Published in final edited form as:
Biometrika. 2014 September ; 101(3): 613–624. doi:10.1093/biomet/asu022.

Estimation of mean response via effective balancing score

Zonghui Hu,
Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Maryland 20892-7609, USA

Dean A. Follmann, and
Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Maryland 20892-7609, USA

Naisyin Wang
Department of Statistics, University of Michigan, Ann Arbor MI 48109-1107, USA

Zonghui Hu: huzo@niaid.nih.gov; Dean A. Follmann: dfollmann@niaid.nih.gov; Naisyin Wang: nwangaa@umich.edu

Summary

We introduce effective balancing scores for estimation of the mean response under a missing at random mechanism. Unlike conventional balancing scores, the effective balancing scores are constructed via dimension reduction free of model specification. Three types of effective balancing scores are introduced: those that carry the covariate information about the missingness, the response, or both. They lead to consistent estimation with little or no loss in efficiency. Compared to existing estimators, the effective balancing score based estimator relieves the burden of model specification and is the most robust. It is a near-automatic procedure which is most appealing when high dimensional covariates are involved. We investigate both the asymptotic and the numerical properties, and demonstrate the proposed method in a study on Human Immunodeficiency Virus disease.

Keywords

Balancing score; Dimension reduction; Missing at random; Nonparametric kernel regression; Prognostic score; Propensity score

1. Introduction

In social and medical studies, the primary interest is usually the mean response, the estimation of which can be complicated by missing observations due to nonresponse, drop out or death. The data observed are triplets $\{(Y_i, \delta_i, X_i), i = 1, \cdots, n\}$, where Y_i is the response, $\delta_i = 1$ if Y_i is observed and $\delta_i = 0$ if Y_i is missing, and X_i is the vector of covariates and always observed. Under the missing at random mechanism (Rosenbaum & Rubin, 1983); that is, $\Pr(\delta = 1 \mid X, Y) = \Pr(\delta = 1 \mid X)$, estimation of $E(Y)$ is mostly developed using the parametric form of the missingness pattern $\pi(X) = \Pr(\delta = 1 \mid X)$ or the response pattern $m(X) = E(Y \mid X)$. Important methods include regression estimation (Rubin, 1987; Schafer, 1997), inverse propensity score estimation (Horvitz & Thompson, 1952), augmented inverse propensity weighting estimation (Robins et al., 1994), and their modified versions such as D'Agostino (1998), Scharfstein et al. (1999), Little & An (2004), Vartivarian & Little (2008), and Cao et al. (2009). A review of most methods can be found in Lunceford & Davidian (2004) and Kang & Schafer (2007). Consistency and efficiency of these estimators rely on correct model specification. Even for the "doubly robust" estimators, either $\pi(X)$ or $m(X)$ needs to be correctly specified for consistency and both correctly specified for efficiency (Robins & Rotnitzky, 1995; Hahn, 1998). When $X \in \mathbb{R}^p$ is high dimensional, model specification is challenging: it is hard for a parametric model to be sufficiently flexible to capture all the important nonlinear and interaction effects yet parsimonious enough to maintain reasonable efficiency.

One family of estimators are built upon the balancing score. According to Rosenbaum & Rubin (1983), a balancing score $b(X)$ has the property $E(Y \mid b(X)) = E(Y \mid b(X), \delta = 1)$. Therefore, $E(Y)$ can be estimated via $b(X)$ over the complete cases $\{(Y_i, \delta_i, X_i) : \delta_i = 1\}$. The most well known balancing scores include the propensity score (Rosenbaum & Rubin, 1983) and the prognostic score (Hansen, 2008). The mean response can be estimated via the balancing score by such nonparametric approaches as stratification (Rosenbaum & Rubin, 1983) and nonparametric regression (Cheng, 1994). Of course, the naive balancing score is X. However, estimation using X as a balancing score is subject to the curse of dimensionality when $X \in \mathbb{R}^p$ is high dimensional (Abadie & Imbens, 2006).

Balancing scores have been estimated through parametric modeling. In comparison to the other estimators, balancing score based estimators are less sensitive to model misspecification, largely due to the nonparametric approaches to utilize the balancing score (Rosenbaum, 2002). One important property of the balancing score based estimator, which has rarely been utilized, is that full parametric modeling is actually unnecessary. For example, if $\pi(x) = f\{b(x)\}$ for some function $b(X)$ and unknown function f, then $E(Y)$ can be estimated via $b(X)$ through stratification or nonparametric regression as subjects with similar values in $b(X)$ have similar values in $\pi(X)$. Provided that we can find such a function $b(X)$, there is no need for the full parametric form of $\pi(X)$.

In this work, we introduce the effective balancing score. Like the propensity score and the prognostic score, the effective balancing score creates a conditional balance between the subjects with response observed and the subjects with response missing. Unlike the conventional balancing scores, estimation of the effective balancing score is free of model specification via the technique of dimension reduction (Li, 1991; Cook & Weisberg, 1991; Cook & Li, 2002; Li & Zhu, 2007; Li & Wang, 2007). The effective balancing score carries all X information about the missingness or the response in the sense $\delta \perp X \mid S$ or $Y \perp X \mid S$, where S stands for the effective balancing score and $\perp$ stands for conditional independence. It thus leads to consistent estimation of $E(Y)$ with little or no loss in efficiency. As a parsimonious summary of X, the effective balancing score is of dimension much smaller than p. Compared with existing methods, the effective balancing score based estimator has the following advantages: (1) It relieves the burden of model specification and is the most robust with potentially optimal efficiency; (2) Through the technique of dimension reduction, the effective balancing score is of low dimension which enables the effective use of stratification and nonparametric regression; (3) It avoids the shortcoming of inverse propensity weighting, i.e., instability caused by estimates of $\pi(X)$ that are close to zero.

2. Effective balancing score

2·1. Effective balancing score

Let $\mathcal{R}$ be the response from $X \in \mathbb{R}^p$. Usually $\mathcal{R}$ relates to X through only a few linear combinations; that is, $\mathscr{R} \perp X|(\beta_1' X, \cdots, \beta_K' X)$ with $\beta_k \in \mathbb{R}^p : k = 1, \cdots, K$ distinctive vectors. Let $B = (\beta_1, \cdots, \beta_K)$ with $\beta_1, \cdots, \beta_K$ orthonormal and K the smallest dimension to satisfy the conditional independence, then B is a basis of the central dimension-reduction space $\mathcal{S}_{\mathcal{R}|X}$ with K the structural dimension (Cook, 1994). The columns of B are arranged in descending order of importance; that is, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K > 0$ where λ_k measures the amount of X information carried by $\beta_k' X$ and is explained in §3.2. In general, K is much smaller than p. If we let $S = B'X$, then $S \in \mathbb{R}^K$ is a parsimonious summary of X: it is of lower dimension than X but carries all X information about $\mathcal{R}$. In this paper, we refer to $B = (\beta_1, \cdots, \beta_K)$ as the effective directions.

