Statistica Sinica **16**(2006), 1169-1192

EFFICIENCIES OF METHODS DEALING WITH MISSING COVARIATES IN REGRESSION ANALYSIS

Cuiling Wang and Myunghee Cho Paik

Yeshiva University and Columbia University

Abstract: Various approaches have been developed to deal with missing covariate problems in regression analysis when the data are missing at random. Among them, three main non-likelihood approaches are through weighting, imputation and conditional likelihood. The imputation method replaces the missing contribution to the estimating function with its conditional expectation. The inverse probability weighting method weights each observed record by the inverse of the observation probability. The conditional method constructs an unbiased estimating function using only complete records by modelling the conditional mean, given that record is observed. In the literature, the efficiencies of these methods have been compared via simulation. In this paper we compare the asymptotic variances and prove some inequalities. We show that in logistic regression the asymptotic variance of the conditional likelihood method is smaller than or equal to that of the inverse probability weighting method. When the fully observed variables are categorical, the imputation method is more efficient than the inverse probability weighting method given that the observation model is correctly specified. We also show that if the missing mechanism is MCAR and the true known probability of observation is used, the asymptotic variance of the inverse probability weighting method is greater than or equal to that of the complete case analysis. We also conduct simulation studies to compare performances in finite samples and later illustrate the methods using data from a stroke study.

Key words and phrases: Efficiency, estimating equation, imputation, inverse probability weighting, logistic regression, missing at random, missing covariate.

1. Introduction

We consider the situation where the conditional mean of outcome Y of a subject given covariates X and Z is of interest. When all data are observed, this often constitutes a regression analysis. We consider the case in which Y and Z are always observed, but X could be missing for some subjects. Throughout the paper the probability of missingness is assumed not to depend on X, i.e., X is missing completely at random or missing at random.

In the presence of a missing covariate, a common practice is complete case (CC) analysis in which records with missing covariates are simply deleted. Under certain missingness mechanisms, the CC analysis would yield biased estimators. There are mainly three non-likelihood based approaches to deal with

this problem: weighting (Zhao and Lipsitz (1992), Robins, Rotnitzky and Zhao (1994) and Zhao, Lipsitz and Lew (1996)), imputation (Reilly and Pepe (1995)) and Paik (1997)) and conditional likelihood (Brelow and Cain (1988) and Wang (1999)). The efficiencies of the estimators from the aforementioned approaches have been compared by simulation studies. Zhao and Lipsitz (1992) have shown that the conditional method performs better than the inverse probability weighting method when the probability of observing is known. The semiparametric efficient estimator proposed by Robins, Rotnitzky and Zhao (1994), an extension of the weighting method, would in theory achieve the semiparametric variance bound. Robins, Rotnitzky and Zhao (1994) showed that the efficiency of the inverse probability weighting estimator can be improved by subtracting an extra term, namely the projection of the backbone estimating function onto the nuisance tangent space formed by nuisance score. However, computation is often intensive, and the efficiency gain over the imputation method is decisive only when the sample size is very large (Paik (2000)). Therefore, the efficiency comparison among the imputation, the inverse probability weighting without efficiency adjustment and the conditional method is still of interest among practitioners. Throughout the paper, we call the method of Zhao and Lipsitz (1992) inverse probability weighting. The efficient version of RRZ is referred to as efficient inverse probability weighting.

In this paper we analytically compare the asymptotic variances of the imputation, inverse probability weighting (IPW) and conditional likelihood methods, and show some new inequalities among the asymptotic variances. We prove that for the logistic model, when the probability of observation is known, the conditional method is more efficient than the IPW method. We also show that when the completely observed variables are categorical, the imputation method has a smaller asymptotic variance than that of the IPW method. Under the same condition, we prove that the imputation method is equivalent to the improved IPW method. Among the IPW estimators based on different models for missingness, we show that the asymptotic variance of the IPW estimators from an over-fitted missingness model is smaller than or equal to that from a simpler missingness model.

We present results from simulation studies. For the most part, the simulation results agree with the theoretical findings. We also illustrate the various methods using data from the Northern Manhattan Stroke Study, which include 3,202 subjects randomly sampled from the northern Manhattan. We analyze two -year stroke incidence as a function of various risk factors. One of the risk factors, left ventricular hypertrophy, is missing for 1,147 subjects because the variable requires taking an ECHO cardiogram. Logistic regression models are applied using the three aforementioned approaches for missing covariates, and the results are compared.

In Section 2, we review the three approaches handling missing covariates. In Section 3, we compare their asymptotic variances. In Section 4 simulation results are shown. We compare the different estimates using stroke study data in Section 5.

2. Methods to Handle Missing Covariates

In this section, we review the three approaches and present the model assumptions and the asymptotic variances associated with the different methods.

2.1. Model and notation

Suppose that the conditional mean of outcome Y given covariate X and Z is $E(Y|X, Z; \beta)$, where β is an unknown finite dimensional parameter vector of interest. Here Y and X are scalars, and Z can be a vector of covariates. We assume that the missing mechanism is missing at random (Rubin (1976)), that is, the probability of observing X may depend on Y and Z, but not on X. Let R be an indicator of X being observed. We also assume that the observation probability is not zero and can be modelled parametrically, say $\pi(X, Y, Z) = \pi(X, Y, Z; \alpha)$, where π is a known, twice-differentiable function and α is a finite-dimensional unknown nuisance parameter. By the MAR assumption, we have $\pi(X, Y, Z) = \pi(Y, Z)$ and $\pi(X, Y, Z; \alpha) = \pi(Y, Z; \alpha)$. Throughout the paper, we use $\pi(Y, Z)$ to generically denote the observation probability, known or unknown, and $\pi(Y, Z; \alpha)$ to denote the parametrically specified observation probability with unknown α. Finally, we assume that (R_i, Y_i, X_i, Z_i), $i = 1, \ldots, n$, are i.i.d.. Let $\mu(X, Z; \beta) = E(Y|X, Z; \beta)$. Then without missing data, the most efficient linear estimating function would be

$$S(Y_i|X_i, Z_i; \beta) = \frac{\partial}{\partial \beta}\{\mu(X_i, Z_i; \beta)\} w(X_i, Z_i; \beta)^{-1}\{Y_i - \mu(X_i, Z_i; \beta)\}, \qquad (2.1)$$

where $w(X_i, Z_i; \beta) = \mathrm{Var}(Y_i|X_i, Z_i; \beta)$.

Since R is Bernoulli with mean $\pi(Y, Z; \alpha)$, the nuisance score from logistic regression is

$$U(R_i|Y_i, Z_i; \alpha) = R_i \frac{\partial}{\partial \alpha} \log\{\pi(Y_i, Z_i; \alpha)\} + (1 - R_i) \frac{\partial}{\partial \alpha} \log\{1 - \pi(Y_i, Z_i; \alpha)\}. \quad (2.2)$$

We occasionally use $S(Y|X, Z; \beta)$ and $U(R|Y, Z; \alpha)$ without the subscript i to denote the contribution from a single observation to the estimating function for β and for α, respectively.

Throughout the paper, we use the following notation: n is the total number of subjects; Y is the outcome variable; X is the covariate that is subject to missing; Z is the fully observed covariates; V is the set of records in which X are

observed; $\bar{V}$ is the set of records in which X are missing; $R = 1$ if X is observed, 0 otherwise; V^{zy} and $\bar{V}^{zy}$ are the subset of completely observed sample and partially observed sample, respectively, with $Y = y$ and $Z = z$; n_{zy}^{v} and $n_{zy}^{\bar{v}}$ are the number of subjects in V^{zy} and $\bar{V}^{zy}$, respectively.

2.2. Complete case analysis

A common treatment of missing data in statistical package is the complete-case (CC) analysis, where the records with missing values are simply deleted. The CC analysis uses the following estimating equation:

$$S_{CC}(\beta) = \sum_{i=1}^{n} R_i S(Y_i|X_i, Z_i; \beta) = 0.$$

When the missing mechanism is missing completely at random (MCAR), that is, the observation probability does not depend on Y and X, but may depend on Z, the estimator from the CC analysis, say $\hat{\beta}_{cc}$, is consistent, and $\sqrt{n}(\hat{\beta}_{cc} - \beta)$ is asymptotically normally distributed with mean zero and variance

$$V_{cc} = \Big[E\{\pi(Z)S(Y|X, Z; \beta)S(Y|X, Z; \beta)^T\}\Big]^{-1}. \tag{2.3}$$

When the observation probability is a constant π_0, the asymptotic variance is $(1/\pi_0)I_v^{-1}$, where $I_v = E\{S(Y|X, Z; \beta)S(Y|X, Z; \beta)^T\}$; but when the observation probability depends on Y, which is the case under MAR, it may yield a biased estimator.

2.3. Imputation

When Y and Z are categorical and X is partially observed, Reilly and Pepe (1995) proposed an imputation method which uses the following estimating equation:

$$S_{im}(\beta) = \sum_{i=1}^{n} R_i S(Y_i|X_i, Z_i; \beta) + \sum_{i=1}^{n} (1 - R_i)\hat{E}\{S(Y_i|X_i, Z_i; \beta)|Y_i, Z_i\} = 0, \tag{2.4}$$

where $\hat{E}\{S(Y|X, Z)|y, z\} = \sum_{j \in V^{zy}} S(Y|X, Z; \beta)/n_{zy}^{v}$. When Z is continuous, the method can be easily extended using a parametric modelling (See Fleiss, Levin and Paik (2003)).

Reilly and Pepe (1995) show that the estimator, $\hat{\beta}_{im}$, is consistent and the asymptotic variance of $\sqrt{n}\hat{\beta}_{im}$ is

$$\begin{aligned} V_{im} = {} & I_v^{-1} E\Big[\frac{1}{\pi(Y, Z)} S(Y|X, Z; \beta)S(Y|X, Z; \beta)^T\Big] I_v^{-1} \\ & - I_v^{-1} E\Big(\frac{1 - \pi(Y, Z)}{\pi(Y, Z)} S(Y|Z; \beta)S(Y|Z; \beta)^T\Big) I_v^{-1}, \end{aligned} \tag{2.5}$$

where $S(Y|Z;\beta) = E\{S(Y|X,Z;\beta)|Y,Z\}$. They show that estimating equation (2.4) can also be expressed as

$$S_{im}(\beta) = \sum_{i=1}^{n} \Big(\frac{n_{Z_iY_i}}{n^v_{Z_iY_i}}\Big) R_i S(Y_i|X_i,Z_i;\beta) = 0, \tag{2.6}$$

which is equivalent to the inverse probability weighting estimating equation when the empirical estimator of $\pi(Y,Z)$ is used.

