u_{K^*})		
Order	u_{K^*}	**M1**	**M2**	**M3**
1	1.5	-80.87	-75.43	-83.10
2	1.5	-80.35	-66.64	-81.74
3	1.5	-80.53	-63.13	-80.53
4	2.0	-80.21	-62.15	-80.24
5	3.0	-79.86	-61.70	-80.39
6	3.0	-80.58	-61.88	-79.80

from all observed Z_2, the SIMEX method reduces part of the bias related to unobserved response Y_2. For analyzing data set with MNAR missingness, one either makes additional assumptions on the distribution of Y_2 or utilizes information from other auxiliary variables that are correlated to Y_2. One benefit of the SIMEX method is that it can use these extra information without altering the response model. The amount of bias reduction depends on the amount of information we collected.

Table 3.15 Table of percentage of bias of SIMEX estimator when missing is MNAR(G3). The simulation number is 1,000, $K = 10$ and $B = 10,000$.

		(K^*, u_{K^*})			
Order	u_{K^*}	**M1**	**M2**	**M3**	**M4**
1	1.5	127.94	90.29	118.55	$48.30^{(1)}$
2	1.5	149.54	93.96	140.42	$36.87^{(1)}$
3	1.5	152.87	94.82	145.90	$37.14^{(1)}$
4	2.0	153.07	95.86	146.91	$35.75^{(1)}$
5	3.0	154.90	94.89	147.81	$37.15^{(1)}$
6	3.0	151.25	93.61	145.83	$33.27^{(1)}$

(1) estimaters with percentage of bias smaller than 65.69, which is the percentage of bias of the naïve estimator $T(Y^{(o,1)})$ from observed data set $Y^{(o,1)}$.

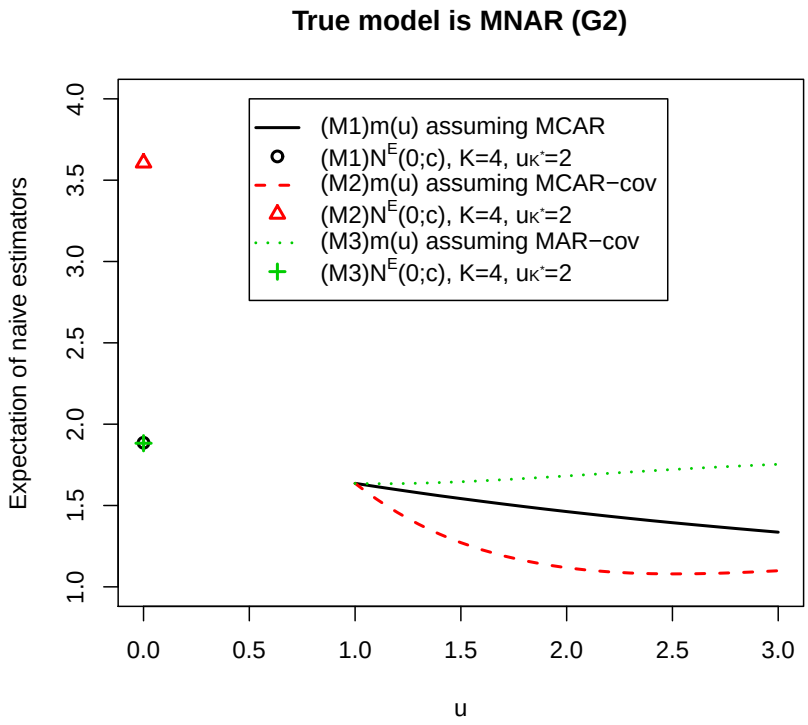

Figure 3.12 The mean function $m(u|Y,R)$ with $B = 10,000$ and $M(0;\hat{c})$ with $K = 4$ and $u_{K^*} = 2$. These are estimated from 1,000 iterations. The missing indicators are generated from MAR model $(G2 : MNAR_1)$.

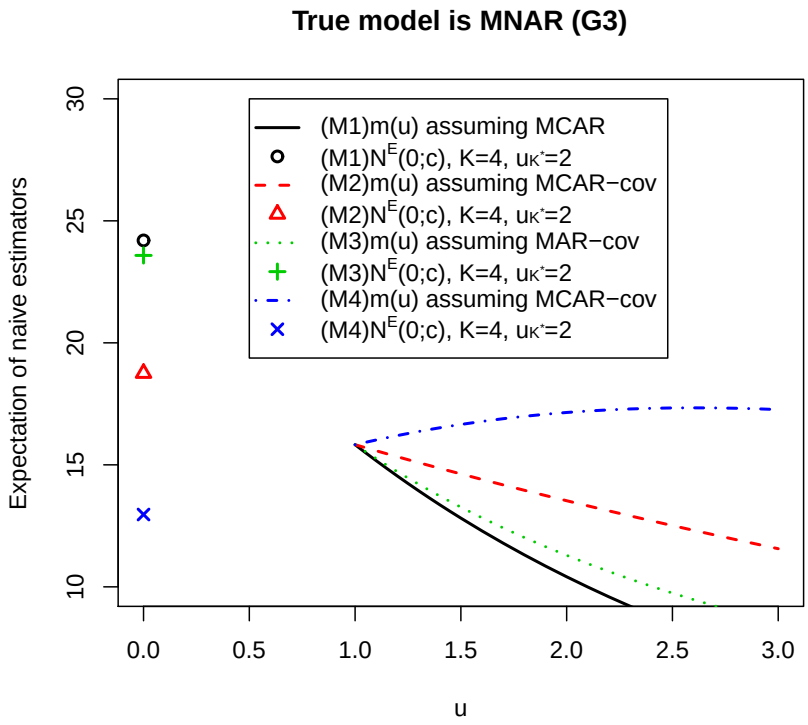

Figure 3.13 The mean function $m(u|Y,R)$ with $B = 10,000$ and $M(0;\hat{c})$ with $K = 4$ and $u_{K^*} = 2$. These are estimated from 1,000 iterations. The missing indicators are generated from MAR model $(G3 : MNAR_2)$.

CHAPTER 4. EXAMPLE: ANALYZE INCOMPLETE BINARY RESPONSE DATA USING THE GEE METHOD UNDER MAR ASSUMPTION

This is an example of a longitudinal study with incomplete binary response data. We use GEE to fit a marginal model when data are incomplete and the missing data mechanism is MAR. The GEE method which is known to be robust to the working correlation structure when data are fully observed. In the presence of missing data, the GEE estimator is biased under the wrong working correlation assumption.

We show that the SIMEX method can reduce the bias and restore the flexibility of the GEE method with regard to the working correlation model without building a new complex model that contains a missing-data mechanism. We also compare the SIMEX method with imputation based and weight based methods. The main assumption behind multiple imputation is that we can appropriately impute missing values multiple times. We show that when the missing-data mechanism is properly specified, the SIMEX method yields good estimates.

4.1 The GEE Method With Presence of Missing Data

The method of generalized estimating equations (GEE) proposed by Liang and Zeger (1986) is an extension of generalized linear models for analyzing correlated response data, which is also called marginal longitudinal generalized linear models (Cantoni et al., 2007). The GEE method is known to yield consistent estimators of regression coefficients and

their variances when the marginal mean structure is correctly specified.

In the presence of missing data, the standard GEE method simply removes incomplete records no matter whether the missingness occurs on the response or covariate variable. The GEE estimators, called the naïve estimators, are consistent with the flexibility of specifying the wrong working correlation. If the missing-data mechanism is MCAR (Liang and Zeger, 1986), which means that skipping a visit or not has to be independent from the responses. If the missing-data mechanism is MAR and the working correlation is correct, the GEE method yields consistent estimators in some special cases. For example, when the response variable is Gaussian distributed and the missing pattern depends arbitrarily on past observations, or when the outcome is binarily distributed and the missing pattern depends on any single previous outcome (Liang and Zeger, 1986).

Some studies have examined the bias and change of efficiency of GEE estimators under MAR assumption. Kenward et al. (1995) compared ML and GEE using all available or completer-only data for longitudinal ordinal data with MAR missing. They especially discuss the distinction between estimating the marginal and joint distribution. They wished to relate covariates to the parameters of marginal distribution of response variables and the GEE method is one of the alternatives to likelihood-based method that has advantages with respect to analyzing a marginal model. They show that ML on completer-only data set and GEE on either all-avialable or the completer-only data set produce misleading conclusions. Touloumi et al. (2001) used a simulation study to show that the bias in the GEE estimates increases with the severity of non-randomness and with increases of the proportion of missing data.

Weighting and Imputation are two commonly used methods in reducing the bias of GEE estimator under MAR. The weighted GEE utilizes the probability of observing of each record or each subject. The missingness may occure on the response or on some covariates. Robins et al. (1995) proposed a method based on a set of weights in the GEE

procedure (called WEE, WGEE or IPWGEE) that yields consistent estimators under MAR even when the working correlation structure is not the true correlation structure. They used the inverse of the probability of observing a record as weight. They showed that the WGEE estimator is consistent and asymptotically normal when the missing pattern is monotone and the probability of missing is consistently estimated. Lipsitz et al. (2000) shows that the WGEE is very close to the ML score equations under the MAR assumption. Preisser et al. (2002) used a simulation study to compare GEE and WGEE for data with repeated binary responses, and with MAR drop-outs. They show that WGEE resulted in smaller bias than GEE with true or identity working correlation in general, but WGEE may yield greater bias than GEE if the missing mechanism model is misspecified.

The imputation-based methods utilize the distribution of unobserved measurement of either the response or a covariate. When the missing models for covariates are MAR and the missing model for drop out is MCAR, there have been studies based on single or multiple imputations. Xie and Paik (1997b) applied multiple imputation (MI) (Rubin, 2004) to the GEE method with misspecified working correlation for correlated binary response model with missing covariates that are MAR. They used a simulation study to compare completer-only, sample average imputation(SA) and MI-GEEs with four different imputation methods: Bayesian bootstrap(BB), approximate Bayesian bootstrap(ABB), fully normal(FN), and mean and variance adjusted hot-deck(MV) imputations. They used an identity working correlation while the true correlation structure is symmetric matrix. They showed that MI-GEEs are fairly unbiased and as efficient as the sample average (SA) imputation method. In their example, the MI-GEEs are robust against moderate misspecification of distribution for imputation. Paik (1997) proposed using single and multiple imputations to sequentially impute missing responses y_{it} for subject i at time t from the smallest t. The estimators from imputation methods are consistent when the missingness is MAR and the missingness model is correctly specified. They

used single mean imputation which imputes the missing response y_{it} by the conditional expectation $E(y_{it}|D_{it}, r_{it} = 0)$ where r_{it} is an indicator variable with $r_{it} = 0$ for variable y_{it} to be missing and D_{it} represents measurements available at time t. D_{it} may include responses from past time points, covariates from current and past time points. When dropout is MAR, $E(y_{it}|D_{it}, r_{it} = 0)$ is estimated by $E(y_{it}|D_{it}, r_{it} = 1)$. They suggested to impute missing data sequentially from $t = t_{i,min} = min\{t; r_{it} = 0\}$ to the end of predetermined visit $t = N$. For $t > t_{i,min}$, unobserved components in D_{it} are replaced by imputed measurements. They proved that when the imputed value is a consistent estimator of the condition mean $E(y_{it}|D_{it}, r_{it} = 0)$, the bias of the GEE estimator of the model coefficients has a large sample normal distribution with mean zero when the sample size goes to infinity. They compared estimators from the single mean imputation with estimators from multiple imputations using the ABB procedure sequentially(MI-ABB). They proved that when both sample size and number of imputation in MI-ABB go to infinity, the estimated coefficients from MI-ABB and single imputation are equivalent. They also concluded that missspecification of the imputation yields estimates that are nearly unbiased.

One benefit of the multiple imputation is the small imputation number. The imputation number for the multiple imputation (MI) method is shown to have good efficiency form only 2 to 10 imputations (Rubin, 2004). Graham et al. (2007) used simulation studies showing that much more imputations are needed to have power that is close to that of the full information maximum likelihood method.

There are also methods that adjust the estimation of covariance in the estimating equations instead of imputing unobserved data or weighting components of estimating equations. Xie and Paik (1997a) proposed a single imputation method on the GEE method for correlated binary response model with missing covariates that are MAR or MAR-cov. Lipsitz et al. (2000) proposed a modefied GEE using Gaussian estimation of the correlation parameters for correlated binary response model with missing responses,

where the standard GEE used an all-available-pairs estimator. This method yields consistent estimators under MCAR and estimators with almost negligible biases under MAR when the working correlation is correctly specified.

The flexibility with respect to the working covariance assumption is a big advantage when data are fully observed or the missing-data mechanism is MCAR but the advantage does not always hold true for MAR. To demonstrate that the SIMEX method can reduce bias when the working correlation is not the true correlation, we analyze an incomplete clinical study data set which has been analyzed by Preisser et al. (2000) using the WGEE method. A simulation study is conducted to compare the biases of several methods using wrong working correlation. Preisser et al. (2002) used a similar simulation study to compare GEE and WGEE using true or identity working correlation under MCAR, MAR or MNAR missingness. We compare bias and coverage rate of 95 percent confidence intervals of GEE, WEE, MI-GEE and SIMEX-GEE estimators. As other weighting, imputing and adjusting estimating equation methods discussed above, we also assume the drop-out missing pattern in this example.

