Published in final edited form as:

Stat Med. 2019 July 10; 38(15): 2783–2796. doi:10.1002/sim.8141.

Best linear inverse probability weighted estimation for two-phase designs and missing covariate regression

Ching-Yun Wang and James Dai

Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, P.O. Box 19024, Seattle, WA 98109-1024, U.S.A.

SUMMARY

The inverse probability weighted estimator is often applied to two-phase designs and regression with missing covariates. Inverse probability weighted estimators typically are less efficient than likelihood-based estimators, but in general are more robust against model misspecification. In this paper, we propose a best linear inverse probability weighted estimator for two-phase designs and missing covariate regression. Our proposed estimator is the projection of the SIPW onto the orthogonal complement of the score space based on a *working regression model* of the observed covariate data. The efficiency gain is from the use of the association between the outcome variable and the available covariates, which is the working regression model. One advantage of the proposed estimator is that there is no need to calculate the augmented term of the augmented weighted estimator. The estimator can be applied to general missing data problems, or two-phase design studies in which the second phase data are obtained in a sub-cohort. The method can also be applied to secondary trait case-control genetic association studies. The asymptotic distribution is derived, and the finite sample performance of the proposed estimator is examined via extensive simulation studies. The methods are applied to a bladder cancer case-control study.

Keywords

Case-control sampling; Inverse selection weighting; Missing at random; Two-phase design

1 Introduction

Estimation of an exposure effect on disease incidence is a frequent objective in biomedical studies. Due to practical considerations, an exposure variable may not be available for all subjects from the study cohort. Incomplete data can be due to various reasons such as noncompliance with data acquisition requests or data may be missing by design to reduce costs. One complication in applying the model in practice is that information on the risk factors may not be available for all subjects involved in the study. Methodology for missing data is often required to address the challenges.

In many studies, covariates data may be missing due to non-response to a survey item. For example, we consider a case-control study of bladder cancer conducted at the Fred Hutchinson Cancer Research Center. This population-based case-control study was designed to address the association between bladder cancer and nutrients.¹ We are interested in covariates smoking pack year and body mass index (BMI, calculated as weight/height²). The

smoking pack year of a participant is defined as the average number of cigarette packs smoked per day multiplied by the years one has been smoking. In this study, smoking pack year is not available for some study subjects since they did not respond to the question of the number of cigarettes smoked per day. Nevertheless, the majority of subjects did have the data on the number of years that they had smoked, namely *smoking year*. Smoking year in this example is a surrogate for smoking pack year. A surrogate variable does not have an effect on the disease incidence given the true covariates. In this problem, a simple approach to effect estimation is the complete-case (CC) analysis, which is to perform the usual analysis using the observations in a subsample with complete data, namely the *complete-case set*, or *validation set*. However, this approach can give rise to bias and reduced statistical efficiency. In addition to the CC estimator, there is a rich body of literature on missing data regression; see Little and Rubin for a review.² By using the terminology of Rubin, data are missing completely at random (MCAR) when the missing data process is independent of both outcomes and covariates, and data are missing at random (MAR) when the missing data process depends on the observed outcomes and covariates.³ In contrast, a nonignorable missing data process is one that depends on unobserved data. Imputation and weighting are 2 important approaches in dealing with missing data problems.^{4,5} Wang, et al. showed that under many situations some inverse selection probability weighted estimators are numerically equivalent to imputation of estimating scores.⁶ The selection probability in the missing data context is defined as the probability of an individual to be in the CC (or validation) set. The idea of the simple inverse probability weighted (SIPW) estimator is to adjust the CC analysis by using the inverse of the selection probability as the weight. The SIPW estimator and some semiparametric estimators were applied to randomized response data with missing covariates in logistic regression.⁷ In addition to weighting, a missing variable can be imputed from the mean conditional on all observed variables, or some of the variables that are related to the missing variable. It is often called the *conditional mean imputation*, *regression imputation*.² The conditional mean imputation can be implemented by establishing another regression model in the complete-case subset, with the missing variable as the outcome and the observed variables as the covariates. However, implementation of the conditional mean imputation may be tedious when a complete-case subset is too small or does not exist. Instead of filling in a single value for each missing value, a multiple imputation procedure imputes multiple sets of data for missing variables. The multiple imputed data sets are then analyzed by using some standard procedures for complete data and combining the results from these analysis to adjust for uncertainty due to missing data. Wu and Wu proposed a three-step hierarchical multiple-imputation method, implemented by Gibbs sampling, for viral dynamic models in acquired immune deficiency syndrome research.⁸ In addition to the well-known methods described above, Chatterjee et al. proposed a pseudoscore (PS) estimator.⁹ The PS estimator approximates the likelihood score in the non-CC subset by calculating the conditional distribution of the missing covariates given the observed covariates, which will be described later in the next section.

In this paper, we propose a new method for regression analysis when the covariates may be MAR. Our proposed estimator is the projection of the SIPW onto the orthogonal complement of the score space based on a *working regression model* of the observed covariate data. The main idea of our approach to obtaining additional information is to make

use of the association between the outcome variable and the available covariates, which is the working regression model mentioned above. A main motivation of the new estimator is similar to the basic idea of seemingly unrelated regression models.¹⁰ The general idea is that in addition to the estimating score of the primary regression, the new estimator would make use of a second regression model from the observed data so that the efficiency of the parameter estimation of the primary regression can be improved from the combined inference. In our problem, our goal is to improve the estimation of the regression coefficients by adding the second (working) regression model based on the observed variables. The paper is organized as follows. In Section 2, we describe a few existing methods, including a full normal multiple imputation estimator, the SIPW estimator and augmented inverse probability weighted estimator. Because the calculation of the augmentation term is somewhat complicated, this leads to our development of best linear estimation in Section 3. In Section 4, we apply the methods to the secondary trait association in case-control genetic studies. In Section 5, results from simulation studies will be presented. Some finite sample issues will be noted. In Section 6, we illustrate the proposed method on the bladder cancer case-control study described above. Some concluding remarks are given in Section 7.

2 Regression Model and Existing Methods

Assume that the total sample size of the study cohort is n . The regression model of interest is the generalized linear model. Let Y_i be the response variable, X_i be the vector of covariates which may be missing among some individuals, Z_i be the vector of covariates which is available for all individuals, $i = 1, \dots, n$. Let W_i be a surrogate vector for X_i , which is relevant to X_i but is independent of Y_i given (X_i, Z_i) . In the bladder cancer example described in the introduction, smoking pack year is the main exposure variable of interest while smoking year is the surrogate variable for pack year. The surrogate assumption seems reasonable since the effect of smoking on bladder cancer is likely determined by pack year: smoking year will not likely affect cancer incidence given that the effect has been determined by pack year. The main interest is to estimate the vector of regression coefficients β in the following regression model

$$E(Y_i | X_i, Z_i) = g(\beta_0 + \beta_1' X_i + \beta_2' Z_i),$$

$i = 1, \dots, n$, where $g(\cdot)$ is a specified function. For example, $g(u) = u$ in linear regression, while $g(u) = (1 + e^{-u})^{-1}$ in logistic regression. Let R_i indicate whether X_i is observed ($R_i = 1$) or not ($R_i = 0$). The CC data set ($R_i = 1$) consists of (Y_i, X_i, Z_i, W_i) , and the non-CC data set ($R_i = 0$) consists of (Y_i, Z_i, W_i) . The selection probability $\pi_i (i = 1, \dots, n)$ is defined as $\text{pr}(R_i = 1 | Y_i, W_i, Z_i)$, where n is the number of all observations. When X is MCAR, π_i is a constant, while X is MAR if π_i depends on the observed data. If X_i is nonignorable missing, then π_i may depend on X_i . In this paper, our methodology development is under the assumption that X is MAR.