Let $\mathcal{R} = \delta$ and B_δ be the effective directions of $\mathcal{S}_{\delta|X}$, then

$$\delta \perp X|B_\delta' X, \quad (1)$$

and we refer to $S_\delta = B_\delta' X$ as the effective propensity score.

Let $\mathcal{R} = Y$ and denote B_Y as the effective directions of $\mathcal{S}_{Y|X}$, then

$$Y \perp X|B_Y' X. \quad (2)$$

Obviously, $B_Y' X$ is a prognostic score satisfying the definition of Hansen (2008). We refer to $S_Y = B_Y' X$ as the effective prognostic score.

Each effective score creates the conditional balance

$$Y \perp \delta|S, \quad (3)$$

where S is either S_δ or S_Y. For $S = S_\delta$, (3) follows similarly as in Theorem 3 of Rosenbaum & Rubin (1983). For $S = S_Y$,

$$\Pr(\delta=1|Y,S)=E\{E(\delta|Y,X)|Y,S\}=E\{E(\delta|X)|Y,S\}=E\{E(\delta|X)|S\},$$

where the second equation is due to missingness at random and the last equation to (2). Since the last expectation is $E(\delta \mid S) = \Pr(\delta = 1 \mid S)$, (3) follows. It is immediate from (3) that both effective scores are balancing scores.

Example 1—Suppose $Y \mid X$ is normal with mean $m(X) = X_1 + \exp(X_2 + X_3) + X_1X_4$ and variance $\sigma^2(X)=X_1^2+X_4^2$, with the probability of observing Y as $\pi(X)=\text{expit}(0.1X_1^2+X_2+X_3)$. Then, the prognostic score is $\{m(X), \sigma^2(X)\}$ and the propensity score is $\pi(X)$. The effective prognostic score is $\{(X_2+X_3)/\sqrt{2}, X_1, X_4\}$ and the effective propensity score is $\{(X_2+X_3)/\sqrt{2}, X_1\}$.

The effective balancing scores may have higher dimensions than their conventional counterparts. However, estimation of the propensity score and the prognostic score requires correct model specification and is subject to the challenges discussed in §1. The effective balancing scores, on the other hand, can be obtained without model specification.

We can also let $\mathcal{R} = (\delta, Y)$ be a bivariate response. Denote B_d as the effective directions for $\mathcal{S}_{(\delta,Y)|X}$, then

$$\delta \perp X|B_d^{'}X \quad \text{and} \quad Y \perp X|B_d^{'}X. \quad (4)$$

In other words, $B_d^{'}X$ carries all X information about both δ and Y, and creates both propensity balance and prognostics balance. We refer to $S_d=B_d^{'}X$ as the effective double balancing score. In Example 1, $S_d=\{(X_2+X_3)/\sqrt{2}, X_1, X_4\}$ is the same as S_Y.

Remark 1—As shown by (1) and (2), so long as either independence in (4) holds, S_d is a balancing score satisfying the conditional balance (3). It is for this reason that we refer to S_d as the effective double balancing score.

In summary, both the effective prognostic score and the effective double balancing score have the properties

$$Y \perp \delta|S \quad \text{and} \quad Y \perp X|S.$$

The first property implies $E(Y \mid S) = E(Y \mid S, \delta = 1)$, which ensures unbiased estimation of $E(Y)$ via S from the complete cases. The second property implies that S carries all X information about the response, which ensures efficient estimation of $E(Y)$ via S. The effective propensity score possesses only the first property and is not as efficient as the other two. We will show in §3 and §4 that S_d can improve over S_Y under certain situations. Without loss of generality, we assume $E(X) = 0$ and $\text{cov}(X) = I_p$ the identity matrix.

2·2. Estimation of effective balancing score

To find the effective balancing scores is to find the effective directions: the effective directions of $\mathcal{S}_{\delta|X}$ for the effective propensity score and the effective directions of $\mathcal{S}_{Y|X}$ for the effective prognostic score. For the effective double balancing score, we need the effective directions of $\mathcal{S}_{(\delta,Y)|X}$.

Remark 2—Under missingness at random, there is the relationship $\mathscr{S}_{(\delta,Y)|X} = \mathscr{S}_{\delta Y|X}$ for continuous response *Y*. The effective directions of $\mathscr{S}_{(\delta,Y)|X}$ can be estimated through the univariate response *δY*. A similar approach applies if *Y* is categorical. See Appendix 1.

There are many dimension reduction methods for estimating the effective directions. The most fundamental methods are the sliced inverse regression (Li, 1991) and the sliced average variance estimation (Cook & Weisberg, 1991). Both methods are developed under the linearity condition; that is, $E(X \mid B'X)$ is a linear function of $B'X$. Many new methods have been developed to improve over these two. To improve estimation efficiency, there are the likelihood based methods of Cook (2007), Cook & Forzani (2008) and Cook & Forzani (2009). To relax the distribution assumption, Li & Dong (2009) and Dong & Li (2010) proposed methods to remove the linearity condition, and Ma & Zhu (2012) successfully applied a semiparametric approach to eliminate all distributional assumptions. These methods lead to root-*n* consistent estimates under proper conditions. As to be shown in Theorem 2, the proposed estimation of $E(Y)$ requires only the effective direction estimates to be root-*n* consistent. In this work, we adopt the fitted principal component method of Cook (2007) in the numerical studies unless stated otherwise. More information about these dimension reduction methods is given in §6.

Remark 3—Under missingness at random, there is the relationship

$$\mathscr{S}_{(\delta,Y)|X}=\text{span}(\mathscr{S}_{\delta|X},\mathscr{S}_{Y|X})$$

following Chiaromonte et al. (2002). That is, the effective directions of $\mathscr{S}_{(\delta,Y)|X}$ include both the effective directions of $\mathscr{S}_{\delta|X}$ and the effective directions of $\mathscr{S}_{Y|X}$.

In addition to the method in Remark 2, Remark 3 suggests a pooling method for the effective directions of $\mathscr{S}_{(\delta,Y)|X}$. Since *δ* and *Y* are mostly related, there is likely overlap between $\mathscr{S}_{\delta|X}$ and $\mathscr{S}_{Y|X}$. Therefore, the pooling method needs to be followed by such a method as Gram-Schmidt's orthogonolization to remove redundancy.

3. Mean response estimation via effective balancing score

In this section, let *S* stand for the effective balancing score and *B* the matrix of effective directions. We first consider *B* as known and later investigate the impact from the estimation of *B*. As $S = B'X$ consists of linear combinations of *X*, it is always observed. As *B* has columns of orthonormal vectors, *S* has the identity covariance matrix. Since *S* carries all *X* information about the missingness or the response, we can use $S \in \mathbb{R}^K$ instead of $X \in \mathbb{R}^p$ for the estimation of $E(Y)$ through stratification or nonparametric regression. In this work, we focus on nonparametric regression.