2.4. Inverse probability weighting

The Inverse Probability Weighting (IPW) method weights each record by the inverse of its probability being observed. When $\pi(Y,Z)$ is known, Zhao and Lipsitz (1992) proposed the IPW method using the estimating equation

$$S_w(\beta) = \sum_{i=1}^{n} \frac{R_i}{\pi(Y_i,Z_i)} S(Y_i|X_i,Z_i;\beta) = 0.$$

Zhao and Lipsitz (1992) showed that the estimator, $\hat{\beta}_w$, is consistent and asymptotically normally distributed, and the asymptotic variance of $\sqrt{n}\hat{\beta}_w$ is

$$V_{wt} = I_v^{-1} E\Big[\frac{1}{\pi(Y,Z)} S(Y|X,Z;\beta) S(Y|X,Z;\beta)^T\Big] I_v^{-1}. \tag{2.7}$$

When $\pi(Y,Z)$ is not known, its estimate can be used. If Y and Z are categorical, we can use the sample mean of R given Y and Z among the completely observed records, which yields the same estimator as in the imputation method. Equivalence between the IPW and the imputation approach when both the imputation and missingness models are saturated is reported by Little (1986), Reilly and Pepe (1995) and Paik (1997). When Y or some components of Z are continuous, we can assume a parametric model for $\pi(Y,Z)$, for example, $\pi(Y,Z) = \pi(Y,Z;\alpha)$, where π is a known function indexed by unknown parameter α. Then β and α can be estimated simultaneously by the following estimating equations (e.g. Zhao, Lipsitz and Lew (1996)):

$$S_w(\alpha,\beta) = \sum_{i=1}^{n} \frac{R_i}{\pi(Y_i,Z_i;\alpha)} S(Y_i|X_i,Z_i;\beta) = 0,$$
$$U(\alpha) = \sum_{i=1}^{n} U(R_i|Y_i,Z_i;\alpha) = 0.$$

Zhao, Lipsitz and Lew (1996) showed that the asymptotic variance of $(\hat{\beta}_w, \hat{\alpha})$ is of a sandwich-type, $I^{-1}\Sigma I^{-1}$, where I is the derivative of $\{S_w(\alpha,\beta), U(\alpha)\}$, and

$\Sigma = \mathrm{Var}\{S_w(\alpha,\beta), U(\alpha)\}$. In the appendix we show the derivation of the following alternative form of the asymptotic variance of $\sqrt{n}(\hat{\beta}_w - \beta)$, which facilitates the comparisons:

$$V_{we} = I_v^{-1} E\Big[\frac{1}{\pi(Y,Z;\alpha)} S(Y|X,Z;\beta) S(Y|X,Z;\beta)^T\Big] I_v^{-1} - I_v^{-1}\Omega_{12} I_\alpha^{-1} \Omega_{12}^T I_v^{-1}, \tag{2.8}$$

where $I_\alpha = E_{YZ}\Big(\dot{\pi}(Y,Z;\alpha)\dot{\pi}(Y,Z;\alpha)^T/[\pi(Y,Z;\alpha)\{1-\pi(Y,Z;\alpha)\}]\Big)$, $\dot{\pi}(Y,Z;\alpha) = \partial/\partial\alpha\{\pi(Y,Z;\alpha)\}$, $\Omega_{12} = E_{YZ}\{S(Y|Z;\beta)\dot{\pi}(Y,Z;\alpha)^T/\pi(Y,Z;\alpha)\}$.

2.5. Improved inverse probability weighting method

Robins, Rotnitzky and Zhao (1994) showed that the efficiency of IPW can be improved by subtracting the projection of the estimating function onto the nuisance tangent space that is the closed span of nuisance scores. The resulting estimating function has the following form:

$$S_{iw} = \sum_{i=1}^{n}\Big(\frac{R_i}{\pi(Y_i,Z_i)} S(Y_i|X_i,Z_i;\beta) - \frac{R_i - \pi(Y_i,Z_i)}{\pi(Y_i,Z_i)} S(Y_i|Z_i;\beta)\Big), \tag{2.9}$$

where $S(Y_i|X_i,Z_i;\beta) = h(X_i,Z_i;\beta)\{Y_i - \mu(X_i,Z_i;\beta)\}$, and $h(X_i,Z_i;\beta) = \partial/\partial\beta\{\mu(X_i,Z_i;\beta)\}w(X_i,Z_i;\beta)^{-1}$. The resulting estimator is the improved augmented IPW estimator. It should be noted that it is the most efficient among those that use $S(Y_i|X_i,Z_i;\beta)$ as the 'kernel' with fixed function h. When h is allowed to be arbitrary, a fully efficient estimator can be obtained, but h should be solved to satisfy (23) in Robins, Rotnitzky and Zhao (1994). In this paper we restrict our attention to the case in which h is fixed, because computation for optimal h is impractically complicated.

Since $S(Y|Z;\beta)$ is usually unknown as the distribution of X is usually unknown, its estimate is used in practice. Zhao, Lipsitz and Lew (1996) discussed modelling $S(Y|Z;\beta)$ parametrically and gave a simulation result for continuous Z. We show in Section 3 that for categorical Y and Z with saturated auxiliary models, the imputation method is equivalent to the improved weighting method.

2.6. Conditional method

Breslow and Cain (1988) proposed the parameter estimator based on the conditional likelihood of Y given X, Z and $R = 1$ when Y is binary. This approach could be extended to the case in which Y is not binary, but a distributional assumption on Y is required. To keep to the main point, we consider the case where Y is binary with probability of logistic form $\mu(X,Z;\beta) = E(Y|X,Z;\beta) = \exp(\beta^T W)/\{1+\exp(\beta^T W)\}$, where $W = (1, Z^T, X)^T$. Let

$\eta(X,Z;\beta) = P(Y=1|X,Z,R=1;\beta)$, $\pi(1,Z) = P(R=1|Y=1,Z)$, and $\pi(0,Z) = P(R=1|Y=0,Z)$. Then $\eta(X,Z;\beta)$, $\mu(X,Z;\beta)$, and $\pi(Y,Z)$ satisfy the following relation:

$$\eta(X,Z;\beta) = \frac{\pi(1,Z)\mu(X,Z;\beta)}{\pi(1,Z)\mu(X,Z;\beta) + \pi(0,Z)\{1-\mu(X,Z;\beta)\}}.$$

When $\pi(Y,Z)$ is known, the estimating equation based on the conditional likelihood is:

$$S_c(\beta) = \sum_{i=1}^{n} R_i W_i \{Y_i - \eta(X_i,Z_i;\beta)\} = 0.$$

The estimator $\hat{\beta}_c$ is consistent for β, and $\sqrt{n}(\hat{\beta}_c - \beta)$ is asymptotically normally distributed with mean 0 and variance $V_{ct} = I_c^{-1}$, where

$$I_c = E_{XZ}\Big[W \frac{\pi(1,Z)\mu(X,Z;\beta)\pi(0,Z)\{1-\mu(X,Z;\beta)\}}{\pi(1,Z)\mu(X,Z;\beta) + \pi(0,Z)\{1-\mu(X,Z;\beta)\}} W^T\Big].$$

When $\pi(Y,Z)$ is not known, it can be parametrically modelled as in the IPW method, and (β,α) can be obtained by solving the following estimating equations simultaneously:

$$S_c(\beta,\alpha) = \sum_{i=1}^{n} R_i W_i \{Y_i - \eta(X_i,Z_i;\beta,\alpha)\} = 0,$$
$$U(\alpha) = \sum_{i=1}^{n} U(R_i|Y_i,Z_i;\alpha) = 0.$$

As discussed by Wang (1999), it can be shown that the estimator $\hat{\beta}_c$ is consistent and $\sqrt{n}(\hat{\beta}_c - \beta)$ is asymptotically normally distributed with mean 0 and variance

$$V_{ce} = I_c^{-1} - I_c^{-1}\Omega_c I_\alpha^{-1}\Omega_c^T I_c^{-1}, \tag{2.10}$$

where

$$\Omega_c = E_{XZ}\Big(W \frac{\pi(1,Z;\alpha)\pi(0,Z;\alpha)\mu(X,Z;\beta)\{1-\mu(X,Z;\beta)\}}{\pi(1,Z;\alpha)\mu(X,Z;\beta) + \pi(0,Z;\alpha)\{1-\mu(X,Z;\beta)\}}$$
$$\Big[\frac{\frac{\partial}{\partial\alpha^T}\{\pi(1,Z;\alpha)\}}{\pi(1,Z;\alpha)} - \frac{\frac{\partial}{\partial\alpha^T}\{\pi(0,Z;\alpha)\}}{\pi(0,Z;\alpha)}\Big]\Big).$$

3. Comparison of Asymptotic Variances

3.1. Summary of results

We outline our results below. Each result is elaborated upon in a subsequent section.

1. Under the MCAR missing mechanism, the asymptotic variance of the IPW estimator using the true observation probability is greater than or equal to that of the CC estimator; equality holds when the observation probability is a constant.
2. When Y and Z are categorical, the asymptotic variance of the imputation method is smaller than or equal to that of the IPW method; equality holds when both the imputation model and the model for $\pi(Y, Z; \alpha)$ of the IPW are saturated.
3. When the fully observed variables Y and Z are categorical, the imputation method is asymptotically equivalent to the improved weighting method.
4. Asymptotically the overfitted models for the observation probability yield the more efficient IPW estimators.
5. When Y is binary and a logistic model for Y is used, using the true $\pi(Y, Z)$, the conditional likelihood method performs asymptotically better than the IPW method in the sense that the difference between the asymptotic variances of the IPW and the conditional estimators is positive semidefinite.

Note that Results 1 and 4 apply to the case of all types of Y and Z, while Results 2 and 3 apply to the case of categorical Y and Z. Result 5 holds for dichotomous Y but all types of Z. For the observation probability $\pi(Y, Z)$ used, Results 1 and 5 apply to the case in which the true and known $\pi(Y, Z)$ is used, while Results 2, 3 and 4 mainly concern the case in which the estimated $\pi(Y, Z)$ is used.

3.2. Comparison between IPW and CC

Since the CC estimator is biased and the IPW estimator adjusts the bias, a comparison of efficiency between the two would not be interesting under MAR. When the missingness mechanism is MCAR, both analyses yield consistent estimators, and we can compare the efficiencies of IPW and CC estimators. In this section, we use notation $\pi(Z)$ instead of $\pi(Y, Z)$ since the observation probability does not depend of Y.

