Variance Estimation after Multiple Imputation

Jiwei Zhao[1,2]
[1]Department of Biostatistics, Yale University School of Public Health, New Haven, CT 06511
[2]Department of Statistics and Actuarial Science, University of Waterloo, Ontario, Canada

Abstract

In complex survey data, we usually encounter appreciable amount of missing values. The missing data mechanism can be missing completely at random (MCAR), missing at random (MAR), or nonignorable missingness. Multiple Imputation (MI, Rubin 1987) is a well-known and well-established procedure to handle missing values and it is an important technique in the literature. In this paper, we consider the regression model with response variable subject to missing values and we concentrate on the variance estimates before and after MI.

At first, we briefly review the results when the missing data mechanism is MAR (Wang and Robins, 1998). Under MAR assumption, the estimates after MI are always less efficient than the ones before MI.

In the following, we focus on the situation when the missing data mechanism is nonignorable. We first propose an estimation procedure before MI under some assumptions on the missing data mechanism. We then conduct MI based on the proposed estimates. However, different from MAR, the variance after MI is not generally necessarily larger than the one before MI. It is possible that MI could increase the estimation efficiency when the missing data mechanism is nonignorable. This is a different phenomenon comparing MAR and nonignorable missing data mechanisms. Finally, intensive simulation studies are conducted to illustrate the finite sample behaviors.

Keywords: Missing data mechanism; Missing at random; Nonignorable missingness; Multiple imputation; Variance estimation.

Introduction

Background

- Appreciable amount of missing values are encountered in real applications
- For example: Korean Labor and Income Panel Survey (KLIPS)
- `http://www.kli.re.kr/klips/en/about/introduce.jsp`
- The variable of interest is the monthly income in 2006, which has about 35% missing values
- Covariates associated are gender, age group, level of education, and the monthly income in 2005

Notations

- Denote Y as the response variable, R as the missing indicator ($R = 1$ observed; $R = 0$ missing)
- Denote the regression model as $p(Y|X; \theta)$, X as the covariates (fully observed)
- Missing data mechanism $p(R = 1|Y, X)$ could be missing at random (MAR) or nonignorable missingness

Literature Review

- Problem of interest: the comparison of variance estimates before and after multiple imputation (MI)
- Under MAR assumption, formulae were provided (Wang and Robins 1998, Robins and Wang 2000)
- Under nonignorable missingness, the situation is complicated...

Multiple Imputation (MI)

- Concentrate on type B estimator (Rubin's nomenclature) from frequentist perspective
- Denote $\hat{\theta}_p$ as the preliminary estimator
- Generate data from $p(Y|X, R = 0, \hat{\theta}_p)$ for missing Y's
- Obtain $\hat{\theta}_{p,m}$ from each complete data set, $m = 1, \dots, M$
- The estimator after MI is defined as $\hat{\theta}_{p,MI} = \frac{1}{M}\sum_{m=1}^{M} \hat{\theta}_{p,m}$

Asymptotic Relative Efficiency (ARE)

- Under some regularity conditions, we have

$$\sqrt{n}(\hat{\theta}_p - \theta) \xrightarrow{d} N(0, V_p),$$
$$\sqrt{n}(\hat{\theta}_{p,MI} - \theta) \xrightarrow{d} N(0, V_{p,MI})$$

- For i-th element of θ, θ_i, the ARE of before MI w.r.t. after MI is defined as

$$\frac{V_{p,MI}^{i,i}}{V_p^{i,i}}$$

Missing At Random (MAR): $p(R = 1|Y, X) = p(R = 1|X)$

Theory

- $p(R = 1|Y, X) = p(R = 1|X)$ implies $p(Y|X, R = 1) = p(Y|X)$
- Denote $\hat{\theta}_{mar}$ as the preliminary estimator, using all observed data
- Information equality $I_c = I_{obs} + I_{mis}$
- Before MI, we have $V_p = I_{obs}^{-1}$
- After MI, we have $V_{p,MI} = I_{obs}^{-1} + M^{-1} I_c^{-1} I_{mis} I_c^{-1}$
- $V_{p,MI} - V_p$ is positive definite
- ARE $\searrow$ 1, as $M \nearrow \infty$