Regression imputation is also called conditional mean imputation, namely that a missing variable is imputed by its conditional mean value given the observed covariates. If the regression model is linear such that $E(Y_i | X_i, Z_i) = \beta_0 + \beta_1' X_i + \beta_2' Z_i$, then it is easily seen that

replacing the missing X_i by $E(X_i|Z_i, W_i)$ will lead to consistent estimation in linear regression when MCAR. In logistic regression, regression imputation is only an approximation since $E\{Y_i|Z_i, W_i\}$ is different from $H(\beta_0 + \beta_2'E(X_i|Z_i, W_i) + \beta_2'Z_i)$, where $H(u) = (1 + e^{-u})^{-1}$. In addition to regression imputation, multiple imputation (MI) has been considered a useful approach which also take into consideration uncertainty due to imputation. Rubin and Schenker studied the full normal MI for observations from a normal distribution.¹¹ In our setting we may assume that covariate X_i given Y_i, Z_i and W_i is linear such that $X_i = \alpha_0 + \alpha_1 Y_i + \alpha_2' Z_i + \alpha_3' W_i + \epsilon_i$, where ϵ_i is normal with mean 0 and variance σ_e^2 . Additional details on full normal MI can be seen in Rubin and Schenker.¹¹ To adjust for the incompleteness of the data, a simple weighted estimator is to take the inverse of the selection probabilities as weights. The simple inverse probability weighted (SIPW) estimator under the regression model is to solve the following weighted estimating equation:

$$\sum_{i=1}^n \frac{R_i}{\pi_i} \phi(Y_i, X_i, Z_i) = 0, \quad (1)$$

where $\phi(Y_i, X_i, Z_i) = (1, X_i', Z_i')' \{Y_i - g(X_i, Z_i)\}$. The selection probabilities π_i , $i = 1, \dots, n$, are known in a two-stage design study, but not the case when the missing data mechanism is by happenstance. Estimating the selection probabilities by parametric or non-parametric methods have been investigated if X is MAR.⁵ See also a related weighted score method in missing outcome regression.¹² For the use of the SIPW estimator, it will be more efficient to use the estimated π values even if the true π 's are known.

The SIPW estimator applies the inverse probability weights to the estimating function for subjects from the CC set. However, it does not directly apply the expected estimating function for subjects from the non-CC set. Therefore, it is natural to include an augmented term to gain efficiency. This idea can be generated to the class of weighted estimators proposed in Robins et al.⁴ One of the simplest estimator among the class is the following augmented inverse probability weighted (AIPW) estimator which solves

$$\sum_{i=1}^n \left[\frac{R_i}{\pi_i} \phi(Y_i, X_i, Z_i) + \left(1 - \frac{R_i}{\pi_i}\right) \widehat{\mathcal{A}}(Y_i, Z_i, W_i) \right] = 0,$$

where $\widehat{\mathcal{A}}(Y_i, Z_i, W_i) = \widehat{E}\{\phi(Y_i, X_i, Z_i) | Y_i, Z_i, W_i\}$. Here, $\mathcal{A}(Y_i, Z_i, W_i)$ is the augmentation term that can be estimated parametrically or nonparametrically.^{5,13} Generally AIPW is more efficient than SIPW, but under some situations they are identical or asymptotically equivalent.^{6,14}

The augmentation term of the AIPW estimator involves the conditional distribution of X given all variables, which can be estimated parametrically or non-parametrically. The conditional expectation of the full data score given the observed data can be calculated by the following:

$$\mathcal{A}(Y_i, Z_i, W_i) = \frac{\int_x \phi(Y_i, x, Z_i) \mathcal{L}(Y_i | X_i, Z_i) \mathcal{L}(X_i | Z_i, W_i)}{\int_x \mathcal{L}(Y_i | X = x, Z_i) \mathcal{L}(X = x | Z_i, W_i)},$$

where $\mathcal{L}(Y_i | X_i, Z_i)$ denotes the likelihood function of Y_i given (X_i, Z_i) . If we assume that X given Z is normally distributed, then the nuisance parameters given above can be estimated consistently. However, calculation of the augmentation term may be tedious under some situations, and could be biased due to model misspecification of the conditional distribution. Along with the main motivation mentioned in the introduction, this is the second motivation for the development of the new method to be described in the next section.

The PS estimator of Chatterjee et al. is based on the idea of estimating the likelihood score among the individuals in the non-CC subset.⁹ Let $q(X, Z, W) = \text{pr}(R = 1 | X, Z, W)$, $h(Y, X, Z, W) = \mathcal{L}(Y | X, Z) / q(X, Z, W)$. The PS estimator solves $\sum_{i=1}^n [\phi(Y_i, X_i, Z_i) + (1 - R_i) \hat{E}\{\phi(Y_i, X_i, Z_i) | Y_i, Z_i, W_i\}] = 0$, where

$$\hat{E}\{\phi(Y_i, X_i, Z_i) | Y_i, Z_i, W_i\} = \frac{\int_x \phi(Y_i, x, Z_i) h(Y_i, x, Z_i, W_i) \mathcal{L}(x | Z_i, W_i, R_i = 1) dx}{\int_x h(Y_i, x, Z_i, W_i) \mathcal{L}(x | Z_i, W_i, R_i = 1) dx} \quad (2)$$

The estimation of the conditional score given the observed data in (2) can be obtained by the empirical averages if Y, Z and W are discrete. One advantage of the PS estimator that it improves efficiency by the use of the conditional distribution of X given $(Z, W, R = 1)$. But, on the other hand, there could be model misspecification in the estimation of (2). When Y is continuous, estimation of $q(X, Z, W)$ may involve model specification of the regression model of Y given (X, Z) . To avoid bias due to misspecification of Y given (X, Z) in (2), the conditional expectation can be estimated consistently with nonparametric smoothers of the estimating score given (Y, Z, W) .

3 Best Linear Inverse Probability Weighted Estimator

As reviewed in the last section, the AIPW estimator is generally more efficient than the SIPW estimator but the implementation could be more technical. The idea now is how to improve the efficiency of the SIPW estimator by including the non-CC sample without complicated calculations. Similar to the concept of seemingly unrelated regression models,¹⁰ our goal is to improve the estimation of β by making use of the available data from the whole cohort with the second (working) regression model of Y given (W, Z) , or Y given Z if in case a surrogate variable of X is not available. The idea is that the new estimator would make use of a second regression model so that the efficiency of the parameter estimation of the primary regression can be improved from the combined inference. In general, the exact association of Y given (W, Z) may not have a simple form. However, we might be able to achieve our goal of improving the estimation of β even though an assumed (*working*) association between Y and W and Z may not be correct. We assume that Y given (W, Z) also follows the same model as Y given (X, Z) , which is a working model, with the vector of

parameters being denoted by $\theta = (\theta_0, \theta_1', \theta_2')$. Let the dimension of β be p and the dimension of θ be p' . In general, p may be different from p' . Let $\hat{\beta}_{sipw}$ denote the SIPW estimator from the CC sample. Let the estimator using covariates (W, Z) based on the whole cohort be denoted by $\hat{\theta}$, and the corresponding SIPW estimator using the CC sample be denoted by $\hat{\theta}_{sipw}$. We intend to propose a method that is the best among the linear combination of $\hat{\beta}_{sipw}$ and $\hat{\theta}_{sipw} - \hat{\theta}$. That is, we aim to construct an estimator that is the most efficient among the class of estimators $\hat{\beta}_{sipw} + \Delta(\hat{\theta}_{sipw} - \hat{\theta})$, for any Δ of dimension $p \times p'$. By the calculation given in the Appendix, the optimal Δ satisfies:

$$\Delta \text{var}\left\{n^{1/2}(\hat{\theta}_{sipw} - \hat{\theta})\right\} = -\text{cov}\left\{n^{1/2}(\hat{\beta}_{sipw} - \beta), n^{1/2}(\hat{\theta}_{sipw} - \hat{\theta})\right\}.$$

The Δ is optimal in terms of minimizing the variance among the class of consistent estimators $\hat{\beta}_{sipw} + \Delta(\hat{\theta}_{sipw} - \hat{\theta})$. Therefore, the optimal estimator can be obtained by calculating the variances and covariances of $\hat{\beta}_{sipw}$ and $\hat{\theta}_{sipw} - \hat{\theta}$. The optimal estimator in general will be more efficient than the SIPW estimator (using the CC sample only).

