Published in final edited form as:

J Stat Plan Inference. 2009 July 1; 139(7): 2341–2350. doi:10.1016/j.jspi.2008.10.024.

HANDLING MISSING DATA BY DELETING COMPLETELY OBSERVED RECORDS

Myunghee Cho Paik and Cuiling Wang

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

Abstract

When data are missing, analyzing records that are completely observed may cause bias or inefficiency. Existing approaches in handling missing data include likelihood, imputation and inverse probability weighting. In this paper, we propose three estimators inspired by deleting some completely observed data in the regression setting. First, we generate artificial observation indicators that are independent of outcome given the observed data and draw inferences conditioning on the artificial observation indicators. Second, we propose a closely related weighting method. The proposed weighting method has more stable weights than those of the inverse probability weighting method (Zhao and Lipsitz, 1992). Third, we improve the efficiency of the proposed weighting estimator by subtracting the projection of the estimating function onto the nuisance tangent space. When data are missing completely at random, we show that the proposed estimators have asymptotic variances smaller than or equal to the variance of the estimator obtained from using completely observed records only. Asymptotic relative efficiency computation and simulation studies indicate that the proposed weighting estimators are more efficient than the inverse probability weighting estimators under wide range of practical situations especially when the missingness proportion is large.

1. INTRODUCTION

When data are missing, analyzing only completely observed records could cause bias or inefficiency. One way of handling missing data is to maximize the observed likelihood obtained by integrating the likelihood for the full data and observation indicators over the missing data (e.g. Little and Rubin, 1987; Laird, 1988). In the non-likelihood framework, approaches such as the imputation and the inverse probability weighting have been proposed. The imputation method replaces the contribution of the estimating function with the missing statistics by its conditional expectation given the observed data (Reilly and Pepe, 1995; Paik, 1996). The inverse probability weighting (Zhao and Lipsitz, 1992) uses only the completely observed records, but weighs each record by the inverse of the probability of observation. The two approaches reflect different viewpoints of the problems involving missing data: the imputation method fills in the missing data with the most plausible values, while the inverse probability weighting blows up the observed records to properly represent the whole data (Fleiss, Levin and Paik, 2003). Correspondingly these two approaches represent different ways of constructing unbiased estimating functions. Recently, Lipsitz et al. (1999) proposed a combination of the two where the authors used weights derived from imputation models. Here

we propose a third approach, which deletes some of completely observed records artificially, thereby motivating a class of estimators. We propose to delete some records so that after deletion, the observation process is independent of the outcome and therefore the complete-record analysis is valid. The implementation of this approach is simpler than that of existing approaches and can be widely disseminated to users. Intuitively, we discard some observed information to undo the harm caused by the missing mechanism and restore the original structure existed in the full data. A similar idea was implemented in survival analysis via artificial censoring to fix dependent censoring (Lin, Robins and Wei, 1996).

Specifically, we propose three estimators. We call the first the deletion estimator, directly reflecting the main idea. That is, we create artificial observation indicators so that the artificially created indicator is conditionally independent of the outcome, and analyze only the records that are ‘artificially’ observed. The artificially created observation indicator is a decreasing function of the observation indicator and artificially observed records constitute a subset of completely observed records. The proposed deletion estimators are consistent when data are missing at random. Although counterintuitive, we show that the deletion estimator has asymptotic variance smaller than or equal to the estimator obtained from the complete-record analysis despite using a smaller number of records, when data are missing completely at random. The second proposed estimator involves a weighting method where the weight is the probability of deletion in the deletion method. The weights of the proposed method are more stable than those of the inverse probability weighting method, and the resulting estimates are more efficient when the missing proportion is high. We also show that the weighting estimate is the limit of the mean of repeatedly computed deletion estimates as the number of replications approaches infinity. Finally, the third estimator is a modified estimator from the second estimator using the argument of Robins et al. (1994) to improve the efficiency. The efficient version of the proposed weighting estimator are shown to perform better than the counterpart of the inverse probability weighting in situations where the proposed weighting estimator performs better than the IPW estimators.

2. Motivation

Let Y denote the outcome and (X, Z) denote covariates. We consider situations where our interest lies in the regression setting and $E(Y|X, Z)$, with a known parametric form, is the quantity of interest. Outcome Y and covariate Z are completely observed and covariate X could be missing. Let R be the observation indicator for X . Throughout the paper, we assume that the data are missing at random (Rubin, 1976), i.e., the observation probability does not depend on the missing variable X itself:

$$P(R|X, Y, Z) = P(R|Y, Z). \quad (1)$$

Our motivation of the proposed method starts from the observation that if R additionally satisfies the condition, $R \perp\!\!\!\perp Y|Z$, we have $E(Y|X, Z, R = 1) = E(Y|X, Z)$, then we can consistently estimate $E(Y|X, Z)$ simply by conducting the Complete-Record (CR) analysis. In practice, however, we cannot control R and cannot force R to satisfy this additional condition. Therefore, a key idea of the proposed method is to generate artificial variable, say R^* , such that $R^* \perp\!\!\!\perp Y|Z$, and to estimate $E(Y|X, Z)$ using only records whose artificial variable R^* equals 1. This approach is simple and can be handled by most of standard software. Then the problem boils down how to generate such R^* .

For a concrete example, consider the case that (Y, Z) are all binary. For brevity, denote $P(R_i = 1|Y_i, Z_i)$ and $P(R_i^* = 1|Y_i, Z_i)$ by $\pi(Y_i, Z_i)$ and $\pi^*(Y_i, Z_i)$, respectively. To satisfy $R^* \perp\!\!\!\perp Y|Z$, we need to generate R^* so that $\pi^*(1, Z_i) = \pi^*(0, Z_i)$, given the sample proportions of observation

$\hat{\pi}(1, Z_i)$ and $\hat{\pi}(0, Z_i)$. Our strategy is to obtain π^* by modifying $\hat{\pi}$. First, note that we can pretend some observed data to be missing, but cannot pretend missing data to be observed. This fact implies that R^* should be bounded by R , i.e., $R_i^* \leq R_i$. That is, if $R_i = 1$, we can set $R_i^* = 0$, but if $R_i = 0$, we should set $R_i^* = 0$. To achieve $\pi^*(1, Z_i) = \pi^*(0, Z_i)$ under the condition $\pi^*(Y, Z) \leq \hat{\pi}(Y, Z)$, we can ‘down-adjust’ as follows. If $\hat{\pi}(1, Z)$ is greater than $\hat{\pi}(0, Z)$, we pick the smaller proportion $\hat{\pi}(0, Z)$, set it equal to both $\pi^*(1, Z)$ and $\pi^*(0, Z)$. This is ‘down-adjusting’ since we equalize the two target probabilities $\pi^*(0, Z)$ and $\pi^*(1, Z)$ by setting equal to the smaller proportion between $\hat{\pi}(0, Z)$ and $\hat{\pi}(1, Z)$. In notation, $\pi^*(Y, Z) = \min_Y \hat{\pi}(Y, Z)$. In this example, for $Y = 0$, there is no difference between the artificial observation status vs. original observation status, i.e., $R_i^* = R_i$ with probability 1. For $Y = 1$, we enforce $\pi^*(1, Z) = \pi^*(0, Z) = \hat{\pi}(0, Z)$. This relationship can be implemented by generating binary R^* with probability being 1 given $R = 1$ as $\hat{\pi}(0, Z)/\hat{\pi}(1, Z)$, the denominator of which nullifies the observation process for $Y = 1$ while the numerator imposes the observation process for $Y = 0$. In summary, R^* takes value 1 with probability $\{\min_{Y \in \{0, 1\}} \hat{\pi}(Y, Z)\}/\hat{\pi}(Y, Z)$. For non-binary Y and Z , we can use the same ‘down-adjusting’ strategy by generating binary R^* with probability of being 1 as $\{\min_{Y \in \{Y_1, \dots, Y_n\}} \hat{\pi}(Y, Z)\}/\hat{\pi}(Y, Z)$.