Simulation Studies

- $p(Y|X; \theta) = N(\alpha + \beta X, \sigma^2), \theta = (\alpha, \beta, \sigma) = (0, 1, 1)$
- Assume

$$p(R = 1|Y, X) = p(R = 1|X) = \frac{\exp(X - 1)}{1 + \exp(X - 1)},$$

where the proportion of observed data is about 50%
- Sample size $N = 300$, and the results are based on 300 simulation runs

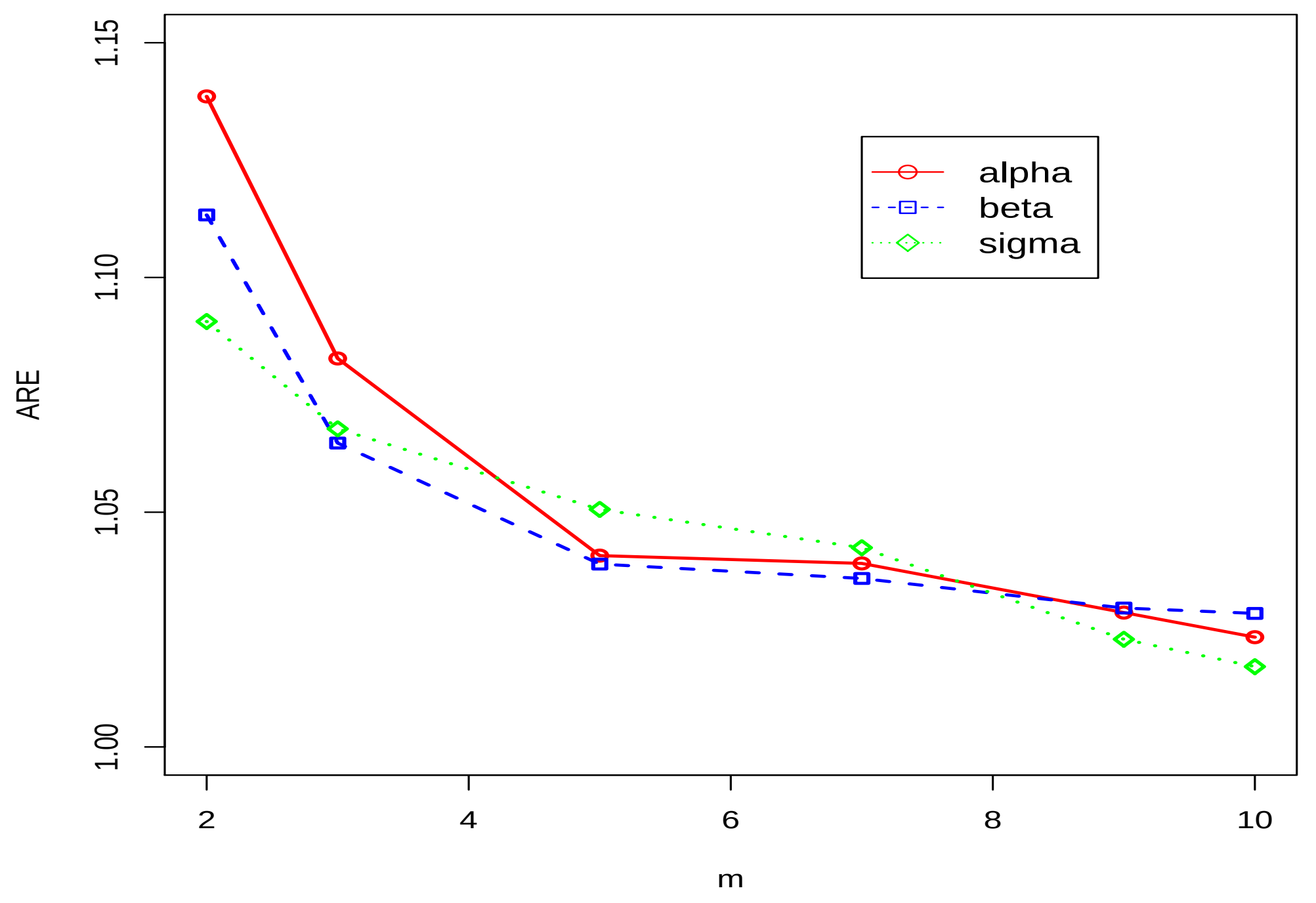

Figure 1: ARE for three parameters under MAR assumption with 50% complete data

Nonignorable Missingness: $p(R = 1|Y, X) = p(R = 1|Y)$

The preliminary estimator

- $p(R = 1|Y, X) = p(R = 1|Y)$ implies

$$p(X|Y, R = 1) = p(X|Y) = \frac{p(Y|X; \theta)p(X)}{\int p(Y|X; \theta)p(X)dX}$$

- The preliminary estimator

$$\hat{\theta}_{non} = \arg\max_{\theta} \prod_{i=1}^{n} \frac{p(Y_i|X_i; \theta)}{\sum_{j=1}^{N} p(Y_i|X_j; \theta)}$$

Nonignorable Missingness: $p(R = 1|Y, X) = p(R = 1|Y)$

Simulation Studies

- $p(Y|X; \theta) = N(\alpha + \beta X, \sigma^2), \theta = (\alpha, \beta, \sigma) = (0, 1, 1)$
- Assume

$$p(R = 1|Y, X) = p(R = 1|Y) = \frac{\exp(Y + \eta)}{1 + \exp(Y + \eta)},$$

where $\eta = 0.5, -1, -2.5$ corresponds to proportion of observed data 25%, 50%, 75% respectively
- Sample size $N = 300$, and the results are based on 300 simulation runs