4.2 Data and Model

The Coronary Artery Risk Development in young Adults (CARDIA) study recruited 5,115 black and white young adults with ages from 17 to 35 in 1986 and recorded their cardiovascular risk factors at 0, 2, 5, 7, 10 years from 1986 (Hughes et al., 1987). The CARDIA data includes self-reported smoking status, age, birth year, education, race and gender. The sample was designed to obtain approximately balanced sample sizes with respect to age, race, gender and education. Preisser et al. (2000) analyzed the first four measurements of 5,078 subjects who had records of smoking status at 1986 (baseline). They removed 578 out of 17,995 records to create a monotone missingness data set. The response variable is whether or not the subject is a smoker. Their goal is to make

inferences on the change in smoking prevalence for each race by sex group in the presence of missing data and intraperson correlation. We compare the results from the SIMEX method with those from the GEE, WGEE, MI-GEE methods.

4.2.1 The response model

Let $X_t = 0, 2, 5, 7$ for $t = 1, 2, 3, 4$, and $X_{gp,i}$ has value $1, 2, 3, 4$ if the ith subject is black males, black females, white males and white females, respectively. The response variable $Y_{it} = 1$ if the status is smoking for the tth visit of the ith subject $(i = 1, \ldots, 5,078)$. The response model is

$$\begin{aligned} logit(\pi_{it}) &= logit(P(Y_{it} = 1)) \\ &= \beta_1 + \beta_2 I(X_{gp,i} = 2) + \beta_3 I(X_{gp,i} = 3) + \beta_4 I(X_{gp,i} = 4) \\ &\quad + \beta_5 X_t + \beta_6 X_t I(X_{gp,i} = 2) \\ &\quad + \beta_7 X_t I(X_{gp,i} = 3) + \beta_8 X_t I(X_{gp,i} = 4) \\ &= \sum_{s=1}^{4} I(X_{gp,i} = s)(\beta_{s0} + \beta_{s1} X_t). \end{aligned} \tag{4.1}$$

4.2.2 The missingness model

Preisser et al. (2000) specified the missingness model as a generalized linear model with linear predictor $logit(\lambda_{it}(\alpha)) = Z_{it}\alpha$. They showed that missing rates were different for subjects who were smoking and nonsmoking during the year of entrance and the missing indicators R_{it} and $Y_{i,t-1}$ were not independent within each time and each gender by race group. Preisser et al. (2000) suggested a process that cumulatively adds significant (deviance reduction) explanatory variables into the missing model. Additionally, they add extra non-significant explanatory variables into the missing model until they have stable weights, which means that weights change little when further explanatory variables are added. For example, if there are nested missing models A, B and C with explanatory variables (W_1), (W_1, W_2) and (W_1, W_2, W_3), respectively. If weights esti-

mated from model A and B are quite different, and weights estimated from model B and C are similar, they suggest model B even though the additional explanatory variable W_2 in model B is not significant.

Table 4.1 lists the explanatory variables $Z^{(1)}$ chosen by the stepwise selection method (AIC) and the explanatory variables $Z^{(2)}$ chosen by Preisser et al. (2000). The stepwise selected model suggests that the increases of probability for data being observed for each increased age unit(in 10 years) are different for each education and race group. The stepwise selected model also suggests that change in the missing rates over time depend on education level. Table 4.2 presents an example of estimating the probabilities for data being observed for a 17-year old subject. The stepwise selected model also yields higher estimated probability for data being observed for subjects with education level 1 and lower for subjects with education level 3 comparing to the Preisser's model.

The stepwise method is convenient but should be used with caution. In this example, the marginal weights from the stepwise selected model are close to the results from Preisser's model. The stepwise selected model yields marginal weights ranging from 1.031 to 2.439. Preisser's model yields marginal weights ranging from 1.03 to 2.282. The marginal weights for the WGEE method are defined as the inverse of $P(R_{it} = 1|Z, Y_{it-1}) = \Pi_{s=2}^{t} P(R_{it} = 1|Z, Y_{it-1}, R_{it-1} = 1)$.

4.3 Results

Table 4.3 list the estimated coefficients β_{s1} in (4.1) for groups $s = 1, 2, 3, 4$. The GEE-1 method used all data including data from skipped and returned subjects. The GEE-2 method deleted returned data to make the missingness pattern be monotone. As Preisser et al. (2000) showed, these two datasets yielded similar estimates from the GEE method. They also used the nearest observed Y or the baseline Y_{i1} to replace Y_{it-1} in the missingness model and concluded that these two datasets yielded similar results from

Table 4.1 Table of coefficients of missing models selected by stepwise method and by Preisser.

	Stepwise ($Z^{(1)}$) AIC: 8649			**Preisser** ($Z^{(2)}$) AIC: 8668.7		
Variables	Coef.	SE	$P(>\|z\|)$	Coef.	SE	$P(>\|z\|)$
Intercept	3.240	0.491	<.001	2.221	0.277	<.001
$Y_{i(t-1)}$	-0.600	0.146	<.001	-0.666	0.454	0.143
Gender=male vs. female	0.108	0.095	0.257	0.101	0.104	0.331
Time=5 vs. 2	-0.555	0.158	<.001	-0.326	0.127	0.010
Time=7 vs. 2	-0.584	0.161	<.001	-0.369	0.130	0.004
Race=black vs. white	0.406	0.440	0.356	-0.594	0.130	<.001
Edu=2(some college) vs. 1	-2.450	0.565	<.001	-0.512	0.091	<.001
Edu=3(high school or less)	-3.251	0.575	<.001	-0.906	0.101	<.001
I(race=black)×I(gender=male)	-0.246	0.120	0.040	-0.255	0.120	0.033
I(race=black)×I(time=5)	0.193	0.152	0.204	0.312	0.144	0.030
I(race=black)×I(time=7)	0.337	0.158	0.033	0.462	0.150	0.002
I(edu=2)×$Y_{i(t-1)}$	0.495	0.175	0.005	0.501	0.181	0.005
I(edu=3)×$Y_{i(t-1)}$	0.444	0.175	0.011	0.439	0.183	0.017
Age(in 10 year units)	0.041	0.186	0.824	0.386	0.099	<.001
Age×I(race=black)	-0.367	0.173	0.034			
Age×I(edu=2)	0.673	0.219	0.002			
Age×I(edu=3)	0.795	0.223	<.000			
I(edu=2)×I(time=5)	0.354	0.190	0.062			
I(edu=2)×I(time=7)	0.372	0.196	0.058			
I(edu=3)×I(time=5)	0.520	0.194	0.007			
I(edu=3)×I(time=7)	0.598	0.203	0.003			
$Y_{i(t-1)}$×Age				-0.036	0.163	0.826
$Y_{i(t-1)}$×I(race=black)				0.072	0.131	0.582
$Y_{i(t-1)}$×I(gender=male)				0.024	0.120	0.843
$Y_{i(t-1)}$×I(time=5)				0.088	0.140	0.529
$Y_{i(t-1)}$×I(time=7)				0.206	0.150	0.169

the WEE method.

Table 4.3 also lists estimates from multiple imputed datasets. The imputation numbers m were 3, 5, 10 and 20. The imputation model for Y_{it} included $Y_{is}, s \neq t$, education, age, race and gender. The R package "mi" imputed each binary $Y_{it}, t = 2, 3, 4$ iteratively until stable. The estimated variance is stable after $m = 10$. The MI estimators are closer to GEE-2 with an exchangeable correlation assumption. Assume that the working correlation is the true correlation, the bias of the GEE estimator under MAR is usually small. There are situations in which the bias can be quiet large. The simulation results

Table 4.2 Table of estimated probabilities of observing a response for a 17-years-old subject who is nonsmoker at time 2, 5, 7 from different gender, race and education groups from two missingness model. The stepwise selected model suggests higher observed rate for subjects with Edu=3 comparing to the estimates from Preisser's model. Preisser's model suggests increasing observation rates for black subjects and decreased observation rates for white subjects over time.

			Stepwise ($Z^{(1)}$)			**Preisser** ($Z^{(2)}$)		
Gender	**Race**	**Edu**	T=2	T=5	T=7	T=2	T=5	T=7
Female	White	1	0.804	0.798	0.806*	0.878*	0.838	0.832
		2	0.881*	0.859	0.857	0.914*	0.885	0.880
		3	0.965*	0.940	0.939	0.947*	0.928	0.925
	Black	1	0.767	0.794	0.824*	0.799	0.796	0.813*
		2	0.857	0.856	0.871*	0.855	0.853	0.866*
		3	0.957*	0.939	0.945	0.908	0.906	0.915*
Male	White	1	0.820	0.815	0.822*	0.888*	0.852	0.846
		2	0.892*	0.871	0.870	0.922*	0.895	0.891
		3	0.968*	0.946	0.944	0.952*	0.934	0.932
	Black	1	0.742	0.771	0.803*	0.773	0.770	0.789*
		2	0.839	0.838	0.855*	0.834	0.832	0.847*
		3	0.950*	0.930	0.937	0.894	0.892	0.902*

*The highest probability over time for each gender, race and education group.

with MNAR assumption in the next section will show the potential large bias when either the working correlation structure is not the true correlation structure or the MAR assumption is not true.

Figure 4.1 shows the estimated expectations of naïve estimators $m(u|y,r)$ with additional missing portions generated from the stepwise selected missing model in solid black dots. The block dots that were the averages of estimators from 1,000 simulation iterations formed a smooth curve. The solid line is the second order polynomial $M(u;c)$ that approximate $m(u|y,r)$ for $1 \leq u \leq 3$.

Two more extrapolation functions from different missingness models are shown in Figure 4.1 for comparison. The dashed lines shows the polynomials that approximate the simulation results $m(u|y,r)$ that are calculated from data with additional missing portions generated from Pressier's missingness model. The intercept of the dashed line

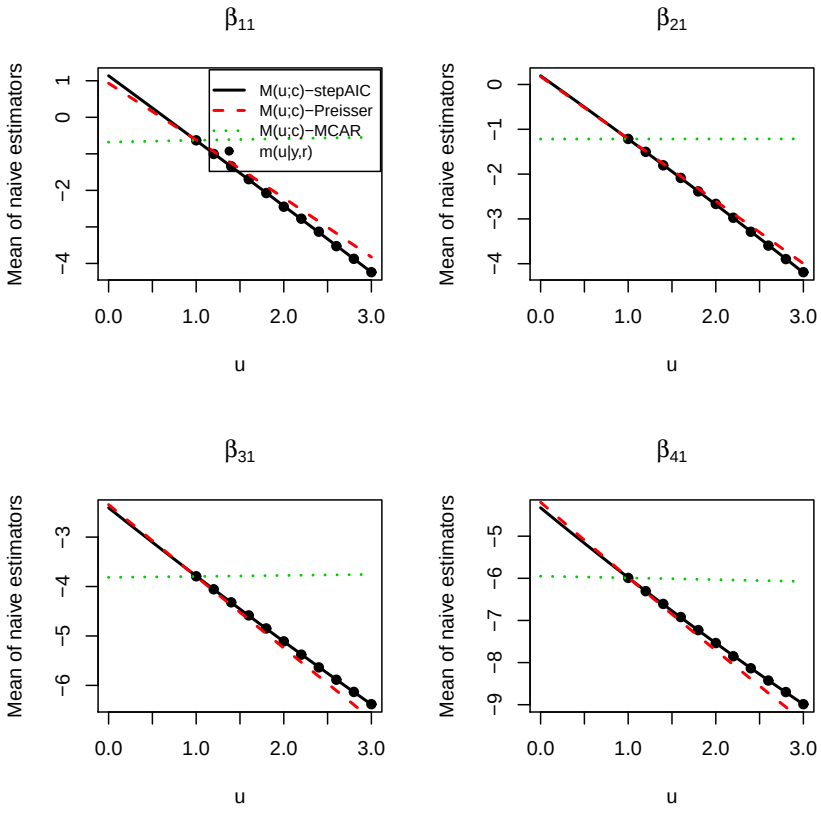

Figure 4.1 The expectation naïve estimators estimated from 1,000 simulation iterations and the second order polynomial approximation.

is very close to the intercept of the black solid line in each plot. The Pressier's missing model yields similar but slightly smaller SIMEX estimators of change of log odds of smoking rate over time for the black group (group=1,2) and higher for the white group (group=3,4).

The dotted lines shows the polynomials that approximate the estimates based on the MCAR assumption. The dotted lines shows the change of expectation of naïve estimators caused by the reduction of sample size and visually shows the effect of MAR assumption on the SIMEX estimators. Both the solid and dashed line suggest the naïve estimators from observed data ($m(1|y,r)$) underestimate the change of log odds ratio for each additional year.

Table 4.4 list the SIMEX estimators with different covariates in the missing model and different values of (K, K^*, u_{K^*}). The number of simulation iteration B is 1,500. The values of (K^*, u_{K^*}) are chosen by the residual plots. The estimated standard deviation of SIMEX estimators are larger than other estimators.

4.3.1 Cross-validation

The 5078 subjects in the smoking data set are randomly partitioned into 2 groups d_1 and d_2. There are 2539 subjects in each group. Let one of these data set be the training data set and the other one be the validation dataset. The probability $P(Y = 1|X)$ is estimated from the training dataset. A threshold of $\hat{Y}$ is selected for each method and each training dataset such that the Euclidean distance,

$$\sqrt{(P(\hat{Y} = 1|Y = 1) - 1)^2 + (P(\hat{Y} = 0|Y = 0) - 1)^2},$$

is minimized. Then, we calculate the sensitivity ($Se = P(\hat{Y} = 1|Y = 1)$) and specificity ($Sp = P(\hat{Y} = 0|Y = 0)$) of the validation dataset. The results are listed in Table 4.5. The values of (sensitivity, specificity) of the SIMEX estimators have smaller distance to (1,1) when $(K, K^*, u_{K^*}) = (1, 2, 1.4)$ or $(K, K^*, u_{K^*}) = (1, 10, 3)$. Two missing models yield similar results for the SIMEX method, but yield different results for the WEE method.