In Section 3, we propose three estimators motivated by the idea of deleting. First is the usual estimator but using the records with $R^* = 1$ only. We call this deletion estimator. The second is the weighting estimator using records with $R = 1$ only with weight $P(R_i^* = 1|Y_i, Z_i, R_i = 1)$. Third is an improved estimator from the second estimator by subtracting projection onto the nuisance tangent space.

3. THREE PROPOSED ESTIMATORS

3.1 Deletion estimator

To formalize the idea presented in Section 2, suppose that $P(R = 1|Y, Z)$ is a known function indexed by unknown parameter α , denoted by $\pi(Y, Z; \alpha) > 0$, and π is a differentiable function of α . Furthermore, under the condition (1), suppose there exists a consistent and asymptotically normally distributed estimator $\hat{\alpha}$, which can be expressed as $n^{\frac{1}{2}}(\hat{\alpha} - \alpha_0) = \sum A_i(\alpha_0) + o_p(1)$, where α_0 is the true value for α , $A_i(\alpha_0)$ ’s are independently distributed with $EA_i(\alpha_0) = 0$. Using

the down-adjusting idea, we set $P(R^* = 1|Y, Z, R = 1) = q(\alpha) = \frac{\pi_M(Z; \alpha)}{\pi(Y, Z; \alpha)}$, where $\pi_M(Z; \alpha) = \min_{Y \in \{Y_1, \dots, Y_n\}} \pi(Y, Z)$. Generating R_i^* so that $R^* = 1$ with probability $q_i(\hat{\alpha})$, the proposed estimator is the one obtained using the records with $R_i^* = 1$ only, or the solution of the following estimating equation.

$$U_n^*(\beta) = \sum R_i^* g(X_i, Z_i; \beta) \{Y_i - \mu(X_i, Z_i; \beta)\} = 0. \quad (2)$$

We show in the next section that the estimating function (2) has zero expectation and its solution $\hat{\beta}^*$ is consistent for the true value β_0 and asymptotically normally distributed.

The construction of the deletion estimator is motivated to fix the bias but could potentially cause a loss of efficiency. However, it turns out that the estimator in fact gains efficiency over the CR estimator when data are missing completely at random. In Appendix D we show that the asymptotic variance of the deletion estimator is smaller than or equal to that of the CR estimator when data are missing completely at random. This may initially sound counterintuitive since $R^* \leq R$, and the complete record analysis utilizes more records than the deletion method. However, the number of deleted records would be small when data are missing

completely at random, and furthermore deletion is executed effectively using information from the observed data, which lead to a more efficient estimator.

3.2. Weighting estimator and its relationship with deletion estimator

A practical weakness of deleting method is that the randomness of R^* produces different estimates each time given the same observed data. To improve upon this feature, one can contemplate an estimating function, $E\{U_n^*(\beta)|R, X_o, Y, Z\}$, where X_o is observed part of X . The conditional expectation replaces R_i^* in the estimating function with its conditional mean $q_i(\hat{\alpha})$ given the observed data as follows:

$$E\{U_n^*(\beta)|R, X_o, Y, Z\} = U_n(\beta, \hat{\alpha}) = \sum R_i q_i(\hat{\alpha}) g(X_i, Z_i; \beta) \{Y_i - \mu(X_i, Z_i; \beta)\} = 0 \quad (3)$$

Equation (3) is the proposed weighting estimating equation, and its solution is the proposed weighting estimator. If the weight $q_i(\hat{\alpha})$ is replaced by $1/\pi(Y_i, Z_i; \hat{\alpha})$, the equation is the Inverse Probability Weighting (IPW) equation (Zhao and Lipsitz, 1992; Robins, Rotnitzky and Zhao, 1994). A clear advantage of the proposed weighting method over IPW is that the weight $q_i(\hat{\alpha})$ is bounded between 0 and 1, while the IPW weight is not. The IPW weight for a particular unit could be dominating since a single $\pi(Y_i, Z_i; \hat{\alpha})$ could be small with outlying covariate value Z_i even when the overall observation probability is moderately large.

We can show that the estimating function (3) has zero expectation and the resulting estimator is consistent and asymptotically normally distributed.

Theorem 1— $n^{\frac{1}{2}}(\hat{\beta} - \beta_0)$ is asymptotically normally distributed with mean 0 and variance $\Gamma_1^{-1}(\beta_0, \alpha_0) \Sigma(\beta_0, \alpha_0) \Gamma_1^{-1}(\beta_0, \alpha_0)$, where

$$\Gamma_1(\beta_0, \alpha_0) = \lim_{n \rightarrow \infty} n^{-1} E\left[-\frac{\partial U_n(\beta, \alpha)}{\partial \beta^T}; \beta_0, \alpha_0\right],$$

and $\Sigma(\beta_0, \alpha_0) = \text{Var}\{n^{-\frac{1}{2}} U_n(\beta_0, \hat{\alpha}); \beta_0, \alpha_0\}$. A consistent estimator for $\Sigma(\beta_0, \alpha_0)$ is given by

$$n^{-1} \sum [R_i q_i(\hat{\alpha}) g(X_i, Z_i; \hat{\beta}) \{Y_i - \mu(X_i, Z_i; \hat{\beta})\} + \Gamma_2(\hat{\alpha}, \hat{\beta}) A_i(\hat{\alpha})]^{\otimes 2},$$

where $A^{\otimes 2} = AA^T$,

$$\Gamma_2(\beta_0, \alpha_0) = \lim_{n \rightarrow \infty} n^{-1} E\left[\frac{\partial U_n(\beta, \alpha)}{\partial \alpha^T}; \beta_0, \alpha_0\right].$$

See Appendix A for proof.

Based on this result, we can establish the asymptotic property of the deletion estimator by expressing $U_n^*(\beta, \hat{\alpha}) = U_n(\beta, \hat{\alpha}) + \{U_n^*(\beta, \hat{\alpha}) - U_n(\beta, \hat{\alpha})\}$.