Result 1. *Under the MCAR missing mechanism, the asymptotic variance of the IPW estimator, using the true observation probability, is greater than or equal to that of the CC estimator. Equality holds when the observation probability is a constant.*

Proof. Under MCAR, $\hat{\beta}_{cc}$ from CC analysis is consistent, and the asymptotic variance of $\sqrt{n}\hat{\beta}_{cc}$ and $\sqrt{n}\hat{\beta}_w$ is V_{cc} (2.3) and V_{wt} (2.7), respectively. Let $a_1 = \{1/\sqrt{\pi(Z)}\}S(Y|X, Z; \beta)$, $a_2 = \sqrt{\pi(Z)}S(Y|X, Z; \beta)$, and $a = a_1 - E(a_1 a_2^T)$

$\{E(a_2a_2^T)\}^{-1}a_2$. Then we see that

$$\begin{aligned} V_{wt} - V_{cc} &= I_v^{-1}E\Big[\frac{1}{\pi(Z)}S(Y|X,Z;\beta)S(Y|X,Z;\beta)^T\Big]I_v^{-1} \\ &\quad -\Big[E\{\pi(Z)S(Y|X,Z;\beta)S(Y|X,Z;\beta)^T\}\Big]^{-1} \\ &= \{E(a_1a_2^T)\}^{-1}E(aa^T)\{E(a_2a_1^T)\}^{-1} \end{aligned}$$

is positive semi-definite. It is zero only when $a = 0$, that is, $\pi(Z)$ is a constant.

Although the IPW method is effective in adjusting the bias when data are missing at random, the IPW estimator using the true $\pi(Z)$ performs no better than the CC estimator under MCAR.

3.3. Comparison of IPW with the imputation method

Result 2. *When Y and Z are categorical, the asymptotic variance of the imputation estimator is smaller than or equal to that of the IPW method. Equality holds when both the imputation model and the model for $\pi(Y,Z;\alpha)$ of the IPW are saturated.*

Proof. Using (2.5) and (2.8), the difference between the asymptotic variances of $\sqrt{n}\hat{\beta}_w$ and $\sqrt{n}\hat{\beta}_{im}$ is

$$V_{we} - V_{im} = I_v^{-1}M_dI_v^{-1}, \tag{3.1}$$

where $M_d = E\Big([\{1-\pi(Y,Z;\alpha)\}/\pi(Y,Z;\alpha)]S(Y|Z;\beta)S(Y|Z;\beta)^T\Big) - \Omega_{12}I_\alpha^{-1}\Omega_{12}^T$. Let $b_1 = S(Y|Z;\beta)\sqrt{\{1-\pi(Y,Z;\alpha)\}/\pi(Y,Z;\alpha)}$, $b_2 = \dot{\pi}(Y,Z;\alpha)^T/[\sqrt{\pi(Y,Z;\alpha)\{1-\pi(Y,Z;\alpha)\}}]$, and $b = a - E(b_1b_2^T)\{E(b_2b_2^T)\}^{-1}b_2$. Then (3.1) can be expressed as

$$\begin{aligned} &I_v^{-1}\Big[E\Big\{\frac{1-\pi(Y,Z;\alpha)}{\pi(Y,Z;\alpha)}S(Y|Z;\beta)S(Y|Z;\beta)^T\Big\} - \Omega_{12}I_\alpha^{-1}\Omega_{12}^T\Big]I_v^{-1} \\ &= I_v^{-1}\Big[E(b_1b_1^T) - E(b_1b_2^T)\{E(b_2b_2^T)\}^{-1}E(b_2b_1^T)\Big]I_v^{-1} \\ &= I_v^{-1}E(bb^T)I_v^{-1}, \end{aligned}$$

which is positive semidefinite. It is zero if and only if $b = 0$, i.e, when $S(Y|Z;\beta) = E(b_1b_2^T)\{E(b_2b_2^T)\}^{-1}\dot{\pi}(Y,Z;\alpha)^T/[\pi(Y,Z;\alpha)\{1-\pi(Y,Z;\alpha)\}]$, or when $S(Y|Z;\beta)$ is a linear transformation of $\dot{\pi}(Y,Z;\alpha)$. This occurs when the model for $\pi(Y,Z;\alpha)$ is saturated if Y and Z are categorical.

Result 3. *When the fully observed variables Y and Z are categorical, the imputation method is asymptotically equivalent to the improved weighting method.*

Proof. We show that the estimator from (2.9), $\hat{\beta}_{iw}$, has the asymptotic variance of $\hat{\beta}_{im}$ at (2.5). Since $S_{iw}(\beta) = \sum_{i=1}^{n}\Big(\{R_i/\pi(Y_i,Z_i)\}S(Y_i|X_i,Z_i;\beta) - [\{R_i - \pi(Y_i,Z_i)\}/\pi(Y_i,Z_i)]S(Y_i|Z_i;\beta)\Big)$, $-(1/n)\partial/\partial\beta\{S_{iw}(\beta)\} \xrightarrow{p} I_v$, and

$$\begin{aligned}\mathrm{Var}\Big[\frac{1}{\sqrt{n}}S_{iw}(\beta)\Big] &= E\Big[\frac{1}{\pi(Y,Z)}S(Y|X,Z;\beta)S(Y|X,Z;\beta)^T\Big]\\ &\quad -E\Big(\frac{1-\pi(Y,Z)}{\pi(Y,Z)}S(Y|Z;\beta)S(Y|Z;\beta)^T\Big).\end{aligned}$$

Thus the asymptotic variance of $\sqrt{n}\hat{\beta}_{iw}$ is

$$\begin{aligned}&I_v^{-1}E\Big[\frac{1}{\pi(Y,Z)}S(Y|X,Z;\beta)S(Y|X,Z;\beta)^T\Big]I_v^{-1}\\ &-I_v^{-1}E\Big(\frac{1-\pi(Y,Z)}{\pi(Y,Z)}S(Y|Z;\beta)S(Y|Z;\beta)^T\Big)I_v^{-1},\end{aligned}$$

which is exactly the same as the variance of the imputation estimator in (2.5).

Furthermore, we show below that when Y and Z are categorical and the empirical estimate of $P(X|Y,Z)$ is used to estimate $S(Y|Z;\beta)$, the improved inverse probability weighting estimating function is the same as that of the imputation method. Let $S^*_{iw}(\beta) = \sum_{i=1}^{n}(\{R_i/\pi(Y_i,Z_i)\}S(Y_i|X_i,Z_i;\beta) - [\{R_i - \pi(Y_i,Z_i)\}/\pi(Y_i,Z_i)]\hat{S}(Y_i|Z_i;\beta))$. If $\hat{S}(Y|Z;\beta) = \hat{E}\{S_\beta(Y|X,Z)|Y,Z\} = \sum_{i\in V^{YZ}} S(Y_i|X_i,Z_i;\beta)/n^v_{YZ}$, then

$$\begin{aligned}&\sum_{i=1}^{n}\frac{R_i-\pi(Y_i,Z_i)}{\pi(Y_i,Z_i)}\hat{S}(Y_i|Z_i;\beta)\\ &=\sum_{i=1}^{n}\frac{R_i}{\pi(Y_i,Z_i)}\sum_{j\in V^{Z_iY_i}}\frac{S(Y_j|X_j,Z_j;\beta)}{n^v_{Z_iY_i}} - \sum_{i=1}^{n}\sum_{j\in V^{Z_iY_i}}\frac{S(Y_j|X_j,Z_j;\beta)}{n^v_{Z_iY_i}}\\ &=\sum_{i\in V}\sum_{j\in V^{Z_iY_i}}\frac{1}{\pi(Y_j,Z_j)}\frac{S(Y_j|X_j,Z_j;\beta)}{n^v_{Z_iY_i}} - \sum_{i=1}^{n}\sum_{j\in V^{Z_iY_i}}\frac{S(Y_j|X_j,Z_j;\beta)}{n^v_{Z_iY_i}}\\ &=\sum_{YZ}\sum_{i\in V^{ZY}}\sum_{j\in V^{Z_iY_i}}\frac{1}{\pi(Y_j,Z_j)}\frac{S(Y_j|X_j,Z_j;\beta)}{n^v_{ZY}} - \sum_{YZ}\frac{n_{ZY}}{n^v_{ZY}}\sum_{j\in V^{ZY}}S(Y_j|X_j,Z_j;\beta)\\ &=\sum_{YZ}\sum_{j\in V^{YZ}}\frac{S(Y_j|X_j,Z_j;\beta)}{\pi(Y_j,Z_j)} - \sum_{YZ}\frac{n_{ZY}}{n^v_{ZY}}\sum_{j\in V^{ZY}}S(Y_j|X_j,Z_j;\beta)\\ &=\sum_{i=1}^{n}\frac{R_i}{\pi(Y_i,Z_i)}S(Y_i|X_i,Z_i;\beta) - \sum_{i=1}^{n}R_i\frac{n_{Y_iZ_i}}{n^v_{Y_iZ_i}}S(Y_i|X_i,Z_i;\beta).\end{aligned}$$

Hence

$$S^*_{iw}(\beta) = \sum_{i=1}^{n} \{ \frac{R_i}{\pi(Y_i, Z_i)} S(Y_i|X_i, Z_i; \beta) - \frac{R_i - \pi(Y_i, Z_i)}{\pi(Y_i, Z_i)} \hat{S}(Y_i|Z_i; \beta) \}$$
$$= \sum_{i=1}^{n} R_i \frac{n_{Y_i Z_i}}{n^v_{Y_i Z_i}} S(Y_i|X_i, Z_i; \beta).$$

This is the estimating equation (2.6) in the imputation method.