In summary, because the difference of basic assumptions between methods, utilizing information from missing model or missing response distribution, the estimators are slightly different. The estimators of WEE and SIMEX method are similar under the same correlation assumption. The two missing models in SIMEX-1 and SIMEX-4 are different, but the differences are small when the missing model is reasonably carefully selected. The estimators of MI-GEE are similar to the GEE estimator with exchangeable correlation structure which is known to have smaller bias when the working correlation is true correlation and the MAR assumption. The difference between MI-GEE and WEE

or SIMEX-GEE shows that the the auxiliary variables do provides information on bias reduction, and emphasizes the benefit of easiness of incorporating auxiliary variables in the WEE or SIMEX-GEE methods. with additional information provided by thek

4.4 Simulation

Here we use a simulation study with smaller sample size and simpler design to demonstrate the use of the SIMEX method when the missingness is MAR or MNAR. Consider a longitudinal binary data wtih $K = 100$ independent subjects and each subject was planned to have $T = 4$ visits. Assume the missingness is MAR or MNAR and the missing pattern is monotone. Simulated data sets are analyzed by the GEE, WGEE, MI and SIMEX method.

4.4.1 The simulation model

Let $I(X_{gp,i} = 1) = 1$ if the ith subject was assigned to group one and $I(X_{gp,i} = 1) = 0$ otherwise. Let π_{it} be the expectation of the response variable Y_{it} of the ith subject at the tth visit. Assume

$$\begin{aligned} logit(\pi_{it}) &= logit(P(Y_{it} = 1)) \\ &= -0.6 + 0.1I(X_{gp,i} = 1) + (-0.2 + 0.2I(X_{gp,i} = 1))X_t \end{aligned}$$

where $X_t = 0, \frac{1}{3}, \frac{2}{3}, 1$. The correlation $cor(Y_{is}, Y_{it})$ is $\rho = 0.6$ for any $s \neq t$. The correlated binary responses are generated by the algorithm proposed by Qaqish (2003). One binary auxiliary variable Z is generated such that the correlation between Y and Z is 0.6.

The missing indicator $R_{it} = 1$ if Y_{it} is observed and $R_{it} = 0$ otherwise. The probability of observing Y_{it} is $P(R_{it} = 1|R_{i,t-1} = 1, Y_{i,t-1}) = \lambda_{it}$ for $t > 1$. Assume

$$\lambda_{it} = \begin{cases} 1 & \text{if } t = 1 \\ \frac{1}{1+e^{-(\alpha_0+\alpha_1 Y_{i,t-1}+\alpha_2 Y_{it})}}, & \text{if } t = 2, 3, \ldots, T \end{cases} \tag{4.2}$$

where $\alpha = (\alpha_0, \alpha_1, \alpha_2)$ is either $(1.7, 1.5, 0)$, which yields MAR missingness, or $(3, -4, 4)$, which yields MNAR missingness. Both α yield observed rates $1, 0.9, 0.8, 0.7$ at time $X_t = 1, 1/2, 2/3, 1$ respectively, if the response is observed at the previous time point.

4.4.2 The fitting model

Suppose that the main object is to estimate the log odds ration of two groups at each time point. The response model is a generalized linear model with a binary response, logit link function and correlation between responses from the same subject. The correlation is assumed the same for any pair of observations from the same subject. The variable X_t is assumed categorical. The linear predictor is

$$\begin{aligned} logit(\pi_{it}) &= logit(P(Y_{it} = 1)) \\ &= \beta_1 + \beta_2 I(X_{gp,i} = 1) + \beta_3 I(X_t = 1/3) + \beta_4 I(X_t = 2/3) + \beta_5 I(X_t = 1) \\ &\quad + (\beta_6 I(X_t = 1/3) + \beta_7 I(X_t = 2/3) + \beta_8 I(X_t = 1)) I(X_{gp,i} = 1) \end{aligned}$$

For WEE and SIMEX-GEE, the working missing model is a generalized linear model with logit link functon. The assumption is

$$\lambda_{it} = \begin{cases} 1, & \text{if } t = 1 \\ \frac{1}{1+e^{-(\alpha_0 + \alpha_1 Y_{i,t-1})}}, & \text{if } t = 2, 3, \ldots, T. \end{cases} \tag{4.3}$$

for data with MAR and

$$\lambda_{it} = \begin{cases} 1, & t = 1 \\ \frac{1}{1+e^{-(\alpha_0 + \alpha_1 Y_{i,t-1} + \alpha_2 Z_{it})}}, & t = 2, 3, \ldots, T. \end{cases} \tag{4.4}$$

for data with MNAR.

For the SIMEX-GEE method, the two extrapolation polynomials we considered are of order 2 with $u = 1, 1.05, 1.10, \ldots, 1.5$ and of order 3 with $u = 1, 1.05, 1.10, \ldots, 2$. These polynomials are estimated from 300 simex iterations. These settings are selected based on plots of simex estimators against the number of iterations from several pretrial simulations.

For the MI-GEE method, the missing values are imputed 3 times. The imputation model for Y_t includes Y_s where $1 \leq s \leq 4$ and $s \neq t$. When the missingness is MNAR, values of the auxiliary variable Z at all four time points are included in the imputation model.

We use the correct working response model to estimate model parameters. The working missing-data mechanism is equal to the true model specified in (4.4).

4.4.3 Simulation results - MAR data

There are 1,000 datasets generated and analyzed by WGEE, MI-GEE and SIMEX-GEE. All estimators calculated from observed dataset are compared with estimators calculated from full and observed datasets by the GEE method (M1:GEE-full and M2:GEE-naïve). We use convergence criteria $1e-4$ for the WGEE algorithm described in Preisser et al. (2000) (M3:WEE). Three imputation models are used for imputation. The first model (M4:MI-1) imputes Y_{it} based on $\{X_{gp,i}, Y_{is}; 1 \leq s \leq 4, s \neq t\}$. The second model (M5:MI-2) imputes Y_{it} based on only $(X_{gp,i}, Y_{i1})$. The third model (M6:MI-3) imputes Y_{it} based on $\{X_{gp,i}, Y_{is}; 1 \leq s \leq 4, s \neq t\}$ but specifies that the Y_{it} are discrete variables in imputation models and transforms the imputed value to one if $Y_{it} > 0.5$ and zero otherwise. The SIMEX method uses 500 iterations in the simulation step. Both WGEE and SIMEX-GEE use the correct missing model to estimate the probability of observing for MAR datasets and add auxiliary variable Z as the explanatory variable for MNAR datasets.

The second column of Table 4.6 shows the percentage of bias of estimated log odds ratio over time. The first four rows are GEE estimators when data are fully observed. The naïve estimators of β_{31} and β_{41} that use partially observed data for GEE estimator without any adjustment both have biases that are about 5 percent higher than the full version. The MI-1 and MI-2 methods both reduce the biases of the naïve estimators. The MI-2 method that imputes unobserved Y_t by models with only Y_1 have relatively

smaller biases.

The SIMEX-1 method with $(K, K^*, u_{K^*}) = (1, 5, 1.5)$ shown in Table 4.7 pulls estimators of β_{21}, β_{31} and β_{41} downward. Figure 4.2 shows the residual plot of β_{41} from the SIMEX-1 method. the curvy trend of the residual indicates that the linear extrapolation function does not fit the mean function $m(u)$ well. In this case, we need to increase both the simulation iteration number (B) and the order of polynomial (K).

Table 4.7 also shows the the SIMEX-2 estimators with $(K, K^*, u_{K^*}) = (1, 3, 1.2)$. The biases of SIMEX-2 estimators have less bias then SIMEX-1 estimators. The process of diagnosing and finding the values of (K, K^*, u_{K^*}) is important to the SIMEX method. The residual plot of the SIMEX-2 estimator of β_{41} (not shown) is still curvy and the standard deviation of $\hat{\beta}_{41}$ (2.546) is significantly larger. Limited by the precision of computation, increase order K and increase iteration number B until the estimators of $m(u)$ converged is suggested for further improvement.

The third column in Table 4.7 shows the standard deviations of SIMEX estimators from 1,000 simulated datasets. The fourth column in Table 4.7 shows the averaged of estimated standard deviation of $\hat{\beta}$. The variance of $\hat{\beta}$ is the SIMEX estimator of $var(\hat{\beta}|Y)$ plus the variance of $m(u|Y, R, \{R_B^{(u_k)}\}_{k=1,\ldots,K^*})$. The standard deviation of $\hat{\beta}_{41}$ is significantly larger than others. The plot of cumulated estimator versus B of several random selected dataset (not shown here) shows that $B = 500$ is not large enough to have converged estimators at the fourth time point. We should increase the simulation number B to reduce the variation of $\hat{\beta}_4$.

4.4.4 Simulation results - MNAR data

The second column of Table 4.8 shows the percentage of bias of estimated log odds ratio over time. Both the WEE method and the SIMEX method utilize auxiliary variable Z which is correlated to incomplete response variable Y. The naïve GEE estimators are higher at the thrid and fourth visits. The WEE estimators are slightly lower at the

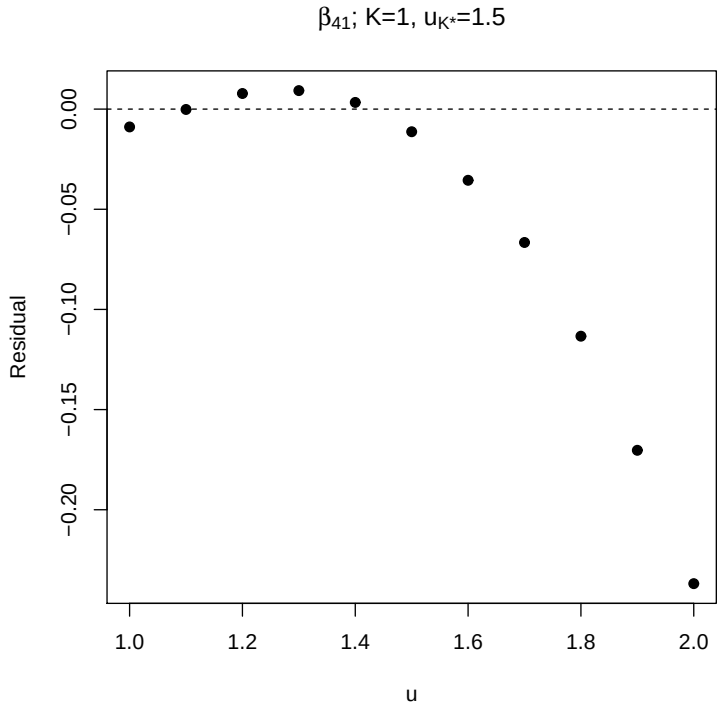

Figure 4.2 Plots of residuals of β_{41} with $(K, K^*, u_{K^*}) = (1, 5, 1.5)$.

third and fourth visits. The MI-GEE estimators are much lower at the second, third and fourth visits.

The second column of Table 4.9 shows the percentage of bias of the SIMEX estimator. The SIMEX estimator is closer to the GEE estimator from fully observed data. The averaged estimated standard deviation SIMEX estimator (at the fourth column) are much higher than the sample standard deviation of the SIMEX estimators (at the third column). That makes the coverage rate higher than 95%. For limited calculation power, the estimated variances and coverage percentages of SIMEX estimators are calculated from only 300 datasets and 30 bootstrap iterations for each dataset. When analyzing a single dataset, the bootstrap iteration should be higher.

Table 4.3 Table of 100 times estimated coefficients $\{\beta_{11}, \ldots, \beta_{41}, \}$, which are changes of log odds of smoking rates for each additional year for group=1(Black-male),2(Black-female),3(White-male) and 4(White-female).

Method	**cov**	$\hat{\beta}_{11}(\hat{\sigma}(\hat{\beta}_{11}))$	$\hat{\beta}_{21}(\hat{\sigma}(\hat{\beta}_{21}))$	$\hat{\beta}_{31}(\hat{\sigma}(\hat{\beta}_{31}))$	$\hat{\beta}_{41}(\hat{\sigma}(\hat{\beta}_{41}))$
GEE-1 [1]	indep	-0.25(0.84)	-0.65(0.72)	-3.37(0.83)	-5.24(0.86)
	exch	1.13(0.71)	-0.24(0.61)	-2.16(0.73)	-4.08(0.75)
	unstr	1.14(0.72)	-0.35(0.61)	-2.28(0.74)	-4.19(0.74)
GEE-2 [2]	indep	-0.63(0.93)	-1.22(0.80)	-3.79(0.89)	-5.99(0.92)
	exch	0.97(0.73)	-0.09(0.64)	-1.99(0.74)	-4.36(0.78)
	unstr	0.91(0.74)	-0.19(0.64)	-2.10(0.74)	-4.53(0.78)
WEE-1 [3]	indep	1.10(0.89)	0.25(0.76)	-2.41(0.86)	-4.41(0.90)
	exch	2.27(0.87)	0.92(0.73)	-0.89(0.78)	-3.11(0.82)
	unstr	1.96(0.93)	0.50(0.77)	-1.34(0.83)	-3.49(0.88)
WEE-2 [4]	indep	0.89(0.81)	0.11(0.69)	-2.31(0.81)	-4.29(0.85)
	exch	2.06(0.86)	0.79(0.72)	-0.80(0.77)	-2.99(0.81)
	unstr	1.74(0.91)	0.34(0.75)	-1.25(0.82)	-3.38(0.87)
MI-1 [5]	indep	1.08(0.86)	-0.13(0.78)	-1.67(0.69)	-4.52(1.06)
	exch	1.08(0.86)	-0.13(0.78)	-1.67(0.69)	-4.52(1.06)
	unstr	1.11(0.86)	-0.25(0.78)	-1.73(0.69)	-4.53(1.06)
MI-2 [6]	indep	0.97(0.76)	0.02(0.76)	-1.80(0.73)	-4.38(0.92)
	exch	0.97(0.76)	0.02(0.76)	-1.80(0.73)	-4.38(0.92)
	unstr	1.00(0.76)	-0.10(0.77)	-1.87(0.73)	-4.39(0.92)
MI-3 [7]	indep	1.03(0.72)	0.09(0.66)	-1.85(0.77)	-4.30(0.86)
	exch	1.03(0.72)	0.09(0.66)	-1.85(0.77)	-4.30(0.86)
	unstr	1.05(0.72)	-0.01(0.67)	-1.91(0.77)	-4.32(0.86)
MI-4 [8]	indep	1.00(0.72)	-0.07(0.66)	-1.86(0.75)	-4.22(0.84)
	exch	1.00(0.72)	-0.07(0.66)	-1.86(0.75)	-4.22(0.84)
	unstr	1.01(0.72)	-0.17(0.66)	-1.92(0.76)	-4.23(0.85)

1. Use the full dataset. A subject may skip a visit and return in the next scheduled visit.
2. The missingness pattern is monotone. The returned visits after skipping the previous visits are deleted.
3. The missingness model is selected by the stepwise selection method.
4. The missingness model is Preisser's missingness model
5. The imputation number is 3.
6. The imputation number is 5.
7. The imputation number is 10.
8. The imputation number is 20.

