Biometrika (2024), **111**, 4, *pp.* 1413–1420 https://doi.org/10.1093/biomet/asae017
Advance Access publication 22 March 2024

Covariate adjustment in randomized experiments with missing outcomes and covariates

By ANQI ZHAO

Fuqua School of Business, Duke University,
100 Fuqua Drive, Durham, North Carolina 27708, U.S.A.
az171@duke.edu

PENG DING

Department of Statistics, University of California, Berkeley,
425 Evans Hall, Berkeley, California 94720, U.S.A.
pengdingpku@berkeley.edu

and FAN LI

Department of Statistical Science, Duke University,
214 Old Chemistry, Box 90251, Durham, North Carolina 27708, U.S.A.
fl35@duke.edu

Summary

Covariate adjustment can improve precision in analysing randomized experiments. With fully observed data, regression adjustment and propensity score weighting are asymptotically equivalent in improving efficiency over unadjusted analysis. When some outcomes are missing, we consider combining these two adjustment methods with the inverse probability of observation weighting for handling missing outcomes, and show that the equivalence between the two methods breaks down. Regression adjustment no longer ensures efficiency gain over unadjusted analysis unless the true outcome model is linear in covariates or the outcomes are missing completely at random. Propensity score weighting, in contrast, still guarantees efficiency over unadjusted analysis, and including more covariates in adjustment never harms asymptotic efficiency. Moreover, we establish the value of using partially observed covariates to secure additional efficiency by the missingness indicator method, which imputes all missing covariates by zero and uses the union of the completed covariates and corresponding missingness indicators as the new, fully observed covariates. Based on these findings, we recommend using regression adjustment in combination with the missingness indicator method if the linear outcome model or missing-completely-at-random assumption is plausible and using propensity score weighting with the missingness indicator method otherwise.

Some key words: Inverse probability weighting; Missingness indicator; Propensity score; Regression adjustment.

1. Covariate adjustment in randomized experiments: a review and open questions

Adjusting for chance imbalance in covariates can improve precision in analysing randomized experiments (Fisher, 1935; Lin, 2013). Consider a randomized controlled trial with two treatment levels of interest, indexed by $z = 1$ for treatment and 0 for control, and a study population of N units, indexed by $i = 1, \ldots, N$. Let $x_i \in \mathbb{R}^J$, $Z_i \in \{1, 0\}$ and $Y_i \in \mathbb{R}$ denote the baseline covariates, treatment assignment and outcome of unit i. Let $Y_i(1) \in \mathbb{R}$ and $Y_i(0) \in \mathbb{R}$ denote the potential

Table 1. *A summary of* $\{\hat{\tau}_{\text{unadj}}, \hat{\tau}_{\text{reg}}, \hat{\tau}_{\text{ps}}\}$ *when all data are observed (column 3) and when outcomes are partially missing (column 4). Let* $x_i' = x_i - \bar{x}$ *denote the centred covariates, where* $\bar{x} = N^{-1}\sum_{i=1}^{N} x_i$. *Let* $\hat{p}_i$ *denote the estimated probability of* Y_i *being observed given* (x_i, Z_i), *and let* $\hat{\pi}_i = Z_i/\hat{e}_i + (1 - Z_i)/(1 - \hat{e}_i)$ *denote the inverse of the estimated probability of the treatment received for unit i with* $\hat{e}_i$ *as the estimated propensity score*

	Regression specification	Weight over $i = 1, \ldots, n$	Weight over $\{i\colon R_i^Y = 1\}$
$\hat{\tau}_{\text{unadj}}$	$\texttt{lm}(Y_i \sim 1 + Z_i)$	1	$\hat{p}_i^{-1}$
$\hat{\tau}_{\text{reg}}$	$\texttt{lm}(Y_i \sim 1 + Z_i + x_i' + Z_i x_i')$	1	$\hat{p}_i^{-1}$
$\hat{\tau}_{\text{ps}}$	$\texttt{lm}(Y_i \sim 1 + Z_i)$	$\hat{\pi}_i$	$\hat{p}_i^{-1}\hat{\pi}_i$

outcomes of unit i under treatment and control, respectively, with $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$. Assume throughout that the N units are an independent and identically distributed sample from some population and that the treatment levels are assigned independently across units with constant treatment probability $\text{pr}(Z_i = 1) = e \in (0, 1)$. Define $\tau_i = Y_i(1) - Y_i(0)$ as the individual treatment effect for unit i and $\tau = E(\tau_i) = E\{Y_i(1)\} - E\{Y_i(0)\}$ as the average treatment effect of interest. We first review below three estimators of τ when (Y_i, x_i, Z_i) are fully observed.