In the CC sample, the SIPW estimator of θ solves the following estimating equation:

$$\sum_{i=1}^n (R_i / \hat{\pi}_i) \begin{pmatrix} 1 \\ W_i \\ Z_i \end{pmatrix} \{Y_i - g(\theta_0 + \theta_1' W_i + \theta_2' Z_i)\} \equiv \sum_{i=1}^n R_i / \hat{\pi}_i \psi(Y_i, W_i, Z_i, \theta) = 0. \quad (3)$$

As discussed, the equation given above is a working model which in general will not hold and hence the limit of the SIPW estimator of θ may not have the same dimension as β . Similarly, $\hat{\theta}$ is the estimator using the observed covariates (W, Z) based on the whole cohort solving the following estimating equation:

$$\sum_{i=1}^n \begin{pmatrix} 1 \\ W_i \\ Z_i \end{pmatrix} \{Y_i - g(\theta_0 + \theta_1' W_i + \theta_2' Z_i)\} \equiv \sum_{i=1}^n \psi(Y_i, W_i, Z_i, \theta) = 0.$$

When the selection probabilities are estimated, the estimating equations (1) and (3) for $\hat{\beta}_{sipw}$ and $\hat{\theta}_{sipw}$ are not summations of independent scores and hence the asymptotic distribution of the proposed BLIPW estimator is slightly complicated. To simplify the presentation, in this rest of section we assume that the selection probabilities are known. The development of the proposed estimator under unknown selection probabilities is provided in Appendix B. By some calculation (Appendix A) it can be shown that the optimal Δ is $G_1^{-1}(\beta) \{C_{13}(\beta, \theta) - C_{12}(\beta, \theta)\} \{C_{22}(\theta) - C_{33}(\theta)\}^{-1} G_2(\theta)$, where $G_1(\beta) = -E\{\phi(Y, X, Z, \beta) / \beta\}$, $G_2(\theta) = -E\{\psi(Y, W, Z, \theta) / \theta\}$, $C_{12}(\beta, \theta) = E[\{\pi(Y, Z, W)\}^{-1} \phi(Y, X, Z, \beta) \psi'(Y, W, Z, \theta)]$, $C_{13}(\beta, \theta) = E[\{\pi(Y, Z, W)\}^{-1} \phi(Y, X, Z, \beta)]$, $C_{22}(\theta) = E[\{\pi(Y, Z, W)\}^{-1} \psi(Y, W, Z, \theta) \psi'(Y, W, Z, \theta)]$, and $C_{33}(\theta) = E[\{\pi(Y, Z, W)\}^{-1} \psi(Y, W, Z, \theta)]$.

$\mathbf{Z}, \boldsymbol{\theta}\}$, $C_{13}(\boldsymbol{\beta}, \boldsymbol{\theta}) = E\{\phi(Y, X, \mathbf{Z}, \boldsymbol{\beta})\psi'(Y, \mathbf{W}, \mathbf{Z}, \boldsymbol{\theta})\}$, $C_{22}(\boldsymbol{\theta}) = E[\{\pi(Y, \mathbf{Z}, \mathbf{W})\}^{-1}\psi(Y, \mathbf{W}, \mathbf{Z}, \boldsymbol{\theta})\psi'(Y, \mathbf{W}, \mathbf{Z}, \boldsymbol{\theta})]$, and $C_{33}(\boldsymbol{\theta}) = E[\psi(Y, \mathbf{W}, \mathbf{Z}, \boldsymbol{\theta})\psi'(Y, \mathbf{W}, \mathbf{Z}, \boldsymbol{\theta})]$.

This suggests that a new estimator, namely the best linear inverse probability weighted (BLIPW) estimator, can be obtained by

$$\hat{\boldsymbol{\beta}}_{sipw} + \mathbf{G}_1^{-1}(\hat{\boldsymbol{\beta}}_{sipw})\{C_{13}(\hat{\boldsymbol{\beta}}_{sipw}, \hat{\boldsymbol{\theta}}) - C_{12}(\hat{\boldsymbol{\beta}}_{sipw}, \hat{\boldsymbol{\theta}})\}\{C_{22}(\hat{\boldsymbol{\theta}}) - C_{33}(\hat{\boldsymbol{\theta}})\}^{-1}\mathbf{G}_2(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_{sipw} - \hat{\boldsymbol{\theta}}).$$

The proposed BLIPW estimator is like a two-step estimator. In the first step, we use the SIPW estimator from the CC set. In the second step, the BLIPW estimator is obtained by the best linear combination of the SIPW estimator and $\hat{\boldsymbol{\theta}}_{sipw} - \hat{\boldsymbol{\theta}}$. In the development of the proposed BLIPW estimator given above, the selection probabilities π_i , $i = 1, \dots, n$ are assumed known. The asymptotic variance for the situation with unknown selection probabilities are given in Appendix B.

4 Secondary Trait Association in Case-Control Genetic Studies

The methods described in the previous sections can be applied to secondary trait associations in case-control studies. Case-control genetic association studies are often nested within large cohorts. In addition to the primary case-control status that drives sample selection for genotyping, the cohorts also typically have extensive measures of covariates, and other intermediate phenotypes, namely *secondary traits*. After interrogation of primary case-control genetic associations, interest often arises in assessing genetic associations with secondary traits. Over the past decade, there has been a proliferation of genome-wide genotyping studies addressing secondary trait analysis of common variants.^{15,16} Reported secondary trait associations include height, and lipid levels, often through meta-analysis of multiple studies. Hence, secondary trait genetic analysis is of great importance. For example, Genomics and Randomized Trial Network (GARNET) is one of several case-control genetic studies nested within the Women Health Initiative (WHI) study. A case-control sample of size 4250 was drawn from WHI hormone therapy trial white (22030) participants. Cases consisted of four diseases: stroke, myocardial infarction, venous thrombotic disease and type 2 diabetes. About 1 million SNPs and data on some eigen vectors are available only in GARNET sample. Cohort data include case-control status, age, BMI, systolic and diastolic blood pressure (secondary trait) measured at baseline and during follow up visits.

We consider a case-control study nested in a cohort of n subjects. All participants in the cohort are ascertained for a dichotomous clinical endpoint D , with $D = 1$ coded for disease (case) and $D = 0$ for no disease (control). Let R be the indicator variable for whether a participant is included in the case-control sample. Let X denote the genetic variant, coded as 0, 1, or 2 depending on the number of variant alleles. Suppose there are also a secondary trait variable Y and a vector of confounding variables $\mathbf{Z}$ to be adjusted for (e.g., age). Typically for rare diseases, all cases and a small proportion of controls are included for genotyping. The interest is in assessing the secondary trait association between the secondary trait (Y) and risk factors $X, \mathbf{Z}$, under regression $E(Y_i | X_i, \mathbf{Z}_i) = g(\beta_0 + \beta_1 X_i + \beta_2^t \mathbf{Z}_i)$, as described earlier. The association of the disease outcome D_i given $(X_i, Y_i, \mathbf{Z}_i)$ is $H(\gamma_0 + \gamma_1 X_i + \gamma_2 Y_i + \gamma_3^t \mathbf{Z}_i)$, in which $H(\cdot)$ is the logistic probability function defined in Section

2. For assessment of genetic association with D , a case-control sample is drawn from the cohort. Here the selection probability is a function of (Y, Z, D) . In this application X is MAR. The methods described in the last few sections can be applied to this application. Tapsoba et al. showed that the SIPW estimator has the advantage of being robust to the model assumption of the association between the disease and the covariates when estimating the regression coefficients of the secondary trait association.¹⁷ In addition to the SIPW, the MI and the proposed BLIPW estimator can also be applied to this setting. But, the disease status variable D will not play the role of a surrogate variable W in the last few sections since it will still be correlated with Y given (X, Z) .

5 Simulation Study

We compared small sample properties of the methods that we described in the last few sections. The complete-case (CC) analysis is based on the analysis using the data from the individuals who do not have any missing data. The full normal MI method was calculated according to the methods described earlier. To calculate the SIPW estimator, we estimated the selection probabilities parametrically. The calculation of the BLIPW was based on the estimating equation described in Section 3. We also calculated the PS and AIPW estimators. In Table 1, we generated X_i , $i = 1, \dots, n$ from a standard normal distribution, where $n = 1000$. Then we generated binary Z with $\text{pr}(Z = 0) = \text{pr}(Z = 1) = 0.5$ such that the correlation coefficient between X and Z is about 0.25. The outcome variables Y_i given X_i and Z_i were generated by linear regression with coefficients $\beta = (0, -0.5, 0.5)$ and $\beta = (0, -1.5, 1.5)$, respectively. The selection probability was generated by $\text{pr}(R = 1 | Y, Z) = H(-0.5Y + 2Z)$. Under this model the missing data probability is 33% when $\beta = (0, -0.5, 0.5)$, and 38% when $\beta = (0, -1.5, 1.5)$, respectively. In the table, “Bias” is the average of $\hat{\beta} - \beta$ from 500 replicates, “SD” denotes the sample standard deviation of the estimates, “ASE” denotes the average of the estimated standard errors of the estimates, and “CP” is the 95% Wald-type confidence interval coverage probability. The CC estimates have bias, while the MI, SIPW, BLIPW, PS and AIPW estimates do not have a bias problem. The MI estimator is in general more efficient than the SIPW, BLIPW, PS and AIPW estimators under this setting when X is MAR and normally distributed. One reason that the MI is better is primarily because the MI is based on a normally distribution assumption of X , and hence it does not have bias and is efficient. The BLIPW estimator is more efficient than the SIPW estimator. The BLIPW estimator for β_1 is slightly more efficient than the PS and AIPW estimators (although they are very close).