3.4. Effect of modelling $\pi(Y, Z; \alpha)$ on efficiency of IPW estimator

Result 4. *Aymptotically, the overfitted models for the missingness probability yield the more efficient IPW estimators.*

In this section we consider IPW estimators obtained under various models for $\pi(Y, Z; \alpha)$, and compare their asymptotic variances. For clarity, denote the IPW estimator using the true $\pi(Y, Z)$ by $\hat{\beta}_w(\pi)$, and using the estimated $\pi(Y, Z)$ by $\hat{\beta}_w(\hat{\pi})$. Consider the case where $\pi(Y, Z)$ is known and only depends on Z, say $\pi(Z)$. This happens in a two-stage study design where the study sub-sample is randomly selected given stratum Z for the second stage. Even when $\pi(Z)$ is known, it can be estimated. Comparing the variance formulae (2.7) and (2.8) we see that the variance of $\hat{\beta}_w$ using the true $\pi(Z)$ is greater than or equal to that using the estimated $\pi(Z)$. So by estimating $\pi(Z)$ the efficiency of $\hat{\beta}_w$ can be improved. However, $\pi(Z)$ can be estimated using various models. We further distinguish the estimated observation probability as follows: we write $\hat{\pi}_Z$ if $\pi(Z)$ is estimated using Z as a covariate, $\hat{\pi}_{YZ}$ if estimated using Y and Z, and finally $\hat{\pi}_S$ if estimated via the saturated model including the interaction term between Y and Z. If the model for the observation probability does not involve Y, Ω_{12} in (2.8) would be 0 and hence the asymptotic variance of $\hat{\beta}(\hat{\pi}_Z)$ is the same as that of using true $\pi(Z)$. If the model includes Y, we have non-zero Ω_{12} and $\mathrm{Var}\{\hat{\beta}_w(\hat{\pi}_{YZ})\} \leq \mathrm{Var}\{\hat{\beta}_w(\hat{\pi}_Z)\} = \mathrm{Var}\{\hat{\beta}_w(\pi)\}$. In addition, by Result 2, we have $\mathrm{Var}\{\hat{\beta}_w(\hat{\pi}_S)\} \leq \mathrm{Var}\{\hat{\beta}_w(\hat{\pi}_{YZ})\}$. This implies that even under MCAR, including Y in the model for the observation probability is a good practice since it improves the efficiency. In summary, asymptotically we have

$$\mathrm{Var}\{\hat{\beta}_w(\hat{\pi}_S)\} \leq \mathrm{Var}\{\hat{\beta}_w(\hat{\pi}_{YZ})\} \leq \mathrm{Var}\{\hat{\beta}_w(\hat{\pi}_Z)\} = \mathrm{Var}\{\hat{\beta}_w(\pi)\}.$$

Note that the IPW estimators we discuss in this section are the ones using the estimated $\pi(Y, Z)$, and the IPW estimator in Section 3.2 is obtained using the true $\pi(Y, Z)$.

3.5. Comparison of IPW and conditional method

Result 5. *When Y is binary and a logistic model for Y is used, the conditional likelihood method using the true $\pi(Y, Z)$ performs asymptotically better than the*

IPW method in the sense that the difference between the asymptotic variances of the IPW estimator and the conditional estimator is positive semidefinite.

Proof. For the logistic model, $S(Y|X, Z; \beta) = W\{Y - \mu(X, Z; \beta)\}$, $I_v = E[W\mu(X, Z; \beta)\{1 - \mu(X, Z; \beta)\}W^T]$, and thus the asymptotic variance of $\sqrt{n}\hat{\beta}_w$ in (2.7) when $\pi(Y, Z)$ is known is

$$\begin{aligned} V_{wt} &= I_v^{-1} E\Big[\frac{1}{\pi(Y,Z)} S(Y|X,Z;\beta) S(Y|X,Z;\beta)^T\Big] I_v^{-1} \\ &= I_v^{-1} E\Big[W\mu(X,Z;\beta)\{1-\mu(X,Z;\beta)\}W^T \\ &\quad \frac{\pi(0,Z)\{1-\mu(X,Z;\beta)\} + \pi(1,Z)\mu(X,Z;\beta)}{\pi(1,Z)\pi(0,Z)}\Big] I_v^{-1}. \end{aligned}$$

To facilitate the comparison between the variances of IPW and conditional estimators, we write
$d = [\pi(1, Z)\mu(X, Z; \beta) + \pi(0, Z)\{1 - \mu(X, Z; \beta)\}]/\{\pi(1, Z)\pi(0, Z)\}$,
$c_1 = W\sqrt{\mu(X, Z; \beta)\{1 - \mu(X, Z; \beta)\}d}$, $c_2 = W\sqrt{\mu(X, Z; \beta)\{1 - \mu(X, Z; \beta)\}/d}$, $c = c_1 - E(c_1c_2^T)\{E(c_2c_2^T)\}^{-1}c_2$. Then $V_{wt} = \{E(c_1c_2^T)\}^{-1}E(c_1c_1^T)\{E(c_1c_2^T)\}^{-1}$, and when $\pi(Y, Z)$ is known, the asymptotic variance of $\sqrt{n}\hat{\beta}_c$ is $V_{ct} = I_c^{-1} = \{E(c_2c_2^T)\}^{-1}$. Thus

$$\begin{aligned} V_{wt} - V_{ct} &= I_v^{-1} E\Big[\frac{1}{\pi(Y,Z)} S(Y|X,Z;\beta) S(Y|X,Z;\beta)^T\Big] I_v^{-1} - I_c^{-1} \\ &= \{E(c_1c_2^T)\}^{-1} E(cc^T)\{E(c_2c_1^T)\}^{-1}, \end{aligned}$$

which is positive semidefinite. Equality holds when $c = 0$, which occurs when $\pi(Y, Z)$ is a constant, and in that case both methods are equivalent to the CC analysis.

Thus in the case that $\pi(Y, Z)$ is known, the conditional method yields estimators with smaller asymptotic variances than the inverse probability weighting method under MAR.

When $\pi(Y, Z)$ is estimated, simulation results (Table 1 and Table 2) show that, using the same model for the observation probability, the variance of the conditional estimate is similar to or smaller than that of the IPW estimate.

As pointed out in Section 2.6, the conditional analysis for non-binary Y requires a distributional assumption, and yields a complicated form of asymptotic variance. The results in this section do not apply in the settings other than logistic regression, such as binary regression with other link functions or linear regression models.

4. Simulation

We conducted simulation studies with 500 replications to examine the performances of the non-likelihood approaches with sample sizes $n = 100$ and

500, representing small and large sample sizes. Two models were considered: logistic and classical linear models. For logistic regression models, we generated Y according to $P(Y = 1|Z, X; \beta) = \text{logit}^{-1}(\beta_0 + \beta_Z Z + \beta_X X)$, where X is a binary variable with $P(X = 1) = 0.5$. For Z, two types were considered: a standard normal variable and a binary variable with $P(Z = 1) = 0.5$. For the binary Z case, $\beta = (\beta_0, \beta_Z, \beta_X) = (-0.5, \log 2, \log 2)$; for the normal Z case, $\beta = (-0.5, 0.2, \log 2)$. For linear models, Y is generated from a normal distribution with mean $\beta_0 + \beta_Z Z + \beta_X X$ and variance 1, where $\beta = (\beta_0, \beta_Z, \beta_X) = (-1, 0.5, 0.5)$, and X and Z are generated as binary variables with $P(X = 1) = P(Z = 1) = 0.5$. For both linear and logistic regression cases, X and Z are generated independently.

For all models we generated the observation indicator with $\pi(Y, Z; \alpha) = P(R = 1|Z, Y; \alpha) = \text{logit}^{-1}(\alpha_0 + \alpha_Z Z + \alpha_Y Y)$. Under MCAR, we set $\alpha = (\alpha_0, \alpha_Z, \alpha_Y) = (-0.5, \log 2, 0)$ for logistic models with binary Z, $(0.2, 0.3, 0)$ for logistic models with normal Z, and $(0, \log 2, 0)$ for linear models. Under MAR, we set $\alpha = (\alpha_0, \alpha_Z, \alpha_Y) = (-0.5, \log 2, 0.5)$ for logistic models with binary Z, $(0.2, 0.3, \log 2)$ for logistic models with normal Z, and $(0, \log 2, 0.2)$ for linear models.

We considered three nested models for $\pi(Y, Z; \alpha)$. The factors included in these models were Z only (Model 0), Z and Y main effects only (Model 1), and Z, Y and their interaction (Model 2). Under MAR, these three models represent under-specified, correctly specified, and over-specified models, respectively; and under MCAR, Model 0 is correctly specified, and Models 1 and 2 are over-specified. The corresponding IPW estimates using the estimated $\pi(Y, Z)$ from these models are denoted by IPW0, IPW1, and IPW2, respectively. Note that under MAR, the estimated $\pi(Y, Z)$ from Model 0 is not consistent, thus IPW0 is not consistent. To calculate the imputation estimates (Impu) and the improved augmented IPW estimates (IPWeff), we estimated $P(X = 1|Z, Y)$ by fitting a logistic model with Z, Y and their interaction term. Note that when Z is normal, this imputation model is mis-specified. The IPW estimates using the true $\pi(Y, Z)$ (IPWt), the complete case estimates (CC) and the maximum likelihood estimates from the full data (Full) were also calculated. For logistic models, we also calculated the conditional estimates using the true $\pi(Y, Z)$ (Cont), and using the estimated $\pi(Y, Z)$ from Model 1 (Con1) and Model 2 (Con2).

The first half of the Tables 1, 2 and 3 present the simulation results under MCAR for logistic models with binary Z, continuous Z, and for linear models, respectively.

First note that under MCAR, all the estimates have negligible bias. In both logistic and linear models, the simulation standard deviations of the IPW estimates using the true observation probability $\pi(Y, Z)$ (IPWt) are bigger than those of the CC estimates, consistent with Result 1 in Section 3.2.

Table 1. Simulation results for the logistic model, binary Z. Parameter $(\beta_0, \beta_Z, \beta_X)$ being estimated in the logistic regression model $P(Y = 1|Z, X) = \text{logit}^{-1}(\beta_0 + \beta_Z Z + \beta_X X)$ is $(-0.5, \log 2, \log 2)$; Z is a binary variable with probability 0.5; the probability of observing X is $P(R = 1|Z, Y) = \text{logit}^{-1}(\alpha_0 + \alpha_Z Z + \alpha_Y Y)$. Sample bias (Bias), which is the difference between the sample mean and the true values, simulation standard deviation (SD), the mean of estimated standard error (MSD), and the proportion of the 95 percent confidence interval containing the true parameter (CP) over 500 replicates are given. The sample size is n, and n_o is the average of the number of observed records in 500 replicates.