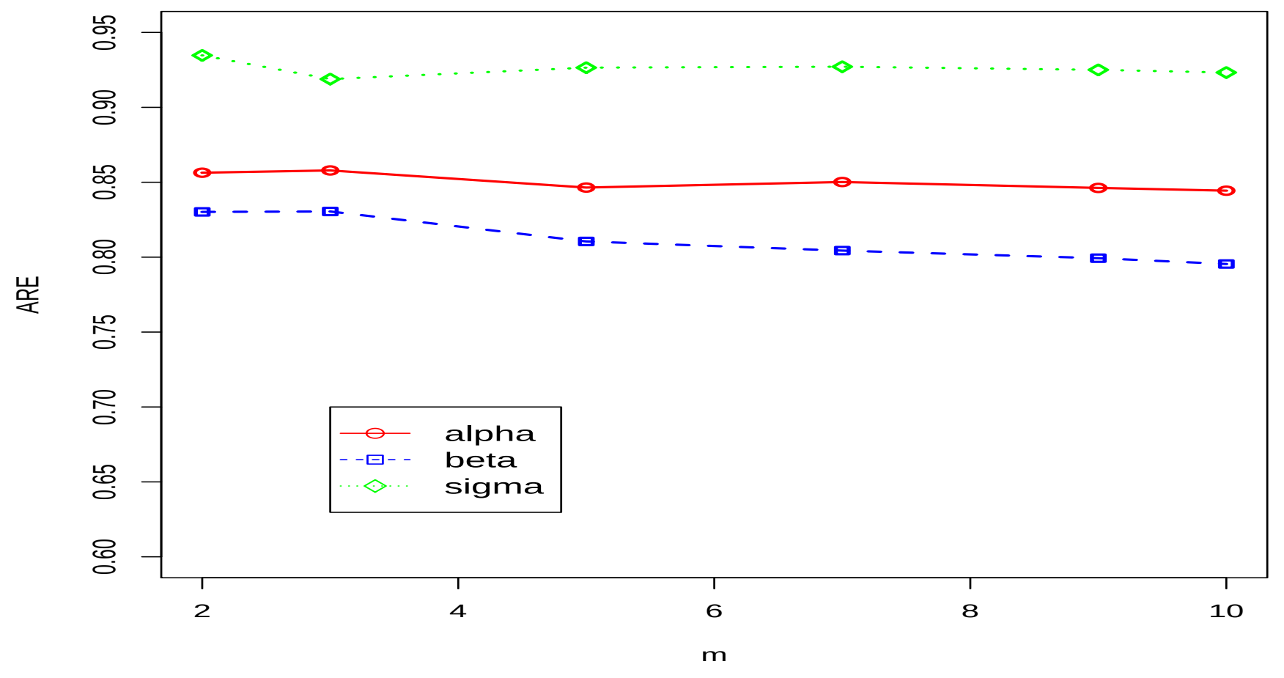

Figure 2: ARE for three parameters under nonignorable missingness with 25% complete data