The data for Table 2 were generated almost the same as these in Table 1 except that the X covariates were from a mixture of two normals with means $(-1, 2)$, variances $(1/4, 1)$, and the mixture percentages were $(2/3, 1/3)$. Under this mechanism, X_i had mean 0 and variance 1. The findings in Table 2 are similar to those in Table 1 except the performance of MI; the MI has biases. The MI approach has been often applied in research, but the caution here is that it may be invalid if X is MAR with skewed covariates. Under this situation, the SIPW, BLIPW, PS and AIPW estimators are consistent. Similar to Table 1, the BLIPW estimator is more efficient than the SIPW estimator. The BLIPW estimator is slightly more efficient than the PS and AIPW estimators (although very close).

In Table 3, we generated Y_i as a binary variable which followed logistic regression for given X_i and Z_i with $\beta = (-\log(2), \log(2), \log(2))$, and $\beta = (-\log(4), \log(4), \log(4))$, respectively. The X covariates were from a mixture of two normals with means $(-1, 2)$, variances $(1/4, 1)$, and the mixture percentages were $(2/3, 1/3)$. We generated $W_i = I[0.5X + 0.87U_i > 0]$, where $U_i (i = 1, \dots, n)$ were from a standard normal distribution. We generated binary $Z_i = I[V_i > 0]$ where $V_i (i = 1, \dots, n)$ were from a standard normal distribution. The disease rate is about 43% under the data generating scheme. The selection probability is $\text{pr}(R = 1 | Y, Z, W) = H(-0.5Y + 0.5Z + W)$, and the CC set is about 62% of the whole cohort under this model. Under this setting, the MI has biases as well. The SIPW, BLIPW, PS and AIPW estimators are consistent. The PS is the most efficient under this setup, but the efficiency of the BLIPW and AIPW are very close to the PS estimator. The PS estimator is more efficient is probably because it is to estimate the estimating score directly based on the likelihood of the observed data. But, on the other hand, in general the PS estimator is not as robust as the BLIPW or AIPW to the distributions involved in the observed data, which will be demonstrated in the next table.

In Table 4, we conducted simulations under secondary trait analysis. The covariates X_i and Z_i were generated by the same distribution as in Table 2. The regression coefficients for Y (binary secondary trait) given X, Z were $\beta = (-\log(2), \log(2), \log(2))$, and $\beta = (-\log(2), \log(4), \log(4))$, respectively. The disease association regression D given (X, Y, Z) were logistic with coefficients $\gamma = (-3, \log(3), \log(3), \log(3))$. We consider the total cohort size of $n = 3,000$. The disease rate ($D = 1$) was about 0.16 based on the parameters given above. When $n = 3,000$, we generated about 200 cases and 200 controls, with X being observed among the case-control sample. The CC and MI estimators have biases since they are not consistent under this setting. The SIPW, BLIPW and AIPW estimators are consistent. The BLIPW and AIPW are more efficient than the SIPW estimator. The AIPW estimator is slightly more efficient than BLIPW estimator, but very close. However, there is complication in the calculation of the PS estimator under this setting. The PS as in (2) will be $E\{\phi(Y_i, X_i, Z_i) | Y_i, Z_i, D_i\}$ but D_i is no longer conditionally independent of Y_i given (X_i, Z_i) . That is, the disease status D_i in the secondary trait setting is not the same as the W_i in (2). Therefore, the PS estimator is biased under the secondary trait setting given that the calculation of the right-hand side of (2) has a model misspecification in the distribution of Y given (X, Z, D) . Additional work is needed to obtain a consistent PS estimator under this secondary trait regression model.

6 DATA ANALYSIS

We applied the methods to the case-control study of bladder cancer described in the Introduction. Eligible subjects were residents of 3 counties of western Washington state who were diagnosed between January 1987 and June 1990 with invasive or noninvasive bladder cancer. This population based case-control study was designed to address the association between bladder cancer and some nutrients. We use the data here for illustrative purposes. Some detailed results can be found in Bruemmer, et al.¹

In this data application, the response variable is the bladder cancer history and the covariate X is the smoking pack year, Z is obesity (BMI is 30 or over). The original study cohort had a

total of 262 cases and 405 controls. In our data application of the methods, our focus is on the effect of smoking among current smokers. After excluding the non-smokers and past smokers, the sample size in the analysis was 251, with 121 bladder cancer cases. There are 72 individuals with missing smoking pack year. In the CC set, there are 121 cases and 58 controls. We consider smoking year (W) as a surrogate variable for smoking pack year (X) in the application. Figure 1 shows the association between smoking pack year and smoking year based the CC data, along with a lowess curve. The correlation coefficient between smoking pack year and smoking year is 0.42 based on the CC data. Smoking year serves reasonably well as a surrogate variable for smoking pack year.

We also checked the missing data mechanism in the application. The selection probabilities for cases are 1 since the cases all have their pack year data. For the controls, we examined smooth curve of the selection probabilities versus smoking year. The selection probabilities for the controls are in general an increasing function of smoking year. Hence, we conducted a logistic regression analysis of R on (Z, W, ZW) (among the controls). The modeling of the selection probabilities appeared reasonable based on the data.

The analysis results are given in Table 5. From the CC analysis, smoking pack year or obesity does not have a significant effect on bladder cancer risk among current smokers. For the effect of smoking pack year, MI, SIPW and BLIPW are significant. The effect of smoking pack year estimated from the PS and AIPW are close to the significant level. One reason that the CC estimator of pack year is weaker than the other estimators is probably because the selection of the CC set is positively associated with the surrogate variable (smoking year). Because of the missingness mechanism, we tend to believe that the estimates from the SIPW, BLIPW, PS and AIPW estimates are more reliable than the CC analysis. In general, the result supports that smoking, measured by packet year, increases the risk of bladder cancer. We did not find a significant association between obesity and bladder cancer in the analysis. This is most likely due to the small sample size.

7 Discussion

We have proposed a new method for regression analysis when covariates may be missing. The new estimator has the advantage of the SIPW estimator and is more efficient, and it does not involve complicated calculations that are needed in carrying out the AIPW estimator. Our BLIPW estimator can be considered as an extension of Chen and Chen, which is consistent when the covariates are MCAR.¹⁸ However, the method of Chen and Chen is not consistent if the selection probabilities depend on the outcome and observed covariates. In addition to the simulation results in the paper, we did investigate the regression imputation (RI) estimator described earlier in Section 2. To keep focus on the presentation, the RI estimates are not included in the tables. From our simulation results, we also noted that in linear regression, when X is MCAR, the RI estimator is valid but it could be inefficient when the slope parameters are large. However, the RI could be biased when covariates are MAR. Therefore, the RI is not a useful method either for large coefficients for the missing covariate, or when the missing data mechanism is not completely at random.

MI has been a popular method for missing data. However, practitioners should be cautious when applying a specific MI method. From the research, we have discussed that there may be situations when it may provide invalid inference. Our simulation results indicated that the full normal MI may have large biases when covariates are MAR and when the covariates are skewed. Qi et al. investigated weighted estimation and MI for censored outcome regression.¹⁹ They also cautioned about model misspecification when applying an MI approach.

The proposed BLIPW estimator can be treated as an improved estimator of SIPW. In Tables 1 and 2, overall the BLIPW estimator is slightly more efficient than the PS and AIPW estimators (although they are very close). When Y is continuous, estimation of the estimating score in (2) may involve model specification of the regression model of Y given (X, Z) . To avoid bias due to misspecification of Y given (X, Z) in (2), we estimated the conditional expectation by nonparametric smoothers of the estimating score given (Y, Z, W) . In both Tables 3 and 4, the AIPW estimator is slightly more efficient than BLIPW estimator, but very close. On the other hand, the BLIPW estimator is relatively easier in the implementation. In Table 3 under logistic regression, the PS is slightly more efficient than the BLIPW estimator. This is probably because for logistic regression the estimating score of the PS estimator makes use of the derivation of the distribution of X given $(Y, Z, W, R = 1)$, as in (2). But, on the other hand, the BLIPW estimator is in general more robust than the PS estimator. This is seen in Table 4 under secondary trait regression in which the PS estimator is biased when there is a model misspecification for Y given (X, Z, D) .