Theorem 2— $n^{\frac{1}{2}}(\widehat{\beta}^* - \beta_0)$ is asymptotically normally distributed with mean 0 and variance $\Gamma_1^{-1}(\beta_0, \alpha_0)\{\sum(\beta_0, \alpha_0) + \Omega(\beta_0, \alpha_0)\}\Gamma_1^{-1}(\beta_0, \alpha_0)$ where $\Gamma_1(\beta_0, \alpha_0)$ and $\Sigma(\beta_0, \alpha_0)$ are defined in Theorem 1, and $\Omega(\beta_0, \alpha_0) = \text{Var}[n^{-\frac{1}{2}}\{U_n^*(\beta, \widehat{\alpha}) - U_n(\beta, \widehat{\alpha})\}; \alpha_0, \beta_0]$. $\Omega(\beta_0, \alpha_0)$ can be consistently estimated by

$$n^{-1} \sum q_i(\widehat{\alpha})\{1 - q_i(\widehat{\alpha})\}g(X_i, Z_i; \widehat{\beta}^*)\{Y_i - \mu(X_i, Z_i; \widehat{\beta}^*)\}^2 g(X_i, Z_i; \widehat{\beta}^*)^T.$$

See Appendix B for proof.

Note that $\Omega(\beta_0, \alpha_0)$ captures the extra variability of the deletion estimator caused by the randomness of R^* given observed data. This leads us to consider yet another related estimator. If we were to repeatedly compute the deletion estimates K times, we would obtain K different deletion estimates, say $\widehat{\beta}_1^*, \dots, \widehat{\beta}_K^*$ due to the randomness of deleting records. Let

$\bar{\beta}_K = K^{-1} \sum_{k=1}^K \widehat{\beta}_k^*$. We show in Appendix C that $n^{\frac{1}{2}}(\widehat{\beta} - \bar{\beta}_K) = o_p(1)$ as n goes to infinity, where $\bar{\beta}_\infty = \lim_{K \rightarrow \infty} \bar{\beta}_K$.

3.3. Efficient weighting estimator

We can express $U_n(\beta, \alpha)$ in a form resembling the class of equations considered by Robins et al. (1994).

$$U_n(\beta, \alpha) = \sum R_i \pi_i(Y_i, Z_i; \alpha)^{-1} h(X_i, Z_i; \beta, \alpha) \{Y_i - \mu(X_i, Z_i; \beta)\} = 0,$$

where $h(X_i, Z_i; \beta, \alpha) = \pi_M(Z_i; \alpha)g(X_i, Z_i; \beta)$. Using the argument of Robins et al. (1994) we can improve the efficiency of the estimator by subtracting the projection of the estimating function onto the nuisance tangent space which is the closed linear span of all random vectors of fixed multiple of the nuisance score:

$$U_n^{\text{eff}}(\beta, \alpha, \phi) = \sum R_i \pi_i(Y_i, Z_i; \alpha)^{-1} h(X_i, Z_i; \beta) \{Y_i - \mu(X_i, Z_i; \beta)\} - \sum \pi(Y_i, Z_i; \alpha)^{-1} \{R_i - \pi(Y_i, Z_i; \alpha)\} \phi(Y_i, Z_i).$$

For the second term of U_n^{eff} to be a projection, one should find the function ϕ that minimizes the norm of U_n^{eff} . Such a function ϕ can be found by decomposing $U_n^{\text{eff}}(\beta, \alpha, \phi)$ into the following two terms:

$$\begin{aligned} U_n^{\text{eff}}(\beta, \alpha, \phi) &= \sum h(X_i, Z_i; \beta) \{Y_i - \mu(X_i, Z_i; \beta)\} \\ &+ \sum \{R_i - \pi(Y_i, Z_i; \alpha)\} \pi(Y_i, Z_i; \alpha)^{-1} [h(X_i, Z_i; \beta) \{Y_i - \mu(X_i, Z_i; \beta)\} - \phi(Y_i, Z_i)] \\ &= C_{1n} + C_{2n} = \sum (C_{1n}^{(i)} + C_{2n}^{(i)}) \end{aligned}$$

Since C_{1n} and C_{2n} are uncorrelated,

$$\begin{aligned} \text{Var}(C_{1n}+C_{2n}) &= \text{Var}(C_{1n}) + E\text{Var}(C_{2n}|X, Y, Z) + \text{Var}E(C_{2n}|X, Y, Z) \\ &= \text{Var}(C_{1n}) + E\left[\frac{1-\pi(Y_i, Z_i; \alpha)}{\pi(Y_i, Z_i; \alpha)}\right] E[h(X_i, Z_i; \beta)\{Y_i - \mu(X_i, Z_i; \beta)\} - \phi_i]^2. \end{aligned}$$

The above expression suggests that the minimum can be achieved when $\phi(Y_i, Z_i)$ equals $\phi_{\text{eff}}^h(Y_i, Z_i) = E[h(X_i, Z_i; \beta)\{Y_i - \mu(X_i, Z_i; \beta)\}|Y_i, Z_i]$.

The solution to the equation $U_n^{\text{eff}}(\beta, \alpha, \phi_{\text{eff}}^h) = 0$ say $\widehat{\beta}(\widehat{\phi}_{\text{eff}}^h)$, is the most efficient estimator given the form of $h(X, Z; \beta)$, and its variance can be estimated by,

$$n^{-1} \widehat{\Gamma}_3^{-1}(\widehat{\beta}(\widehat{\phi}_{\text{eff}}^h), \widehat{\alpha}) \left\{ \sum (\widehat{C}_{1n}^{(i)} + \widehat{C}_{2n}^{(i)})^{\otimes 2} \right\} \widehat{\Gamma}_3^{-1}(\widehat{\beta}(\widehat{\phi}_{\text{eff}}^h), \widehat{\alpha}), \text{ where}$$

$$\Gamma_3(\beta_0, \alpha_0) = \lim_{n \rightarrow \infty} n^{-1} E \left\{ - \frac{\partial U_n^{\text{eff}}(\beta, \alpha, \phi_{\text{eff}}^h)}{\partial \beta^T} \middle| \beta_0, \alpha_0 \right\}$$

When we replace $h(X, Z; \beta)$ with $g(X, Z; \beta)$, the resulting estimator is an efficient version of the IPW estimator, say $\widehat{\beta}_{\text{IPW}}(\widehat{\phi}_{\text{eff}}^g)$. However, neither $\widehat{\beta}(\widehat{\phi}_{\text{eff}}^h)$ nor $\widehat{\beta}_{\text{IPW}}(\widehat{\phi}_{\text{eff}}^g)$ is fully efficient. To obtain a fully efficient estimator, one has to replace $h(X, Z; \beta)$ with an arbitrary function of (X, Z) say, $h^*(X, Z; \beta)$ and solve for $h^*(X, Z)$ satisfying integral equation (23) of Robins et al. (1994) and find its corresponding ϕ , say $\phi_{\text{eff}}^{h^*} = E\{h^*(X, Z; \beta)(Y - \mu)|Y, Z\}$. It is hard to analytically compare the efficiencies of the estimators derived from g and h functions. Note that ϕ is a function of an unknown quantity involving $E[h(X_i, Z_i; \beta)|Y_i, Z_i]$ and $E[h(X_i, Z_i; \beta)\mu(X_i, Z_i; \beta)|Y_i, Z_i]$. In GLM with canonical link function, ϕ is a function of $E(X_i|Y_i, Z_i)$, $E[\mu(X_i, Z_i; \beta)|Y_i, Z_i]$, and $E[X_i \mu(X_i, Z_i; \beta)|Y_i, Z_i]$. We suggest estimating unknown quantities in ϕ by parametric modelling as suggested by Zhao, Lipsitz and Lew (1996).