3·1. Nonparametric regression via effective balancing score

Let $m(S) = E(Y \mid S)$ be the conditional mean response given the effective balancing score, then $E(Y) = E\{m(S)\}$ can be estimated through the estimation of $m(S)$. To obviate model specification, we estimate $m(\cdot)$ by nonparametric kernel regression (Silverman, 1986)

$$\hat{m}(s)=\sum_{i=1}^{n}\delta_i y_i \mathscr{K}_H(s_i-s)/\sum_{i=1}^{n}\delta_i \mathscr{K}_H(s_i-s), \quad (5)$$

where $S_i = B'X_i$, $\mathscr{K}_H(u) = \det(H)^{-1}\mathscr{K}(H^{-1}u)$ for $u = (u_1, \cdots, u_K)$ with H the bandwidth matrix and $\mathscr{K}(\cdot)$ the kernel function. Since S has identity covariance, we take $H = h_n I_K$ with h_n a scalar bandwidth (Härdle et al., 2004). We then estimate $E(Y)$ by

$$\hat{\mu}=n^{-1}\sum_{i=1}^{n}\hat{m}(s_i). \quad (6)$$

We refer to $\hat{\mu}$ as the nonparametric regression via effective balancing score estimator, or briefly the nonparametric balancing score estimator. By the result of Devroye & Wagner (1980), $m(\hat{s})$ converges in probability to $E(\delta Y \mid s)/E(\delta \mid s)$. It is immediate from (3) that $E(\delta Y \mid s) = E(\delta \mid s)E(Y \mid s)$. Therefore, $m(\hat{s})$ converges in probability to $m(s)$, and consequently (6) to $\mu = E(Y)$.

Theorem 1—Under the regularity conditions, the nonparametric balancing score estimator $\hat{\mu}$ is asymptotically normally distributed. If as $n \to \infty$, $h_n \to 0$ and $nh_n^K \to \infty$, then

$$n^{1/2}(\hat{\mu}-\mu) \to N(0,\sigma^2)$$

with

$$\sigma^2=\mathrm{var}(Y)+E\left[\{\pi(S)^{-1}-1\}\mathrm{var}(Y|S)\right].$$

where $\pi(S) = E(\delta \mid S)$.

For $S = S_Y$ or $S = S_d$, due to (2) and (4), we have $Y \perp X \mid S$ and thus $\mathrm{var}(Y \mid S) = \mathrm{var}(Y \mid X)$. It follows that

$$\sigma^2=\mathrm{var}(Y)+E\left[\{\pi(X)^{-1}-1\}\mathrm{var}(Y|X)\right],$$

, which is the optimal efficiency for the semiparametric estimators of $E(Y)$, see Hahn (1998). This means that the nonparametric balancing score estimation via S_Y or S_d is both consistent and optimally efficient. For $S = S_\delta$, as $\mathrm{var}(Y \mid S) \geq \mathrm{var}(Y \mid X)$, the optimal efficiency may not be reached.

Theorem 2—With B replaced by its root-n consistent estimate $\hat{B}$, the nonparametric balancing score estimators have the same asymptotic properties as in Theorem 1.

Proof of Theorem 1 and 2 are given in the Appendix. Due to Theorem 2, we will use S_δ, S_Y, and S_d for the effective balancing scores whether B is known or estimated.

3·2. Dimension of effective balancing score

To determine the dimension of the effective balancing score is to determine K, the number of effective directions. A simple approach is the sequential permutation test of Cook & Yin (2001).

The dimension of the effective balancing score affects performance of the proposed estimator through nonparametric regression (5). Following Theorem 1, the impact from nonparametric regression is asymptotically negligible for $h_n \sim n^{-\alpha}$ with $0 < \alpha < 1/K$. For larger K, selection of h_n is more constrained as α falls in a narrower range. More specifically, nonparametric regression introduces bias $h_n^2 \mathscr{B}$ and variance $(n^2 h_n^K)^{-1} \mathscr{V}$ to $\hat{\mu}$, see Appendix 2. The mean squared error of $\hat{\mu}$ is minimized at $h_{\text{opt}} \sim n^{-2/(K+4)}$. At h_{opt}, the asymptotic variance is $n^{-1} \sigma^2 + n^{-8/(K+4)} \mathcal{V}$. If $K \leq 3$, $\hat{\mu}$ is root-n consistent and the variance from nonparametric regression is asymptotically negligible. If $K = 4$, $\hat{\mu}$ is root-n consistent but the variance from nonparametric regression is not asymptotically negligible. If $K > 4$, $\hat{\mu}$ converges slower than $n^{-1/2}$. Ideally, we would like S of dimension no more than 3 to reach the minimum mean squared error, root-n consistency, and negligible impact from nonparametric regression. Note that without dimension reduction; that is, $S = X \in \mathbb{R}^p$, the proposed estimator reduces to the nonparametric regression estimation of Cheng (1994) which can perform poorly for large p.

We compare the three effective balancing scores. The effective double balancing score and effective prognostic score improve over the effective propensity score, as S_Y and S_d lead to more efficient estimation than S_δ as shown by Theorem 1. The effective double balancing score can improve over the effective prognostic score when $\mathcal{S}_{Y|X}$ is more than three-dimensional but $\mathcal{S}_{\delta|X}$ is less than three-dimensional. Here is a hypothetical example. Suppose $\mathcal{S}_{Y|X}$ has 5 effective directions, $\mathcal{S}_{\delta|X}$ has one effective direction, and $B_d = (B_\delta, B_Y)$ has the effective propensity direction B_δ as the most important. To maintain conditional balance (3), S_Y needs to be of dimension 5. For S_d, we can use only the first three components: while the first component ensures conditional balance and thus consistency, the other two components enhance efficiency.

We can use $S^* = (\beta_1' X, \beta_2' X, \beta_3' X)$ in case of $K > 3$, which shows generally good performance in numerical studies. Most dimension reduction methods estimate β_k's as the eigenvectors of a kernel matrix, and the corresponding eigenvalue λ_k reflects the amount of X information carried by $\beta_k' X$, see §6. When the first three components carry enough X information in the sense that $(\lambda_1 + \lambda_2 + \lambda_3)/(\lambda_1 + \cdots, \lambda_K)$ is no less than 0.90, S^* leads to good estimation. We refer to S^* as the dimension further reduced effective score. If $K > 3$ and the first three components carry a low percentage of X information, which rarely happens in practice, we can use generalized additive modeling for the estimation of $E(Y \mid S)$. That is, instead of the multivariate kernel regression (5), $E(Y \mid S)$ is estimated through the additive model $Y = g_1(\beta_1' X) + \cdots + g_K(\beta_K' X) + \varepsilon$, where each g_k is nonparametric and estimated by smoothing on a single coordinate, see Hastie & Tibshirani (1986). Though the generalized additive model is a bit restrictive by assuming the additivity, it relieves the curse of dimensionality that hinders multivariate kernel regression when K is big.