A simple unbiased estimator of τ is the difference in means of the outcomes between the two treatment groups, commonly referred to as the unadjusted estimator and denoted $\hat{\tau}_{\text{unadj}}$. It is numerically equal to the coefficient of Z_i from the ordinary-least-squares fit of the *unadjusted regression* of Y_i on $(1, Z_i)$, denoted $\texttt{lm}(Y_i \sim 1 + Z_i)$ under the R convention.

Regression adjustment and propensity score weighting are two ways to adjust for chance imbalances in covariates. First, the interacted regression $\texttt{lm}\{Y_i \sim 1 + Z_i + (x_i - \bar{x}) + Z_i(x_i - \bar{x})\}$, where $\bar{x} = N^{-1}\sum_{i=1}^{N} x_i$, gives a covariate-adjusted variant of the unadjusted regression (Tsiatis et al., 2008; Lin, 2013; Negi & Wooldridge, 2021). The ordinary-least-squares coefficient of Z_i defines a regression-adjusted estimator of τ, denoted $\hat{\tau}_{\text{reg}}$.

Next, let $e_i = \text{pr}(Z_i = 1 \mid x_i)$ denote the propensity score of unit i (Rosenbaum & Rubin, 1983). The propensity score weighting approach to covariate adjustment weights observations by functions of an estimate of e_i (Williamson et al., 2014). In our setting, e_i is known and equals e for all units. Nevertheless, standard results suggest that we can still estimate e_i using a working model as a means to improve efficiency; see, e.g., Hahn (1998), Hirano et al. (2003) and Shen et al. (2014). Specifically, let $\hat{e}_i$ be the maximum likelihood estimate of e_i based on the logistic regression of Z_i on $(1, x_i)$, denoted $\texttt{glm}(Z_i \sim 1 + x_i)$ under the R convention. We can estimate τ by the coefficient of Z_i from the weighted-least-squares fit of the unadjusted regression $\texttt{lm}(Y_i \sim 1 + Z_i)$, where we weight unit i by $\hat{e}_i^{-1}$ if $Z_i = 1$ and by $(1 - \hat{e}_i)^{-1}$ if $Z_i = 0$, summarized as $\hat{\pi}_i = Z_i/\hat{e}_i + (1 - Z_i)/(1 - \hat{e}_i)$. We denote the resulting estimator by $\hat{\tau}_{\text{ps}}$, where subscript ps stands for propensity score weighting. Other propensity score weights such as overlap weighting can also be used (Zeng et al., 2021).

The $\{\hat{\tau}_{\text{unadj}}, \hat{\tau}_{\text{reg}}, \hat{\tau}_{\text{ps}}\}$ together define three regression estimators of τ with fully observed data, summarized in the first three columns of Table 1. Under mild regularity conditions, they are all consistent and asymptotically normal (Tsiatis et al., 2008; Lin, 2013; Williamson et al., 2014; Negi & Wooldridge, 2021), with $\hat{\tau}_{\text{reg}}$ and $\hat{\tau}_{\text{ps}}$ being asymptotically equivalent in improving precision over $\hat{\tau}_{\text{unadj}}$ (Shen et al., 2014; Zeng et al., 2021); see Theorem S1 in the Supplementary Material.

Missing data are common in practice and pose challenges to inference. Assuming missingness only in covariates, Zhao & Ding (2024) proposed using the interacted regression with missingness indicators for covariates included as additional covariates, and showed that the resulting inference guarantees asymptotic efficiency over unadjusted analysis. Despite the vast literature on missing data and that on covariate adjustment, there lacks theoretical guidance on covariate adjustment with missingness in both outcomes and covariates in randomized experiments. Many important questions such as the following remain open. How do we conduct covariate adjustment in the presence of missing outcomes? Does the resulting inference ensure consistency and efficiency gain over unadjusted analysis? Does the asymptotic equivalence between regression adjustment and propensity score weighting still hold? Can the missingness indicator method of Zhao & Ding (2024) be extended to the presence

of missing outcomes? Chang et al. (2023) discussed some of these issues and proposed an estimator without theoretical investigation. This paper provides theoretical answers to these questions and proposes two easy-to-implement estimators. We begin with the case of missingness in only outcomes in § 2 and then extend to the case of missingness in both covariates and outcomes in § 3.