As we mentioned earlier, the proposed BLIPW estimator is motivated by the idea of improving the estimation of the regression coefficients by adding the second (working) regression model of the observed variables. The efficiency gain is from the use of the association between the outcome variable and the available covariates and surrogates, which is the working regression model. Another viewpoint is that we could treat the BLIPW estimator as a control variable approach,²⁰ in which W plays the role of a control variable. The efficiency of the BLIPW estimator is improved via the control variable approach which exploits information about the errors in estimates of known quantities to reduce the error of an estimate of an unknown quantity.

Acknowledgments and Data Availability Statement

This research was partially supported by US National Institutes of Health grants CA53996 (Wang and Dai), GM100573 (Wang), HL114901 (Wang and Dai), HL130483 (Wang), MH105857 (Wang), and a travel award from the Mathematics Research Promotion Center of National Science Council of Taiwan (Wang). The data that support the findings of this study are available from Dr. Emily White. Restrictions apply to the availability of these data. Data are available from Dr. Ching-Yun Wang with the permission of Dr. Emily White.

Appendix A:: Asymptotic Variance of BLIPW Estimator with known selection probabilities

We now show that the BLIPW estimator is the most efficient among the class of $\hat{\beta}_{sipw} - \Delta(\hat{\theta}_{sipw} - \hat{\theta})$, for any of dimension $p \times p'$. We assume that the selection probabilities are known for simplicity in this subappendix. The optimal satisfies:

$$\frac{\partial}{\partial \Delta} \text{var} \left[n^{1/2} \left\{ \left(\hat{\beta}_{sipw} - \beta \right) + \Delta \left(\hat{\theta}_{sipw} - \hat{\theta} \right) \right\} \right] = 0.$$

By direct calculation, the optimal Δ satisfies the following:

$$\Delta \text{var} \left\{ n^{1/2} \left(\hat{\theta}_{sipw} - \hat{\theta} \right) \right\} = - \text{cov} \left\{ n^{1/2} \left(\hat{\beta}_{sipw} - \beta \right), n^{1/2} \left(\hat{\theta}_{sipw} - \hat{\theta} \right) \right\}.$$

It can be shown that

$$\text{var} \left\{ n^{1/2} \left(\hat{\theta}_{sipw} - \hat{\theta} \right) \right\} = G_2^{-1}(\theta) \{ C_{22}(\theta) - C_{33}(\theta) \} \{ G_2^{-1}(\theta) \}',$$

where $G_2(\theta)$, $C_{22}(\theta)$ and $C_{33}(\theta)$ are given in Section 3. It can also be shown that

$$\text{cov} \left\{ n^{1/2} \left(\hat{\beta}_{sipw} - \beta \right), n^{1/2} \left(\hat{\theta}_{sipw} - \hat{\theta} \right) \right\} = G_1^{-1}(\beta) \{ C_{12}(\beta, \theta) - C_{13}(\beta, \theta) \} \{ G_2^{-1}(\theta) \}',$$

where $G_1(\beta)$, $C_{12}(\beta, \theta)$ and $C_{13}(\beta, \theta)$ are given in Section 3. Hence, the optimal Δ is $G_1^{-1}(\beta) \{ C_{13}(\beta, \theta) - C_{12}(\beta, \theta) \} \{ C_{22}(\theta) - C_{33}(\theta) \}^{-1} G_2(\theta) \equiv \Delta_{opt}$. We now calculate the variance of $n^{1/2} \left(\hat{\beta}_{sipw} - \beta \right) - \Delta_{opt} n^{1/2} \left(\hat{\theta}_{sipw} - \hat{\theta} \right)$. It can be shown by standard calculations that $\text{var} \left\{ n^{1/2} \left(\hat{\beta}_{sipw} - \beta \right) \right\} = G_1^{-1}(\beta) C_{11}(\beta) \{ G_1^{-1}(\beta) \}'$. Therefore, by straight-forward calculations it can be shown that

$$\begin{aligned} \text{var} \left\{ n^{1/2} \left(\hat{\beta}_{blipw} - \beta \right) \right\} &= G_1^{-1}(\beta) C_{11}(\beta) \{ G_1^{-1}(\beta) \}' - G_1^{-1}(\beta) \\ &\quad \{ C_{13}(\beta, \theta) - C_{12}(\beta, \theta) \} \left[\{ C_{22}(\theta) - C_{33}(\theta) \}^{-1} \right]' \{ C_{13}(\beta, \theta) - C_{12}(\beta, \theta) \} \{ G_1^{-1}(\beta) \}'. \end{aligned}$$

Appendix B:: Asymptotic Variance of BLIPW Estimator with estimated selection probabilities

When the selection probability π is estimated, we write $\hat{\beta}_{sipw}$ as $\hat{\beta}_{sipw}(\hat{\pi})$, and $\hat{\theta}_{sipw}$ as $\hat{\theta}_{sipw}(\hat{\pi})$. The optimal Δ would minimize the variance of

$$n^{1/2} \left[\left\{ \left(\hat{\beta}_{sipw}(\hat{\pi}) - \beta \right) + \Delta \left(\hat{\theta}_{sipw}(\hat{\pi}) - \hat{\theta} \right) \right\} \right].$$

By similar calculation as in Appendix A, the optimal Δ can be shown to be

$$\Delta^* = \left(\text{cov}[n^{1/2}\hat{\beta}_{sipw}(\hat{\pi}) - \hat{\beta}], n^{1/2}(\hat{\theta} - \theta) \right) - \text{cov}\left[n^{1/2}\{\hat{\beta}_{sipw}(\hat{\pi}) - \beta\}, n^{1/2}\{\hat{\theta}_{sipw}(\hat{\pi}) - \theta\}\right] \\ \times \left(\text{var}\left[n^{1/2}\{\hat{\theta}_{sipw}(\hat{\pi}) - \theta\}\right] \right)^{-1} \quad (4)$$

Let the selection probability π be a function of a vector of parameters γ , and the maximum likelihood estimator of γ is denoted by $\hat{\gamma}$. Let $\mathcal{B} = (\beta', \gamma')'$, and $\Theta = (\theta', \gamma')'$. Let $\Phi(Y_i, X_i, Z_i, W_i, \mathcal{B})$ be the vector of the stack estimating equations for $\hat{\beta}_{sipw}$ and $\hat{\gamma}$ using the observed data from the i th subject. That is, $\Phi(Y_i, X_i, Z_i, W_i, \mathcal{B})$ consists of $(R_i/\pi_i)\phi(Y_i, X_i, Z_i, \beta)$ and the estimating score for γ . Let $\Psi(Y_i, Z_i, W_i, \Theta)$ be the vector of the stack estimating equations for $\hat{\theta}_{sipw}$ and $\hat{\gamma}$ using the observed data from the i th subject. That is, $\Psi(Y_i, Z_i, W_i, \Theta)$ consists of $(R_i/\pi_i)\psi(Y_i, Z_i, W_i, \theta)$ and the estimating score for γ . Also, let $\Psi^*(Y_i, Z_i, W_i, \Theta)$ be the vector of the stack estimating equations for $\hat{\theta}$ and $\hat{\gamma}$ using the observed data from the i th subject. That is, $\Psi^*(Y_i, Z_i, W_i, \Theta)$ consists of $\psi(Y_i, Z_i, W_i, \theta)$ and the estimating score for γ .

The development of the BLIPW when the selection probabilities are estimated is similar to Appendix A but we now use the stack estimating equations to take into account variability due to the estimation of π . Let $\mathcal{G}_1(\mathcal{B}) = -E\{\partial\Phi(Y, X, Z, W, \mathcal{B})/\partial\mathcal{B}\}$,

$\mathcal{G}_2(\Theta) = -E\{\partial\Psi(Y, Z, W, \Theta)/\partial\Theta\}$, $\mathcal{G}_3(\Theta) = -E\{\partial\Psi^*(Y, Z, W, \Theta)/\partial\Theta\}$. Also, let

$\mathcal{G}_{12}(\mathcal{B}, \Theta) = E\{\Phi(Y, X, Z, W, \mathcal{B})\Psi'(Y, Z, W, \Theta)\}'$,

$\mathcal{G}_{13}(\mathcal{B}, \Theta) = E\{\Phi(Y, X, Z, W, \mathcal{B})\Psi^*(Y, Z, W, \Theta)\}'$, $\mathcal{G}_{22}(\Theta) = E\{\Psi(Y, Z, W, \Theta)\Psi'(Y, Z, W, \Theta)\}'$,

$\mathcal{G}_{33}(\Theta) = E\{\Psi^*(Y, Z, W, \Theta)\Psi^*(Y, Z, W, \Theta)\}'$, $\mathcal{G}_{23}(\Theta) = E\{\Psi(Y, Z, W, \Theta)\Psi^*(Y, Z, W, \Theta)\}'$,

$\mathcal{G}_{32}(\Theta) = E\{\Psi^*(Y, Z, W, \Theta)\Psi'(Y, Z, W, \Theta)\}'$.