Table 4.4 Table of 100 times SIMEX estimators of coefficients $\{\beta_{11}, \ldots, \beta_{41}, \}$, which are changes of log odds of smoking rates for each additional year for group=1(Black-male),2(Black-female),3(White-male) and 4(White-female). The simulation number $B = 1,500$. The working covariance is independent.

Method	(K, K^*, u_{K^*})	$\hat{\beta}_{11}(\hat{\sigma}(\hat{\beta}_{11})^1$	$\hat{\beta}_{21}(\hat{\sigma}(\hat{\beta}_{21}))$	$\hat{\beta}_{31}(\hat{\sigma}(\hat{\beta}_{31}))$	$\hat{\beta}_{41}(\hat{\sigma}(\hat{\beta}_{41}))$
SIMEX-1 [2]	(1,2,1.4)	1.12(1.02)	0.25(0.75)	-2.45(1.06)	-4.46(1.11)
SIMEX-2 [2]	(2,5,2)	1.02(1.11)	0.26(0.82)	-2.43(1.08)	-4.47(1.14)
SIMEX-3 [2]	(3,10,3)	1.05(1.09)	0.16(0.81)	-2.46(1.08)	-4.45(1.14)
SIMEX-4 [3]	(1,2,1.4)	0.99(1.06)	0.21(0.75)	-2.35(1.06)	-4.35(1.11)
SIMEX-5 [3]	(2,5,2)	0.90(1.15)	0.22(0.82)	-2.25(1.08)	-4.35(1.13)
SIMEX-6 [3]	(3,10,3)	0.83(1.12)	0.17(0.81)	-2.29(1.08)	-4.36(1.13)

1. The $(P2a1)$ and $(P2a2)$ in Equation 2.18 are calculated from 20 bootstrap iteration and the derivative is approximated by the slope of a linear funtion with $\delta = 0.05$ and $B = 400$ for each bootsrap sample.
2. The missingness model is selected by the stepwise selection method.
3. The missingness model is selected by Preisser(2000).

Table 4.5 Table of averaged sensitivities ($Se = P(\hat{Y} = 1|Y = 1)$ specificities ($Sp = P(\hat{Y} = 0|Y = 0)$) and Euclidean distances of (Se, Sp) from $(1, 1)$. The smaller distance is better. The working correlations are assumed independent.

Method	(K, K^*, u_{K^*})	**Sensitivity**	**Specificity**	**distance to (1,1)**
GEE-1 [1]		0.57	0.54	0.69
GEE-2 [2]		0.57	0.54	0.63
WEE-1 [3]		0.42	0.68	0.63
WEE-2 [4]		0.57	0.54	0.63
MI-3 [5]		0.52	0.49	0.70
SIMEX-1 [6]	(1,2,1.4)	0.49	0.69	0.59
SIMEX-2 [6]	(1,5,2)	0.51	0.31	0.85
SIMEX-3 [6]	(1,10,3)	0.49	0.69	0.59
SIMEX-4 [7]	(1,2,1.4)	0.49	0.69	0.59
SIMEX-5 [7]	(1,5,2)	0.51	0.31	0.85
SIMEX-6 [7]	(1,10,3)	0.49	0.69	0.59

1. Use the original dataset with arbitrary missing pattern.
2. Use the monotone dataset where the measurements from returned visits after skipped previous visits are deleted.
3. The explanatory variables of the missingness model are selected by the stepwise selection method.
4. The explanatory variables of the missingness model is selected by Preisser(2000).
5. The imputation number is 10.
6. Simulation number is 1500. The missingness model is selected by the stepwise selection method.
7. Simulation number is 1500. The missingness model is Preisser's missingness model.

Table 4.6 Table of mean, standard deviation and percentage of bias of estimated log odds ratios, estimated asymptotic standard deviation and coverage rates of 95% confidence intervals from $1,000$ simulated datasets with MAR.

Method	**Cov**	**t**	$\bar{\hat{\beta}}$	$100\frac{\hat{\beta}-\beta}{\beta}$	$sd(\hat{\beta})$	$\bar{\hat{\sigma}}(\hat{\beta}))$	CR[1]
GEE-full[2]	Indep	1	0.107	6.853	0.418	0.421	0.955
		2	0.178	7.083	0.410	0.423	0.961
		3	0.248	6.256	0.431	0.426	0.949
		4	0.298	-0.803	0.433	0.430	0.955
GEE-naive[3]	Indep	1	0.107	6.853	0.418	0.421	0.955
		2	0.178	6.724	0.458	0.465	0.964
		3	0.263	12.515	0.522	0.515	0.958
		4	0.367	22.365	1.769	0.576	0.958
WEE[4]	Indep	1	0.107	6.505	0.418	0.416	0.953
		2	0.179	7.202	0.461	0.463	0.963
		3	0.262	12.480	0.523	0.514	0.958
		4	0.317	5.682	0.598	0.578	0.957
MI-1[5]	Indep	1	0.107	6.853	0.418	0.421	0.955
		2	0.173	3.863	0.440	0.457	0.973
		3	0.260	11.296	0.488	0.489	0.969
		4	0.339	12.945	1.204	0.557	0.968
MI-2[6]	Indep	1	0.107	6.853	0.418	0.421	0.955
		2	0.169	1.169	0.437	0.455	0.966
		3	0.256	9.819	0.492	0.494	0.975
		4	0.323	7.660	0.836	0.576	0.973
MI-3[7]	Indep	1	0.107	6.853	0.418	0.421	0.955
		2	0.168	1.010	0.428	0.452	0.970
		3	0.236	1.119	0.451	0.475	0.980
		4	0.270	-9.984	0.471	0.503	0.979

1. CR $= P(\beta \in 95\%CI)$.
2. Use the full dataset.
3. Use the observed dataset with monotone missing pattern.
4. This row summarize only 999 simulated dataset, since the standard deviation estimator is not available for one dataset.
5. The number of imputation for MI is 3. The imputation model for Y_t includes all other response variables $\{Y_s; s \neq t\}$ for each $t = 2, 3, 4$.
6. The number of imputation for MI is 3. The imputation model for Y_t includes only Y_1 for each $t = 2, 3, 4$.
7. The number of imputation for MI is 3. The imputation model for Y_t includes all other response variables $\{Y_s; s \neq t\}$ for each $t = 2, 3, 4$ but Y_t is assumed continuous and the cut off point is an arbitrary selected number 0.5.

Table 4.7 Table of mean, standard deviation and percentage of bias of estimated log odds ratios, estimated asymptotic standard deviation and coverage rates of 95% confidence intervals from 1,000 simulated datasets with MAR. The number of iterations for SIMEX $B = 500$.

Method	(K, K^*, u_{K^*})	**t**	$\bar{\hat{\beta}}$	$100\frac{\hat{\beta}-\beta}{\beta}$	$sd(\hat{\beta})$	$\bar{\hat{\sigma}}(\hat{\beta}))^2$	$CR^1[2]$
SIMEX-1[2]	(1,5,1.5)	1	0.107	6.853	0.418	0.617	0.993
		2	0.176	5.593	0.458	1.294	0.976
		3	0.240	2.792	0.513	1.574	0.982
		4	0.208	-30.531	2.468	2.095	0.964
SIMEX-2[3]	(1,2,1.2)	1	0.107	6.853	0.418	0.617	0.993
		2	0.176	5.465	0.459	1.294	0.976
		3	0.248	6.265	0.520	1.574	0.982
		4	0.301	0.211	2.546	2.094	0.964

1. CR = $P(\beta \in 95\%CI)$.
2. These two columns are calculated from only 300 datasets and 30 bootstrap iterations for each dataset.

Table 4.8 Table of mean, standard deviation and percentage of bias of estimatimated log odds ratios, estimated asymptotic standard deviation and coverage rates of 95% confidence intervals from $1,000$ simulated datasets with MNAR.

Method	**Cov**	**t**	$\bar{\hat{\beta}}$	$100\frac{\bar{\hat{\beta}}-\beta}{\beta}$	$sd(\hat{\beta})$	$\bar{\hat{\sigma}}(\hat{\beta}))$	CR[1]
GEE-full[2]	Indep	1	0.107	6.853	0.418	0.421	0.955
		2	0.178	7.083	0.410	0.423	0.961
		3	0.248	6.256	0.431	0.426	0.949
		4	0.298	-0.803	0.433	0.430	0.955
GEE-nave[3]	Indep	1	0.107	6.853	0.418	0.421	0.955
		2	0.175	4.793	0.429	0.442	0.962
		3	0.267	14.315	0.480	0.476	0.954
		4	0.357	18.854	0.523	0.516	0.953
	Exch[4]	1	0.036	-63.956	0.503	0.469	0.932
		2	0.089	-46.725	0.487	0.473	0.947
		3	0.150	-35.579	0.487	0.475	0.951
		4	0.208	-30.661	0.498	0.488	0.938
WEE[5]	Indep	1	0.107	6.505	0.418	0.416	0.953
		2	0.179	7.202	0.461	0.463	0.963
		3	0.262	12.480	0.523	0.514	0.958
		4	0.317	5.682	0.598	0.578	0.957
MI-1[6]	Indep	1	0.107	6.853	0.418	0.421	0.955
		2	0.157	-5.914	0.413	0.428	0.968
		3	0.221	-5.145	0.439	0.452	0.977
		4	0.288	-3.930	0.454	0.473	0.971
MI-2[7]	Indep	1	0.107	6.853	0.418	0.421	0.955
		2	0.156	-6.341	0.412	0.425	0.970
		3	0.222	-4.770	0.442	0.447	0.967
		4	0.272	-9.249	0.464	0.472	0.965
MI-3	Indep	1	0.107	6.853	0.418	0.421	0.955
		2	0.153	-8.478	0.409	0.424	0.969
		3	0.214	-8.287	0.422	0.439	0.971
		4	0.267	-11.001	0.426	0.449	0.968

1. CR $= P(\beta \in 95\%CI)$.
2. Use the full dataset. A subject may skip a visit and return in the next scheduled visit.
3. Use the observed dataset with monotone missing pattern.
4. Extreme large ($> e + 10$) estimators from 3 dataset are removed from these summaries.
5. One dataset that does not have estimators of standard deviations is removed.
6. The number of imputation for MI is 3. The imputation model for Y_t includes all other response variables $\{Y_s; s \neq t\}$ for each $t = 2, 3, 4$.
7. The number of imputation for MI is 3. The imputation model for Y_t includes only Y_1.
8. The number of imputation for MI is 3. The imputation model for Y_t includes all other response variables $\{Y_s; s \neq t\}$ for each $t = 2, 3, 4$ but Y_t is assumed continuous and the cut off point is an arbitrary selected number 0.5.

Table 4.9 Table of mean, standard deviation and percentage of bias of estimated log odds ratios, estimated asymptotic standard deviation and coverage rates of 95% confidence intervals from $1,000$ simulated datasets with MNAR. The number of iteration for SIMEX $B = 500$.

Method	(K, K^*, u_{K^*})	**t**	$\bar{\hat{\beta}}$	$100\frac{\hat{\beta}-\beta}{\beta}$	$sd(\hat{\beta})$	$\bar{\hat{\sigma}}(\hat{\beta}))^2$	CR^1[2]
SIMEX^2	(1, 5, 1.5)	1	0.107	6.853	0.418	0.6200.995	
		2	0.172	2.972	0.429	0.9680.966	
		3	0.254	8.909	0.475	1.3430.995	
		4	0.293	-2.372	0.569	1.6230.991	
SIMEX^2	(1, 2, 1.2)	1	0.107	6.853	0.418	0.6200.995	
		2	0.172	3.040	0.429	0.9680.966	
		3	0.254	9.138	0.478	1.3430.995	
		4	0.318	5.931	0.523	1.6220.991	

1. CR $= P(\beta \in 95\%CI)$.
2. These two columns are calculated from only 300 datasets and 30 bootstrap iterations for each dataset.

CHAPTER 5. DISCUSSION

The SIMEX method is a general simulation-based approach that can be used to adjust for bias when the missing data mechanisms are appropriately specified. The procedure includes modeling missing probabilities at recording levels and regularly modeling the response distributions without considering the missing mechanisms. The previous chapters explained the idea of extending SIMEX for measurement error model to missing data problems, proposed an algorithm for finding the SIMEX estimator, and discussed analytical and practical properties of the SIMEX estimator.