2. Covariate adjustment with missing outcomes

2.1. *Regression estimators with missing outcomes*

We first extend § 1 to the presence of missing outcomes. Assume throughout the rest of the paper that x_i and Z_i are fully observed for all units, whereas Y_i is missing for some units. Let $R_i^Y \in \{1, 0\}$ be the indicator of Y_i being observed for unit i, with $R_i^Y = 1$ if Y_i is observed and $R_i^Y = 0$ otherwise. Recall from Table 1 that, when all data are observed, $\hat{\tau}_{\text{unadj}}$, $\hat{\tau}_{\text{reg}}$ and $\hat{\tau}_{\text{ps}}$ are the coefficients of Z_i from the least-squares fits of the unadjusted regression, the interacted regression and the unadjusted regression over all units, respectively, with weights $\pi_{i,\text{unadj}} = 1$, $\pi_{i,\text{reg}} = 1$ and $\pi_{i,\text{ps}} = \hat{\pi}_i$ for unit i. In the presence of missing outcomes, let $p_i = \text{pr}(R_i^Y = 1 \mid x_i, Z_i)$ denote the probability of Y_i being observed given (x_i, Z_i), and let $\hat{p}_i$ be an estimate of p_i. By the inverse probability of observation weighting (Seaman & White, 2013), we can instead fit the corresponding regression over units with observed outcomes, indexed by $\{i\colon R_i^Y = 1\}$, with weight $\pi'_{i,\diamond} = \hat{p}_i^{-1}\pi_{i,\diamond}$ ($\diamond = \text{unadj}, \text{reg}, \text{ps}$) for unit i. This generalizes $\{\hat{\tau}_{\text{unadj}}, \hat{\tau}_{\text{reg}}, \hat{\tau}_{\text{ps}}\}$ to the presence of missing outcomes, summarized in the last column of Table 1. The definitions of the $\hat{\tau}_\diamond$ when all data are observed are special cases with $R_i^Y = 1$ and $\hat{p}_i = p_i = 1$ for all i. We hence use the same notation to denote the generalized estimators with missing outcomes to highlight the connection. The generalized $\hat{\tau}_{\text{ps}}$ is a double-weighted estimator, where we use $\hat{p}_i^{-1}$ and $\hat{\pi}_i$ to address missing outcomes and covariate adjustment, respectively (Chang et al., 2023; Negi, 2024).

2.2. *Asymptotic theory*

We now establish the asymptotic properties of the generalized $\{\hat{\tau}_{\text{unadj}}, \hat{\tau}_{\text{reg}}, \hat{\tau}_{\text{ps}}\}$. To begin with, the following assumption specifies the outcome missingness mechanism.

Assumption 1. Suppose that

(i) $R_i^Y \perp\!\!\!\perp \{Y_i(1), Y_i(0)\} \mid (x_i, Z_i)$;
(ii) $p_i = \text{pr}(R_i^Y = 1 \mid x_i, Z_i) = \{1 + \exp(-U_i^{\mathrm{T}}\beta^*)\}^{-1}$, where $U_i = U(x_i, Z_i)$ is a known vector function of (x_i, Z_i) and β^* is the unknown parameter, and we construct $\hat{p}_i$ by the logistic regression $\texttt{glm}(R_i^Y \sim U_i)$ over $i = 1, \ldots, N$.

Assumption 1(i) ensures that R_i^Y is independent of Y_i conditioning on the fully observed (x_i, Z_i). The outcome is hence missing at random in the sense that whether an outcome is missing is independent of the value of the outcome conditional on the observables. Assumption 1(ii) further specifies the functional form of the outcome missingness mechanism. We focus on the logistic missingness model because of its prevalence in practice. We conjecture that similar results hold for general missingness models and relegate the formal theory to future research. We use U_i to denote the regressor vector in the true outcome missingness model under Assumption 1(ii). In practice, the true value of U_i is often unknown. Common choices for fitting a working model include $U_i = (1, x_i^{\mathrm{T}})^{\mathrm{T}}$, $U_i = (1, x_i^{\mathrm{T}}, Z_i)^{\mathrm{T}}$ and $U_i = (1, x_i^{\mathrm{T}}, Z_i, Z_i x_i^{\mathrm{T}})^{\mathrm{T}}$.

Theorem 1 below generalizes the theory of covariate adjustment with fully observed data to the presence of missing outcomes and gives the asymptotic distributions of the generalized $\{\hat{\tau}_{\text{unadj}}, \hat{\tau}_{\text{reg}}, \hat{\tau}_{\text{ps}}\}$. Let $\tilde{Y}_i = e^{-1}Y_i(1) + (1-e)^{-1}Y_i(0)$, and denote by $\text{proj}(\tilde{Y}_i \mid 1, x_i) = E(\tilde{Y}_i) + \text{cov}(\tilde{Y}_i, x_i)\{\text{cov}(x_i)\}^{-1}\{x_i - E(x_i)\}$ the linear projection of $\tilde{Y}_i$ on $(1, x_i)$.