Then it can be seen that

$$\text{cov}\left\{n^{1/2}(\hat{\mathcal{B}}_{sipw} - \mathcal{B}), n^{1/2}(\hat{\Theta} - \Theta)\right\} = \mathcal{G}_1^{-1}(\mathcal{B})\mathcal{G}_{13}(\mathcal{B}, \Theta)\left\{\mathcal{G}_3^{-1}(\Theta)\right\}'.$$

In the variance-covariance matrix given above, we use the notation

$\left[\mathcal{G}_1^{-1}(\mathcal{B})\mathcal{G}_{13}(\mathcal{B}, \Theta)\left\{\mathcal{G}_3^{-1}(\Theta)\right\}'\right]_{\beta, \theta}$ to denote the upper-left portion, which can be easily seen

to be $\text{cov}\left[n^{1/2}\{\hat{\beta}_{sipw}(\hat{\pi}) - \beta\}, n^{1/2}(\hat{\theta} - \theta)\right]$. Similarly, it can be shown that

$$\text{cov}\left[n^{1/2}\{\hat{\beta}_{sipw}(\hat{\pi}) - \beta\}, n^{1/2}\{\hat{\theta}_{sipw}(\hat{\pi}) - \theta\}\right] = \left[\mathcal{G}_1^{-1}(\mathcal{B})\mathcal{G}_{12}(\mathcal{B}, \Theta)\left\{\mathcal{G}_2^{-1}(\Theta)\right\}'\right]_{\beta, \theta}.$$

By direct calculation, we have that

$$\begin{aligned} \text{var}\left[n^{1/2}\left\{\hat{\boldsymbol{\theta}}_{sipw}(\hat{\pi}) - \boldsymbol{\theta}\right\}\right] &= \text{var}\left[n^{1/2}\left\{\hat{\boldsymbol{\theta}}_{sipw}(\hat{\pi}) - \boldsymbol{\theta}\right\}\right] + \text{var}\left\{n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\right\} \\ &- \text{cov}\left[n^{1/2}\left\{\hat{\boldsymbol{\theta}}_{sipw}(\hat{\pi}) - \boldsymbol{\theta}\right\}, n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\right] - \text{cov}\left[n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), n^{1/2}\left\{\hat{\boldsymbol{\theta}}_{sipw}(\hat{\pi}) - \boldsymbol{\theta}\right\}\right] \end{aligned}$$

Similar to the variance-covariance calculation given above, we have

$$\text{var}\left[n^{1/2}\left\{\hat{\boldsymbol{\theta}}_{sipw}(\hat{\pi}) - \boldsymbol{\theta}\right\}\right] = \left[\mathcal{E}_2^{-1}(\boldsymbol{\Theta})\mathcal{E}_{22}(\boldsymbol{\Theta})\left\{\mathcal{E}_2^{-1}(\boldsymbol{\Theta})\right\}'\right]_{\boldsymbol{\theta}, \boldsymbol{\theta}},$$

$$\text{var}\left\{n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\right\} = \left[\mathcal{E}_3^{-1}(\boldsymbol{\Theta})\mathcal{E}_{33}(\boldsymbol{\Theta})\left\{\mathcal{E}_3^{-1}(\boldsymbol{\Theta})\right\}'\right]_{\boldsymbol{\theta}, \boldsymbol{\theta}},$$

$$\text{cov}\left[n^{1/2}\left\{\hat{\boldsymbol{\theta}}_{sipw}(\hat{\pi}) - \boldsymbol{\theta}\right\}, n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\right] = \left[\mathcal{E}_2^{-1}(\boldsymbol{\Theta})\mathcal{E}_{23}(\boldsymbol{\Theta})\left\{\mathcal{E}_3^{-1}(\boldsymbol{\Theta})\right\}'\right]_{\boldsymbol{\theta}, \boldsymbol{\theta}},$$

and

$$\text{cov}\left[n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), n^{1/2}\left\{\hat{\boldsymbol{\theta}}_{sipw}(\hat{\pi}) - \boldsymbol{\theta}\right\}\right] = \left[\mathcal{E}_3^{-1}(\boldsymbol{\Theta})\mathcal{E}_{32}(\boldsymbol{\Theta})\left\{\mathcal{E}_2^{-1}(\boldsymbol{\Theta})\right\}'\right]_{\boldsymbol{\theta}, \boldsymbol{\theta}}.$$

Hence the optimal Δ^* in (4) can be estimated by using the variance-covariance components discussed above, such that

$$\begin{aligned} \Delta^* &= \left\{ \left[\mathcal{E}_1^{-1}(\mathcal{B})\mathcal{E}_{13}(\mathcal{B}, \boldsymbol{\Theta})\left\{\mathcal{E}_3^{-1}(\boldsymbol{\Theta})\right\}' \right]_{\boldsymbol{\beta}, \boldsymbol{\theta}} - \left[\mathcal{E}_1^{-1}(\mathcal{B})\mathcal{E}_{12}(\mathcal{B}, \boldsymbol{\Theta})\left\{\mathcal{E}_2^{-1}(\boldsymbol{\Theta})\right\}' \right]_{\boldsymbol{\beta}, \boldsymbol{\theta}} \right\} \times \\ &\left(\left[\mathcal{E}_2^{-1}(\boldsymbol{\Theta})\mathcal{E}_{22}(\boldsymbol{\Theta})\left\{\mathcal{E}_2^{-1}(\boldsymbol{\Theta})\right\}' \right]_{\boldsymbol{\theta}, \boldsymbol{\theta}} + \left[\mathcal{E}_3^{-1}(\boldsymbol{\Theta})\mathcal{E}_{33}(\boldsymbol{\Theta})\left\{\mathcal{E}_3^{-1}(\boldsymbol{\Theta})\right\}' \right]_{\boldsymbol{\theta}, \boldsymbol{\theta}} - \left[\mathcal{E}_2^{-1}(\boldsymbol{\Theta})\mathcal{E}_{23}(\boldsymbol{\Theta})\left\{\mathcal{E}_3^{-1}(\boldsymbol{\Theta})\right\}' \right]_{\boldsymbol{\theta}, \boldsymbol{\theta}} - \left[\mathcal{E}_3^{-1}(\boldsymbol{\Theta})\mathcal{E}_{32}(\boldsymbol{\Theta})\left\{\mathcal{E}_2^{-1}(\boldsymbol{\Theta})\right\}' \right]_{\boldsymbol{\theta}, \boldsymbol{\theta}} \right)^{-1} \end{aligned}$$

Based on the calculation results given above, we can proceed to the asymptotic variance of the BLIPW estimator,

$$\begin{aligned} &\text{var}\left[n^{1/2}\left\{\hat{\boldsymbol{\beta}}_{blipw}(\hat{\pi}) - \boldsymbol{\beta}\right\}\right] \\ &= \text{var}\left[n^{1/2}\left\{\left(\hat{\boldsymbol{\beta}}_{sipw}(\hat{\pi}) - \boldsymbol{\beta}\right)\right\} + \Delta^* n^{1/2}\left\{\hat{\boldsymbol{\theta}}_{sipw}(\hat{\pi}) - \boldsymbol{\theta}\right\}\right] \\ &= \left[\mathcal{E}_1^{-1}(\mathcal{B})\mathcal{E}_{11}(\mathcal{B})\left\{\mathcal{E}_1^{-1}(\mathcal{B})\right\}' \right]_{\boldsymbol{\beta}, \boldsymbol{\beta}} + \Delta^* \left[\mathcal{E}_2^{-1}(\boldsymbol{\Theta})\mathcal{E}_{21}(\boldsymbol{\Theta}, \mathcal{B})\left\{\mathcal{E}_1^{-1}(\mathcal{B})\right\}' \right]_{\boldsymbol{\theta}, \boldsymbol{\beta}} \\ &- \Delta^* \left[\mathcal{E}_3^{-1}(\boldsymbol{\Theta})\mathcal{E}_{31}(\boldsymbol{\Theta}, \mathcal{B})\left\{\mathcal{E}_1^{-1}(\mathcal{B})\right\}' \right]_{\boldsymbol{\theta}, \boldsymbol{\beta}} \end{aligned}$$