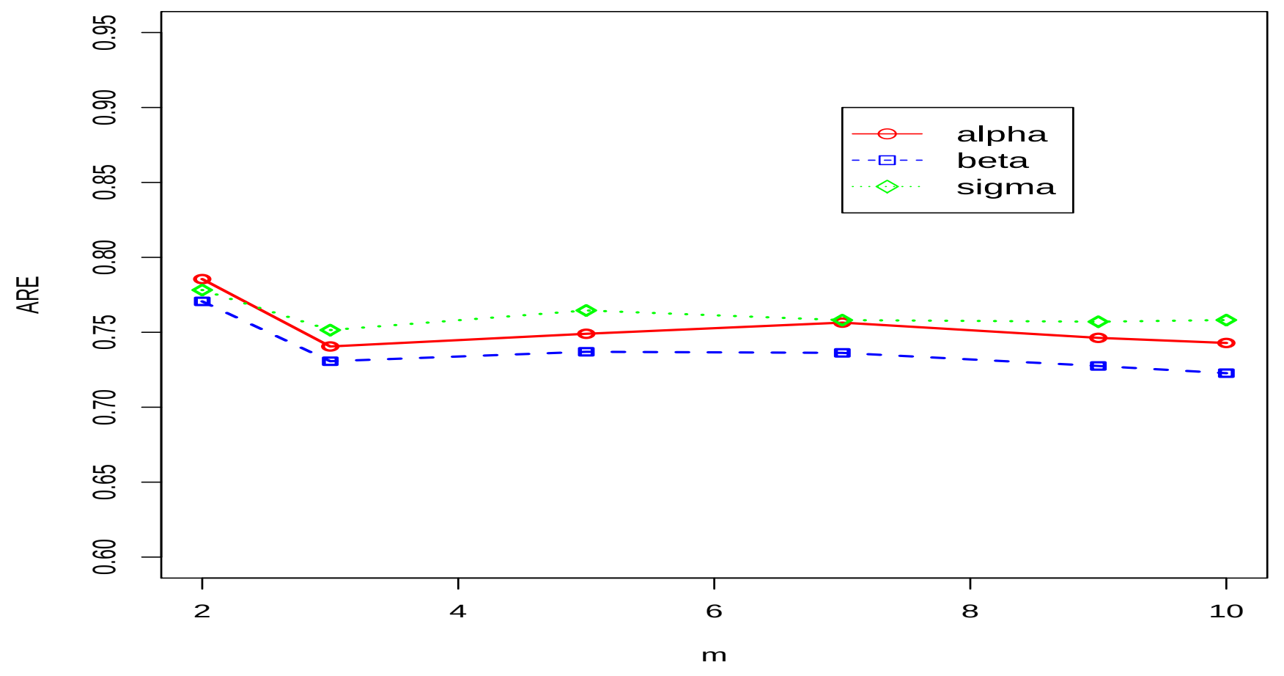

Figure 3: ARE for three parameters under nonignorable missingness with 50% complete data