Theorem 1. *Assume complete randomization with $Z_i \perp\!\!\!\perp \{Y_i(1), Y_i(0), x_i\}$ and that Assumption 1 holds. Under standard regularity conditions, as $N \to \infty$, we have $\sqrt{N}(\hat{\tau}_\diamond - \tau) \to \mathcal{N}(0, v_\diamond)$ in distribution for $\diamond = \text{unadj}, \text{reg}, \text{ps}$, where*

(i) $v_{\text{ps}} = v_{\text{unadj}} - e(1-e)\,\text{var}\{\text{proj}(\tilde{Y}_i \mid 1, x_i)\} \leqslant v_{\text{unadj}}$;
(ii) v_{reg} *can be either greater or less than* v_{unadj} *depending on the data-generating process. As two special cases, we have* $v_{\text{reg}} \leqslant v_{\text{ps}} \leqslant v_{\text{unadj}}$ *if*

(a) Y_i *is missing completely at random with* $p_i = p \in (0, 1)$ *or*
(b) *the outcome model* $E(Y_i \mid x_i, Z_i = z) = E\{Y_i(z) \mid x_i\}$ *is linear in* x_i *for* $z = 0, 1$.

We relegate the explicit expressions of v_{unadj} and v_{reg} to Theorem S1 in the Supplementary Material. When $p_i = 1$ for all i, the three asymptotic variances reduce to those in the standard theory for fully observed data with $v_{\text{reg}} = v_{\text{ps}} \leqslant v_{\text{unadj}}$. Theorem 1 has two implications. First, it ensures the consistency and asymptotic normality of $\{\hat{\tau}_{\text{unadj}}, \hat{\tau}_{\text{reg}}, \hat{\tau}_{\text{ps}}\}$ in the presence of missing outcomes. Second, it clarifies the relative efficiency of $\{\hat{\tau}_{\text{unadj}}, \hat{\tau}_{\text{reg}}, \hat{\tau}_{\text{ps}}\}$ and highlights a key deviation from the theory when all outcomes are observed: regression adjustment by the interacted specification no longer guarantees efficiency gain in the presence of missing outcomes, but propensity score weighting still does. The asymptotic equivalence between the two methods for improving precision therefore breaks down.

More specifically, Theorem 1(i) ensures that adjustment by propensity score weighting reduces the asymptotic variance by $e(1-e)\,\text{var}\{\text{proj}(\tilde{Y}_i \mid 1, x_i)\}$. This expression does not depend on p_i such that the reduction is the same as the reduction when outcomes are fully observed. Observe that adding more covariates to x_i never reduces the variance of $\text{proj}(\tilde{Y}_i \mid 1, x_i)$. Adjusting for more covariates by propensity score weighting hence never hurts the asymptotic efficiency of the resulting $\hat{\tau}_{\text{ps}}$. This underpins the extension to the case with missingness in both covariates and outcomes in § 3 below.

On the other hand, Theorem 1(ii) suggests that $\hat{\tau}_{\text{reg}}$ does not ensure efficiency gain over the unadjusted estimator $\hat{\tau}_{\text{unadj}}$ unless the outcomes are missing completely at random or the true outcome model is linear in x_i. The latter condition echoes the standard result in semiparametric efficiency theory. In particular, Robins et al. (2007) pointed out that $\hat{\tau}_{\text{reg}}$ can be written as a classic augmented inverse propensity-score-weighted estimator with a linear outcome model that corresponds to the interacted regression; see Proposition S2 in the Supplementary Material. Standard theory ensures that it achieves semiparametric efficiency if both the missingness model and the outcome model are correctly specified (Tsiatis, 2006).

3. Covariate adjustment with missingness in both covariates and outcomes

3.1. *Overview and recommendation*

We now extend to the case of missingness in both covariates and outcomes. Recall that $x_i \in \mathbb{R}^J$ is the vector of baseline covariates that are fully observed for all units. Assume that, in addition to x_i, we also have K partially observed covariates, denoted by $w_i = (w_{i1}, \ldots, w_{iK}) \in \mathbb{R}^K$ for $i = 1, \ldots, N$. Of interest is how we may use this additional information to further improve inference.