$$\begin{aligned}
&= \left[\mathcal{E}_1^{-1}(\mathcal{B}) \mathcal{E}_{11}(\mathcal{B}) \left\{ \mathcal{E}_1^{-1}(\mathcal{B}) \right\}' \right]_{\beta, \beta} \\
&- \left(\left[\mathcal{E}_1^{-1}(\mathcal{B}) \mathcal{E}_{13}(\mathcal{B}, \Theta) \left\{ \mathcal{E}_3^{-1}(\Theta) \right\}' \right]_{\beta, \theta} - \left[\mathcal{E}_1^{-1}(\mathcal{B}) \mathcal{E}_{12}(\mathcal{B}, \Theta) \left\{ \mathcal{E}_2^{-1}(\Theta) \right\}' \right]_{\beta, \theta} \right) \times \\
&\left(\left[\mathcal{E}_2^{-1}(\Theta) \mathcal{E}_{22}(\Theta) \left\{ \mathcal{E}_2^{-1}(\Theta) \right\}' \right]_{\theta, \theta} + \left[\mathcal{E}_3^{-1}(\Theta) \mathcal{E}_{33}(\Theta) \left\{ \mathcal{E}_3^{-1}(\Theta) \right\}' \right]_{\theta, \theta} - \left[\mathcal{E}_2^{-1}(\Theta) \mathcal{E}_{23}(\Theta) \left\{ \mathcal{E}_3^{-1}(\Theta) \right\}' \right]_{\theta, \theta} - \left[\mathcal{E}_3^{-1}(\Theta) \mathcal{E}_{32}(\Theta) \left\{ \mathcal{E}_2^{-1}(\Theta) \right\}' \right]_{\theta, \theta} \right)^{-1} \\
&\times \\
&\left(\left[\mathcal{E}_3^{-1}(\Theta) \mathcal{E}_{31}(\Theta, \mathcal{B}) \left\{ \mathcal{E}_1^{-1}(\mathcal{B}) \right\}' \right]_{\theta, \beta} - \left[\mathcal{E}_2^{-1}(\Theta) \mathcal{E}_{21}(\Theta, \mathcal{B}) \left\{ \mathcal{E}_1^{-1}(\mathcal{B}) \right\}' \right]_{\theta, \beta} \right)
\end{aligned}$$

The asymptotic variance of the BLIPW estimator given above also shows that the BLIPW estimator is more efficient than the SIPW estimator when the selection probabilities are estimated.

References

- [1]. Bruemmer B, White E, Vaughan T and Cheney C (1996), Nutrient Intake in Relationship to Bladder Cancer Among Middle aged Men and Women. *American Journal of Epidemiology*, 144, 485–495. [PubMed: 8781464]
- [2]. Little RJA and Rubin DB (2002). *Statistical Analysis with Missing Data*, John Wiley & Sons, 2nd Edition New York.
- [3]. Rubin DB (1976). Inference and missing data. *Biometrika* 63, 581–592.
- [4]. Robins JM, Rotnitzky A and Zhao LP (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89, 846–866.
- [5]. Qi L, Wang CY and Prentice RL (2005). Weighted estimators for proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, 100, 1250–1263.
- [6]. Wang CY, Lee SM and Chao EC (2007). Numerical Equivalence of Imputing Scores and Weighted Estimators in Regression Analysis with Missing Covariates. *Biostatistics*, 8, 468–473. [PubMed: 16971377]
- [7]. Hsieh SH, Lee SM, and Shen PS (2009). Semiparametric analysis of randomized response data with missing covariates in logistic regression. *Computational Statistics & Data Analysis*. 53, 2673–2692.
- [8]. Wu H and Wu L (2002). Identification of significant host factors for HIV dynamics models by nonlinear mixed-effect models. *Statistics in Medicine*, 21, 753–771. [PubMed: 11870815]
- [9]. Chatterjee N, Chen Y-H, and Breslow N (2003). A Pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association*, 98, 158–168.
- [10]. Conniffe D (1985). Estimating regression equations with common explanatory variables but unequal numbers of observations. *Journal of Econometrics*, 27, 179–196.
- [11]. Rubin DB and Schenker N (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366374.
- [12]. Hu XJ, Schroeder RJ, Wang WC, and Boyett JM (2007). Pseudoscore-based estimation from biased observations, *Statistics in Medicine*, 26, 2836–52. [PubMed: 17133648]
- [13]. Wang CY and Chen HY (2001). Augmented inverse probability weighted estimator for Cox missing covariate regression. *Biometrics*, 57, 414–419. [PubMed: 11414564]
- [14]. Wang S and Wang CY (2001). Asymptotic comparisons of kernel assisted estimators in missing covariate regression. *Statistics and Probability Letters*, 55, 439–449.

- [15]. Lettre G, Jackson AU, Gieger C et al. (2008). Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet.* 40, 584–591. [PubMed: 18391950]
- [16]. Monsees GM, Tamimi RM, Kraft P (2009). Genome-wide association scans for secondary traits using case-control sample. *Genet Epidemiol.*, 33, 717–728. [PubMed: 19365863]
- [17]. Tapsoba JD, Kooperberg C, Reiner A, Wang CY and Dai JY (2014). Robust estimation for secondary-trait association in case-control genetic studies. *American Journal of Epidemiology*, 179, 1264–72. [PubMed: 24723002]
- [18]. Chen YH, and Chen H (2000). A unified approach to regression analysis under double sampling design. *Journal of the Royal Statistical Society, Ser. B*, 62, 449–460.
- [19]. Qi L, Wang YF, He Y (2010). A comparison of multiple imputation and fully augmented weighted estimators for Cox regression with missing covariates. *Statistics in Medicine*, 29, 2592–604. [PubMed: 20806403]
- [20]. Rubinstein RY and Marcus R (1985). Efficiency of multivariate control variates in Monte Carlo simulation. *Operations Research*, 33, 661–677.

Figure 1:
Scatterplot of smoking pack year versus smoking year and a smoothing lowess curve

Table 1:

Linear regression with MAR, normal covariates

		CC	MI	SIPW	BLIPW	PS	AIPW
$\beta_0 = 0$	Bias	-24.0	-0.5	-0.7	-0.7	-2.0	-1.3
	SD	6.4	5.3	5.5	5.4	5.4	5.4
	ASE	6.3	5.2	5.3	5.1	5.0	5.0
	CP	4.0	94.8	95.0	94.2	91.6	92.4
$\beta_1 = -0.5$	Bias	1.8	0.0	-0.2	0.1	0.3	0.5
	SD	3.8	3.9	4.5	4.1	4.2	4.2
	ASE	3.9	3.9	4.4	4.0	3.9	3.9
	CP	94.0	95.4	94.4	94.4	93.4	92.6
$\beta_2 = 0.5$	Bias	16.0	0.4	0.7	0.5	1.7	1.0
	SD	8.5	7.3	7.4	7.3	7.4	7.4
	ASE	8.0	7.1	7.2	7.1	7.0	7.0
	CP	47.6	95.0	94.4	93.4	93.4	93.4
$\beta_0 = 0$	Bias	-23.8	-0.3	-0.7	-0.4	-2.0	-1.3
	SD	6.4	6.3	6.8	6.5	6.4	6.5
	ASE	6.5	6.4	6.7	6.5	6.2	6.2
	CP	4.6	94.8	94.4	94.4	92.8	93.2
$\beta_1 = -1.5$	Bias	5.9	-0.1	-0.1	0.0	1.1	1.1
	SD	4.2	3.8	5.2	4.3	4.5	4.5
	ASE	4.2	3.9	5.1	4.1	4.1	4.1
	CP	72.4	93.0	94.6	94.0	92.6	92.4
$\beta_2 = 1.5$	Bias	11.5	0.3	0.8	0.3	1.4	0.7
	SD	8.5	8.3	8.7	8.6	8.6	8.6
	ASE	8.2	8.2	8.5	8.2	8.1	8.1
	CP	71.8	94.4	94.4	94.2	93.8	93.4

NOTE: Bias, SD, ASE and CP have been multiplied by 100. CC is the complete-case analysis, MI is the full normal multiple imputation, SIPW is the simple inverse selection probability estimator, BLIPW is the best linear inverse selection probability weighted estimator, PS is the pseudoscore estimator, and AIPW is the augmented inverse selection probability estimator. The sample size $n = 1000$. The selection probability is $\text{pr}(R = 1 | Y, Z) = H(-0.5Y + 2Z)$, and under this model the missing data probability is 33% when $\beta = (0, -0.5, 0.5)$, and 38% when $\beta = (0, -1.5, 1.5)$, respectively. The results are from 500 simulation replicates.