4. ASYMPTOTIC RELATIVE EFFICIENCY

Given the fact that we can not establish any inequalities between the asymptotic variances of WTEF and IPWEF, an important practical question remains as to when to use deletion family estimators instead of inverse probability estimators. Obviously, when any one π_i is small, the weights for IPW or IPWEF could be unstable and result in non-convergence or inefficient estimators. This could happen when the overall probability is small, or when the observation probability depends on continuous variable whose values can be extreme, or when the effect of the covariate on the observation probability is large.

To systematically investigate the aforementioned conjecture, the asymptotic relative efficiency of WTEF over IPWEF is computed in various situations and summarized in Figure 1. Two columns of graphs show the efficiencies of $\widehat{\beta}_X$ and $\widehat{\beta}_Z$, respectively for binary X and Z . The pattern for the asymptotic relative efficiency of $\widehat{\beta}_0$ is similar to that of $\widehat{\beta}_X$. Within each graph, five curves are shown for different values of α_Y . Note that $\alpha_Y = 0$ corresponds to missing completely at random and a large value of $|\alpha_Y|$ corresponds to a severe degree of departure from missing- completely-at-random case. Two rows display the efficiencies under different overall missing proportions controlled by $\alpha_0 = -1$ and $\alpha_0 = 1$, respectively. Note that large negative values of α_0 and α_Z correspond to a large overall proportion of missingness. The graphs show that in most cases the efficiency of WTEF over IPWEF increases for large negative values of α_Z . This confirms our conjecture that the proposed estimate performs better than the IPWEF when $\widehat{\pi}(Y_i, Z_i)$ is small. Also we see that the efficiency of WTEF is increased as α_0 decreases. The graphs on the third column feature situations when $\widehat{\beta}_Z$ of IPWEF is more efficient than that of WTEF. When both α_Z and α_Y are negative, IPWEF has a smaller asymptotic variance than the proposed one. But even in the worst case, the efficiency is around 0.8.

5. SIMULATION

We conducted simulation studies with 500 replications for logistic regression models and classical linear models. Two sample sizes, 2000 and 500 were used and only the results for $n = 500$ is shown. For all models we generated observation indicators with $P(R=1|Y,Z) = \text{logit}^{-1}(\alpha_0 + \alpha_Y Y + \alpha_Z Z)$. For the case in which data are missing completely at random (MCAR), we set $\alpha = (-2, 0, 1)$ for the logistic model, and $\alpha = (-1.5, 0, 1)$ for the linear model. For the case of MAR, we set $\alpha = (-2, 1, 1)$ for the logistic model, and $\alpha = (-1.5, 0.3, 1)$ for the linear model. We generated two types of (X, Z) , first as standard normal variables and second as binary indicators. In all cases X and Z are generated independently. Given X and Z , Y is generated as a binary indicator with probability $\mu(X, Z; \beta) = P(Y = 1|X, Z) = \text{logit}^{-1}(X + Z)$ for logistic models, and as a normal variable with mean $X + Z$ and variance 1 for linear models.

The performance of eight estimates are given from full data (Full), complete records (CR), IPW using the correct missingness model (IPW1), IPW using the overfitted model with the interaction term between Y and Z (IPW2), efficient version of IPW (IPWEF), the proposed deletion method (Del), the proposed weighting method (WT), and the efficient version of the proposed weighting method (WTEF). In computing the projection term for IPWEF and WTEF when (Y, Z) are all binary, sample means are computed for each of four categories, and when at least one element of (Y, Z) is continuous, linear or logistic models are used depending on the type of X with linear predictor $Y + Z + Y * Z$.

For all estimates, the bias of the point estimates, simulation mean square errors, and the average number of records used in each method are shown in Table 1. Table 2 reports the coverage probabilities.

We focus on three comparisons: (i) CR estimate vs. the deletion estimate (Del); (ii) IPW (IPW1 and IPW2) vs. the proposed weighting (WT); and (iii) IPWEF vs. WTEF. First under MCAR, note that the bias of the eight estimates are negligible. In logistic models, the deletion estimates (Del) of β_0 and β_Z are more efficient than those from the complete-record analysis despite using a smaller number of records when Z is continuous or dichotomous. Note that this simulation result using a sample size of 500 agrees with asymptotic results stated in Section 3.1 that the variance of the deletion estimators are smaller than or equal to the CR estimators. On the other hand, the deletion estimate of β_X is slightly less efficient than the complete record estimate of β_X . In linear models, the efficiency gain of the deletion method under MCAR is much smaller than in logistic models because the minimum of $\pi(Y, Z; \alpha)$ is taken over the entire range of Y and thus a higher proportion of records is deleted. The deletion method provides valid estimates in both models under MAR, while the CR analysis is biased.

In both logistic and linear models, the proposed weighting estimates (WT) are generally more efficient than the IPW estimates obtained by correctly specifying missingness models (IPW1) or the IPW estimates obtained by specifying overfitted missingness models (IPW2). The efficiency of IPW1 is substantially poorer than the WT estimates. The advantage of the proposed weighting estimate (WT) over IPW is apparent when Z is continuous. Under this condition, $\pi(Y_i, Z_i; \hat{\alpha})$ is small for extreme values of Z_i , and the inverse probability becomes large and unstable. The efficiency gain of the proposed weighting method (WT) is most notable in linear regression with continuous Z .

The two efficient versions, IPWEF and WTEF, show smaller variances than their original versions as anticipated. However, note that for the case in which X and Z are dichotomous in logistic models, and where auxiliary models are saturated, there is no improvement over the original versions, IPW2 and WT. Between the IPWEF and WTEF, the performance is comparable in logistic regression models: IPWEF has a slight advantage in β_Z and WTEF has

advantages in β_X . In linear models with continuous Z , WTEF estimates have substantially smaller variances than the IPWEF estimates.

Coverage probabilities of the eight methods are shown in Table 2. The coverage probabilities for consistent estimates are reasonable overall, which demonstrates that the asymptotic variances behave well in situations considered. Under MAR, the CR estimates are biased and the coverage probabilities are far from their nominal value. Although the deletion estimators are consistent, their coverage probabilities are less than its target 95% in linear models, because the number of records used in the analysis is small. Although not shown, our simulations show that the coverage probabilities become within the 95% confidence intervals of the nominal value when the overall sample size is 2000.

6. CONVERGENCE ISSUE

We have shown in Section 4 that asymptotic relative efficiency is generally better in proposed weighting estimator than the IPW estimators. Another reason to prefer the weighting estimator is that weighting estimates suffer much less from non-convergence problem than IPW estimates as the missing proportion is increased. For example, under the simulation situation shown in Table 1 ($\alpha_0 = -2$, $\alpha_X = 0$, $\alpha_Z = 1$), convergence problems did not occur. However, when α_Z is changed to -1 , IPW2 did not converge 42 times out of 500 simulation runs whereas the weighting estimates converged all 500 times. Furthermore, when α_Z is increased to -1.5 , IPW2 did not converge 68 times out of 500, while the weighting estimates did not converge 6 times. Similar trend is observed when α_0 or α_Y is decreased.

7. CONCLUDING REMARKS

We have proposed three estimators based on the idea of deleting completely observed records. The deletion estimator serves as a conceptual device for the other two estimators, but may not be attractive for practical use due to randomness of the artificial observation indicator. The weighting and the improved weighting estimators are viable alternatives to the inverse probability weighting estimators. When some of the predicted observation probabilities are small, the proposed weighting estimators suffer much less from non-convergence problems and are more efficient than the inverse probability weighting estimators. While discarding some completely observed data to handle missing data may sound paradoxical, it proves to be effective when is done in an informative way.