3·3. Estimation procedure

Step 1. Estimate the effective directions B and determine the dimension K;

Step 2. If $K \leq 3$, compute the effective balancing score $S = B'\hat{X}$; if $K > 3$, let $S^{*}=(\beta_1' X, \beta_2' X, \beta_3' X)$ be the dimension further reduced effective score;

Step 3. Estimate $E(Y)$ by nonparametric regression via the effective balancing score S or the dimension further reduced effective score S^{*}.

For bandwidth selection, the optimal bandwidth is $h_{opt} \sim n^{-2/(K+4)}$ which minimizes the mean squared error of $\hat{\mu}$ and can be estimated by the plug in method (Fan & Marron, 1992), see Appendix 2. This optimal bandwidth is smaller than the conventional bandwidth $h_n \sim n^{-1/(K+4)}$, which is optimal for the estimation of conditional mean $m(S)$ (Härdle et al., 2004). At the conventional bandwidth, though the proposed estimator does not attain the minimal mean squared error, the bias and variance from nonparametric regression are asymptotically negligible. Therefore, when the sample size is large, we can use the conventional bandwidth which is easier to determine (Sheather & Jones, 1991).

For variance estimation, we can use the asymptotic variance formula in Theorem 1. The asymptotic variance leaves out the negligible terms; that is, the variability introduced by the estimation of the effective directions and the nonparametric regression of $m(S)$. We recommend bootstrap for variance estimation: bootstrap n samples from the original triplets $\{(Y_i, X_i, \delta_i) : i = 1, \cdots, n\}$ with replacement; compute the nonparametric balancing score estimate $\mu^{(\hat{b})}$ over the bootstrapped data $\{(Y_i, X_i, \delta_i)^{(b)} : i = 1, \cdots, n\}$; repeat these two steps many times and use the sample variance of $\mu^{(\hat{b})}$ as the estimate of $\text{var}(\hat{\mu})$. The bootstrap estimate includes all sources of variation.

4. Numerical Studies

We investigate the numerical performance of the proposed estimators: $\hat{\mu}_{\delta}$ uses the effective propensity score, $\hat{\mu}_Y$ uses the effective prognostic score, and $\hat{\mu}_d$ uses the effective double balancing score. Also computed are the commonly used model based estimators: the parametric regression estimation $\hat{\mu}_{reg}$, the inverse propensity weighted estimator $\hat{\mu}_{ipw}$, and the augmented inverse propensity weighted estimator $\hat{\mu}_{aipw}$. In the model based estimations, we use linear regression for $m(X)$ and linear logistic regression for $\pi(X)$. In all simulations, 200 datasets with $n = 200$ or $n = 1000$ are used.

In simulation 1, $X = (X_1, \cdots, X_{10})$ has components of independent $N(0, 1)$, $\pi = \text{expit}(X_1)$ and $Y = 3X_1 + 5X_2 + \varepsilon$ with ε of independent $N(0, 1)$. Estimation results are in Table 1. With $m(X)$ linear and $\pi(X)$ logistic linear, both working models are correct for the model based estimations. We see the nonparametric balancing score estimators have comparable performance to the model based estimators. Due to adoption of the nonparametric procedures, additional bias and variation are introduced to the proposed estimators.

However, the additional bias and variation diminish as sample size gets large. The estimators $\hat{\mu}_Y$ and $\hat{\mu}_d$ reach the optimal efficiency, and $\hat{\mu}_\delta$ is less efficient. The last observation agrees with the discussion following Theorem 1.

In simulation 2, $X = (X_1, \cdots, X_{10})$ has components of independent $N(0, 1)$, $\pi =$ expit{exp(X_2)} and $Y = (X_1 + X_3) - 10X_2^4 + 5\exp(X_4 + X_5) - X_3(X_4 + X_5) + \varepsilon$ with ε of independent $N(0, 1)$. Estimation results are in Table 2. As $m(X)$ is nonlinear and $\pi(X)$ is log-logistic, the working models are incorrect and we see large bias in the model based estimators. The nonparametric balancing score estimators show negligible bias and good efficiency.

In this simulation, $\mathcal{S}_{\delta|X}$ has one effective direction, $\mathcal{S}_{Y|X}$ has four effective directions, and $\mathcal{S}_{(\delta,Y)|X}$ has four effective directions. For $\hat{\mu}_Y$ and $\hat{\mu}_d$, we use only the first three components of the estimated S_Y and S_d. Among its first three components, S_d has X_2 as information conveyer for δ and the other two components as primary information conveyers for Y. Therefore, the dimension reduced score S_d^* still maintains the conditional balance (3) and leads to consistent estimate with good efficiency. For S_Y, its first three components carry around 93% X information about Y. The dimension reduced score S_Y^* does not maintain the conditional balance, but it conveys enough X information for the proposed estimation: $\hat{\mu}_Y$ has much smaller bias and is more stable than the model based estimators. This simulation also shows that $\hat{\mu}_d$ can outperform $\hat{\mu}_\delta$ and $\hat{\mu}_Y$: it outperforms the former in efficiency and the latter in consistency.

Dimension reduction methods are mostly developed under certain distributional assumptions. It is thus worth investigating robustness of the proposed estimation to the distribution assumptions under which the effective directions are estimated. For this purpose, we have the following simulation. In simulation 3, $Z_1, \cdots, Z_4$ are independent $N(0, 1)$, $\pi =$ expit($-Z_1 + 0.5Z_2 - 0.25Z_3 - 0.1Z_4$), and $Y = 210 + 4Z_1 + 2Z_2 + 2Z_3 + Z_4 + \varepsilon$. Suppose the covariates actually observed are $X_1 = \exp(Z_1/2)$, $X_2 = Z_2/(1 + \exp(Z_1))$, $X_3 = (Z_1Z_3/25 + 0.6)^3$, $X_4 = (Z_3 + Z_4 + 20)^2$, $X_5 = X_3X_4$, and $X_6, \cdots, X_{10}$ of independent uniform (0, 1). This setup mimics that of Kang & Schafer (2007). Here we use the sliced inverse regression to estimate the effective directions, even though X does not satisfy the linearity condition, to explore robustness.

Estimation results are in Table 3. Here we see that the proposed estimation is quite robust to mild violation of the linearity condition. This is not surprising, as sliced inverse regression is not sensitive to the linearity condition (Li, 1991). The effective balancing scores are all 4-dimensional, and we use the dimension reduced effective scores in the proposed estimation. Though the dimension reduced effective scores lose some X information, the proposed estimators still outperform the model based estimators. The inverse propensity weighting estimator $\hat{\mu}_{ipw}$ has huge bias and variability, demonstrating the instability associated with inverse propensity weighting. The doubly robust estimator $\hat{\mu}_{aipw}$ has poor performance, exemplifying the drawback of doubly robust estimators whose performance relies on model specification.