The assumptions for the simulation step in the SIMEX method include

- The missingness is MAR or MNAR with additional information, e.g. auxiliary variables or missing pattern group.
- There are enough samples for the simulation step.
- Probabilities of observing each record can be consistently estimated. Missing indicators given observed Y and covariates are assumed independent or can be factorized such that the simulation of $R^{(u|1)}$ is possible.
- Sufficient simulation iterations, B, such that all the estimated $m(u; y, r)$ converged.

The assumptions for the extrapolation step in the SIMEX method include

- The naïve estimator is consistent when the data are fully observed.
- The number of estimated mean of naïve estimators K^* is greater than the order of polynomial extrapolation function K.

- The order K of the extrapolation polynomial $M(u;c)$ is large enough such that $M(u;c)$ approximates $m(u|y,r)$ well.

Those simple assumptions of the SIMEX method made it easy to generally apply to a variety of statistical methods.

The SIMEX method uses a smooth extrapolation function to approximate the conditional mean function $m(u|y,r)$ which is smooth everywhere but at $u=1$. After taking expectation according to (y,r), both lines are smooth and the expectation of extrapolation function is promised to converge to $m(u)$, which is the expectation of $m(u|y,r)$, for $u>0$ when order K goes to infinity.

In general, the extrapolation steps are variable. The extrapolation of an arbitrary smooth function is not promised in general. The expectation of naïve estimators is analytic and can be approximated well by polynomials due to the special form of factorized probability of observing. Although the expectation functions $m(u|y)$ is smooth and analytic, the extrapolation method should still be use with caution. The SIMEX estimator may have larger bias than the naïve estimator if the missingness model is very wrong or the order K is too large so that the end of the extrapolation function becomes too far away from the true extrapolation function.

A major difference between the original measurement error SIMEX and the missing data SIMEX method is that we incorporate structures for the missing data mechanism and structure of the missing model. Another difference is that we define the extrapolation function on real numbers instead on complex numbers. That concept of approximating a smooth function $m(u|y)$ is easier to explain. Additionally, the convergence of Taylor polynomials of $m(u|\tilde{y})$ is promised under our assumptions on missing data mechanism. Therefore, the concept of extrapolation fits the mean function of the missing data analysis.

One benefit of the SIMEX method is the missingness model and the response model can be dealt with separately. Therefore, the response model is kept simple and inference

can be made easier. The missingness model can be a mixture model or a totally different type from the response model. Additionally, auxiliary variables in the missingness model may improve the probability estimation. This benefit is similar to most inverse probability weighted methods. Both the SIMEX and the weighting approaches needs to estimate the missing probability. The SIMEX method is relatively less sensitive to small probabilities of observing ($p(R = 1)$) while the weighting is sensitive to small probabilities since the WEE approach requires the inverses of observing probability as weights.

Another benefit of the SIMEX method is the intuitive and practical simplicity . The concept of the SIMEX method is easy to understand, and the visual presentation of the mean function explains the existence of bias of naïve estimators. The researcher only needs the skill of finding a consistent estimator for the full dataset and the skill of using the least squares method to implement the SIMEX algorithm. The simulation steps that estimate the trend of the bias by increasing the missing probabilities also provide a visual display for the effect of missingness.

Note that the SIMEX method assumes fixed yet unknown missing values and does not need to worry about the distribution of missing values. The full likelihood based methods and other imputation methods usually treat the unobserved portion as random variables. The SIMEX method don't make assumptions on the unobserved portion of data but needs assumptions of a consistently estimated missing model. The imputation methods make assumption on the unobserved data, but make no assumption on the missing model. The SIMEX method takes significantly more calculation time than the multiple imputation method. The SIMEX method may need to find the naïve estimator with simulated missing indicators hundreds or even thousands of times. The multiple imputation methods usually need to analyze the response model less than 20 times.

The simulation steps can have problems when the average of the missing probabilities is large and the sample size is relatively small. When the original probabilities of missing are high, the remaining sample size, $\sum_{i=1}^{n} R^{(u_k|1)}$, will be small and the uncertainty

associated with using extrapolation is also higher. In this situation, other strategies for dealing missing data may provide better results.

The extrapolation steps can have problems when the order of extrapolation polynomial gets larger. The higher order polynomial can fit a smooth curve better but the shape beyond the range of u_k can be unpredictable. We use the residual plot as in Cook and Stefanski (1994) to avoid really wild shapes and to choose an extrapolation that fits the mean function well for $1 \leq u \leq u_{K^*} + 1$. Finding the values of (K, K^*, u_{K^*}) that yields a extrapolation function $M(u; c)$ that approximates $m(u|y, r)$ for $1 \leq u \leq u_{K^*}$ is important for bias reduction.

Other issues include the model selection, diagnose and finding variance estimate. Based on the simulation result, the variance estimator of the SIMEX estimator is higher than its true variance. The portion of the vaiance estimator labeled $(P2a)$ in (2.18) that involves bootstrap methods may be improved by increasing the number of bootstrap iteration.

A. MORE DETAILS OF m FUNCTIONS

A.1 Properties of $m(u|y, r^{(1)}) \equiv \mathbb{E}_{R^{(u)}|\mathcal{P},R^{(1)}}(T(y|R^{(u|1)})), u \geq 1$

Consider the situation that one specific $\{R_i^{(1)}; i = 1, \ldots, n\}$ is observed at $u = 1$. In Chapter 2, the distribution and expectation of the random process at one specific $u > 1$ given observed $R^{(1)}$ has been discussed and used in the simulation step. Here, we discuss the first two derivatives of the function of condition expectaions.

The conditional probability of $R_i^{(u)} = 1$ given $R_i^{(1)}$ is described in Equation 2.6. For each $r^{(u|1)} \in \Omega\left(R^{(u|1)}\right)$, the corresponding observed and missing sets of indexes are

$$\begin{aligned} \mathcal{I}^o(R^{(u|1)}) &= \{i : i = 1, \ldots, n, r_i^{(u|1)} = 1\} \\ &= \{i : i = 1, \ldots, n, r_i^{(1)} = 1 \text{ and } r_i^{(u-1)} = 1\} \end{aligned} \tag{A.1}$$

and

$$\begin{aligned} \mathcal{I}^m(R(u|1)) &= \{i : i = 1, \ldots, n, r_i^{(u|1)} = 0\} \\ &= \{i : i = 1, \ldots, n, r_i^{(1)} = 0 \text{ or } r_i^{(u-1)} = 0\}, \end{aligned} \tag{A.2}$$

respectively.

Let $r^{(1)}$ denote the observed missing indicators, $y^{(o,1)}$ denotes the observed response values, and $y^{(o,u|1)} = \{y_i; i \in \mathcal{I}^o(R^{(u|1)})\}$ denotes the subset of $y^{(o,1)}$ selected by $R^{(u|1)}$. Let $m(u|y, r)$ be the conditional mean of naïve estimator, $E_{R^{(u|1)}|Y,R^{(1)}}T(y, r^{(u|1)})$, for $u > 1$.

Given fixed sets of (y, r) which are particular sample realizations of $(Y, R^{(1)})$, let $\mathcal{I}^o(R^{(1)})$ and $\mathcal{I}^m(R^{(1)})$ denote sets of indexes of the observed and missing records corresponding to $r^{(1)}$. Let $R^{(0)}$ be a length n vector and the value of any elements of $R^{(0)}$ is one.

$R^{(0)}$ is the starting point of the random process $R^{(u)}$. The estimator $T(y|R^{(0)}) = T(y) = \hat{\theta}$ is a consistent estimator of θ. The estimator T converges to θ in probability,

$$\mathbb{E}(T(Y)) = \int_{\Omega(Y)} T(y) f(y|x;\theta) dy \to \theta \text{ as } n \to \infty.$$

We use $\hat{\theta}_{\text{naïve}} = T(y, r) = T(y^{(o,1)})$ to denote the naïve estimator based on only the observed portion $y^{(o)} = \{y_i; i \in \mathcal{I}^o(R^{(1)})\}$ By utilizing only $y^{(o)}$, the conditional expectation becomes

$$\begin{aligned} \mathbb{E}(T(Y|r)) &= \mathbb{E}_{Y^{(o)}|r}(T(y^{(o)}, r)) \\ &= \int_{\Omega(Y^{(o)})} T(y^{(o)}, r) \int_{\Omega(Y^{(m)})} \frac{f(y, r|x, z; \theta, \eta)}{f(r|x, z; \theta, \eta)} dy^{(m)} dy^{(o)}. \end{aligned}$$

The expectation of the naïve estimator depends on r. The marginal expectation of T, $\mathbb{E}_{Y,R}(T(Y, R))$, depends on the coefficients of distribution of R. The dimension of the vector of coefficients for R may be greater than one and the way it influences in the marginal expectation can be complicated. This makes the inspection of relationships between the marginal expectation and the patterns of missingness difficult.

The conditional expectation of those simulated naïve estimators $T(y, R^{(u|1)})$ are

$$\begin{aligned} m(u|y^{(o,1)}, r^{(1)}) &\equiv \mathbb{E}_{R^{(u|1)}|r^{(1)},\mathcal{P}} \left(T(y^{(o,1)}, R^{(u|1)})\right) \\ &= \sum_{r^{(u|1)} \in \Omega(R^{(u|1)})} T(y^{(o,1)}|r^{(u|1)}) f(r^{(u|1)}|r^{(1)}, \mathcal{P}), \end{aligned} \tag{A.3}$$

for $u > 1$. For each $r^{(u|1)} \in \Omega(R^{(u|1)})$, $T(y^{(o,1)}|r^{(u|1)})$ is a consistent for u and the probability of $R^{(u|1)} = r$,

$$f(r|r^{(1)}, \mathcal{P}) = \Pi_{i=1}^{n} \pi^{(u-1),r_i} (1 - \pi^{(u-1)})^{1-r_i}$$

, is a smooth function of u. Since $m(u|y, r)$ is a finite summation of smooth function of u for $u > 1$, $m(u|y, r)$ is a smooth function of u for $u > 1$ given $0 < p_i \leq 1$ for all $i = 1, \ldots, n$.

A.1.1 The first two derivatives

The first derivative of $m(u|y,r)$ for $u>1$ is

$$\begin{aligned}
\frac{d}{du}m(u|y,r) &= \frac{d}{du}\sum_{r^{(u|1)}\in\Omega(R^{(u|1)})} T(y,r^{(u|1)})f(r^{(u|1)}|r^{(1)},\mathcal{P}) \\
&= \sum_{r^{(u|1)}\in\Omega(R^{(u|1)})} T(y,r^{(u|1)})\frac{d}{du}f(r^{(u|1)}|r^{(1)},\mathcal{P}) \\
&= \sum_{r^{(u|1)}\in\Omega(R^{(u|1)})} T(y,r^{(u|1)})\sum_{i\in\mathcal{I}^o(R^{(1)})}\frac{\partial f(r^{(u|1)}|r^{(1)},\mathcal{P}^{(u)})}{\partial p_i^{u-1}}\frac{dp_i^{u-1}}{du} \\
&= \sum_{r^{(u|1)}\in\Omega(R^{(u|1)})} T(y,r^{(u|1)})\sum_{i\in\mathcal{I}^o(R^{(1)})}\frac{\partial f(r^{(u|1)}|r^{(1)},\mathcal{P})}{\partial p_i^{u-1}}p_i^{u-1}\log(p_i).
\end{aligned}$$

By the assumption described in (2.7),

$$\begin{aligned}
\frac{\partial f(r^{(u|1)}|r^{(1)},\mathcal{P})}{\partial p_i^{u-1}} &= \frac{\partial \Pi_{j\in I^{(o,1)}} p_j^{(u-1)r_j^{(u|1)}}(1-p_j^{u-1})^{1-r_j^{(u|1)}}}{\partial p_i^{u-1}} \\
&= \begin{cases} f(r^{(u|1)}|r^{(1)},\mathcal{P})\frac{1}{p_i^{u-1}} & \text{if } r_i^{(u|1)}=1 \\ f(r^{(u|1)}|r^{(1)},\mathcal{P})\frac{1}{1-p_i^{u-1}} & \text{if } r_i^{(u|1)}=0 \end{cases}
\end{aligned}$$

Therefore, the first derivative of $m(u|y,r)$ has the following form

$$\begin{aligned}
\frac{d}{du}m(u|y^{(o,1)},r^{(1)}) &= \sum_{r^{(u|1)}\in\Omega(R^{(u|1)})} T(y,r^{(u|1)})f(r^{(u|1)}|r^{(1)},\mathcal{P}) \\
&\quad\times\left(\sum_{i\in\mathcal{I}^o(R^{(u|1)})\cap\mathcal{I}^o(R^{(1)})}\log(p_i)-\sum_{i\in\mathcal{I}^m(R^{(u|1)})\cap\mathcal{I}^o(R^{(1)})}\frac{p_i^{u-1}}{1-p_i^{u-1}}\log(p_i)\right) \\
&= \sum_{r^{(u|1)}\in\Omega(R^{(u|1)})} T(y,r^{(u|1)})f(r^{(u|1)}|r^{(1)},\mathcal{P}) \\
&\quad\times\left(\sum_{i\in\mathcal{I}^o(R^{(1)})}\log(p_i)-\sum_{i\in\mathcal{I}^m(R^{(u|1)})\cap\mathcal{I}^o(R^{(1)})}\frac{\log(p_i)}{1-p_i^{u-1}}\right),
\end{aligned}$$

which is finite under the assumption that $T(y,r^{(u|1)})$ is finite for all $r^{(u|1)}\in\Omega(R^{(u|1)})$.