APPENDIX A: Asymptotic distribution of β

Since $q(\alpha)$ is a differentiable function of α , we can write

$$\begin{aligned} n^{-\frac{1}{2}} U_n(\widehat{\beta}, \widehat{\alpha}) &= n^{-\frac{1}{2}} \{ U_n(\beta_0, \alpha_0) + \frac{\partial U_n(\beta, \alpha)}{\partial \beta^T} (\widehat{\beta} - \beta_0) + \frac{\partial U_n(\beta, \alpha)}{\partial \alpha^T} (\widehat{\alpha} - \alpha_0) \} + o_p(1) \\ &= n^{-\frac{1}{2}} \{ U_n(\beta_0, \alpha_0) - n \Gamma_1(\beta_0, \alpha_0) (\widehat{\beta} - \beta_0) + \Gamma_2(\beta_0, \alpha_0) \sum A_i(\alpha_0) + o_p(1) \} \end{aligned}$$

where

$$\Gamma_1(\beta_0, \alpha_0) = \lim_{n \rightarrow \infty} E \{ n^{-1} - \frac{\partial U_n(\beta, \alpha)}{\partial \beta^T} ; \beta_0, \alpha_0 \}, \quad \Gamma_2(\beta_0, \alpha_0) = \lim_{n \rightarrow \infty} E \{ n^{-1} \frac{\partial U_n(\beta, \alpha)}{\partial \alpha^T} ; \beta_0, \alpha_0 \},$$

$$\frac{\partial U_n(\beta, \alpha)}{\partial \alpha^T} = \sum R_i \frac{\partial q_i(\alpha)}{\partial \alpha^T} g(X_i, Z_i; \beta) \{Y_i - \mu(X_i, Z_i; \beta)\},$$

$$\frac{\partial q(\alpha)}{\partial \alpha^T} = \frac{\pi_M(Z_i; \alpha)}{\pi(Y_i, Z_i; \alpha)} \left\{ \frac{\pi'_M(Z_i; \alpha)}{\pi_M(Z_i; \alpha)} - \frac{\pi'(Y_i, Z_i; \alpha)}{\pi(Y_i, Z_i; \alpha)} \right\}, \text{ and}$$

π' denotes derivative of π with respect to α . Denoting the i th contribution to $U_n(\beta, \alpha)$ by $U_n^{(i)}(\beta, \alpha)$,

$$\begin{aligned} n^{\frac{1}{2}}(\widehat{\beta} - \beta_0) &= \Gamma_1^{-1}(\beta_0, \alpha_0) \{n^{-\frac{1}{2}} U_n(\beta_0, \alpha_0) + \Gamma_2(\beta_0, \alpha_0) \sum A_i(\alpha_0)\} + o_p(1) \\ &= \sum_{i=1}^n \Gamma_1^{-1}(\beta_0, \alpha_0) \{n^{-\frac{1}{2}} U_n^{(i)}(\beta_0, \alpha_0) + \Gamma_2(\beta_0, \alpha_0) A_i(\alpha_0)\} + o_p(1) \\ &= \sum H_i(\beta_0, \alpha_0) + o_p(1) \end{aligned}$$

First, we can show that $U_n^{(i)}(\beta_0, \alpha_0)$ has mean zero:

$$\begin{aligned} E\{E_{\mathbb{Q}[Y, X_0, Z]} U_n^{(i)}(\beta, \alpha)\} &= E[\pi(Z_i, Y_i; \alpha) q_i(\alpha) g(X_i, Z_i; \beta) \{Y_i - \mu(X_i, Z_i; \beta)\}] \\ &= E[\pi_M(Z_i; \alpha) g(X_i, Z_i; \beta) \{Y_i - \mu(X_i, Z_i; \beta)\}] = 0. \end{aligned}$$

In addition, by assumption made in Section 2, $A_i(\alpha_0)$'s have mean zero and are independent. Then, $\sum H_i$ is a sum of independently distributed random vectors with mean zero with finite variance, and $\widehat{\beta}$ is asymptotically normally distributed.

APPENDIX B: Asymptotic distribution of $\widehat{\beta}^*$

Using Taylor's expansion, we have $U_n^*(\widehat{\beta}^*) = 0 = U_n^*(\beta_0) + \frac{\partial U_n^*(\beta)}{\partial \beta^T} (\widehat{\beta}^* - \beta_0) + o_p(1)$. First, it is easy to verify that

$$\lim_{n \rightarrow \infty} E\{n^{-1} - \frac{\partial U_n^*(\beta)}{\partial \beta^T}; \beta_0, \alpha_0\} = \Gamma_1(\beta_0, \alpha_0).$$

Then $n^{\frac{1}{2}}(\widehat{\beta}^* - \beta_0) = n^{-\frac{1}{2}} \Gamma_1^{-1}(\beta_0, \alpha_0) U_n^*(\beta_0) + o_p(1)$. Rewriting $U_n^*(\beta) = U_n(\beta, \widehat{\alpha}) + U_n^*(\beta) - U_n(\beta, \widehat{\alpha})$, and using the fact that $n^{-\frac{1}{2}} U_n(\beta, \widehat{\alpha}) = n^{-\frac{1}{2}} U_n(\beta, \alpha_0) + \Gamma_2(\beta_0, \alpha_0) \sum A_i(\alpha_0) + o_p(1)$, we have

$$\begin{aligned} n^{\frac{1}{2}}(\widehat{\beta}^* - \beta_0) &= \Gamma_1^{-1}(\beta_0, \alpha_0) [n^{-\frac{1}{2}} U_n(\beta_0, \alpha_0) + \Gamma_2(\beta_0, \alpha_0) \sum A_i(\alpha_0) + n^{-\frac{1}{2}} \{U_n^*(\beta_0) - U_n(\beta_0, \widehat{\alpha})\}] + o_p(1) \\ &= T_{1n} + T_{2n} + o_p(1) = \sum_{i=1}^n (T_{1n}^{(i)} + T_{2n}^{(i)}) + o_p(1), \end{aligned}$$

where $T_{1n}^{(i)} = n^{-\frac{1}{2}} U_n^{(i)}(\beta_0, \alpha_0) + \Gamma_2 A_i(\alpha_0)$, $T_{1n} = \sum T_{1n}^{(i)}$ and $T_{2n}^{(i)} = \{U_n^{*(i)}(\widehat{\beta}) - U_n^{(i)}(\beta_0, \widehat{\alpha})\} = \{R_i^* - q_i(\widehat{\alpha})\} g(X_i, Z_i; \beta_0) \{Y_i - \mu(X_i, Z_i; \beta_0)\}$, and $T_{2n} = \sum T_{2n}^{(i)}$. Note that $E_{R \sim \{R, Y, X_o, Z\}}[\{R_i^* - q_i(\widehat{\alpha})\} | R, Y, X_o, Z; \alpha_0] = 0$, and $E(T_{2n}^{(i)} | R, Y, X_o, Z; \alpha_0) = 0$. Furthermore, $E\{Cov(T_{1n}^{(i)}, T_{2n}^{(i)} | R, Y, X_o, Z)\} = 0$ and $Cov\{E(T_{1n}^{(i)} | R, Y, X_o, Z), E(T_{2n}^{(j)} | R, Y, X_o, Z)\} = 0$. Therefore $\sum (T_{1n}^{(i)} + T_{2n}^{(i)})$ represents a sum of independent random vectors with mean 0 and finite variance. We also find that

$$\begin{aligned} Var(T_{1n} + T_{2n}) &= E\{Var(T_{1n} + T_{2n} | R, Y, X_o, Z)\} + Var\{E(T_{1n} + T_{2n} | R, Y, X_o, Z)\} \\ &= E\{Var(T_{2n} | R, Y, X_o, Z)\} + Var(T_{1n}). \end{aligned}$$