In summary, the proposed estimators have comparable performance to the model based estimators when the parametric models are correctly specified, and outperform the model based estimators otherwise. The proposed estimators also show roughly root-*n* consistency. When the effective balancing score is more than three dimensional, its first three components lead to good estimate.

5. Application

We demonstrate the proposed estimation by an Human Immunodeficiency Virus study, where 820 infected patients received combination antiretroviral therapy and had baseline characteristics measured prior to therapy, see Matthews et al. (2011). The baseline characteristics included weight, body mass index, age, CD4 counts, HIV viral load, hemoglobin, platelet, SGPT, and albumin. We are interested in the CD4 counts 96 weeks post therapy. Due to drop out and death, around 50% patients were lost to follow-up at 96 weeks. It is plausible to assume missing at random; that is, whether a patient stayed in the study depended on his/her baseline characteristics. In this study, X is the vector of baseline characteristics and Y is the CD4 counts at 96 weeks. Our interest is the mean CD4 count $E(Y)$.

We first fit the response pattern $m(X) = E(Y \mid X)$ by linear regression and the propensity score $\pi(X) = \Pr(\delta = 1 \mid X)$ by linear logistic regression. Figure 1 shows poor fit of $\hat{m}(X)$ and $\hat{\pi}(X)$. With X of dimension 9, it is nearly impossible to try out all possible higher order terms for $m(X)$ and $\pi(X)$. This casts doubt on the reliability of model based estimators. We turn to the effective balancing scores for the estimation of $E(Y)$.

The estimates of $E(Y)$ are in Table 4, where the standard deviations are estimated by bootstrap with 200 replications. In the proposed estimation, the effective propensity score $S_{\delta;}$ is 1-dimensional, the effective prognostic score S_Y and the effective double balancing score S_d are 2-dimensional, where the dimensions are determined by the sequential permutation test (Cook & Yin, 2001).

Diagnostic analysis indicates overlapping of $\mathcal{S}_{\delta|X}$ and $\mathcal{S}_{Y|X}$. More specifically, the first effective direction of $\mathcal{S}_{\delta|X}$ and that of $\mathcal{S}_{Y|X}$ are close, both close to the first effective direction of $\mathcal{S}_{(\delta,Y)|X}$. As the first effective direction of $\mathcal{S}_{Y|X}$ conveys about 70% X information about Y, the three nonparametric balancing score estimates are quite close. The inverse propensity weighting estimator $\hat{\pi}_{ipw}$ shows big bias and variability due to the poor fit of $\hat{\pi}(X)$ and the sensitivity associated with inverse weighting. In spite of the poor fit of $\hat{m}(X)$, the regression estimator $\hat{\mu}_{reg}$ seems to have little bias. This is because the bias of $\hat{\mu}_{reg}$ is $n^{-1}\sum_{i=1}^{n}\{\hat{m}(X_i)-m(X_i)\}$, and averaging over the samples can sometimes mitigate the point-wise bias in $\hat{m}(X)$.

6. Discussion

Most dimension reduction methods recover $B = (\beta_1, \cdots, \beta_K)$ as the eigenvectors of a kernel matrix. The sliced inverse regression takes $\text{cov}\{E(X \mid \mathcal{R})\}$ as the kernel matrix and the sliced average variance estimation uses $E[\{I - \text{cov}(X \mid \mathcal{R})\}^2]$, both estimated through slicing the response $\mathcal{R}$. The eigenvectors corresponding to the K largest eigenvalues are the estimates.

Both methods give root-n consistent estimates under the linearity condition, which is satisfied if X has an elliptically symmetric distribution. The sliced average variance additionally assumes cov($X \mid B'X$) to be constant.

The principal fitted component method of Cook (2007) is an extension of the sliced inverse regression. The method first finds a basis function $F_y = \{f_1(y), \cdots, f_r(y)\}$ for the inverse regression $X \mid Y$, and then estimates the effective directions through $P_F X$, the projection of X onto the subspace spanned by F_y. Though derived from normal likelihood function, the method is not tied to normality. It has “double robustness” in the sense that root-n consistency is attained under either normality or F_y is well correlated to $E(X \mid Y)$. Appropriate selection of F_y allows more effective utilization of the inverse regression information than the sliced inverse regression. Approaches for finding F_y include the inverse response plot of X versus Y (Cook, 1998), spline basis, and inverse slicing. When the inverse regression $X \mid Y$ has isotropic errors, estimates of $\beta_1, \cdots, \beta_K$ are simply the K largest eigenvectors of cov($P_F X$). Cook (2007), Cook & Forzani (2008) and Cook & Forzani (2009) give details about this method under various scenarios. Ding & Cook (2013) further extends this method to matrix-valued covariates.

Recently, Ma & Zhu (2012) proposed the semiparametric dimension reduction method. It is the only method that requires no distributional assumptions for root-n consistency. The estimation of $B = (\beta_1, \cdots, \beta_K)$ is from an estimating equation derived from a semiparametric influence function. By appropriately defining the terms in the influence function, this semiparametric method includes many dimension reduction methods as special cases. For example, one estimating equation takes the form

$$E\left([E(X|Y)-E\{E(X|Y)|B^TX\}]\{X-E(X|B^TX)\}^T\right)=0,$$

which reduces to the sliced inverse regression under the linearity condition. Consistency is achieved if either $E(\cdot \mid Y)$ or $E(\cdot \mid B^T X)$ is correctly specified, and nonparametric regression is proposed for estimating the two conditional means to circumvent model specification. This method can also handle categorial covariates so long as at least one covariate is continuous. This is a powerful method for dimension reduction but involves intensive computation.

As mentioned in §2, any root-n consistent dimension reduction method is good for finding the effective directions in the proposed method. We can pick a method of our convenience so long as the distributional assumptions are satisfied.

References

Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. Econometrica. 2006; 74:235–267.

Cao W, Tsiatis A, Davidian M. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. Biometrika. 2009; 96:732–734.

Cheng PE. Nonparametric estimation of mean functionals with data missing at random. Journal of the American Statistical Association. 1994; 89:81–87.

Chiaromonte F, Cook DR, Li B. Sufficient dimension reduction in regressions with categorical predictors. The annals of Statistics. 2002; 30:475–497.

Cook DR, Li B. Dimension reduction for conditional mean in regression. The annals of Statistics. 2002; 30:455–474.

Cook RD. On the interpretation of regression plots. Journal of American Statistical Association. 1994; 89:177–189.

Cook, RD. Regression graphics: ideas for studying regressions through graphics. New York: Wiley; 1998.

Cook RD. Fisher lecture: dimension reduction in regression. Statistical Science. 2007; 22:1–26.

Cook RD, Forzani L. Principal fitted components for dimension reduction in regression. Statistical Science. 2008; 23:485–501.

Cook RD, Forzani L. Likelihood-based sufficient dimension reduction. Journal of the American Statistical Association. 2009; 104:197–208.