		Full	CC	IPWt	IPW0	IPW1	IPW2	Cont	Con1	Con2	IPWff	Impu
$(\alpha_0, \alpha_Z, \alpha_Y) = (-0.5, \log 2, 0)$(MCAR) $n = 100$, $n_o = 46.5$												
β_0	Bias	-0.01	-0.04	-0.05	-0.05	-0.05	-0.06	-0.04	-0.04	-0.04	-0.06	-0.06
	SD	0.375	0.658	0.667	0.675	0.584	0.493	0.658	0.563	0.471	0.493	0.493
	MSD	0.368	0.611	0.607	0.606	0.544	0.466	0.611	0.541	0.459	0.466	0.466
	CP	0.97	0.96	0.94	0.95	0.96	0.96	0.96	0.96	0.97	0.96	0.96
β_Z	Bias	-0.00	0.02	0.02	0.03	0.02	0.02	0.02	0.01	0.01	0.02	0.02
	SD	0.435	0.706	0.711	0.713	0.695	0.467	0.706	0.688	0.459	0.467	0.467
	MSD	0.426	0.668	0.667	0.668	0.660	0.461	0.668	0.658	0.457	0.461	0.461
	CP	0.95	0.94	0.94	0.94	0.95	0.95	0.94	0.95	0.95	0.95	0.95
β_X	Bias	0.04	0.10	0.11	0.12	0.12	0.12	0.10	0.10	0.10	0.12	0.12
	SD	0.416	0.666	0.681	0.694	0.696	0.699	0.666	0.666	0.666	0.699	0.699
	MSD	0.426	0.659	0.669	0.669	0.669	0.669	0.659	0.654	0.652	0.669	0.669
	CP	0.96	0.97	0.97	0.96	0.96	0.96	0.97	0.96	0.96	0.96	0.96
$n = 500$,$n_o = 230.3$												
β_0	Bias	-0.01	0.01	0.01	0.01	0.00	-0.00	0.01	0.00	-0.00	-0.00	-0.00
	SD	0.158	0.264	0.265	0.265	0.231	0.202	0.264	0.231	0.201	0.202	0.202
	MSD	0.161	0.256	0.257	0.257	0.229	0.197	0.256	0.227	0.195	0.197	0.197
	CP	0.96	0.94	0.94	0.94	0.96	0.94	0.94	0.96	0.94	0.94	0.94
β_Z	Bias	0.02	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.02	0.02	0.02
	SD	0.192	0.283	0.283	0.283	0.278	0.196	0.283	0.278	0.196	0.196	0.196
	MSD	0.186	0.281	0.281	0.281	0.279	0.192	0.281	0.279	0.192	0.192	0.192
	CP	0.94	0.96	0.96	0.96	0.96	0.95	0.96	0.96	0.95	0.95	0.95
β_X	Bias	0.00	0.00	-0.00	-0.00	-0.00	-0.00	0.00	0.00	0.00	-0.00	-0.00
	SD	0.186	0.292	0.297	0.297	0.297	0.297	0.292	0.292	0.292	0.297	0.297
	MSD	0.186	0.278	0.282	0.283	0.283	0.283	0.278	0.277	0.277	0.283	0.283
	CP	0.95	0.94	0.95	0.95	0.94	0.94	0.94	0.94	0.94	0.94	0.94
$(\alpha_0, \alpha_Z, \alpha_Y) = (-0.5, \log 2, 0.5)$(MAR) $n = 100$,$n_o = 53.1$												
β_0	Bias	-0.01	0.27	-0.01	0.26	-0.03	-0.03	-0.01	-0.03	-0.03	-0.03	-0.03
	SD	0.375	0.595	0.600	0.611	0.524	0.458	0.595	0.511	0.444	0.458	0.458
	MSD	0.368	0.563	0.559	0.561	0.502	0.440	0.563	0.501	0.436	0.440	0.440
	CP	0.97	0.93	0.94	0.92	0.95	0.95	0.95	0.96	0.96	0.95	0.95
β_Z	Bias	-0.00	-0.07	0.01	-0.07	0.01	0.01	0.01	0.02	0.01	0.01	0.01
	SD	0.435	0.653	0.655	0.658	0.638	0.452	0.653	0.635	0.448	0.452	0.452
	MSD	0.426	0.627	0.626	0.627	0.616	0.450	0.627	0.615	0.448	0.450	0.450
	CP	0.95	0.94	0.94	0.94	0.95	0.95	0.94	0.95	0.95	0.95	0.95
β_X	Bias	0.04	0.08	0.09	0.10	0.10	0.10	0.08	0.08	0.08	0.10	0.10
	SD	0.416	0.629	0.644	0.659	0.656	0.659	0.629	0.629	0.629	0.659	0.659
	MSD	0.426	0.623	0.628	0.633	0.630	0.630	0.623	0.619	0.617	0.630	0.630
	CP	0.96	0.96	0.96	0.96	0.96	0.95	0.96	0.96	0.96	0.95	0.95
$n = 500$,$n_o = 263.2$												
β_0	Bias	-0.01	0.29	0.01	0.29	-0.00	-0.01	0.01	-0.00	-0.01	-0.01	-0.01
	SD	0.158	0.242	0.242	0.242	0.216	0.190	0.242	0.216	0.190	0.190	0.190
	MSD	0.161	0.238	0.239	0.239	0.214	0.188	0.238	0.213	0.187	0.188	0.188
	CP	0.96	0.75	0.96	0.76	0.96	0.94	0.96	0.96	0.94	0.94	0.94
β_Z	Bias	0.02	-0.07	0.01	-0.07	0.01	0.02	0.01	0.01	0.02	0.02	0.02
	SD	0.192	0.270	0.270	0.270	0.265	0.196	0.270	0.265	0.196	0.196	0.196
	MSD	0.186	0.265	0.265	0.265	0.262	0.190	0.265	0.262	0.190	0.190	0.190
	CP	0.94	0.95	0.96	0.95	0.96	0.94	0.96	0.96	0.94	0.94	0.94
β_X	Bias	0.00	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	SD	0.186	0.278	0.282	0.283	0.282	0.282	0.278	0.278	0.278	0.282	0.282
	MSD	0.186	0.264	0.267	0.268	0.267	0.267	0.264	0.264	0.264	0.267	0.267
	CP	0.95	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94

Table 2. Simulation result for the logistic model, normal Z. Parameter $(\beta_0, \beta_Z, \beta_X)$ being estimated in the logistic regression model $P(Y = 1|Z, X) = \text{logit}^{-1}(\beta_0 + \beta_Z Z + \beta_X X)$ is $(-0.5, 0.2, \log 2)$; Z is a standard normal variable; the probability of observing X is $P(R = 1|Z, Y) = \text{logit}^{-1}(\alpha_0 + \alpha_Z Z + \alpha_Y Y)$. Sample bias (Bias), which is the difference between the sample mean and the true values, simulation standard deviation (SD), the mean of the estimated standard error (MSD), and the proportion of the 95 percent confidence interval containing the true parameter (CP) over 500 replicates are given. The sample size is n, and n_o is the average of the number of observed records in 500 replicates.

		Full	CC	IPWt	IPW0	IPW1	IPW2	Cont	Con1	Con2	IPWff	Impu
$(\alpha_0, \alpha_Z, \alpha_Y) = (0.2, 0.3, 0)$(MCAR) $n = 100$, $n_o = 54.6$												
β_0	Bias	-0.03	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.03	-0.03	-0.03
	SD	0.327	0.452	0.457	0.458	0.407	0.408	0.452	0.401	0.400	0.405	0.402
	MSD	0.304	0.423	0.424	0.424	0.378	0.375	0.423	0.375	0.372	0.374	0.374
	CP	0.94	0.94	0.94	0.94	0.95	0.95	0.94	0.96	0.96	0.96	0.96
β_Z	Bias	0.01	0.01	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01
	SD	0.217	0.311	0.317	0.324	0.320	0.244	0.311	0.306	0.238	0.225	0.223
	MSD	0.217	0.308	0.305	0.302	0.300	0.227	0.308	0.298	0.230	0.224	0.224
	CP	0.97	0.97	0.95	0.93	0.94	0.95	0.97	0.95	0.97	0.96	0.97
β_X	Bias	0.07	0.07	0.08	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07
	SD	0.444	0.616	0.620	0.618	0.618	0.621	0.616	0.616	0.616	0.623	0.618
	MSD	0.422	0.586	0.590	0.589	0.589	0.588	0.586	0.583	0.582	0.590	0.589
	CP	0.95	0.95	0.95	0.94	0.94	0.94	0.95	0.95	0.95	0.94	0.94
$n = 500$,$n_o = 274.3$												
β_0	Bias	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
	SD	0.130	0.181	0.183	0.183	0.162	0.160	0.181	0.160	0.158	0.160	0.160
	MSD	0.132	0.179	0.180	0.180	0.160	0.159	0.179	0.159	0.158	0.159	0.159
	CP	0.96	0.94	0.94	0.94	0.95	0.95	0.94	0.96	0.96	0.95	0.95
β_Z	Bias	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	SD	0.096	0.131	0.133	0.133	0.132	0.101	0.131	0.130	0.101	0.098	0.098
	MSD	0.093	0.128	0.130	0.130	0.129	0.097	0.128	0.127	0.097	0.095	0.095
	CP	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
β_X	Bias	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
	SD	0.183	0.248	0.250	0.250	0.250	0.251	0.248	0.248	0.248	0.251	0.250
	MSD	0.184	0.250	0.252	0.252	0.252	0.252	0.250	0.249	0.249	0.252	0.252
	CP	0.96	0.95	0.95	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95
$(\alpha_0, \alpha_Z, \alpha_Y) = (0.2, 0.3, \log 2)$(MAR) $n = 100$,$n_o = 61.9$												
β_0	Bias	-0.03	0.24	-0.02	0.24	-0.02	-0.02	-0.02	-0.02	-0.02	-0.03	-0.02
	SD	0.327	0.430	0.435	0.438	0.392	0.393	0.430	0.387	0.386	0.388	0.386
	MSD	0.304	0.388	0.390	0.390	0.353	0.351	0.388	0.351	0.349	0.353	0.351
	CP	0.94	0.87	0.95	0.86	0.95	0.94	0.95	0.95	0.95	0.95	0.95
β_Z	Bias	0.01	-0.03	0.01	-0.04	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	SD	0.217	0.299	0.301	0.307	0.301	0.238	0.299	0.295	0.235	0.223	0.222
	MSD	0.217	0.285	0.284	0.282	0.278	0.221	0.285	0.276	0.223	0.222	0.222
	CP	0.97	0.96	0.96	0.94	0.95	0.95	0.97	0.96	0.97	0.97	0.97
β_X	Bias	0.07	0.06	0.06	0.06	0.06	0.07	0.06	0.06	0.06	0.07	0.06
	SD	0.444	0.584	0.587	0.591	0.590	0.590	0.584	0.584	0.584	0.590	0.588
	MSD	0.422	0.547	0.549	0.550	0.549	0.548	0.547	0.545	0.544	0.552	0.549
	CP	0.95	0.94	0.94	0.93	0.93	0.94	0.94	0.94	0.94	0.94	0.94
$n = 500$,$n_o = 310.3$												
β_0	Bias	-0.01	0.25	-0.01	0.25	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
	SD	0.130	0.167	0.167	0.168	0.147	0.145	0.167	0.146	0.145	0.145	0.145
	MSD	0.132	0.166	0.166	0.166	0.150	0.150	0.166	0.150	0.149	0.150	0.150
	CP	0.96	0.66	0.95	0.66	0.95	0.96	0.95	0.95	0.95	0.96	0.96
β_Z	Bias	0.01	-0.04	0.01	-0.04	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	SD	0.096	0.127	0.129	0.129	0.127	0.100	0.127	0.125	0.100	0.098	0.098
	MSD	0.093	0.119	0.120	0.121	0.119	0.095	0.119	0.117	0.095	0.094	0.094
	CP	0.94	0.93	0.94	0.92	0.94	0.93	0.94	0.93	0.94	0.94	0.94
β_X	Bias	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
	SD	0.183	0.233	0.233	0.234	0.233	0.233	0.233	0.233	0.233	0.233	0.233
	MSD	0.184	0.234	0.235	0.236	0.235	0.235	0.234	0.234	0.234	0.235	0.235
	CP	0.96	0.95	0.95	0.95	0.95	0.95	0.95	0.94	0.95	0.95	0.95

Table 3. Simulation results for the linear model. The parameter $(\beta_0, \beta_Z, \beta_X)$ being estimated in the classical linear model $E(Y|Z,X) = \beta_0 + \beta_Z Z + \beta_X X$, is $(-1, 0.5, 0.5)$; the probability of observing X is $P(R=1|Z,Y) = \text{logit}^{-1}(\alpha_0+\alpha_Z Z+\alpha_Y Y)$. Sample bias (Bias), which is the difference between the sample mean and the true values, simulation standard deviation (SD), the mean of estimated standard error (MSD), and the proportion of the 95 percent confidence interval containing the true parameter (CP) over 500 replicates are given. The sample size is n, and n_o is the average of the number of observed records in 500 replicates.