Therefore $n^{\frac{1}{2}}(\widehat{\beta}_* - \beta_0)$ is asymptotically normally distributed with mean 0 and variance $E\{Var(T_{2n} | R, Y, X_o, Z)\} + Var(T_{1n})$. A consistent estimator of $Var(T_{1n})$ is given in Appendix A and $E\{Var(T_{2n} | R, Y, X_o, Z)\}$ can be consistently estimated by

$$\sum q_i(\widehat{\alpha}) \{1 - q_i(\widehat{\alpha})\} g(X_i, Z_i; \widehat{\beta}_*) \{Y_i - \mu(X_i, Z_i; \widehat{\beta}_*)\}^2 g(Y_i, Z_i; \widehat{\alpha})^T$$

APPENDIX C: Relationship of deletion estimator and weighting estimator

The proof is similar to that of theorem 2 of Reilly and Pepe (1997). Denote the artificial observation indicator for the i^{th} unit and the k^{th} replication by $R_{i,k}^*$, and the deletion estimator from the k^{th} replication by $\widehat{\beta}_k^*$, the estimating function can be written as

$$U_{n,k}^*(\beta, \widehat{\alpha}) = \sum_{i=1}^n R_{i,k}^* g(X_i, Z_i; \beta) (Y_i - \mu_i).$$

From Taylor expansion, we have

$$\sqrt{n}(\widehat{\beta}_k^* - \beta_0) = \Gamma_{1,k}(\widehat{\beta}_k^*, \widehat{\alpha})^{-1} \frac{U_{n,k}^*(\beta_0, \widehat{\alpha})}{\sqrt{n}},$$

where $\widehat{\beta}_k^* \in (\widehat{\beta}_k^*, \beta_0)$. Then the mean of K deletion estimators can be expressed as

$$\begin{aligned} \sqrt{n}(\bar{\beta}_K - \beta_0) &= \frac{1}{K} \sum_{k=1}^K \Gamma_{1,k}(\widehat{\beta}_k^*, \widehat{\alpha})^{-1} \frac{U_{n,k}^*(\beta_0, \widehat{\alpha})}{\sqrt{n}} \\ &= \frac{1}{K} \sum_{k=1}^K \left\{ \Gamma_{1,k}(\widehat{\beta}_k^*, \widehat{\alpha})^{-1} - \Gamma_1^{-1}(\beta_0, \widehat{\alpha}) \right\} \frac{U_{n,k}^*(\beta_0, \widehat{\alpha})}{\sqrt{n}} + \frac{1}{K} \sum_{k=1}^K \Gamma_1^{-1}(\beta_0, \widehat{\alpha})^{-1} \frac{U_{n,k}^*(\beta_0, \widehat{\alpha})}{\sqrt{n}} \\ &= \Phi_1(K, n) + \Phi_2(K, n). \end{aligned}$$

Given the observed data and n , let $\beta_\infty = \lim_{K \rightarrow \infty} \bar{\beta}_K$. Then

$$\sqrt{n}(\bar{\beta}_\infty - \beta_0) = \lim_{K \rightarrow \infty} \Phi_1(K, n) + \lim_{K \rightarrow \infty} \Phi_2(K, n) = \Phi_1^n + \Phi_2^n.$$

First, observe that $\Phi_1^n \rightarrow 0$ as $n \rightarrow \infty$. Then, since $K^{-1} \sum_{k=1}^K R_{i,k}^* \xrightarrow{P} q_i(\bar{\alpha})$, we have

$$\begin{aligned} \Phi_2^n &= \lim_{K \rightarrow \infty} \frac{1}{K} \Gamma^{-1}(\beta_0, \bar{\alpha}) \sum_{k=1}^K \frac{U_{n,k}^*(\beta_0, \bar{\alpha})}{\sqrt{n}} = \frac{1}{\sqrt{n}} \Gamma^{-1}(\beta_0, \bar{\alpha}) \sum_{i=1}^n \lim_{K \rightarrow \infty} \sum_{k=1}^K R_{i,k}^* g(X_i, Z_i; \beta_0) (Y_i - \mu_i) \\ &= \frac{1}{\sqrt{n}} \Gamma^{-1}(\beta_0, \bar{\alpha}) \sum_{i=1}^n q_i(\bar{\alpha}) g(X_i, Z_i; \beta_0) (Y_i - \mu_i) = \Gamma^{-1}(\beta_0, \bar{\alpha}) \frac{U_n(\beta_0, \bar{\alpha})}{\sqrt{n}}. \end{aligned}$$

$$\text{Thus } \sqrt{n}(\bar{\beta}_\infty - \beta_0) = \Gamma^{-1}(\beta_0, \bar{\alpha}) \frac{U_n(\beta_0, \bar{\alpha})}{\sqrt{n}} + o_p(1), \text{ and } \sqrt{n}(\hat{\beta} - \bar{\beta}_\infty) = o_p(1).$$

APPENDIX D: Asymptotic variance inequality between deletion estimator and CR estimator

To proceed, we need notation for the estimating function for β without missing data, $S(\beta) = \Sigma S(Y_i/X_i, Z_i; \beta)$ and estimating function for the nuisance parameter α , $U(\alpha) = \Sigma U_i(\alpha)$. As shown in Section 3, the asymptotic variance of $\hat{\beta}^*$ is $\Gamma_1^{-1}(\sum + \Omega)\Gamma_1^{-1}$, where $\Gamma_1 = \{E\pi_M(Z)S(Y/X, Z; \beta)S(Y/X, Z; \beta)^T\}$,

$$\begin{aligned} \Sigma &= \Sigma_i - D \\ &= E\left\{\frac{\pi_M(Z)^2}{\pi(Y,Z)} S(Y/X, Z; \beta)S(Y/X, Z; \beta)^T\right\} \\ &\quad - E\{RqS(Y/X, Z; \beta)U(\alpha)^T\}\{EU(\alpha)U(\alpha)^T\}^{-1}E\{RqU(\alpha)S(Y/X, Z; \beta)^T\}, \end{aligned}$$

$$\text{and } \Omega = E\{Rq(1 - q)S(Y/X, Z; \beta)S(Y/X, Z; \beta)^T(Y - \mu)^2\}.$$

Under MCAR, $\pi(Y, Z) = \pi(Z) = \pi_M(Z)$ thus $q = q(Z) = 1$, hence

$$\Gamma_1 = E\{\pi(Z)S(Y/X, Z; \beta)S(Y/X, Z; \beta)^T\},$$

$$\sum_i = E\{\pi(Z)S(Y/X, Z; \beta)S(Y/X, Z; \beta)^T\} = \Gamma_1,$$

$D = E\{RS(Y/X, Z; \beta)U(\alpha)^T\}\{EU(\alpha)U(\alpha)^T\}^{-1}E\{RU(\alpha)S(Y/X, Z; \beta)^T\}$, and $\Omega = 0$. Thus under MCAR, the asymptotic variance of deletion estimator $\hat{\beta}^*$, is $\Gamma_1^{-1}(\Gamma_1 - D)\Gamma_1^{-1}$. Denoting the CR estimator by $\hat{\beta}_{CR}$, the difference of the asymptotic variances is $\text{Var}(\hat{\beta}_{CR}) - \text{Var}(\hat{\beta}^*) = \Gamma_1^{-1}D\Gamma_1^{-1}$, which is positive semidefinite. This proves that the CR estimator has no smaller asymptotic variance than that of the deletion estimator.