The second derivative of $m(u|y^{(o,1)}, r^{(1)})$ for $u > 1$ is

$$\begin{aligned}
&\frac{d^2}{du^2} m(u|y^{(o,1)}, r^{(1)}) \\
= &\sum_{r^{(u|1)} \in \Omega(r^{(u|1)})} T(y^{(o,1)}|r^{(u|1)}) f(r^{(u|1)}|r^{(1)}, \mathcal{P}) \\
&\times \left(\left(\sum_{i \in \mathcal{I}^o(R^{(1)})} \log(p_i) - \sum_{i \in \mathcal{I}^m(R^{(u|1)}) \cap \mathcal{I}^o(R^{(1)})} \frac{\log(p_i)}{1 - p_i^{u-1}} \right)^2 - \sum_{i \in \mathcal{I}^m(R^{(u|1)}) \cap \mathcal{I}^o(R^{(1)})} \frac{p_i^{u-1} \log(p_i)}{(1 - p_i^{u-1})^2} \right)
\end{aligned} \tag{A.4}$$

The second derivative of $m(u|y, r^{(u|1)}$ exists for any $u > 1$.

For each $r^{(u|1)} \in \Omega(R^{(u|1)})$, the element $T(y, r^{(u|1)}) f(r^{(u|1)}|r^{(1)}, \mathcal{P})$ in $m(U|y, r)$ is continuous for $u \geq 1$ and has finite first derivative. The first derivative is also a continuous function of u and its derivative is also a continuous function of u. For each $r^{(u|1)} \in \Omega(R^{(u|1)})$ and for each $i = 1, \ldots, n$, the element $T(y, r^{(u|1)}) f(r^{(u|1)}|r^{(1)}, \mathcal{P}) log(p_i)$ or $T(y, r^{(u|1)}) f(r^{(u|1)}|r^{(1)}, \mathcal{P}) \frac{log(p_i)}{1-p_i^{u-1}}$ in $\frac{d}{du} m(U|y, r)$ is also continuous and has finite first derivative for $u > 1$. The higher order derivatives can be found recursively, and all the derivatives exist. Therefore, the function $m(u|y, r^{(u|1)})$ is smooth for $u > 1$.

The value of $m(u|y, r)$ at $u = 1$ is the naïve estimator that is calculated from the observed dataset $Y^{(o,1)}$. For $u \to 1^+$, $\lim_{u \to 1^+} P(R_i^{u-1} = 1) = 1$, $\forall i \in \mathcal{I}^o(R^{(1)})$. By the assumption described in (2.7), $\lim_{u \to 1^+} P(R_i^{u-1} = 1, \forall i \in \mathcal{I}^o(R^{(1)})) = 1$ and the conditional expectation is

$$\begin{aligned}
\lim_{u \to 1^+} m(u|y, r^{(1)}) &= \lim_{u \to 1^+} \sum_{r \in \Omega(r^{(u|1)})} T(y|r) f(r|r^{(1)}, \mathcal{P}^{(u-1)}) \\
&= T(y^{(o,1)}) \\
&= m(1|y, r^{(1)}).
\end{aligned} \tag{A.5}$$

Therefore, the function $m(u|y, r^{(u|1)}$ is continuous for $u \geq 1$.

Since we are interested in things that happen around $u = 0$, we would like to get information as close to $u = 0$ as possible. Therefore, we further discuss the derivatives

for $u = 1+ = \lim_{\delta \to 0} 1 + \delta$, which yeilds the Taylor expantion at 1+. The first derivative at $u \to 1^+$ is

$$\begin{aligned}
\lim_{u \to 1^+} \frac{d}{du} m(u|y, r^{(1)}) &= \lim_{u \to 1^+} \sum_{r \in \Omega(r^{(u|1)})} T(y|r) f(r|r^{(1)}, \mathcal{P}) \\
&\quad \times \left(\sum_{i \in \mathcal{I}^o(R^{(u|1)}) \cap \mathcal{I}^o(R^{(1)})} \log(p_i) - \sum_{i \in \mathcal{I}^m(R^{(u|1)}) \cap \mathcal{I}^o(R^{(1)})} \frac{p_i^{u-1}}{1 - p_i^{u-1}} \log(p_i) \right) \\
&= T(y|r) \sum_{i \in \mathcal{I}^o(R^{(1)})} \log(p_i) - \sum_{i \in \mathcal{I}^o(R^{(1)})} T(y^{(o,1,-i)}) \log(p_i) \\
&= \sum_{i \in \mathcal{I}^o(R^{(1)})} \log(p_i) \left(T(y^{(o,1)}) - T(y^{(o,1,-i)}) \right) \qquad \text{(A.6)}
\end{aligned}$$

where $y^{(o,1,-i)} \equiv \{y_i^{(o,1)}\}_{i \in I^{(o,1)} \setminus \{i\}}$. The second derivative at $u \to 1^+$ is

$$\begin{aligned}
\lim_{u \to 1^+} \frac{d^2}{du^2} m(u|y^{(o,1)}, r^{(1)}) &= T(y^{(o,1)}) \left(\sum_{i \in \mathcal{I}^o(R^{(1)})} \log(p_i) \right)^2 \\
&\quad + \sum_{i \in \mathcal{I}^o(R^{(1)})} T(y^{(o,1,-i)}) \log(p_i) \left(\log(p_i) - 2 \left(\sum_{i \in \mathcal{I}^o(R^{(o)})} \log(p_i) \right) \right) \\
&\quad + 2 \sum_{i \in \mathcal{I}^o(R^{(1)})} \sum_{j \in \mathcal{I}^o(R^{(1)}) \setminus \{i\}} T(y^{(o,1,-i,-j)}) \log(p_i) \log(p_j) \qquad \text{(A.7)}
\end{aligned}$$

where $y^{(o,1,-i,-j)} \equiv \{y_i^{(o,1)}\}_{i \in I^{(o,1)} \setminus \{i,j\}}$. The first or second order Taylor series expansion at $u = 1+$ can be calculated from A.6 and A.7.

Table A.1 demonstrates calculations of the elements needed for m and $\frac{d}{du}m$ for $n = 3$, $r^{(1)} = (0, 1, 1)$ and let $n_T = 1$ which is the minimum sample size needed for calculating T. Note that $r^{(u|1)} = (0, 0, 0) \notin \Omega(R^{(u|1)})$. If $p_2 = p_3 = 0.7$ and $u = 2$, $f((0, 0, 0)|r^{(1)}, \mathcal{P}) = 0.09$ and $\frac{d}{du} f((0, 0, 0)|r^{(1)}, \mathcal{P}) = 0.15$. When the sample size n is larger, both $P(r^{(u|1)} \notin \Omega(R^{(u|1)})|r^{(1)}, \mathcal{P})$ and $\frac{d}{du} P(r^{(u|1)}) \notin \Omega(R^{(u|1)})|r^{(1)}, \mathcal{P})$ are nearly zero. For example, if $p_i = 0.7$, $n = 10$ and $u = 2$, then $f(\sum_{i=1}^n r_i^{(2|1)} = 0|r^{(1)}, \mathcal{P}) = 6 \times 10^{-6}$ and $\frac{d}{du} f(\sum_{i=1}^n r_i^{(2|1)} = 0|r^{(1)}, \mathcal{P}) = 5 \times 10^{-5}$. Therefore, for the Taylor series polynomial, we simplify the process by using derivatives of $f(r^{(u|1)}|r^{(1)}, \mathcal{P})$ instead of $f(r^{(u|1)}|r^{(1)}, \mathcal{P}, \sum_{i=1}^n r_i^{(u|1)} \geq n_T)$ where n_T is the minimum number of records such that the naïve estimator exists.

Table A.1 Example of finding derivatives of $f(r^{(u|1)}|\mathcal{P}, r^{(1)})$.

$r^{(u\|1)}$	$y^{(o,u\|1)}$	$f(r^{(u\|1)}\|r^{(1)},\mathcal{P})$	$\frac{d}{du}f(r^{(u\|1)}\|r^{(1)},\mathcal{P})$
$(0,1,1)$	y_2, y_3	$p_2^{u-1}p_3^{u-1}$	$p_2^{u-1}p_3^{u-1}log(p_2) + p_2^{u-1}p_3^{u-1}log(p_3)$
$(0,1,0)$	y_2	$p_2^{u-1}(1-p_3^{u-1})$	$p_2^{u-1}(1-p_3^{u-1})log(p_2) - p_2^{u-1}p_3^{u-1}log(p_3)$
$(0,0,1)$	y_3	$(1-p_2^{u-1})p_3^{u-1}$	$-p_2^{u-1}p_3^{u-1}log(p_2) + (1-p_2^{u-1})p_3^{u-1}log(p_3)$
$(0,0,0)$	-	$(1-p_2^{u-1})(1-p_3^{u-1})$	$-p_2^{u-1}(1-p_3^{u-1})log(p_2) - (1-p_2^{u-1})p_3^{u-1}log(p_3)$

A.2 Properties of $m(u|y, r^{(1)})$ for $0 \leq u < 1$

This section discuss the properties of $m(u|y, r^{(1)})$ for $0 < u < 1$. Since the expectation of $m(u|y, r^{(u)})$ with respect to distribution of $(Y, R^{(1)})$ is the marginal mean function $m(u)$ for $0 < u < 1$, the properties discussed here complete the description of $m(u|y, r^{(u)})$.

Note that different from the full likelihood based methods and the multiple imputation method, the SIMEX method considers $y^{(m)}$ as fixed. The random process $R^{(u)}$ is conducted on a fixed $y = (y^{(o)}, y^{(m)})$. Although the SIMEX method never requires the values of $y^{(m)}$, the values of $y^{(m)}$ are still considerd fixed. Therefore, the values of $y^{(m)}$ are assumed fixed in this section.

Consider the situation that one specific $\{R_i^{(1)}; i = 1, \ldots, n\}$ is observed at $u = 1$. In Chapter 2, the distribution and expectation of the random process at one specific $u > 1$ given observed $R^{(1)}$ has been discussed and used in the simulation step. Here, we show the distribution and expectation of the random process at one specific $u \in (0, 1)$ given observed $R^{(1)}$ and also the derivatives of $m(u|y, r)$.

Let $u \in (0, 1)$ and $u^* = 1 - u$. Let $R_i^{(u)}$ and $R_i^{(u^*)}$ be two independent random variables defined in Equation (2.3). The random variable $R_i^{(1)} \equiv R_i^{(u)} \times R_i^{(u*)} = 1$ if both $R_i^{(u)} = 1$ and $R_i^{(u^*)} = 1$.

Let $R_i^{(u|1)}$ represent the conditional random varaible $R^{(u)}$ given $R^{(1)}$. The conditional distribution of $R_i^{(u)} = 1$ given $R_i^{(1)}$

$$\pi_i^{(u|1)} \equiv P(R_i^{(u)} = 1|R_i^{(1)}, \mathcal{P}) = \begin{cases} \dfrac{p_i^u(1-p_i^{u^*})}{1-p_i^1} & \text{if } R_i^{(1)} = 0, \\ 1 & \text{if } R_i^{(1)} = 1. \end{cases} \tag{A.8}$$

Note that for simpilicity, the situation of a monotone missing pattern is not considered in this section. When the missing pattern is monotone, the condistional distribution described in A.8 should be altered to match the missing pattern. For example, let $R^{(1)} = (R_1^{(1)}, R_2^{(1)}, R_3^{(1)}, R_4^{(1)}) = (1,0,0,0)$ are missing indicators of the same subject at time 1 to 4, two examples of the conditional probabilities given $R^{(1)}$ are $P(R^{(u|1)} = (1,1,0,1)) = 0$ and $P(R^{(u|1)} = (1,1,0,0)) = \pi_2^{(u|1)}(1 - \pi_3^{(u|1)})$.

For each $r^{(u|1)} \in \Omega\left(R^{(u|1)}\right)$, the corresponding observed and missing sets of indexes are

$$\begin{aligned}
\mathcal{I}^o(R^{(u|1)}) &= \{i : i = 1, \ldots, n, R_i^{(u|1)} = 1\} \\
&= \{i : i = 1, \ldots, n, R_i^{(1)} = 1 \text{ and } R_i^{(u-1)} = 1\}, \\
\mathcal{I}^m(R^{(u|1)}) &= \{i : i = 1, \ldots, n, R_i^{(u|1)} = 0\} \\
&= \{i : i = 1, \ldots, n, R_i^{(1)} = 0 \text{ or } R_i^{(u-1)} = 0\}.
\end{aligned}$$

Assume that the values of $\mathcal{P}$ are fixed or consistently estimated without involveing unobserved $Y^{(m,1)}$. The conditional expectation of $T(Y|R^{(u)})$ for $0 \leq u < 1$ is

$$\begin{aligned}
m(u|y, r^{(1)}) &\equiv \mathbb{E}_{R^{(u|1)}|r^{(1)},y}(T(Y|r^{(u|1)})) \\
&= \sum_{r^{(u|1} \in \Omega(R^{(u|1)})} T(y^{(o,u|1)}) f(r|r^{(1)}, \mathcal{P}).
\end{aligned}$$

The conditional expectation at $u \to 1^-$ is the naïve estimator $m(1|y, r) = T(y^{(o,1)})$. The conditional expectation at $u = 0$ is $m(0|y, r) = T(y)$.