		Full	CC	IPWt	IPW0	IPW1	IPW2	IPWeff	Impu
$(\alpha_0, \alpha_Z, \alpha_Y) = (0, \log 2, 0)$(MCAR) $n = 100$, $n_o = 54.4$									
β_0	Bias	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02
	SD	0.177	0.245	0.247	0.247	0.225	0.209	0.210	0.209
	MSD	0.175	0.243	0.235	0.235	0.215	0.195	0.195	0.195
	CP	0.94	0.94	0.92	0.92	0.93	0.92	0.91	0.92
β_Z	Bias	-0.00	0.00	0.00	0.00	0.00	-0.01	-0.01	-0.01
	SD	0.199	0.261	0.261	0.260	0.260	0.210	0.207	0.207
	MSD	0.202	0.269	0.261	0.261	0.259	0.204	0.205	0.206
	CP	0.95	0.95	0.95	0.95	0.95	0.94	0.94	0.94
β_X	Bias	-0.01	-0.03	-0.03	-0.03	-0.02	-0.02	-0.02	-0.02
	SD	0.202	0.267	0.271	0.272	0.275	0.281	0.282	0.278
	MSD	0.202	0.266	0.261	0.260	0.261	0.261	0.258	0.259
	CP	0.95	0.95	0.94	0.94	0.94	0.93	0.92	0.93
$n = 500$,$n_o = 291.3$									
β_0	Bias	-0.00	-0.00	-0.00	-0.00	-0.00	-0.01	-0.01	-0.01
	SD	0.075	0.108	0.108	0.108	0.099	0.089	0.088	0.088
	MSD	0.077	0.107	0.107	0.107	0.098	0.088	0.087	0.087
	CP	0.95	0.96	0.96	0.96	0.96	0.94	0.95	0.95
β_Z	Bias	-0.00	-0.01	-0.01	-0.01	-0.01	-0.00	-0.00	-0.00
	SD	0.087	0.117	0.117	0.117	0.116	0.088	0.088	0.088
	MSD	0.089	0.119	0.118	0.118	0.117	0.091	0.091	0.091
	CP	0.97	0.96	0.95	0.95	0.95	0.96	0.96	0.96
β_X	Bias	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01
	SD	0.094	0.123	0.124	0.124	0.124	0.124	0.122	0.122
	MSD	0.089	0.117	0.118	0.118	0.118	0.118	0.115	0.115
	CP	0.94	0.94	0.95	0.95	0.95	0.94	0.94	0.94
$(\alpha_0, \alpha_Z, \alpha_Y) = (0, \log 2, 0.2)$(MAR) $n = 100$,$n_o = 55.8$									
β_0	Bias	0.01	0.13	0.03	0.13	0.03	0.03	0.02	0.02
	SD	0.177	0.253	0.263	0.255	0.235	0.215	0.211	0.205
	MSD	0.175	0.254	0.245	0.242	0.221	0.197	0.196	0.197
	CP	0.94	0.92	0.92	0.90	0.92	0.92	0.93	0.93
β_Z	Bias	-0.00	-0.04	-0.00	-0.04	-0.01	-0.01	-0.00	-0.00
	SD	0.199	0.271	0.277	0.272	0.274	0.212	0.209	0.209
	MSD	0.202	0.276	0.270	0.267	0.267	0.206	0.207	0.207
	CP	0.95	0.95	0.95	0.95	0.94	0.93	0.95	0.95
β_X	Bias	-0.01	-0.03	-0.03	-0.03	-0.03	-0.02	-0.02	-0.02
	SD	0.202	0.273	0.285	0.281	0.291	0.297	0.296	0.284
	MSD	0.202	0.272	0.270	0.267	0.269	0.268	0.264	0.266
	CP	0.95	0.94	0.93	0.93	0.93	0.91	0.91	0.93
$n = 500$,$n_o = 278.9$									
β_0	Bias	-0.00	0.10	-0.00	0.10	-0.00	-0.00	-0.00	-0.00
	SD	0.075	0.110	0.114	0.111	0.103	0.091	0.087	0.087
	MSD	0.077	0.111	0.113	0.111	0.102	0.091	0.088	0.087
	CP	0.95	0.85	0.94	0.84	0.95	0.95	0.95	0.95
β_Z	Bias	-0.00	-0.04	-0.00	-0.04	-0.00	-0.00	-0.00	-0.00
	SD	0.087	0.120	0.122	0.120	0.120	0.090	0.089	0.089
	MSD	0.089	0.121	0.123	0.120	0.121	0.092	0.092	0.092
	CP	0.97	0.95	0.95	0.95	0.95	0.96	0.96	0.96
β_X	Bias	0.00	-0.00	0.01	-0.00	0.01	0.01	0.01	0.01
	SD	0.094	0.123	0.127	0.125	0.129	0.129	0.123	0.122
	MSD	0.089	0.119	0.123	0.120	0.123	0.123	0.119	0.119
	CP	0.94	0.95	0.96	0.95	0.96	0.95	0.95	0.95

In assessing the effect of over-specification of the observation models, the estimates using the most over-fitted models (IPW2) overall show smaller variance than those using the second most over-fitted models (IPW1), which in turn show smaller variances than the estimates using the correctly specified models (IPW0), or the estimates using the true model assuming that it is known (IPWt). Note that under MCAR, IPW0 gives valid estimates, and is as efficient as the IPWt, consistent with Result 4 in Section 3.4. However, for some cases with $n = 100$, with small observation probability (e.g., $\alpha_0 = -1$ for the logistic model, binary Z, MCAR case, not shown in the table), the estimates from more complicated observation models are unstable due to non-convergence problems. This echoes the speculation of a referee that the benefit of the over-fitted observation models may not be present when the estimates of the observation model are unstable for finite sample sizes.

The imputation estimate generally has no bigger variance than all the IPW estimates. Note that when all variables are categorical, the imputation estimate (Impu) and the IPW2 estimate become identical. And the Impu is the same as the improved augmented IPW estimates (IPWeff), consistent with Results 2 and 3 in Section 3.3. When Z is continuous, the imputation estimates generally have smaller variances than the IPW estimates. These trends are notable in the estimates for β_0 and β_Z, but not for β_X, in which case, the efficiencies of the two approaches are very similar.

Note that for logistic models under MCAR, the conditional method using the true $\pi(Z)$ reduces to the complete case method. The tables show that the conditional estimates (Cont) have smaller variances than the IPWt, consistent with Result 5 in Section 3.5.

The second half of the Tables 1, 2 and 3 show the results under MAR for logistic models with binary Z, continuous Z, and for linear models, respectively.

First, note that under MAR, the CC estimates and the IPW0 estimates using the under-specified observation model are biased. Other than these, the tables show similar results as in the MCAR case, confirming the theoretical findings shown in Results 2-5. Although we do not have analytic comparison of efficiencies between the conditional and the IPW estimators when $\pi(Y, Z)$ is estimated, comparisons between Con1 and IPW1, Con2 and IPW2, the conditional and the IPW estimates that use the same estimated $\pi(Y, Z)$, suggest that the conditional estimates have comparable, or smaller, variances than the corresponding IPW estimates.

5. Analysis of Data from a Stroke Study

The Northern Manhattan Stroke Study (NOMASS) is a prospective study whose main goal is to identify the risk factors for stroke (Sacco, Boden-Albala,

Gan, Chen, Kargman, Shea, Paik and Hauser (1998)). We analyzed two years of follow-up data for 3,202 subjects who were stroke-free at the beginning of follow-up. The outcome of the analysis is an indicator of stroke or stroke-related death during the first two years. Fifty seven subjects had a positive outcome. Covariates included hypertension (HTN, 74%), diabetes (22%), moderate alcohol drinking (MALCOHOL, 33%), smoking status (non-smoker (47%), former smoker (38%) and current smoker (15%)), age (cut off at 65, 65% older than 65), gender (37% male) , race (white 21% , black 25%, Hispanic 54%) and education (45% completed high school). Another important covariate, abnormal left ventricular hypertrophy (ABNORMLV) was observed only for 2,055 subjects, and was missing for 1,147 subjects.

We fitted logistic models handling missing covariates by the IPW, imputation and conditional approaches. For the probability of observing ABNORMLV, we used logistic regression models. Also, the imputation models for ABNORMLV were specified as logistic models. As one of the covariates of the model for the probability of observation, we used the stroke indicator, allowing the case of MAR.

Table 4 shows the estimates and standard errors (SE) of auxiliary models, namely, the observation models and imputation models. In the observation model, we found that the subjects with stroke or stroke death are more likely to have missing ABNORMLV values, indicating the data are missing at random, rather than missing completely at random. The estimates from the two nested observation models are similar. Model R2 represents an over-fitted model. For imputation models, gender, age, race, hypertension, and diabetes are strong predictors of ABNORMLV. Model X2 has smoking variables as additional covariates, but the effect of smoking is weak. Model X2 represents an over-fitted imputation model.

Table 5 lists the estimates from the complete case analysis (CC), the inverse probability weighting estimates using the estimated probability of observation from model R1 (IPW1) and model R2 (IPW2), the imputation estimates using imputation model X1 (Impu1) and model X2 (Impu2), and the improved augmented IPW (IPWeff). For the improved augmented IPW estimate, we used the estimated probability of observation from model R1 and the imputation model X1. Various models yielded qualitatively similar results, except that the age variable is significant under the imputation methods and the improved augmented IPW (IPWeff), and is nonsignificant under other methods. In addition, the imputation estimates and IPWeff reveal a positive effect of hypertension and the diminished risk of race (black). For all models, the odds of having a stroke within two years among diabetics is more than 2.4 times than for those without diabetes. The results also show that smoking (former smoker or current smoker)

and old age are risk factors for stroke, while higher education and moderate alcohol drinking are moderate protective factors for stroke.