To this end, we recommend using the missingness indicator method to address missing covariates (Zhao & Ding, 2024) and then constructing the regression-adjusted and propensity-score-weighted estimators based on the augmented covariate vectors from the missingness indicator method; see Algorithm 1 below. We show in § 3.2 below that the resulting estimators preserve the theoretical properties in Theorem 1. Accordingly, we recommend using regression adjustment when the linear outcome model or missing-completely-at-random assumption is plausible and using propensity score weighting otherwise. The results combine the theory in § 2 on missing outcomes and that of Zhao & Ding (2024) on missing covariates, and offer a full picture of covariate adjustment with missing outcomes and covariates.

Let $R_i^w = (R_{i1}^w, \ldots, R_{iK}^w) \in \{1, 0\}^K$ represent the missingness in w_i, with $R_{ik}^w = 1$ if w_{ik} is observed and $R_{ik}^w = 0$ if w_{ik} is missing. Let $w_i^0 \in \mathbb{R}^K$ denote an imputed variant of w_i, where we impute all missing elements with zero. Note that w_i^0 is in fact the elementwise product, or intuitively the interaction, between w_i and R_i^w regardless of the actual values of the missing elements in w_i. The concatenation of (x_i, w_i^0, R_i^w), denoted $x_i^{\text{mim}} \in \mathbb{R}^{J+2K}$, gives the vector of fully observed covariates under the missingness indicator method, which imputes all missing covariates by zero and augments

the completed covariates by the corresponding missingness indicators. We use superscript mim to signify the missingness indicator method.

Observe that x_i^{mim} summarizes all observed information in (x_i, w_i). Renew $e_i = \text{pr}(Z_i = 1 \mid x_i^{\text{mim}}) = e$ and $p_i = \text{pr}(R_i^Y = 1 \mid x_i^{\text{mim}}, Z_i)$ as the propensity score and the probability of having an observed outcome given x_i^{mim}. Algorithm 1 below states the procedure for constructing the recommended estimators, denoted $\hat{\tau}_{\text{reg}}(x_i^{\text{mim}})$ and $\hat{\tau}_{\text{ps}}(x_i^{\text{mim}})$, as variants of $\hat{\tau}_{\text{reg}}$ and $\hat{\tau}_{\text{ps}}$ after replacing x_i with x_i^{mim} as the new fully observed covariate vector; cf. Table 1.

Algorithm 1. Construction of the recommended estimators with missing outcomes and covariates.

1. Construct $x_i^{\text{mim}} = (x_i, w_i^0, R_i^w)$ as the new fully observed covariate vector.
2. Estimate p_i from the prespecified outcome missingness model, denoted by $\hat{p}_i$. When the outcome missingness model is unknown, compute $\hat{p}_i$ from the logistic regression $\texttt{glm}(R_i^Y \sim 1 + x_i^{\text{mim}} + Z_i + x_i^{\text{mim}} Z_i)$ over $i = 1, \ldots, N$.
3. When the linear outcome model or missing-completely-at-random assumption is plausible, compute $\hat{\tau}_{\text{reg}}(x_i^{\text{mim}})$ as the coefficient of Z_i from the weighted-least-squares fit of the interacted regression $\texttt{lm}\{Y_i \sim 1 + Z_i + (x_i^{\text{mim}} - \bar{x}^{\text{mim}}) + Z_i(x_i^{\text{mim}} - \bar{x}^{\text{mim}})\}$ over $\{i\colon R_i^Y = 1\}$, where we weight unit i by $\hat{p}_i^{-1}$.
4. Otherwise, estimate e_i from the logistic regression $\texttt{glm}(Z_i \sim 1 + x_i^{\text{mim}})$ over $i = 1, \ldots, N$, denoted by $\hat{e}_i$, and compute $\hat{\tau}_{\text{ps}}(x_i^{\text{mim}})$ as the coefficient of Z_i from the weighted-least-squares fit of the unadjusted regression $\texttt{lm}(Y_i \sim 1 + Z_i)$ over $\{i\colon R_i^Y = 1\}$, where we weight unit i by $\hat{p}_i^{-1}\{Z_i/\hat{e}_i + (1 - Z_i)/(1 - \hat{e}_i)\}$.