Cook RD, Weisberg S. Discussion of “sliced inverse regression for dimension reduction”. Journal of American Statistical Association. 1991; 86:328–332.

COOK RD, Yin XR. Dimension reduction and visualization in discriminant analysis. Australian & New Zealand Journal of Statistics. 2001; 43:147–177.

D’Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Statistics in Medicine. 1998; 17:2265–2281. [PubMed: 9802183]

Devroye LP, Wagner TJ. Distribution-free consistency results in nonparametric discrimination and regression function estimations. The Annuals of Statistics. 1980; 8:231–239.

Ding, S.; Cook, RD. Statistica Sinica. 2013. Dimension folding pca and pfc for matrix-valued predictors. To appear

Dong Y, Li B. Dimension reduction for non-elliptically distributed predictors: second-order moments. Biometrika. 2010; 97:279–294.

Fan J, Marron JS. Best possible constant for bandwidth selection. The Annals of Statistics. 1992; 20:2057–2070.

Hahn J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica. 1998; 66:315–331.

Hansen BB. The prognostic analogue of the propensity score. Biometrika. 2008; 95:481–488.

Härdle, W.; Müller, M.; Sperlich, S.; Werwatz, A. Nonparametric and semiparametric models. Berlin Heidelberg: Springer-Verlag; 2004.

Hastie T, Tibshirani R. Generalized additive models. Statistical Science. 1986; 1:297–318.

Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. Journal of American Statistical Association. 1952; 47:663–685.

Kang DY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science. 2007; 22:523–539.

Li B, Dong Y. Dimension reduction for non-elliptically distributed predictors. The Annals of Statistics. 2009; 37:1272–1298.

Li B, Wang S. On directional regression for dimension reduction. Journal of the American Statistical Association. 2007; 102:997–1008.

Li KC. Sliced inverse regression for dimension reduction. Journal of American Statistical Association. 1991; 86:316–327.

Li Y, Zhu LX. Asymptotics for sliced average variance estimation. The Annals of Statistics. 2007; 35:41–69.

Little R, An H. Robust likelihood-based analysis of multivariate data with missing values. Statistica Sinica. 2004; 14:949–968.

Lunceford J, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects. Statistics in Medicine. 2004; 23:2937–2960. [PubMed: 15351954]

Ma Y, Zhu L. A semi parametric approach to dimension reduction. Journal of the American Statistical Association. 2012; 107:168–179. [PubMed: 23828688]

Matthews GV, Manzini P, Hu Z, Khabo P, Maja P, Matchaba G, Sangweni P, Metcalf J, Pool N, Orsega S, Emery S. STUDY TEAM PI. Impact of lamivudine on hiv and hepatitis b virus-related outcomes in hiv/hepatitis b virus individuals in a randomized clinical trial of antiretroviral therapy in southern africa. AIDS. 2011; 25:1727–1735. [PubMed: 21716078]

Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. Journal of American Statistical Association. 1995; 90:122–129.

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.

Rosenbaum, PR. Observational Studies. 2. New York: Springer; 2002.

Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 60:211–213.

Rubin, DB. Multiple imputation for nonresponse in surveys. New York: Wiley; 1987.

Ruppert D, Wand MP. Multivariate locally weighted least squares regression. Annals of Statistics. 1994; 22:1346–1370.

Schafer, JL. Analysis of incomplete multivariate data. London: Chapman and Hall; 1997.

Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semi-parametric nonresponse models. Journal of American Statistical Association. 1999; 94:1096–1120.

Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society, Series B. 1991; 53:683–690.

Silverman, BW. Density estimation for statistics and data analysis, Vol 26. of monographs on statistics and applied probability. London: Chapman and Hall; 1986.

Vartivarian, S.; Little, RJA. Proceedings of the Survey Research Methods Section, American Statistical Association. American Statistical Association; 2008. On the formation of weighted adjustment cells for unit nonresponse.

Appendix

Appendix 1. Proof for =

Denote B_d as the basis for $\mathcal{S}_{(\delta,Y)|X}$ and B_{d*} as the basis for $\mathcal{S}_{\delta Y|X}$. From (4),

$$\delta \perp X|B_d^{'}X, \quad Y \perp X|B_d^{'}X.$$

It follows that $\delta Y \perp X|B_d^{'}X$ and $\mathcal{S}_{\delta Y|X} \subset \mathcal{S}_{(\delta,Y)|X}$.

Note that

$$\Pr(\delta{=}0|X){=}\Pr(\delta Y{=}0|X){-}\Pr(Y{=}0,\delta{=}1|X){=}\Pr(\delta Y{=}0|X),$$

where the second equation is due to $\Pr(Y = 0 \mid X) = 0$ for Y continuous. Since B_{d*} is the basis for $\mathcal{S}_{\delta Y|X}$, the right hand side of the above equation is a function of $B_{d*}^{'}X$. Thus $\delta \perp X|B_{d*}^{'}X$ and $\mathcal{S}_{\delta|X} \subset \mathcal{S}_{(\delta Y)|X}$.

Note that

$$\Pr(Y \le y, \delta{=}1|X){=}\Pr(\delta Y \le y|X){-}\Pr(\delta{=}0|X)I(y \le 0), \quad \text{and}$$
$$\Pr(Y \le y, \delta{=}1|X){=}\Pr(Y \le y|X)\Pr(\delta{=}1|X).$$

The second equation is true due to missing at random. In the above equations, Pr($\delta Y \le y \mid X$) is a function of $B_{d*}^{'}X$ as B_{d*} is the basis for $\mathcal{S}_{\delta Y|X}$, Pr($\delta = 0 \mid X$) and Pr($\delta = 1 \mid X$) are functions of $B_{d*}^{'}X$ as $\mathcal{S}_{(\delta)|X} \subset \mathcal{S}_{(\delta Y)|X}$. Therefore, Pr($Y \le y \mid X$) is a function of $B_{d*}^{'}X$. It follows that $Y \perp X|B_{d*}^{'}X$ and $\mathcal{S}_{Y|X} \subset \mathcal{S}_{(\delta Y)|X}$.

As $\mathcal{S}_{\delta|X} \subset \mathcal{S}_{(\delta Y)|X}$ and $\mathcal{S}_{Y|X} \subset \mathcal{S}_{(\delta Y)|X}$, it follows from Remark 2 that $\mathcal{S}_{(\delta,Y)|X} \subset \mathcal{S}_{(\delta Y)|X}$.

If Y is categorical, we can perform a shift transformation $Y^* = Y + c$ such that $Y^* > 0$. It follows that $\mathcal{S}_{(\delta,Y)|X} = \mathcal{S}_{(\delta,Y^*)|X} = \mathcal{S}_{\delta Y^*|X}$.