Table 4. Simulation results for the logistic model, binary Z. Parameter $(\beta_0, \beta_Z, \beta_X)$ being estimated in the logistic regression model $P(Y = 1|Z, X) = \text{logit}^{-1}(\beta_0 + \beta_Z Z + \beta_X X)$ is $(-0.5, \log 2, \log 2)$; Z is a binary variable with probability 0.5; the probability of observing X is $P(R = 1|Z, Y) = \text{logit}^{-1}(\alpha_0 + \alpha_Z Z + \alpha_Y Y)$. Sample bias (Bias), which is the difference between the sample mean and the true values, simulation standard deviation (SD), the mean of estimated standard error (MSD), and the proportion of the 95 percent confidence interval containing the true parameter (CP) over 500 replicates are given. The sample size is n, and n_o is the average of the number of observed records in 500 replicates.

	Probability of observation model				Imputation model			
	Model R1		Model R2		Model X1		Model X2	
	Biasimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	0.521	0.151	0.228	0.181	-1.747	0.165	-1.783	0.173
Male	0.223	0.083	0.217	0.083	0.601	0.097	0.588	0.100
Age65	-0.526	0.123	-0.516	0.124	0.323	0.101	0.323	0.102
Black	0.065	0.135	0.097	0.137	0.520	0.145	0.520	0.146
Hisp	0.421	0.121	0.564	0.134	-0.064	0.127	-0.057	0.128
Hedu			0.254	0.090				
MALCOHOL	0.823	0.169	0.828	0.170				
Smoker1	0.095	0.159	0.086	0.160			0.075	0.106
Smoker2	-0.440	0.171	-0.431	0.171			0.064	0.146
HTN			0.102	0.089	0.798	0.118	0.799	0.118
Diabetes			0.076	0.093	0.320	0.113	0.316	0.113
Stroke	-0.559	0.273	-0.566	0.275	0.683	0.406	0.676	0.407
Age65*Smoker1	0.001	0.186	0.010	0.186				
Age65*Smoker2	0.378	0.223	0.369	0.223				
Black*MALCOHOL	-0.760	0.229	-0.778	0.230				
Hisp*MALCOHOL	-0.582	0.208	-0.596	0.209				

In terms of precision, the imputation estimate has smaller standard errors than the IPW estimates. The standard errors of the imputation and the improved augmented IPW (IPWeff) estimate are similar. Using the same specified probability of observation model, the conditional estimates generally show smaller standard errors than the IPW estimates. Note that Model R1 for the probability of observation is nested in Model R2, but since the additional terms in model R2 involve only the covariates and not the outcome of the main analysis, the efficiency of IPW2 is not improved over IPW1. The imputation estimates from the two imputation models are quite similar.

The results from this data analysis are mostly consistent with the asymptotic results in Section 3. Although the estimators using the over-fitted missingness

models have smaller variances asymptotically for the IPW methods, the example does not demonstrate an advantage of IPW2 over IPW1. This is possibly due to small cell counts in some variables.

Table 5. Estimated coefficients and standard errors for the probability of having a stroke within two years of follow-up, $\text{logit}(\mu_{it}) = \beta_0 + \beta_1 Male_i + \beta_2 Age65_i + \beta_3 Black_i + \beta_4 Hisp_i + \beta_5 Hedu_i + \beta_6 HTN_i + \beta_7 Diabetes_i + \beta_8 Malcohol_i + \beta_9 Smoker1_i + \beta_{10} Smoker2_i + \beta_{11} Abnormlv_i$. In each cell, the first row is the point estimate, the second row is the standard error; * indicates significant at 0.05 level.

param	CC	IPW1	IPW2	Con1	Con2	Impu1	Impu2	IPWeff
Intercept	-6.007*	-5.550*	-5.486*	-5.777*	-5.736*	-5.567*	-5.563*	-5.542*
	0.956	1.116	1.118	1.102	1.098	0.709	0.708	0.711
Male	-0.229	-0.317	-0.342	-0.258	-0.257	0.002	0.004	0.020
	0.426	0.407	0.409	0.399	0.399	0.308	0.308	0.308
Age65	0.783	0.862	0.873	0.837	0.836	1.226*	1.227*	1.238*
	0.482	0.498	0.500	0.491	0.490	0.426	0.426	0.423
Black	0.545	0.429	0.389	0.572	0.568	-0.149	-0.148	-0.130
	0.626	0.635	0.641	0.622	0.622	0.389	0.389	0.389
Hisp	0.118	-0.110	-0.167	0.091	0.072	-0.256	-0.257	-0.261
	0.672	0.734	0.748	0.708	0.707	0.436	0.436	0.436
Hedu	-0.042	-0.163	-0.205	-0.041	-0.075	-0.361	-0.362	-0.354
	0.468	0.502	0.508	0.495	0.494	0.355	0.355	0.356
HTN	-0.010	0.098	0.102	-0.008	-0.021	0.394	0.395	0.417
	0.521	0.517	0.519	0.513	0.513	0.403	0.403	0.403
Diabetes	1.240*	1.140*	1.155*	1.238*	1.228*	0.883*	0.883*	0.897*
	0.400	0.413	0.412	0.408	0.408	0.281	0.281	0.281
MALCOHOL	-0.316	-0.225	-0.198	-0.355	-0.355	-0.531	-0.531	-0.540
	0.459	0.442	0.445	0.450	0.450	0.339	0.339	0.336
Smoker1	0.419	0.413	0.405	0.406	0.406	0.123	0.118	0.120
	0.461	0.482	0.483	0.476	0.475	0.329	0.329	0.329
Smoker2	0.891	0.842	0.820	0.913	0.913	0.467	0.463	0.473
	0.556	0.548	0.544	0.553	0.552	0.399	0.399	0.400
ABNORMLV	0.729	0.599	0.580	0.730	0.731	0.696	0.690	0.574
	0.409	0.406	0.405	0.406	0.406	0.404	0.403	0.411

Appendix. Derivation of Variance of IPW and Conditional Estimators

We show asymptotic variance formulas that are expressed differently from the original references to facilitate comparisons.

A.1. Derivation of (2.8) when $\pi(Y, Z)$ is estimated

When $\pi(Y, Z)$ is not known, and a parametric model $\pi(Y, Z; \alpha)$, is assumed, (β, α) can be obtained by solving

$$\begin{pmatrix} S_w(\alpha, \beta) \\ U(\alpha) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n \frac{R_i}{\pi(Y_i, Z_i; \alpha)} S(Y_i | X_i, Z_i; \beta) \\ \sum_{i=1}^n U(R_i | Y_i, Z_i; \alpha) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

where $S(Y_i|X_i, Z_i; \beta)$ and $U(R_i|Y_i, Z_i; \alpha)$ are defined as in (2.1) and (2.2), respectively. Let

$$I(\beta, \alpha) = \begin{pmatrix} -\frac{\partial}{\partial \beta^T}\{S_w(\alpha, \beta)\} & -\frac{\partial}{\partial \alpha^T}\{S_w(\alpha, \beta)\} \\ -\frac{\partial}{\partial \beta^T}\{U(\alpha)\} & -\frac{\partial}{\partial \alpha^T}\{U(\alpha)\} \end{pmatrix},$$
$$\Sigma_w = \begin{pmatrix} E\{S_w(\alpha, \beta)S_w(\alpha, \beta)^T\} & E\{S_w(\alpha, \beta)U(\alpha)^T\} \\ E\{U(\alpha)S(\beta)^T\} & E\{U(\alpha)U(\alpha)^T\} \end{pmatrix}.$$

We have $-(1/n)\partial/\partial\beta^T\{S_w(\alpha, \beta)\} \xrightarrow{p} I_v$, and $-(1/n)\partial/\partial\alpha^T\{S_w(\alpha, \beta)\} = (1/n)$ $\sum_{i=1}^n R_i S(Y_i|X_i, Z_i; \beta)\partial/\partial\alpha^T\{\pi(Y_i, Z_i; \alpha)\}/\{\pi(Y_i, Z_i; \alpha)^2\} \xrightarrow{p}$ $E_{YZ}[S(Y_i|Z_i; \beta)\partial/\partial\alpha^T\{\pi(Y_i, Z_i; \alpha)\}/\pi(Y_i, Z_i; \alpha)] = \Omega_{12}$.

On the other hand, $-(1/n)\partial/\partial\beta\{U(\alpha)\} = 0$, and $-(1/n)\partial/\partial\alpha^T\{U(\alpha)\} =$ $-(1/n)\sum_{i=1}^n \partial/\partial\alpha^T\{U(R_i|Y_i, Z_i; \alpha)\} \xrightarrow{p}$ $E\Big(\partial/\partial\alpha\{\pi(Y_i, Z_i; \alpha)\}\partial/\partial\alpha^T\{\pi(Y_i, Z_i; \alpha)\}/[\pi(Y_i, Z_i; \alpha)\{1 - \pi(Y_i, Z_i; \alpha)\}]\Big) = I_\alpha$. Then

$$-\Big(\frac{1}{n}\Big)I(\beta, \alpha) \xrightarrow{p} \begin{pmatrix} I_v & \Omega_{12} \\ 0 & I_\alpha \end{pmatrix}.$$

Also, since

$$\begin{aligned} &E\Big\{\frac{1}{n}S_w(\alpha, \beta)S_w(\alpha, \beta)^T\Big\} \\ &\quad = E\Big[\frac{R_i}{\pi(Y_i, Z_i; \alpha)^2}S(Y_i|X_i, Z_i; \beta)S_\beta^T(Y_i|X_i, Z_i)\Big] \\ &\quad = E\Big[\frac{1}{\pi(Y_i, Z_i; \alpha)}E\{S(Y_i|X_i, Z_i; \beta)S_\beta^T(Y_i|X_i, Z_i)|Y_i, Z_i\}\Big] \\ &\quad = \Omega_w, \\ &E\Big\{\frac{1}{n}S_w(\alpha, \beta)U(\alpha)^T\Big\} \\ &\quad = E\Big[\frac{R_i}{\pi(Y_i, Z_i; \alpha)}S(Y_i|X_i, Z_i; \beta)U(R_i|Y_i, Z_i; \alpha)^T\Big] \\ &\quad = E\Big[S_\beta(Y_i|Z_i)\frac{\partial}{\partial\alpha^T}\frac{\pi(Y_i, Z_i; \alpha)}{\pi(Y_i, Z_i; \alpha)}\Big] \\ &\quad = \Omega_{12}, \end{aligned}$$

and $E[(1/n)E\{U(\alpha)U(\alpha)^T\}] = E\{U(R_i|Y_i, Z_i; \alpha)U(R_i|Y_i, Z_i; \alpha)^T\} = I_\alpha$, we have

$$\mathrm{Var}\Big[\frac{1}{\sqrt{n}}\{S_w(\alpha, \beta), U(\alpha)\}^T\Big] = \frac{1}{n}\Sigma_w = \begin{pmatrix} \Omega_w & \Omega_{12} \\ \Omega_{12}^T & I_\alpha \end{pmatrix}.$$