Despite the apparent oversimplification by imputing all missing covariates with zero in forming x_i^{mim}, the resulting $\hat{\tau}_{\text{reg}}(x_i^{\text{mim}})$ and $\hat{\tau}_{\text{ps}}(x_i^{\text{mim}})$ are invariant over a general class of imputation schemes. Specifically, consider a covariatewise imputation strategy where, for $k = 1, \ldots, K$, we impute all missing values in the kth partially observed covariate by a common value $c_k \in \mathbb{R}$ (Zhao & Ding, 2024). Let $c = (c_1, \ldots, c_K)$ represent the imputation scheme. The resulting imputed variant of w_i equals $w_i^c = (w_{i1}^c, \ldots, w_{iK}^c)^{\text{T}} \in \mathbb{R}^K$ with $w_{ik}^c = w_{ik}$ if w_{ik} is observed and $w_{ik}^c = c_k$ otherwise. This defines a general class of imputed variants of w_i that includes w_i^0 as a special case with $c_k = 0$ for all k. Another common choice of c_k is the average of the observed values in $(w_{ik})_{i=1}^N$. Let $x_i^{\text{mim}}(c) = (x_i, w_i^c, R_i^w) \in \mathbb{R}^{J+2K}$ denote a variant of x_i^{mim} where we use the more general w_i^c in place of w_i^0.

Proposition 1. *Assume that in Algorithm 1 we replace all x_i^{mim} by $x_i^{\text{mim}}(c)$. The resulting estimators are invariant to the choice of the imputed values and equal $\hat{\tau}_{\text{reg}}(x_i^{\text{mim}})$ and $\hat{\tau}_{\text{ps}}(x_i^{\text{mim}})$ for all $c \in \mathbb{R}^K$.*

3.2. *Asymptotic justification of the recommended estimators in Algorithm 1*

Theorem 2 below states the asymptotic properties of $\hat{\tau}_{\text{reg}}(x_i^{\text{mim}})$ and $\hat{\tau}_{\text{ps}}(x_i^{\text{mim}})$ in Algorithm 1. For comparison, let x_i^{sub} be a subvector of x_i^{mim}, and let $\hat{\tau}_{\text{ps}}(x_i^{\text{sub}})$ denote a variant of $\hat{\tau}_{\text{ps}}(x_i^{\text{mim}})$ with x_i^{mim} replaced by x_i^{sub} in step 3 of Algorithm 1. A common choice of x_i^{sub} is $x_i^{\text{sub}} = x_i$, where we use only fully observed covariates in constructing $\hat{e}_i$.

Theorem 2. *Assume complete randomization with $Z_i \perp\!\!\!\perp \{Y_i(1), Y_i(0), x_i, w_i, R_i^w\}$ and that Assumption 1 holds with all x_i replaced by x_i^{mim}. Under standard regularity conditions, as $N \to \infty$,*

(i) *Theorem 1 holds with $(\hat{\tau}_{\text{reg}}, \hat{\tau}_{\text{ps}})$ replaced by $\{\hat{\tau}_{\text{reg}}(x_i^{\text{mim}}), \hat{\tau}_{\text{ps}}(x_i^{\text{mim}})\}$ and $\hat{\tau}_{\text{unadj}}$ renewed based on the renewed definition of $\hat{p}_i$ from step 2 of Algorithm 1;*
(ii) *the asymptotic variance of $\hat{\tau}_{\text{ps}}(x_i^{\text{sub}})$ is greater than or equal to that of $\hat{\tau}_{\text{ps}}(x_i^{\text{mim}})$.*

Other than being independent of Z_i, Theorem 2 does not require further assumptions on the missingness mechanism of w_i. Therefore, Theorem 2 holds even if w_i is missing not at random, which departs from the standard literature of missing covariates under the missing-at-random assumption (Robins et al., 1994).

In addition, the independence between R_i^w and Z_i ensures that $x_i^{\text{mim}} = (x_i, w_i^0, R_i^w)$ is effectively a fully observed pretreatment covariate vector, generalizing x_i. Theorem 2 requires that Assumption 1