Appendix 2. Proof of Theorem 1

Theorem 1 is developed under the following regularity conditions:

1. The kernel function satisfies: $\int \mathbf{u}\, \mathcal{K}(\mathbf{u})d\mathbf{u} = 0$, $\int \mathbf{u}\mathbf{u}^T \mathcal{K}(\mathbf{u})d\mathbf{u} = \gamma_{\mathcal{K}} I_2$, and $\int \mathcal{K}^2(\mathbf{u})d\mathbf{u} = \tau_{\mathcal{K}}$, with $\gamma_{\mathcal{K}} < \infty$ and $\tau_{\mathcal{K}} < \infty$.
2. $\pi(x)$ is bounded away from 0.
3. The density of x is bounded away from 0.

We write $n^{1/2}(\hat{\mu} - \mu)$ as

$$n^{1/2}(\hat{\mu}{-}\mu){=}n^{1/2}A_n{+}n^{1/2}B_n{+}n^{1/2}C_n,$$

with

$$A_n{=}n^{-1}\sum_{i=1}^{n} m(S_i){-}\mu,$$
$$B_n{=}n^{-1}\sum_{i=1}^{n} E\{\hat{m}(S_i){-}m(S_i)|\mathscr{O}_i\},$$
$$C_n{=}n^{-1}\sum_{i=1}^{n} \hat{m}(S_i){-}m(S_i){-}E\{\hat{m}(S_i){-}m(S_i)|\mathscr{O}_i\},$$

where $\mathscr{O}_i = \{(X_j, Y_j, \delta_j) : j \ne i\}$. It is obvious that $n^{1/2}A_n$ converges in distribution to $N(0, \text{var}\{m(S)\})$.

By (5),

$$\hat{m}(S_i){=}n^{-1}\sum_{j=1}^{n}\delta_j Y_j \mathscr{K}_H(S_i{-}S_j)/n^{-1}\sum_{j=1}^{n}\delta_j \mathscr{K}_H(S_i{-}S_j),$$

with $n^{-1}\sum_{j=1}^{n}\delta_j\mathscr{K}_H(S_i-S_j)=\pi(S_i)f(S_i)+o_p(1)$ and *f*(*s*) the density of *S*. It follows that

$$B_n=n^{-1}\sum_{j=1}^{n}\delta_j E[\frac{\mathscr{K}_H(S_i-S_j)\{Y_j-m(S_i)\}}{\pi(S_i)f(S_i)}|\mathscr{O}_i]\{1+o_p(1)\}.$$

Similar to the argument for Theorem 2.1 of Cheng (1994), it can be shown that $\sqrt{n}(B_n-B_n^*)=o_p(1)$ with

$$B_n^*=n^{-1}\sum_{j=1}^{n}\delta_j\frac{Y_j-m(S_j)}{\pi(S_j)}.$$

Due to conditional independence (3), $n^{1/2}B_n^*$ converges in distribution to N (0, E\{var(Y | S)/π(S)\}).

For C_n, $E(C_n)$ = 0 and $nE(C_n^2)\leq E[\{\hat{m}(S)-m(S)\}^2]=O(\{\mathrm{tr}(HH^T)\}^2+\{n\det(H)\}^{-1})$, thus $C_n = o_p(n^{-1/2})$. As A_n and B_n^* are independent, $n^{1/2}(\hat{\mu} - \mu)$ is asymptotically normal of mean 0 and variance var\{$m(S)$\} + E\{var(Y | S)/π(S)\}.

Following Ruppert & Wand (1994), the negligible terms involving H are

$$\begin{array}{rl} E(B_n) & =E\{\hat{m}(S)-m(S)\}=\frac{1}{2}\mathrm{tr}\{H^T\Delta_m(S)H\},\\ n\{\mathrm{var}(C_n)\} & =E[\{\hat{m}(S)-m(S)\}^2]=\{n\det(H)\}^{-1}E\left[\mathrm{var}(Y|S)\{f(S)\pi(S)\}^{-1}\right]\tau_{\mathscr{K}}. \end{array}$$

With $H = h_nI_K$, $E(B_n)=h_n^2\mathscr{B}$ and $n\{\mathrm{var}(C_n)\}=(nh_n^K)^{-1}\mathscr{V}$,

$$\begin{array}{rl} \mathscr{B} & =\frac{1}{2}\mathrm{tr}\{\Delta_m(S)\},\\ \mathscr{V} & =E\left[\mathrm{var}(Y|S)\{f(S)\pi(S)\}^{-1}\right]\tau_{\mathscr{K}}. \end{array}$$

In the above expressions, $\Delta_m(S)=E\{\mathscr{H}_m(S)+2\nabla_m^T(S)\nabla_{\pi f}/\pi f(S)\}\gamma_{\mathscr{K}}$, where $\nabla_m(s)$ and $\mathcal{H}_m(s)$ stand for the gradient and the Hessian matrix of $m(s)$, respectively, $\nabla_{\pi f}/\pi f(s)$ = \{ $\pi(s)\nabla_f(s)$ + $\nabla_\pi(s)f(s)$\}/\{ $\pi(s)$/\{ $\pi(s)f(s)$\} with $\nabla_\pi(s)$ and $\nabla_f(s)$ the gradients of $\pi(s)$ and $f(s)$, respectively.

The mean squared error is

$$E\{(\hat{\mu}-\mu)^2\}=h_n^4\mathscr{B}^2+(n^2h_n)^{-K}\mathscr{V}+n^{-1}\sigma^2.$$

The optimal bandwidth, which minimizes the mean squared error, is

$$h_{opt}=\{K\mathscr{V}/(4\mathscr{B}^2)\}^{1/(K+4)}n^{-2/(K+4)},$$

which can be estimated by the plug-in method.

Appendix 3. Proof of Theorem 2

Denote the proposed estimator under $\hat{B}$, the root-n estimate of B, as $\hat{\hat{\mu}}$ which is given as in (6) except that

$$\hat{m}(S_i)=n^{-1}\sum_{j=1}^{n}\delta_j Y_j \mathscr{K}_H(\hat{S}_i-\hat{S}_j)/n^{-1}\sum_{j=1}^{n}\delta_j \mathscr{K}_H(\hat{S}_i-\hat{S}_j),$$

with $\hat{S} = \hat{B}X$.

The difference between $\hat{\hat{\mu}}$ and $\hat{\mu}$ comes from that between $\mathscr{K}_H(S_i - S_j)$ and $\mathscr{K}_H(\hat{S}_i-\hat{S}_j)$. With $H = h_n I_K$, $\mathscr{K}_H(S_i-S_j)=h_n^{-K}\mathscr{K}\{(S_i-S_j)/h_n\}$ and $\mathscr{K}_H(\hat{S}_i-\hat{S}_j)=h_n^{-K}\mathscr{K}\{(\hat{S}_i-\hat{S}_j)/h_n\}$. The latter can be further written as

$$h_n^{-K}\mathscr{K}\left(\frac{S_i-S_j}{h_n}+\frac{(\hat{B}-B)(X_i-X_j)}{h_n}\right),$$

At optimal bandwidth $h_n \sim n^{-2/(K+4)}$ and $\hat{B} - B = O_p(n^{-1/2})$, the second term inside the kernel function is $O\{n^{-K/(2K+8)}\} = o_p(n^{-1/2})$. It follows that $\sqrt{n}(\hat{\hat{\mu}}-\hat{\mu})\sim o_p(1)$, and $\hat{\hat{\mu}}$ is asymptotically equivalent to $\hat{\mu}$.