References

Breslow NE, Cain KC. Logistic regression for two-stage case-control data. *Biometrika* 1988;75:11–20.

- Fleiss, JL.; Levin, B.; Paik, MC. Statistical Methods for Rates and Proportions. John Wiley & Sons; New York: 2003.
- Laird NM. Missing data in longitudinal studies. *Statistics in Medicine* 1988;7:305–315. [PubMed: 3353609]
- Little, R.; Rubin, D. Statistical Analysis with Missing Data. John Wiley & Sons; New York: 1987.
- Lin DY, Robins JM, Wei LJ. Comparing two failure time distributions in the presence of dependent censoring. *Biometrika* 1996;83:381–393.
- Lipsitz SR, Ibrahim JG, Fitzmaurice GM. Likelihood methods for incomplete longitudinal binary responses with incomplete categorical covariates. *Biometrics* 1999;55:214–223. [PubMed: 11318157]
- Paik MC. Quasi-likelihood regression models with missing covariates. *Biometrika* 1996;83:825–834.
- Reilly M, Pepe M. A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* 1995;82:299–314.
- Reilly M, Pepe M. The relationship between hot-deck multiple imputation and weighted likelihood. *Statistics in Medicine* 1997;16:5–19. [PubMed: 9004380]
- Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 1994;89:846–866.
- Rubin DB. Inference and missing data. *Biometrika* 1976;63:581–92.
- Wang Y-G. Estimating equations with nonignorably missing response data. *Biometrics* 1999;55:984–989. [PubMed: 11315039]
- Zhao L, Lipsitz S. Designs and analysis of two-stage studies. *Statistics in Medicine* 1992;11:769–782. [PubMed: 1594816]
- Zhao LP, Lipsitz SR, Lew D. Regression analysis with missing covariate data using estimating equations. *Biometrics* 1996;52:1165–1182. [PubMed: 8962448]

Figure 1.
Asymptotic Relative Efficiency of Weighting estimators vs. IPW estimators

Table 1

Simulation results based on 500 replications for eight estimators

Logit model MCAR $N=500$ $a_0 = -2$ $Z \sim \text{Bern}(0.5)$

Method	N	BIAS			MSE		
		β_0	β_x	β_z	β_0	β_x	β_z
Full	500	0.00081	0.01545	0.03165	0.0270	0.0521	0.0555
CR	192	0.01934	0.02882	0.01968	0.0980	0.1360	0.1346
Del	179	-0.00598	0.02496	0.04232	0.0559	0.1514	0.0789
IPW1	192	0.01114	0.02296	0.02153	0.0766	0.1579	0.1358
IPW2	192	-0.00359	0.02241	0.04140	0.0538	0.1612	0.0666
WT	192	-0.00779	0.02956	0.04089	0.0479	0.1377	0.0659
IPWEF	192	-0.00359	0.02241	0.04140	0.0538	0.1612	0.0666
WTEF	192	-0.00780	0.02956	0.04089	0.0479	0.1377	0.0659

Logit model MAR $N=500$ $a_0 = -2$ $Z \sim \text{Bern}(0.5)$

Method	N	BIAS			MSE		
		β_0	β_x	β_z	β_0	β_x	β_z
Full	500	-0.00172	0.03808	0.00016	0.0285	0.0491	0.0399
CR	274	0.67198	0.03902	-0.26063	0.5378	0.1271	0.1936
Del	189	-0.01569	0.04220	0.02281	0.0668	0.1355	0.0639
IPW1	274	0.02256	0.03018	-0.01382	0.0618	0.1328	0.1148
IPW2	274	0.00601	0.03073	0.00537	0.0464	0.1337	0.0424
WT	274	-0.00013	0.04186	0.00762	0.0449	0.1281	0.0431
IPWEF	274	0.00600	0.03073	0.00538	0.0464	0.1337	0.0424
WTEF	274	-0.00014	0.04186	0.00763	0.0449	0.1281	0.0431

Logit model MCAR $N=500$ $a_0=-2$ $Z \sim N(0, 1)$

Logit model MCAR $N=500$ $\alpha_0 = -2$ $Z \sim \text{Bern}(0.5)$

Method	N	BIAS			MSE		
		β_0	β_X	β_Z	β_0	β_X	β_Z
Method	N	β_0	β_X	β_Z	β_0	β_X	β_Z
Full	500	0.000213	-0.00318	0.01070	0.0232	0.040	0.0124
CR	111	0.032485	0.00473	0.02839	0.1286	0.239	0.0792
Del	100	0.010974	0.01280	0.00765	0.0784	0.275	0.0520
IPW1	111	-0.000974	0.02302	0.09129	0.1053	0.396	0.1778
IPW2	111	0.001121	0.02735	0.06641	0.1056	0.403	0.1339
WT	111	0.011963	0.00652	0.00226	0.0691	0.239	0.0387
IPWEF	111	-0.004622	0.03423	0.03617	0.1125	0.466	0.0349
WTEF	111	0.001892	0.00680	0.02235	0.0678	0.238	0.0281

Logit model, MAR $N=500$ $\alpha_0=-2$, $Z \sim N(0, 1)$

Method	N	BIAS			MSE		
		β_0	β_X	β_Z	β_0	β_X	β_Z
Method	N	β_0	β_X	β_Z	β_0	β_X	β_Z
Full	500	-0.00499	0.0185	0.00836	0.0182	0.0443	0.0165
CR	128	0.78337	0.0758	-0.11042	0.7322	0.2847	0.1318
Del	77	-0.02025	0.0716	0.01231	0.0989	0.3575	0.0702
IPW1	128	-0.00898	0.1132	0.08954	0.0894	0.4189	0.2041
IPW2	128	-0.01155	0.1194	0.03092	0.0787	0.4223	0.1024
WT	128	-0.00122	0.0781	-0.00168	0.0619	0.2979	0.0485
IPWEF	128	-0.02788	0.1351	0.01242	0.0801	0.4492	0.0395
WTEF	128	-0.01197	0.0790	0.01895	0.0612	0.2977	0.0430