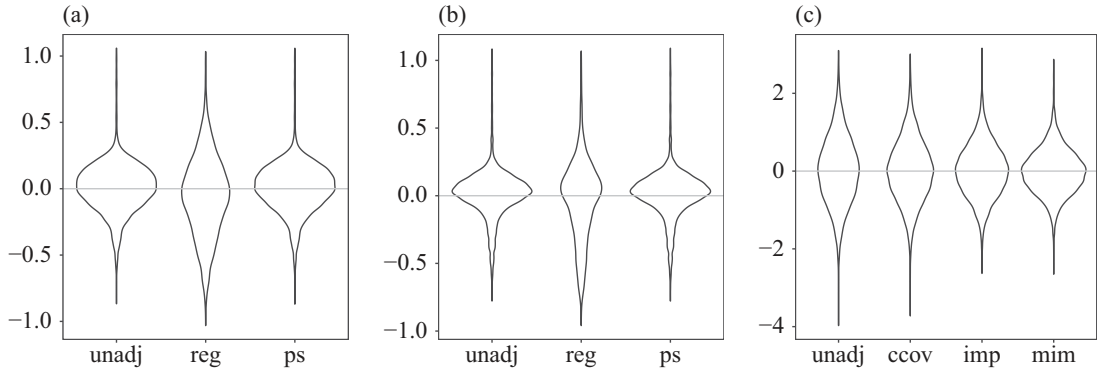

Fig. 1. Violin plots over 10^4 independent replications. (a),(b) Deviations of $\{\hat{\tau}_{\text{unadj}}, \hat{\tau}_{\text{reg}}, \hat{\tau}_{\text{ps}}\}$ from τ with (a) $e = 0.2$ and (b) $e = 0.5$. (c) Deviations of $\hat{\tau}_{\text{unadj}}$ (unadj), $\hat{\tau}_{\text{ps}}(x_i)$ (ccov), $\hat{\tau}_{\text{ps}}(x_i^{\text{imp}})$ (imp) and $\hat{\tau}_{\text{ps}}(x_i^{\text{mim}})$ (mim) from τ.

holds with all x_i replaced by x_i^{mim}, and thereby ensures that the outcome is missing at random with $R_i^Y \perp\!\!\!\perp Y_i \mid (x_i^{\text{mim}}, Z_i)$. This is the weakest form of the missing-at-random assumption based on the observed information in (x_i, w_i, Z_i) (Rosenbaum & Rubin, 1984). An alternative, more standard form of the missing-at-random assumption is $R_i^Y \perp\!\!\!\perp Y_i \mid (x_i, Z_i)$. This is a more restrictive condition because it does not allow the missingness in outcomes to depend on the missingness pattern of the covariates, as represented by R_i^w, or the observed values in w_i.

Theorem 2(i) ensures that $\hat{\tau}_{\text{ps}}(x_i^{\text{mim}})$ is asymptotically more efficient than $\hat{\tau}_{\text{unadj}}$, while $\hat{\tau}_{\text{reg}}(x_i^{\text{mim}})$ is asymptotically more efficient than both $\hat{\tau}_{\text{unadj}}$ and $\hat{\tau}_{\text{ps}}(x_i^{\text{mim}})$ when the outcomes are either linear in covariates or missing completely at random. Theorem 2(ii) ensures that $\hat{\tau}_{\text{ps}}(x_i^{\text{mim}})$ is asymptotically more efficient than all alternative propensity-score-weighted estimators that use only a subset of x_i^{mim} for estimating the propensity score, including $\hat{\tau}_{\text{ps}}(x_i)$ that uses only x_i. Recall that w_i^0 is a variant of w_i by imputing all missing covariates by zero. From Theorem 2, this rather basic imputation guarantees efficiency gain irrespective of the true values of the missing covariates, illustrating the quick wins that can be achieved by adjusting for partially observed covariates.

As a comparison, the approach reviewed by Seaman & White (2013) only applies inverse probability weighting to units with fully observed outcomes and covariates, and requires correct specification of both the outcome and covariate missingness models. Accordingly, it requires the covariates to be missing at random. Our proposed method, in contrast, places no restriction on the covariate missingness mechanism and does not require the specification of the covariate missingness model.

4. Simulation and a real-data example

4.1. *Regression adjustment does not guarantee efficiency gain*

We first illustrate the possibly worse precision of $\hat{\tau}_{\text{reg}}$ in finite samples. Assume missingness in only outcomes. We generate $\{x_i, Y_i(1), Y_i(0), Z_i, R_i^Y\}_{i=1}^N$ as independent realizations of $x_i \sim \text{Un}(-10, 10)$, $Y_i(1) = \sin(x_i)$, $Y_i(0) = -\cos(x_i)$, $Z_i \sim \text{Ber}(e)$ and $R_i^Y \sim \text{Ber}(p_i)$, where $p_i = \{1 + \exp(-1 - 2x_i)\}^{-1}$. Panels (a) and (b) of Fig. 1 show the distributions of the deviations of $\{\hat{\tau}_{\text{unadj}}, \hat{\tau}_{\text{reg}}, \hat{\tau}_{\text{ps}}\}$ from τ over 10 000 independent replications at $N = 1000$ and $e = 0.2, 0.5$. The regression adjusted $\hat{\tau}_{\text{reg}}$ has worse precision than $\hat{\tau}_{\text{unadj}}$ in both cases. Similar patterns are observed for other choices of potential outcomes and combinations of N and e; we omit the results to avoid repetition.