Let $q_i^{(u)} = \frac{p_i^u - p_i}{1 - p_i}$. By the equation (A.8),

$$f(r^{(o,u|1)}|r^{(1)}, \mathcal{P}) = \left(\prod_{i \in \mathcal{I}^m(R^{(1)}) \cap \mathcal{I}^o(R^{(u|1)})} q_i^{(u)} \right) \left(\prod_{i \in \mathcal{I}^m(R^{(1)}) \cap \mathcal{I}^o(R^{(u|1)})} (1 - q_i^{(u)}) \right)$$

The first derivative of $m(u|y, r^{(1)})$ for $u < 1$ is

$$\begin{aligned}\frac{d}{du}m(u|y,r^{(1)}) &= \sum_{\Omega(R^{(u|1)})} T(y^{(o,u|1)}) \sum_{i\in\mathcal{I}^m(R^{(1)})} \frac{\partial}{\partial q_i^{(u)}} f(r^{(u|1)}|r^{(1)}, P^{(u)}) \frac{dq_i^{(u)}}{du} \\ &= \sum_{r\in\Omega(R^{(u|1)})} T(y^{(o,u|1)}) f(r|r^{(1)}, \mathcal{P}) \\ &\quad \times \left(\sum_{i\in\mathcal{I}^m(R^{(1)})\cap\mathcal{I}^o(R^{(u|1)})} \frac{p_i^u \log(p_i)}{p_i^u - p_i} - \sum_{i\in\mathcal{I}^m(R^{(1)})\cap\mathcal{I}^o(R^{(u|1)})} \frac{p_i^u \log(p_i)}{1-p_i^u} \right). \end{aligned} \tag{A.9}$$

When $u \to 1^-$, $q_i^{(u)} \to 0$. The first derivative of $m(u|y^{(o,1)}, r^{(1)})$ for $u \to 1^-$ is

$$\begin{aligned}\lim_{u\to 1^-} \frac{d}{du}m(u|y,r^{(1)}) &= \left(T(y^{(o,1)}) \sum_{i\in\mathcal{I}^m(R^{(1)})} \frac{-p_i log(p_i)}{1-p_i} + \sum_{i\in\mathcal{I}^m(R^{(1)})} T(\{y^{(o,1)}, y_i\}) \frac{p_i log(p_i)}{1-p_i} \right) \\ &= \sum_{i\in\mathcal{I}^m(R^{(1)})} \frac{p_i log(p_i)}{1-p_i} (T(y^{(o,1)}, Y_i) - T(y^{(o,1)})). \end{aligned} \tag{A.10}$$

Then, we look at the difference d_m between differences, $\frac{d}{du}m(1^+|y^{(o,1)}, r^{(1)})$ in (A.6) and $\frac{d}{du}m(1^-|y^{(o,1)}, r^{(1)})$ in (A.10),

$$\begin{aligned} d_m &= \sum_{i\in\mathcal{I}^m(R^{(1)})} \frac{p_i \log(p_i)}{1-p_i} \left(T(y^{(o,1)}, y_i) - T(y^{(o,1)}) \right) \\ &\quad - \sum_{i\in\mathcal{I}^o(R^{(1)})} \log(p_i) \left(T(y^{(o,1)}) - T(y^{(o,1,-i)}) \right). \end{aligned} \tag{A.11}$$

In general, the value of d_m is not necessarily zero. The first derivative of $m(u|y, r^{(1)})$ at $u = 1$ usually does not exist. The function $m(u|y, r^{(1)})$ is continuous for $u > 0$. The function $m(u|y, r^{(1)})$ is smooth for $u > 1$ and $0 < u < 1$ but not smooth at $u = 1$.

A.3 Properties of $m(u|y) \equiv \mathbb{E}_{R^{(u)}|\mathcal{P}}(T(y|R^{(u)})), u \geq 0$

The conditional expectation $m(u|y, R)$ that integrates out the randomness from the simulation step has been discussed in Appendix A.1 and A.2. This section contains discussions about the conditional expectation that integrates out the randomness from

the missing data mechanism. Define the mean function $m(u|y)$ as

$$\begin{aligned} m(u|y) &\equiv \mathbb{E}_{R^{(1)}|Y}(m(u|Y, R^{(1)})) \\ &= \mathbb{E}_{R^{(1)}|Y}(\mathbb{E}_{\{(R^{(u_k|1)}\}|R^{(1)},Y)}(T(y^{(o,u|1)}))) \\ &= \mathbb{E}_{R^{(1)},\{R^{(u_k|1)}\}|Y}(T(y^{(o,u|1)})) \\ &= \mathbb{E}_{R^{(u)}|Y}(T(y^{(o,u)})). \end{aligned}$$

Although the function $m(u|y, R)$ is not smooth at $u = 1$, we can show that the function $m(u|y) = E(m(u|y, r))$ is a smooth function for $u > 0$ by showing that its derivatives exist for $u > 0$. Therefore, the expectation of the smooth extrapolation function (with respect to $f_{R^{(1)}|Y}$) can approximate the expectation function $m(u|y)$ well for $u > 0$. Then, when taking expectation on both smooth functions with respect to f_Y, both functions are smooth and analytic (Appendix A.4), and the marginal expectation of the SIMEX estimator converges to the expectation of $T(Y)$ when the order of extrapolation function goes to infinity.

The first derivative of $m(u|y)$ is

$$\begin{aligned} \frac{d}{du}m(u|y) &= \frac{d}{du}\sum_{r^{(u)}\in\Omega(R^{(u)})} T(y^{(o,u)})f(r^{(u)}|\mathcal{P}) \\ &= \sum_{r^{(u)}\in\Omega(R^{(u)})} T(y^{(o,u)})\frac{d}{du}f(r^{(u)}|\mathcal{P}) \\ &= \sum_{r^{(u)}\in\Omega(R^{(u)})} T(y^{(o,u)})\sum_{i=1}^{n}\frac{\partial f(r^{(u)}|\mathcal{P})}{\partial p_i^u}\frac{dp_i^u}{du} \\ &= \sum_{r^{(u)}\in\Omega(R^{(u)})} T(y^{(o,u)})\sum_{i=1}^{n}\frac{\partial f(r^{(u)}|\mathcal{P})}{\partial p_i^u}p_i^u\log(p_i) \end{aligned}$$

By the assumption described in (2.7) and the chain rule,

$$\begin{aligned} \frac{\partial f(r^{(u)}|\mathcal{P})}{\partial p_i^u} &= \frac{\partial \Pi_{j=1}^n p_j^{ur_j^{(u)}}(1-p_j^u)^{1-r_j^{(u)}}}{\partial p_i^u} \\ &= \begin{cases} f(r^{(u)}|\mathcal{P}^u)\frac{1}{p^u} & \text{if } r_i^{(u)} = 1 \\ f(r^{(u)}|\mathcal{P}^u)\frac{1}{1-p^u} & \text{if } r_i^{(u)} = 0 \end{cases} \end{aligned} \tag{A.12}$$

So the first derivative of $m(u|y)$ is

$$\frac{d}{du}m(u|y) = \sum_{r^{(u)}\in\Omega(R^{(u)})} T(y|r^{(u)})f(r^{(u)}|\mathcal{P})\left(\sum_{i\in\mathcal{I}^o(R^{(u)})}\log(p_i) - \sum_{i\in\mathcal{I}^o(R^{(u)})}\frac{p_i^u}{1-p_i^u}\log(p_i)\right)$$

Since we have assumed that $p_i > 0$ for all $i = 1,\ldots,n$, the first derivative of $m(u|y)$ exists for $u > 0$. The term $1 - p_i^u$ can be canceled since there is the same term in $f(r|\mathcal{P}^{(u)})$, so it won't cause a problem when $u \to 0$. For easier calculation, we add $\sum_{i\in\mathcal{I}^m(R^{(u)})}\log(p_i) - \sum_{i\in\mathcal{I}^m(R^{(u)})}\log(p_i)$ to the parentheses and rewrite the first derivative as

$$\frac{d}{du}m(u|y) = \sum_{r\in\Omega(R^{(u)})} T(y|r)f(r|\mathcal{P}^{(u)})\left(\sum_{i=1}^{n}\log(p_i) - \sum_{i\in\mathcal{I}^m(R^{(u)})}\frac{\log(p_i)}{1-p_i^u}\right).$$

The second derivative of $m(u|y)$ is

$$\begin{aligned}\frac{d^2}{du^2}m(u|y) &= \sum_{r\in\Omega(R^{(u)})} T(y|r)f(r|\mathcal{P}^{(u)}) \\ &\times \left(\left(\sum_{i=1}^{n}\log(p_i) - \sum_{i\in\mathcal{I}^m(R^{(u)})}\frac{\log(p_i)}{1-p_i^u}\right)^2 - \sum_{i\in\mathcal{I}^m(R^{(u)})}\frac{d}{du}\frac{\log(p_i)}{1-p_i^u}\right) \\ &= \sum_{r\in\Omega(R^{(u)})} T(y|r)f(r|\mathcal{P}^{(u)}) \\ &\times \left(\left(\sum_{i=1}^{n}\log(p_i) - \sum_{i\in\mathcal{I}^m(R^{(u)})}\frac{\log(p_i)}{1-p_i^u}\right)^2 - \sum_{i\in\mathcal{I}^m(R^{(u)})}\frac{p_i^u(\log(p_i))^2}{(1-p_i^u)^2}\right).\end{aligned}$$

Similar to previous arguments in Appendix A, the higher order derivatives of $m(u|y)$ can be calculated recursively for each element in $\frac{d}{du}m(u|y)$ and all derivatives exist. Therefore, the function $m(u|y)$ is smooth and continuous for $u > 1$.

Here we further look at the first two derivatives of $m(u|y)$ when $u \to 0$. When $u \to 0^+$, the first derivative is

$$\begin{aligned}\lim_{u\to 0^+}\frac{d}{du}m(u|y) &= T(y)\sum_{i=1}^{n}\log(p_i) - \sum_{i=1}^{n}T(y^{(-i)})\log(p_i) \\ &= \sum_{i=1}^{n}\big(T(y) - T(y^{(-i)})\big)\log(p_i) \qquad \text{(A.13)}\end{aligned}$$

where $y^{(-i)} \equiv \{y_i\}_{i\in\{1,\ldots,n\}\setminus\{i\}}$. The second derivative at $u \to 0^+$ is

$$\begin{aligned}\lim_{u\to 0^+} \frac{d^2}{du^2} m(u|y) &= T(y)\left(\sum_{i=1}^{n}\log(p_i)\right)^2 \qquad (A.14)\\ &\quad + \sum_{i=1}^{n} T(y^{(-i,-j)})\log(p_i)\left(\log(p_i) - 2\sum_{j=1}^{n}\log(p_j)\right)\\ &\quad + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} T(y^{(-i,-j)})\log(p_i)\log(p_j) \qquad (A.15)\end{aligned}$$

where $y^{(-i,-j)} \equiv \{y_i\}_{i\in\{1,\ldots,n\}\setminus\{i,j\}}$.

A.4 Properties of $m(u) \equiv \mathbb{E}_{Y,R^{(u)}}(T(Y, R^{(u)}))$

The conditional mean of the naïve estimator given one specific set of observed responses y has been discussed in section A.3 This section discusses the derivatives of the marginal expectation

$$\begin{aligned} m(u) &\equiv E_Y(m(u|y))\\ &= E_Y E_{R^{(u)}|Y}(T(Y, R^{(u)}))\\ &= E_{Y,R^{(u)}}(T(Y, R^{(u)})),\end{aligned}$$

which is the marginal expectation of the naïve estimators for $u > 0$. We need to show that the function $m(u)$ is also a smooth function for $u > 0$, Then, $m(u)$ is equal to the marginal expectation of the extrapolation function when the order K of the extrapolation functio goes to infinity.

The jth derivative of $m(u)$ is

$$\frac{d^j}{du^j} m(u) = \frac{d^j}{du^j} E_Y\left(m(u|y)\right)$$

Assume that the interchange of integral and derivative yields the same result, then

$$\frac{d^j}{du^j} m(u) = E_Y\left(\frac{d^j}{du^j} m(u|y)\right)$$

By (A.12), the first derivative of $m(u)$ is

$$\begin{aligned}
&\frac{d}{du}m(u) \\
=& \sum_{r\in\Omega(R^{(u)})}\int_Y \left[T(y,r)f(r|\mathcal{P}^{(u)})\left(\sum_{i\in\mathcal{I}^o(R^{(u)})}\log(p_i) - \sum_{i\in\mathcal{I}^m(R^{(u)})}\frac{p_i^u}{1-p_i^u}\log(p_i)\right)\right] f(y)dy \\
=& \sum_{r\in\Omega(R^{(u)})}\sum_{i=1}^{n}\int_Y \left[T(y,r)f(r|\mathcal{P}^{(u)})\log(p_i)\right. \\
&\times\left.\left[\left(I(i\in\mathcal{I}^o(R^{(u)})) - I(i\in\mathcal{I}^m(R^{(u)}))\frac{p_i^u}{1-p_i^u}\right)\right] f(y)dy. \\
=& \sum_{r\in\Omega(R^{(u)})}\sum_{i=1}^{n}\int_Y \left[T(y,r)f(r|\mathcal{P}^{(u)})\log(p_i)\left(n - I(i\in\mathcal{I}^m(R^{(u)}))\frac{1}{1-p_i^u}\right)\right] f(y)dy.
\end{aligned} \tag{A.16}$$

The derivative of $m(u)$ in (A.16) does not directly shows the smoothness of $m(u)$. We will show the smoothness of $m(u)$ by the fact that the expectation of the extrapolation function $N(u) = E_{Y^{(1)}|Y}(M(u))$ converges uniformly to $m(u|y)$ as the order K of the polynomial $M(u)$ goes to infinity. Figure 2.2 shows the relationship between $M(u)$, $N(u)$ and $N^E(u)$.