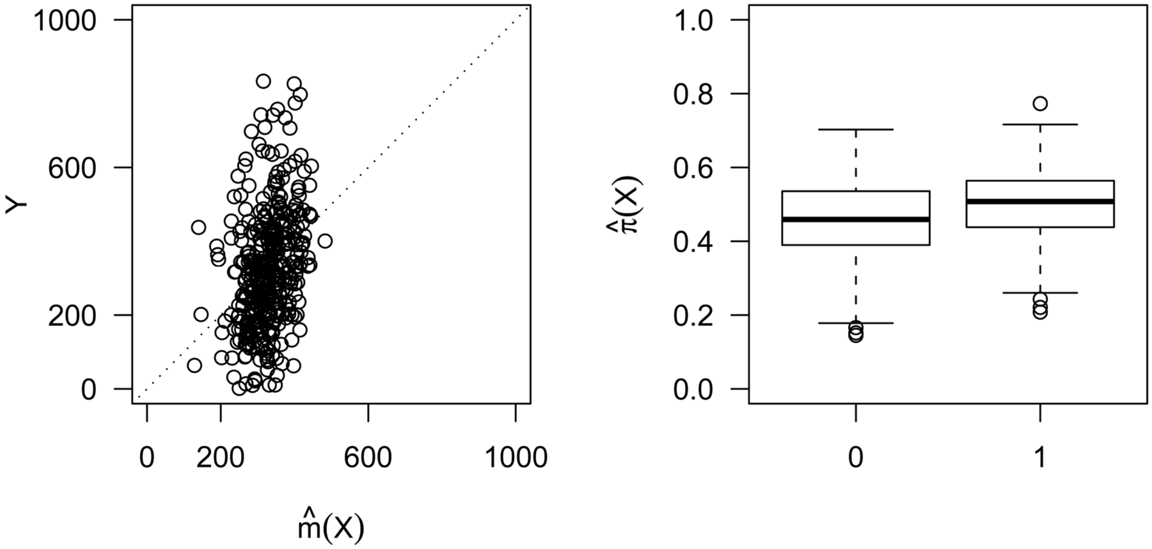

Fig. 1.

Parametric fit to the response and the missingness. On the left is the observed response versus the fitted response from linear regression. On the right is the box plot of the fitted propensity score from linear logistic regression: 0 for the subjects with *Y* missing and 1 for the subjects with *Y* observed.

Table 1

Results for simulation 1: Monte Carlo bias (Bias), standard deviation (SD), root mean squared error (RMSE), and the estimated standard deviation (ESD) and coverage percentage (CP) of the 95% confidence interval from bootstrap with 200 replications.

		$\hat{\mu}_{\delta}$	$\hat{\mu}_{Y}$	$\hat{\mu}_{d}$	$\hat{\mu}_{reg}$	$\hat{\mu}_{ipw}$	$\hat{\mu}_{aipw}$
n = 200	Bias	0.04	0.09	0.09	−0.04	0.00	−0.03
	SD	0.57	0.43	0.45	0.42	0.58	0.43
	RMSE	0.57	0.44	0.46	0.43	0.58	0.43
	ESD	0.63	0.45	0.48	0.43	0.61	0.43
	CP	96.5	94.0	95.5	95.0	95.0	95.0
n = 1000	Bias	0.02	0.04	0.05	0.00	0.02	0.00
	SD	0.23	0.20	0.19	0.19	0.23	0.19
	RMSE	0.23	0.20	0.20	0.19	0.23	0.19
	ESD	0.24	0.20	0.20	0.19	0.24	0.19
	CP	96.5	96.0	95.5	96.5	96.0	95.5

Table 2

Results for simulation 2: Monte Carlo bias (Bias), standard deviation (SD), root mean squared error (RMSE), and the estimated standard deviation (ESD) and coverage percentage (CP) of the 95% confidence interval from bootstrap with 200 replications.

		$\hat{\mu}_{\delta}$	$\hat{\mu}_{Y}$	$\hat{\mu}_{d}$	$\hat{\mu}_{reg}$	$\hat{\mu}_{ipw}$	$\hat{\mu}_{aipw}$
n = 200	Bias	0.08	0.14	0.06	3.12	−7.72	−7.39
	SD	8.80	8.47	7.91	8.98	24.45	23.83
	RMSE	8.80	8.47	7.91	9.50	25.64	24.95
	ESD	9.20	8.85	8.35	7.45	12.69	16.96
	CP	93.5	92.0	92.5	78.0	93.5	94.0
n = 1000	Bias	0.06	0.11	0.04	1.91	−10.74	−10.88
	SD	3.86	3.38	3.32	4.06	10.67	10.45
	RMSE	3.86	3.38	3.32	4.49	15.14	15.09
	ESD	3.83	3.67	3.56	3.91	7.49	7.52
	CP	95.0	95.5	95.0	88.0	80.5	82.0

Table 3

Results for simulation 3: Monte Carlo bias (Bias), standard deviation (SD), root mean squared error (RMSE).

		$\hat{\mu}_{\delta}$	$\hat{\mu}_{Y}$	$\hat{\mu}_{d}$	$\hat{\mu}_{reg}$	$\hat{\mu}_{ipw}$	$\hat{\mu}_{aipw}$
n = 200	Bias	−0.17	0.08	−0.12	0.33	247.21	−2.70
	SD	0.50	0.42	0.42	0.46	184.3	3.21
	RMSE	0.53	0.43	0.44	0.57	308.35	4.20
n = 1000	Bias	−0.14	0.07	−0.09	0.27	261.72	−4.38
	SD	0.20	0.19	0.19	0.23	110.06	6.48
	RMSE	0.24	0.2	0.21	0.35	283.92	7.82

Table 4

Estimates of mean CD4 counts at 96 weeks: the proposed estimator $\hat{\mu}_{\delta}$, $\hat{\mu}_{Y}$, and $\hat{\mu}_{d}$, the inverse propensity weighting estimator $\hat{\mu}_{ipw}$, the regression estimator $\hat{\mu}_{reg}$, and the augmented inverse propensity weighting estimator $\hat{\mu}_{aipw}$.

	$\hat{\mu}_{\delta}$	$\hat{\mu}_{Y}$	$\hat{\mu}_{d}$	$\hat{\mu}_{reg}$	$\hat{\mu}_{ipw}$	$\hat{\mu}_{aipw}$
Estimate	322.7	323.3	323.0	322.4	680.6	328.5
SD	8.6	8.8	8.8	8.4	33.2	8.9