Then $\sqrt{n}(\hat{\beta}_w - \beta, \hat{\alpha} - \alpha)^T$ has an asymptotic normal distribution with mean 0 and variance

$$\begin{pmatrix} I_v & \Omega_{12} \\ 0 & I_\alpha \end{pmatrix}^{-1} \begin{pmatrix} \Omega_w & \Omega_{12} \\ \Omega_{12}^T & I_\alpha \end{pmatrix} \begin{pmatrix} I_v & 0 \\ \Omega_{12}^T & I_\alpha \end{pmatrix}^{-1}.$$

Since $\begin{pmatrix} I_v & \Omega_{12} \\ 0 & I_\alpha \end{pmatrix}^{-1} = \begin{pmatrix} I_v^{-1} & -I_v^{-1}\Omega_{12}I_\alpha^{-1} \\ 0 & I_\alpha^{-1} \end{pmatrix}$,
it follows that the asymptotic variance of $\sqrt{n}(\hat{\beta}_w - \beta)$ is

$$I_v^{-1} E\Big[\frac{1}{\pi(Y,Z;\alpha)} S(Y|X,Z;\beta) S(Y|X,Z;\beta)^T\Big] I_v^{-1} - I_v^{-1}\Omega_{12}I_\alpha^{-1}\Omega_{12}^T I_v^{-1}.$$

A.2. Derivation of variance of conditional estimator when $\pi(Y, Z)$ is known

The conditional likelihood given $R = 1$ is

$$L_c(\beta) = \prod_{i=1}^{n} \Big[\eta(X_i, Z_i; \beta)^{Y_i}\{1 - \eta(X_i, Z_i; \beta)\}^{1-Y_i}\Big]^{R_i},$$

where $\eta(X_i, Z_i; \beta) = E(Y_i|X_i, Z_i, R_i = 1; \beta)$.

The score function is

$$S_c(\beta) = \sum_{i=1}^{n} R_i W_i \{Y_i - \eta(X_i, Z_i; \beta)\},$$

and the information matrix is

$$\begin{aligned} I_c &= -E\Big[\frac{1}{n}\frac{\partial}{\partial\beta}\{S_c(\beta)\}\Big] \\ &= E\Big(W_i \eta(X_i, Z_i; \beta)\{1 - \eta(X_i, Z_i; \beta)\} \\ &\quad \Big[\pi(1, Z_i)\mu(X_i, Z_i; \beta) + \pi(0, Z_i)\{1 - \mu(X_i, Z_i; \beta)\}\Big]\Big) \\ &= E\Big[W_i \frac{\pi(1, Z_i)\mu(X_i, Z_i; \beta)\pi(0, Z_i)\{1 - \mu(X_i, Z_i; \beta)\}}{\pi(1, Z_i)\mu(X_i, Z_i; \beta) + \pi(0, Z_i)\{1 - \mu(X_i, Z_i; \beta)\}} W_i^T\Big]. \end{aligned}$$

Thus $\sqrt{n}\hat{\beta}_c$ has the asymptotic variance I_c^{-1}.

A.3. Derivation of (2.10) when $\pi(Y, Z)$ is estimated

As in the weighting method, when $\pi(Y, Z)$ is not known and a parametric model $\pi(Y, Z; \alpha)$ is assumed, (β, α) can be obtained by solving

$$\begin{pmatrix} S_c(\alpha, \beta) \\ U(\alpha) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} R_i W_i \{Y_i - \eta(X_i, Z_i; \beta, \alpha)\} \\ \sum_{i=1}^{n} U(R_i|Y_i, Z_i; \alpha) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Let

$$I(\beta,\alpha)=\begin{pmatrix} -\frac{\partial}{\partial\beta}\{S_c(\alpha,\beta)\} & -\frac{\partial}{\partial\alpha}\{S_c(\alpha,\beta)\} \\ -\frac{\partial}{\partial\beta}\{U(\alpha)\} & -\frac{\partial}{\partial\alpha}\{U(\alpha)\} \end{pmatrix},$$

$$\Sigma_c=\begin{pmatrix} E\{S_c(\alpha,\beta)S_c(\alpha,\beta)^T\} & E\{S_c(\alpha,\beta)U(\alpha)^T\} \\ E\{U(\alpha)S_c(\alpha,\beta)^T\} & E\{U(\alpha)U(\alpha)^T\} \end{pmatrix}.$$

Since $-(1/n)\partial/\partial\beta\{S_c(\alpha,\beta)\} \xrightarrow{p} I_c$,

$$\begin{aligned} &-\Big(\frac{1}{n}\Big)\frac{\partial}{\partial\alpha}\{S_c(\alpha,\beta)\} \\ &=\frac{1}{n}\sum_{i=1}^{n} R_iW_i\frac{\partial}{\partial\alpha}\{\eta(X_i,Z_i;\beta,\alpha)\} \\ &\xrightarrow{p} E\Big[R_iW_i\frac{\partial}{\partial\alpha}\{\eta(X_i,Z_i;\beta,\alpha)\}\Big] \\ &=E\Big(W_i\frac{\pi(1,Z_i;\alpha)\pi(0,Z_i;\alpha)\mu(X_i,Z_i;\beta)\{1-\mu(X_i,Z_i;\beta)\}}{\pi(1,Z_i;\alpha)\mu(X_i,Z_i;\beta)+\pi(0,Z_i;\alpha)\{1-\mu(X_i,Z_i;\beta)\}} \\ &\quad\Big[\frac{\frac{\partial}{\partial\alpha^T}\{\pi(1,Z_i;\alpha)\}}{\pi(1,Z_i;\alpha)}-\frac{\frac{\partial}{\partial\alpha^T}\{\pi(0,Z_i;\alpha)\}}{\pi(0,Z_i;\alpha)}\Big]\Big) \\ &=\Omega_c, \end{aligned}$$

$-(1/n)\partial/\partial\beta\{U(\alpha)\}=0$, and $-(1/n)\partial/\partial\alpha\{U(\alpha)\} \xrightarrow{p} I_\alpha$, we have

$$-\Big(\frac{1}{n}\Big)I(\beta,\alpha) \xrightarrow{p} \begin{pmatrix} I_c & \Omega_c \\ 0 & I_\alpha \end{pmatrix}.$$

Also, since $E\{(1/n)S_c(\alpha,\beta)S_c(\alpha,\beta)^T\}=I_c$, $E\{(1/n)S_c(\alpha,\beta)U(\alpha)^T\}=E[R_iW_i\{Y_i-\eta(X_i,Z_i;\beta,\alpha)\}U(R_i|Y_i,Z_i;\alpha)^T]=\Omega_c$, and $E[(1/n)E\{U(\alpha)U(\alpha)^T\}]=I_\alpha$, we have

$$\mathrm{Var}\Big[\frac{1}{\sqrt{n}}\{S_c(\alpha,\beta),U(\alpha)\}^T\Big]=\frac{1}{(n)\Sigma_c}=\begin{pmatrix} I_c & \Omega_c \\ \Omega_c^T & I_\alpha \end{pmatrix}.$$

Then $\sqrt{n}(\hat\beta_c-\beta,\hat\alpha-\alpha)^T$ has asymptotic variance

$$\begin{pmatrix} I_c & \Omega_c \\ 0 & I_\alpha \end{pmatrix}^{-1}\begin{pmatrix} I_c & \Omega_c \\ \Omega_c^T & I_\alpha \end{pmatrix}\begin{pmatrix} I_c & 0 \\ \Omega_c^T & I_\alpha \end{pmatrix}^{-1}.$$

Since $\begin{pmatrix} I_c & \Omega_c \\ 0 & I_\alpha \end{pmatrix}^{-1}=\begin{pmatrix} I_c^{-1} & -I_c^{-1}\Omega_cI_\alpha^{-1} \\ 0 & I_\alpha^{-1} \end{pmatrix}$,
it follows that the asymptotic variance of $\sqrt{n}\hat\beta_c$ is $I_c^{-1}-I_c^{-1}\Omega_cI_\alpha^{-1}\Omega_c^TI_c^{-1}$.

References

Breslow, N. E. and Cain, K. C. (1988). Logistic regression for two-state case-control data. *Biometrika* **75**, 11-20.

Fleiss, J. L., Levin, B. and Paik, M. C. (2003). *Statistical Methods for Rates and Proportions*. 3rd edition. Wiley Series in Probability and Statistics.

Little, R. J. A. (1986). Survey nonresponse adjustments for estimates of means. *Internat. Statist. Rev.* **54**, 139-157.

Paik, M. C. (1997). The generalized estimating equation approach when data are not missing completely at random. *J. Amer. Math. Soc.* **92**, 1320-1329.

Paik, M. C. (2000). Methods for missing covariates in logistic regression. *Comm. Statist. Simulation Comput.* **29**, 1-19.

Reilly, M. and Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* **82**, 299-314.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Math. Soc.* **89**, 846-866.

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

Sacco, R. L., Boden-Albala, B., Gan, R., Chen, X., Kargman, D. E., Shea, S., Paik, M. C. and Hauser, W. A. (1998). Stroke incidence among White, Black, and Hispanic residents of the same community: The Northern Manhattan Stroke Study. *Amer. J. Epidemiology* **147**, 259-268.

Wang, Y. (1999). Estimating equations with nonignorably missing response data. *Biometrics* **55**, 984-989.

Zhao, L. P. and Lipsitz, S. (1992). Designs and analysis of two-stage studies. *Statist. Medicine* **11**, 769-782.

Zhao, L. P., Lipsitz, S. and Lew, D. (1996). Regression analysis with missing covariate data using estimating equations. *Biometrics* **52**, 1165-1182.

Department of Epidemiology and Population Health, Albert Einstein College of Medicine of Yeshiva University, Jack and Pearl Resnick Campus, 1300, Morris Park Avenue, Belfer Building, Room 1303E, Bronx, NY 10461, U.S.A.

E-mail: cuwang@aecom.yu.edu

Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 West 168th Street, 6th Floor, New York, New York 10032, U.S.A.

E-mail: mcp@biostat.columbia.edu

(Received August 2004; accepted May 2005)