For simplicity, we have used simplified notation $N^E(u) = E_Y(N(u))$ for a given order K and response Y. Here, we use $N^E(u;K)$ and $N(u;K,Y)$ to denote the values of the Kth order polynomial that approximate $m(u|Y)$ and $m(u)$, respectively, for a given Y. The function $N(u;K,Y)$ is a polynomial that converges uniformly to $m(u|y)$ when $K\to\infty$ in a finite range of u, say $u\in[0,u_{K^*}]$. Therefore, the integral and limit can be swiched,

$$\begin{aligned}
\lim_{K\to\infty} N^E(u;K) &= \lim_{K\to\infty}\int_y N(u;K,y)f(y)dy \\
&= \int_y \lim_{K\to\infty} N(u;K,y)f(y)dy \\
&= \int_y m(u|y)f(y)dy \\
&= m(u),
\end{aligned}$$

for $u \in [0, u_{K^*}]$. The polynomial $N^E(u; K)$ converges to $m(u)$ when $K \to \infty$. The function $m(u)$ is continuous for $u \in [0, u_{K^*}]$. Again, the polynomial $N^E(u; K, Y)$ uniformly converges to $m(u)$, therefore, the derivative of $N^E(u; K, Y)$ converges to derivative of $m(u)$. The function $m(u)$ is smooth and continuous for $u \in [0, u_{K^*}]$ ((Clark, 2009)).

B. SUPPORTING MATERIALS FOR CHAPTER 4

B.1 Weighted generalized estimating equations (WGEE)

Robins et al. (1995) proposed the approach known as WGEE that results in consistent estimators when the missing model is correctly specified. However, it performs worse (in terms of bias) than GEE if the missing model is wrong.

The weight w_{i,t_j} for the ith subject at time t_j, $j = 1, ..., J$ is given by

$$\hat{w}_{i,t_j} = (\hat{p}_{i,t_j})^{-1} = P(R_{i,t_j} = 1|Y_i^{(obs)}, z_{i,\leq t_j}, \hat{\phi})^{-1}$$

$$W_i = diag(R_{i,t_1}\hat{w}_{i,t_1}, ..., R_{i,t_J}\hat{w}_{i,t_J}).$$

If the missing pattern is monotonic, the weight is

$$\hat{w}_{i,t_j} = (1 \times \hat{p}_{i,2} \times ... \times \hat{p}_{i,t_j})^{-1} = P(R_{i,t_j} = 1|Y_{i,\leq t_j}, z_{i,\leq t_j}, \hat{\phi})^{-1}.$$

The ordinary GEE estimator for θ is obtained by solving the set of estimating equations

$$\sum_{i=1}^{n} D_i^{(obs)}(X^{(obs)}, \theta)(V^{(obs)})^{-1}[Y_i^{(obs)} - E(Y_i^{(obs)})] = 0$$

where $D_i^{(obs)}$ is the first derivative matrix of a mean model with respect to parameters and $V_i^{(obs)}$ is the variance structure matrix of the mean model. The WGEE estimator for θ on the other hand is obtained by solving a set of similar estimating equations but that now include the weights:

$$\sum_{i=1}^{n} D_i^{(obs)}(X^{(obs)}, \theta)(V^{(obs)})^{-1}W_i[Y_i^{(obs)} - E(Y_i^{(obs)})] = 0.$$

B.2 Multiple Imputation (MI)

Rubin (2004) discussed the Bayesian theory underpinning the method of multiple imputation. Here, missing observations are estimated using samples from the posterior distribution of the missing data and yield an "approximately valid Bayesian inference". In the context of MI, we write

$$\begin{aligned}
\mathbb{E}_{Y|R}(T|R) &= \int_{\Omega(Y)} T f(Y|R,x,z;\theta,\phi)dY \\
&= \int_{\Omega(Y)} T f(Y^{(mis)}|Y^{(obs)},R,x,z;\theta,\phi) f(Y^{(obs)}|R,x,z;\theta,\phi)dY \\
&= \mathbb{E}_{Y^{(obs)}|R}(\mathbb{E}_{Y^{(mis)}|Y^{(obs)},R}(T|Y^{(obs)},R)).
\end{aligned}$$

To find the conditional expectation with respect to $f(Y^{(mis)}|Y^{(obs)},R,x,z;\theta,\phi)$, we need to sample from the conditional distribution

$$\begin{aligned}
f(Y^{(mis)}|Y^{(obs)},R,x,z;\theta,\phi) &= \frac{f(Y^{(mis)},R|Y^{(obs)},x,z;\theta,\phi)}{f(R|Y^{(obs)},x,z;\theta,\phi)} \\
&= \frac{f(Y^{(mis)}|Y^{(obs)},x,z;\theta,\phi) f(R|Y^{(mis)},Y^{(obs)},z;\phi)}{f(R|Y^{(obs)},x,z;\theta,\phi)}.
\end{aligned}$$

If missing is ignorable, then are expressions greatly simplify so that

$$\begin{aligned}
f(Y^{(mis)}|Y^{(obs)},R,x,z;\theta,\phi) &= \frac{f(Y^{(mis)}|Y^{(obs)},x,z;\theta,\phi) f(R|Y^{(obs)},z;\phi)}{f(R|Y^{(obs)},z;\phi)} \\
&= f(Y^{(mis)}|Y^{(obs)},x;\theta).
\end{aligned}$$

This says that assumptions about the structure (distribution, parameters) of the missing data need to be made.

B.3 List of explanatory variables in the missing model in Preisser et al.(2000)

- $Z_{race,i} = 1$ if the race of the ith individual is black and $X_{race} = 0$ otherwise,
- $Z_{sex,i} = 1$ if the gender is male and zero otherwise,

- $Z_{group,i} = 1$ if the ith individual is black male,
- $Z_{age,i} = age/10$ is the age at 1986 in 10 year units,
- $Z_{edu1,i} = 1$ if the education level is high school or less,
- $Z_{edu2,i} = 1$ if the educations level is some college,
- $Z_{year5,it} = 1$ if $X_{it} = 5$,
- $Z_{year7,it} = 1$ if $X_{it} = 7$,
- $Z_{lagy,it} = Y_{i(t-1)}$,
- $Z_{race,i} \times Z_{year,it}$,
- $(Z_{race,i} + Z_{edu1,i} + Z_{edu2,i} + Z_{sex,i} + Z_{age,i} + Z_{year,it}) \times Z_{lagy,it}$.

BIBLIOGRAPHY

Agresti, A. (1990), *Categorical Data Analysis*, Wiley-Interscience.

Athreya, K. B. and Lahiri, S. N. (2006), *Measure Theory and Probability Theory*, Springer.

Baker, S. G., Ko, C.-W., and Graubard, B. I. (2003), "A sensitivity analysis for non-randomly missing categorical data arising from a national health disability survey," *Biostatistics*, 4, 41–56.

Cantoni, E., Field, C., Flemming, J. M., and Ronchetti, E. (2007), "Longitudinal variable selection by cross-validation in the case of many covariates," *Statistics in Medicine*, 26, 919–930.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006), *Measurement Error in Nonlinear Models: A Modern Perspective*, Chapman and Hall/CRC.

Carroll, R. J. and Stefanski, L. A. (1997), "Asymptotic Theory for the Simex Estimator in Measurement Error Models," in *Advances in Statistical Decision Theory and Applications*, Birkhäuser Boston, chap. 10, pp. 151–164.

Clark, P. L. (2009), "Sequences and series of functions II: power series," .

Cook, J. R. and Stefanski, L. A. (1994), "Simulation-extrapolation estimation in parametric measurement error models," *Journal of the American Statistical Association*, 89, 1314–1328.

Dahlquist, G. (1974), *Numerical Methods*, Dover Publications, pp. 101–103.

Daniels, M. J. and Hogan, J. W. (2008), *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*, Chapman & Hall.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood estimation from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Series B*, 39, 1–38.

Gelman, A., Hill, J., Su, Y.-S., Yajima, M., and Pittau, M. G. (2012), *R: Missing-data Imputation and Model Checking*.

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007), "How many imputations are really needed? Some practical clarifications of multiple imputation theory," *Prevention Science*, 8, 206–213.

He, W., Yi, G. Y., and Xiong, J. (2007), "Accelerated failure time models with covariates subject to measurement error," *Statistics in medicine*, 26, 4817–4832.

Hedeker, D. and Gibbons, R. D. (1997), "Application of random-effect pattern-mixture models for missing data in longitudinal studies," *Psychoiogical methods*, 2, 64–78.

Hughes, G. H., Cutter, G., Donahue, R., Friedman, G. D., Hulley, S., Hunkeler, E., Jacobs, D. R., Liu, K., Orden, S., Pirie, P., Tucker, B., and Wagenknecht, L. (1987), "Recruitment in the Coronary Artery Risk Development in Young Adults (CARDIA) Study," *Controlled Clin1ical Trials*, 8, 68S–73S.

Kenward, M. G., Lesaffre, E., and Molenberghs, G. (1995), "An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random," *Biometrics*, 50, 945–953.

Küchenhoff, H., Lederer, W., and Lesaffre, E. (2007), "Asymptotic variance estimation for the misclassification SIMEX," *Computational Statistics & Data Analysis*, 51, 6197–6211.

Küchenhoff, H., Mwalili, S. M., and Lesaffre, E. (2006), "A general method for dealing with misclassification in regression: The misclassification," *Biometrics*, 62, 85–96.

Lederer, W. and Kchenhoff, H. (2009), *simex: SIMEX- and MCSIMEX-Algorithm for measurement error models*, r package version 1.4.

Lederer, W. and Küchenhoff, H. (2006), "A short introduction to the SIMEX and MC-SIMEX," *R news*, 6, 26–31.

Li, X., Song, X., and Gray, R. H. (2002), "Comparison of the missing-indicator method and conditional logistic regression in 1:m Matched Case-Control Studies with Missing Exposure Values," *American Journal of Epidemiology*, 159, 603–610.

Liang, K. and Zeger, S. (1986), "Longitudinal data analysis using generalized linear models," *Biometrika*, 73, 13–22.

Lin, X. L. and Carroll, R. J. (1999), "SIMEX variance component tests in generalized linear mixed measurement error models," *Biometrics*, 55, 613–619.

Lipsitz, S. R., Molenberghs, G., Fitzmaurice, G. M., and Ibrahim, J. (2000), "GEE with gaussian estimation of the correlations when data are incomplete," *Biometrics*, 56, 528–536.

Little, R. (1988), "A test of missing completely at random for multivariate data with missing values," *Journal of the American Statistical Association*, 83, 1198–1202.

— (1993), "Pattern-mixture models for multivariate incomplete data," *Journal of the American Statistical Association*, 88, 125–134.

— (1995), "Modeling the dropout mechanism in repeated-measures studies," *Journal of the American Statistical Association*, 90, 1112–1121.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data, Second Edition*, Wiley-Interscience.

Molenberghs, G. and Kenward, M. G. (2007), *Missing Data in Clinical Studies*, John Wiley & Sons.

Paik, M. C. (1997), "The Generalized Estimating Equations Approach When Data Are Not Missing Completely at Random," *Journal of the American Statistical Association*, 92, 1320–1329.

Park, T. and Davis, C. S. (1993), "A test of missing data mechanism for repeated categorical data," *Biometrics*, 49, 631–638.

Park, T. and Lee, S.-Y. (1997), "A test of missing completely at random for longitudinal data with mising observations," *Statistics in medicine*, 16, 1859–1871.

Preisser, J. S., Galecki, A. T., Lohman, K. K., and Wagenknecht, L. E. (2000), "Analysis of smoking trends with incomplete longitudinal binary responses," *Journal of the American Statistical Association*, 95, 1021–1031.

Preisser, J. S., Lohman, K. K., and Rathouz, P. J. (2002), "Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random," *Statistics in Medicine*, 21, 3035–3054.

Qaqish, B. F. (2003), "A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations," *Biometrika*, 90, 455–463.

Qu, Y. and Lipkovich, I. (2009), "Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach," *Statistics in Medicine*, 28, 1402–1414.

Robbins, M. W. and White, T. K. (2011), "Farm commodity payments and imputation in the agricultural reource management survey," *American Journal of agricultural economics*, 93, 606–612.

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

Rubin, D. (1976), "Inference with missing data," *Biometrika*, 63, 581–592.

— (2004), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons, Inc.

Slate, E. H. and Bandyopadhyay, D. (2009), "An investigation of the MC-SIMEX method with application to measurement error in periodontal outcomes," *Statistics in medicine*, 28, 3523–3538.

Staudenmayer, J. and Ruppert, D. (2004), "Local polynomial regression and simulation-extrapolation," *Journal of the Royal Statistical Society*, 66, 17–30.

Stefanski, L. A. and Bay, J. M. (1996), "Simulation extrapolation deconvolution of finite population cumlative distribution function estimators," *Biometrika*, 83, 407–417.

Stefanski, L. A. and Cook, J. R. (1995), "Simulation-extrapolation: the measurement error jacknife," *Journal of the American Statistical Association*, 90, 1247–1256.

Subramanian, S. (2009), "The mutiple imputations based Kaplan-Meier estimator," *Statistics and Probability Letters*, 79, 1906–1914.

Touloumi, G., Babiker, A. G., Pocock, S. J., and Darbyshire, J. H. (2001), "Impact of missing data due to drop-outs on estimators for rates of change in longitudinal studies: a simulation study," *Statistics in Medicine*, 21, 3035–3054.

Wang, N., Lin, X., Gutierrez, R. G., and Carroll, R. J. (1998), "Bias analysis and SIMEX approach in generalized linear mixed measurement error models," *Journal of the American Statistical Association*, 93, 249–261.

Wu, C.-F. (1981), "Asymptotic Theory of Nonlinear Least Squares Estimation," *The Annals of Statistics*, 9, 501–513.

Xie, F. and Paik, M. C. (1997a), "Generalized estimating equation model for binary outcomes with missing covariates," *Biometrics*, 53, 1458–1466.

— (1997b), "Multiple imputation methods for the missing covariates in generalized estimating equation," *Biometrics*, 53, 1538–1546.

Xiong, J., He, W., and Yi, G. Y. (2010), *simexaft: simexaft*, r package version 1.0.3.

Yan, J., Hjsgaard, S., and Halekoh, U. (2012), *R: Generalized Estimating Equation Package*.