Linear model MCAR $N=500$ $\alpha_0 = -1.5$ $Z \sim \text{Bern}(.5)$

Logit model MCAR $N=500$ $\alpha_0 = -2$ $Z \sim \text{Bern}(0.5)$

Method	N	BIAS			MSE		
		β_0	β_X	β_Z	β_0	β_X	β_Z
Method	N	β_0	β_X	β_Z	β_0	β_X	β_Z
Full	500	-0.00537	0.00225	0.00603	0.00468	0.00182	0.00875
CR	141	0.00807	0.00851	0.00130	0.02571	0.00658	0.03606
Del	111	-0.00562	0.01307	0.00324	0.02448	0.01072	0.03488
IPW1	141	0.00326	0.00921	0.00250	0.02192	0.00736	0.03666
IPW2	141	0.00849	0.01168	-0.00593	0.01473	0.00766	0.02226
WT	141	0.00849	0.01163	-0.00580	0.01469	0.00716	0.02218
IPWEF	141	0.00991	0.01568	-0.00730	0.01400	0.00503	0.02135
WTEF	141	0.00988	0.01380	-0.00719	0.01399	0.00492	0.02130

Linear model MAR
 $N=500$ $\alpha_0 = -1.5$ $Z \sim \text{Bern}(.5)$

Method	N	BIAS			MSE		
		β_0	β_X	β_Z	β_0	β_X	β_Z
Method	N	β_0	β_X	β_Z	β_0	β_X	β_Z
Full	500	-0.00640	0.00165	0.00424	0.0044	0.0020	0.0086
CR	159	0.22426	-0.01645	-0.05741	0.0755	0.0073	0.0403
Del	67.9	0.00255	-0.00120	-0.01139	0.0914	0.0186	0.1230
IPW1	159	-0.00877	-0.00198	0.01241	0.0248	0.0094	0.0438
IPW2	159	0.00278	-0.00122	-0.00377	0.0167	0.0097	0.0279
WT	159	0.00176	0.00030	-0.00216	0.0168	0.0087	0.0279
IPWEF	159	-0.00282	0.00861	0.00027	0.0138	0.0056	0.0244
WTEF	159	-0.00295	0.00749	0.00063	0.0137	0.0052	0.0244

Linear model MCAR $N=500$ $\alpha_0 = -1.5$ $Z \sim N(0, 1)$

Logit model MCAR $N=500$ $\alpha_0 = -2$ $Z \sim \text{Bern}(0.5)$

Method	N	BIAS			MSE		
		β_0	β_x	β_z	β_0	β_x	β_z
<hr/>							
Method	N	β_0	β_x	β_z	β_0	β_x	β_z
<hr/>							
Full	500	-0.00023	-0.00345	-0.00017	0.0019	0.0024	0.0020
CR	110	-0.00324	-0.00235	-0.00198	0.0143	0.0106	0.0136
Del	104	0.01079	-0.01431	-0.01812	0.0112	0.0111	0.0124
IPW1	110	0.00223	-0.00340	-0.00509	0.0129	0.0149	0.0248
IPW2	110	0.00309	-0.00401	-0.00721	0.0115	0.0146	0.0206
WT	110	0.00106	-0.00277	-0.00347	0.0091	0.0106	0.0102
IPWEF	110	-0.00097	0.02637	0.00094	0.0099	0.0174	0.0159
WTEF	110	0.00018	0.00869	-0.00275	0.0084	0.0064	0.0087

Linear model MAR
 $N=500$ $\alpha_0 = -1.5$ $Z \sim N$
(0, 1)

Method	N	BIAS			MSE		
		β_0	β_X	β_Z	β_0	β_X	β_Z
Full	500	-0.00286	-0.00089	0.00519	0.0021	0.0021	0.0022
CR	124	0.22911	-0.02213	-0.07032	0.0664	0.0077	0.0157
Del	35	0.00213	-0.04259	0.02422	0.1336	0.0824	0.0539
IPW1	124	0.00595	-0.00892	-0.00348	0.0192	0.0174	0.0327
IPW2	124	0.00620	-0.00965	-0.00367	0.0182	0.0170	0.0300
WT	124	-0.00463	-0.01044	0.00826	0.0156	0.0094	0.0138
IPWEF	124	-0.00977	0.03179	0.00963	0.0151	0.0355	0.0299
WTEF	124	-0.00999	0.00438	0.01125	0.0122	0.0058	0.0106

Table 2

Coverage probabilities based on 500 replications for eight estimators

Logit model $N=500$ $Z \sim \text{Bern}(0.5)$

Method	MCAR			MAR		
	β_0	β_X	β_Z	β_0	β_X	β_Z
Full	0.944	0.946	0.934	0.948	0.948	0.948
CR	0.958	0.966	0.944	0.518	0.964	0.954
Del	0.972	0.970	0.964	0.958	0.954	0.958
IPW1	0.956	0.962	0.950	0.968	0.956	0.974
IPW2	0.950	0.960	0.952	0.952	0.954	0.944
WT	0.960	0.962	0.948	0.950	0.960	0.944
IPWEF	0.950	0.960	0.952	0.952	0.954	0.944
WTEF	0.960	0.962	0.948	0.950	0.960	0.944

Logit model $N=500$ $Z \sim N(0, 1)$

Method	MCAR			MAR		
	β_0	β_X	β_Z	β_0	β_X	β_Z
Full	0.944	0.950	0.952	0.964	0.952	0.964
CR	0.926	0.952	0.940	0.440	0.956	0.960
Del	0.938	0.954	0.968	0.966	0.960	0.954
IPW1	0.944	0.954	0.934	0.956	0.956	0.960
IPW2	0.938	0.950	0.952	0.950	0.954	0.962
WT	0.932	0.950	0.948	0.952	0.958	0.960
IPWEF	0.942	0.950	0.954	0.960	0.960	0.956
WTEF	0.940	0.950	0.956	0.960	0.956	0.958

Linear model $N=500$ $Z \sim \text{Bern}(0.5)$

Logit model $N=500$ $Z \sim \text{Bern}(0.5)$

Method	MCAR			MAR		
	β_0	β_x	β_z	β_0	β_x	β_z
Method	β_0	β_x	β_z	β_0	β_x	β_z
Full	0.940	0.956	0.940	0.958	0.956	0.972
CR	0.958	0.956	0.958	0.536	0.928	0.918
Del	0.942	0.952	0.960	0.928	0.934	0.926
IPW1	0.934	0.952	0.962	0.954	0.928	0.960
IPW2	0.944	0.942	0.958	0.942	0.930	0.954
WT	0.944	0.946	0.958	0.944	0.946	0.952
IPWEF	0.952	0.956	0.956	0.954	0.950	0.954
WTEF	0.948	0.968	0.956	0.954	0.954	0.954

Linear model $N=500$ $Z \sim N(0, 1)$

Method	MCAR			MAR		
	β_0	β_x	β_z	β_0	β_x	β_z
Method	β_0	β_x	β_z	β_0	β_x	β_z
Full	0.958	0.950	0.968	0.952	0.942	0.938
CR	0.942	0.942	0.940	0.338	0.928	0.944
Del	0.948	0.954	0.938	0.942	0.920	0.924
IPW1	0.954	0.932	0.938	0.954	0.942	0.940
IPW2	0.952	0.934	0.932	0.952	0.946	0.938
WT	0.948	0.934	0.934	0.952	0.938	0.936
IPWEF	0.954	0.960	0.946	0.946	0.962	0.942
WTEF	0.950	0.960	0.944	0.948	0.964	0.932