Table 2:

Linear regression with MAR, X is mixture of normals

		CC	MI	SIPW	BLIPW	PS	AIPW
$\beta_0 = 0$	Bias	-23.5	-1.2	0.1	0.0	-1.3	-0.6
	SD	5.8	5.0	5.2	4.9	4.9	4.9
	ASE	6.3	5.1	5.3	4.9	4.9	4.9
	CP	3.0	94.4	96.2	95.8	92.8	94.4
$\beta_1 = -0.5$	Bias	1.9	1.0	0.1	0.3	0.4	0.6
	SD	3.8	3.6	4.1	3.9	3.9	4.0
	ASE	3.7	3.7	4.1	3.9	3.8	3.8
	CP	91.6	94.4	95.2	94.6	94.4	94.2
$\beta_2 = 0.5$	Bias	15.2	0.5	-0.6	-0.7	0.5	-0.2
	SD	7.8	6.9	7.0	6.8	6.9	6.9
	ASE	8.0	7.0	7.1	6.9	6.9	6.9
	CP	55.2	95.2	94.8	95.6	94.2	95.0
$\beta_0 = 0$	Bias	-23.8	-4.9	-0.0	-0.1	-2.0	-0.8
	SD	6.2	6.0	6.4	5.7	5.7	5.7
	ASE	6.5	6.2	6.5	5.9	5.8	5.8
	CP	4.4	87.0	96.2	95.6	94.2	95.2
$\beta_1 = -1.5$	Bias	5.5	7.3	0.0	0.4	0.8	0.7
	SD	3.9	3.6	4.2	3.7	4.0	4.0
	ASE	3.8	3.5	4.2	3.6	3.7	3.7
	CP	69.4	45.8	94.0	93.8	92.0	92.6
$\beta_2 = 1.5$	Bias	11.0	1.5	-0.5	-0.6	1.1	-0.1
	SD	8.2	7.9	8.1	7.6	7.7	7.7
	ASE	8.3	7.9	8.3	7.8	7.8	7.8
	CP	75.2	94.0	96.8	96.0	95.6	95.6

NOTE: See the foot note of Table 1 for notation. The sample size $n = 1000$. The selection probability is $\text{pr}(R = 1 | Y, Z) = H(-0.5Y + 2Z)$, and under this model the missing data probability is 33% when $\beta = (0, -0.5, 0.5)$, and 39% when $\beta = (0, -1.5, 1.5)$, respectively. The results are from 500 simulation replicates.

Table 3:

Logistic regression with MAR, X is mixture of normals

		CC	MI	SIPW	BLIPW	PS	AIPW
$\beta_0 = -\ln(2)$	Bias	-21.9	0.0	-0.4	-0.4	-0.5	-0.5
	SD	13.9	10.7	12.7	10.9	10.7	10.8
	ASE	13.6	10.4	12.3	9.8	10.2	10.3
	CP	64.2	94.4	95.8	92.4	93.8	93.6
$\beta_1 = \ln(2)$	Bias	1.8	3.6	-0.3	-0.1	-0.2	-0.2
	SD	10.3	9.2	10.4	10.2	10.1	10.2
	ASE	10.2	9.8	10.3	10.0	9.8	10.1
	CP	94.6	94.0	94.4	94.2	94.0	94.8
$\beta_2 = \ln(2)$	Bias	5.1	0.2	-0.2	0.0	0.1	0.1
	SD	18.4	14.8	18.6	14.8	14.7	14.7
	ASE	17.9	14.1	18.0	13.5	13.8	14.0
	CP	93.8	94.6	95.4	93.0	93.8	94.0
$\beta_0 = -\ln(2)$	Bias	-21.7	4.7	-0.5	-0.2	-0.3	-0.4
	SD	15.2	12.1	14.0	12.6	12.5	12.6
	ASE	14.9	12.2	13.6	11.7	11.6	12.0
	CP	68.6	92.6	93.8	93.6	93.0	94.0
$\beta_1 = -\ln(4)$	Bias	0.8	9.3	1.5	-1.2	-0.8	-1.1
	SD	13.7	11.4	14.0	13.7	13.5	13.7
	ASE	13.5	12.8	13.6	13.4	13.1	13.4
	CP	94.4	90.8	95.4	94.4	94.4	94.0
$\beta_2 = \ln(4)$	Bias	6.0	-1.5	0.8	0.6	0.3	0.5
	SD	20.8	17.4	20.8	17.8	17.6	17.7
	ASE	20.3	17.2	20.5	16.7	16.6	17.0
	CP	94.0	94.2	95.0	93.4	93.2	94.2

NOTE: See the foot note of Table 1 for notation. The sample size $n = 1000$. The selection probability is $\text{pr}(R = 1 | Y, Z, W) = H(-0.5Y + 0.5Z + W)$, and under this model there are 38% of subjects with missing X data. The results are from 500 simulation replicates.

Table 4:

Secondary Trait Regression with binary secondary trait

		CC	MI	SIPW	BLIPW	PS	AIPW
$\beta_0 = -\ln(2)$	Bias	16.8	-3.1	-1.6	-1.4	8.3	-1.3
	SD	18.5	9.2	16.3	9.2	7.6	9.0
	ASE	17.3	9.8	16.0	9.3	7.4	9.2
	CP	81.6	96.2	94.2	95.6	80.2	95.6
$\beta_1 = \ln(2)$	Bias	0.5	-1.9	-2.1	-2.1	25.1	-1.7
	SD	11.3	11.0	13.5	13.2	10.8	13.1
	ASE	11.1	11.7	13.5	13.2	9.2	13.2
	CP	95.8	95.2	95.0	94.6	27.2	95.2
$\beta_2 = \ln(2)$	Bias	12.1	5.7	1.6	0.9	-11.8	1.2
	SD	23.8	13.1	26.9	13.0	11.3	12.7
	ASE	23.5	13.7	25.5	13.3	10.5	13.2
	CP	93.6	94.8	94.8	95.4	76.8	97.0
$\beta_0 = -\ln(2)$	Bias	16.4	3.0	-1.6	-2.1	9.2	-1.7
	SD	19.7	12.1	18.2	12.5	10.8	12.4
	ASE	18.8	12.9	17.9	12.6	9.0	12.5
	CP	84.4	94.8	95.2	96.2	74.6	96.2
$\beta_1 = \ln(4)$	Bias	0.6	1.4	-2.9	-2.8	24.4	-2.2
	SD	15.1	15.0	18.4	18.1	15.6	18.0
	ASE	15.1	15.6	18.1	17.8	13.1	17.8
	CP	94.0	92.8	94.6	94.6	53.0	94.4
$\beta_2 = \ln(4)$	Bias	13.9	5.9	1.5	2.2	-13.9	1.6
	SD	28.2	19.0	31.9	20.3	18.7	20.0
	ASE	27.4	19.5	30.8	20.0	14.0	19.8
	CP	93.0	94.8	94.0	94.8	73.4	94.6

NOTE: See the foot note of Table 1 for notation. The number of controls in the CC set is denoted by n_{C0} , and the number of cases in the CC set is denoted by n_{C1} . The sample size $n = 3000$, $n_{C0} = n_{C1} = 200$. The PS estimator in this setting has bias due to the estimation of the conditional expectation of the full data score given (Y, Z, D) in (2). The results are from 500 simulation replicates.

Table 5:

Bladder cancer data analysis

		CC	MI	SIPW	BLIPW	PS	AIPW
Intercept	$\hat{\beta}_0$	0.194	-0.867	-0.850	-0.947	-0.652	-0.635
	SE	0.387	0.331	0.341	0.320	0.356	0.319
Pack year (divided by 10)	$\hat{\beta}_1$	0.103	0.160	0.147	0.161	0.115	0.113
	SE	0.071	0.065	0.066	0.064	0.068	0.060
Obesity	$\hat{\beta}_1$	0.261	0.344	0.229	0.173	0.328	0.283
	SE	0.516	0.423	0.463	0.446	0.426	0.434

Note: CC is the complete-case analysis, MI is the full normal multiple imputation, SIPW is the simple inverse selection probability estimator, BLIPW is the best linear inverse selection probability weighted estimator, PS is the pseudoscore estimator, and AIPW is the augmented inverse selection probability estimator. The coefficient estimates (and SEs) for smoking pack year listed in the table have been multiplied by 10. The analysis is based on current smokers ($n = 251$, $n_V = 179$).