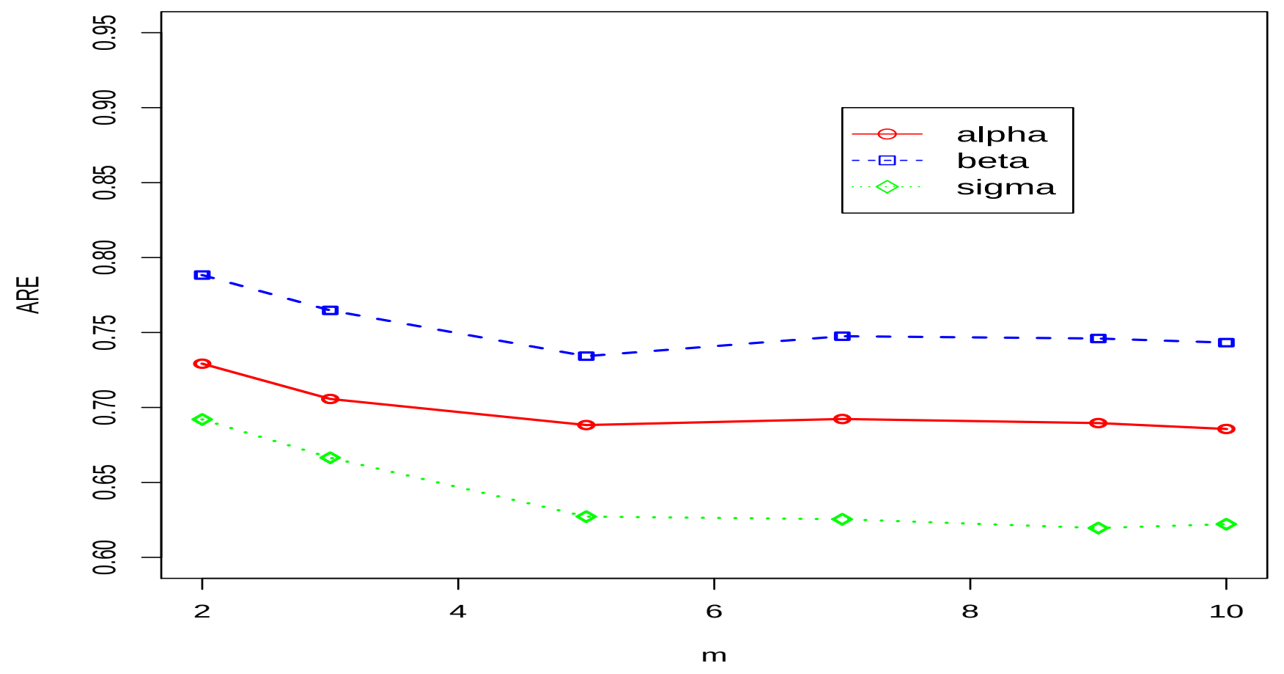

Figure 4: ARE for three parameters under nonignorable missingness with 75% complete data

Concluding Remarks

- The estimator after MI is always less efficient than the one before MI for MAR; however, for nonignorable missingness, MI could increase the estimation efficiency
- For nonignorable missingness, as m varies, the ARE becomes stable sooner than the MAR case
- As the proportion of observed data increases, ARE becomes smaller, which means the estimator is getting more efficient
- This paper only proposes one kind of preliminary estimator for nonignorable missingness, and how to propose various preliminary estimators, including the efficient one, is under further investigation
- The theoretical foundation of variance estimation after MI for nonignorable missingness is under further investigation

Created by Jiwei Zhao
Email: `jiwei2012zhao@gmail.com`