4.2. *Efficiency gain by propensity score weighting*

We next illustrate the efficiency gain by propensity score weighting, along with the benefits of adjusting for partially observed covariates. Consider a treatment-control experiment with $N = 500$ units and treatment probability $e = 0.2$. Let $(\xi_i)_{i=1}^N$ be independent Ber(0.4) to divide the units into two latent classes. Assume that we have $J = 1$ fully observed covariates $x_i \sim \mathcal{N}(\xi_i, 1)$ and $K = 9$ partially observed covariates $w_i = (w_{i1}, \ldots, w_{iK})^{\text{T}} \sim \mathcal{N}(\xi_i 1_K, I_K)$. We generate the missingness indicators

Table 2. *Point estimates and estimated variances of the unadjusted, regression-adjusted and propensity-score-weighted estimators based on x_i and x_i^{mim}, respectively. The variances are estimated by the bootstrap over 10^4 independent replications*

	$\hat{\tau}_{\text{unadj}}(x_i^{\text{mim}})$	$\hat{\tau}_{\text{reg}}(x_i^{\text{mim}})$	$\hat{\tau}_{\text{ps}}(x_i^{\text{mim}})$	$\hat{\tau}_{\text{unadj}}(x_i)$	$\hat{\tau}_{\text{reg}}(x_i)$	$\hat{\tau}_{\text{ps}}(x_i)$
Point estimate	−7.03	−6.07	−5.47	−4.78	−4.66	−4.60
Estimated variance	6.26	7.50	4.72	5.69	5.85	5.63

and potential outcomes as $R_{ik}^w \sim \text{Ber}\{0.5\xi_i + 0.95(1-\xi_i)\}$ for $k = 1, \ldots, K$, $R_i^Y \sim \text{Ber}(p_i)$, where $p_i = \{1+\exp(-1-2x_i)\}^{-1}$, and $Y_i(z) \sim \mathcal{N}\{\mu_i(z), 1\}$, where $\mu_i(z) = 3\xi_i + (x_i + \sum_{k=1}^K w_{ik})\gamma_{z|\xi_i} + 3\sum_{k=1}^K R_{ik}^w$ with $(\gamma_{1|1}, \gamma_{0|1}) = (1, -1)$ and $(\gamma_{1|0}, \gamma_{0|0}) = (0.5, -0.5)$. The data-generating process ensures that the covariates are missing not at random with units with $\xi_i = 1$ having both a higher chance of missing covariates and on average greater values of covariates.

Let $\hat{\tau}_{\text{ps}}(x_i)$ denote a variant of $\hat{\tau}_{\text{ps}}(x_i^{\text{mim}})$ where we use only the fully observed x_i in constructing $\hat{e}_i$. Let $\hat{\tau}_{\text{ps}}(x_i^{\text{imp}})$ denote a variant of $\hat{\tau}_{\text{ps}}(x_i^{\text{mim}})$ where we use only the union of x_i and w_i^0 in constructing $\hat{e}_i$. Figure 1(c) shows the distributions of the deviations of $\hat{\tau}_{\text{unadj}}$, $\hat{\tau}_{\text{ps}}(x_i)$, $\hat{\tau}_{\text{ps}}(x_i^{\text{imp}})$ and $\hat{\tau}_{\text{ps}}(x_i^{\text{mim}})$ from τ over 10 000 independent replications. The results are coherent with the asymptotic theory in Theorem 2, with $\hat{\tau}_{\text{ps}}(x_i^{\text{mim}})$ being the most precise.

4.3. *A real-data example*

We now apply the proposed method to the Best Apnea Interventions for Research trial of Bakker et al. (2016). A total of 169 patients were recruited and randomized with equal probability to active treatment and control. One outcome of interest is the 24-h systolic blood pressure measured at six months, which is missing for 45 patients.

As an illustration, we consider four baseline covariates, namely, age, gender, baseline apnea-hypopnea index and baseline diastolic blood pressure, for estimating the outcome missingness model and covariate adjustment. The first three covariates are fully observed, and the last covariate is missing for eight patients. Table 2 summarizes the point estimates and estimated variances of the unadjusted, regression-adjusted and propensity-score-weighted estimators based on the fully observed covariates and the augmented covariates under the missingness indicator method, respectively. The variances are estimated by the bootstrap over 10 000 independent replications. The results are coherent with the asymptotic theory, with the combination of propensity score weighting and missingness indicator method ($\hat{\tau}_{\text{ps}}(x_i^{\text{mim}})$) resulting in the smallest bootstrap variance. The two regression-adjusted estimators $\hat{\tau}_{\text{reg}}(x_i^{\text{mim}})$ and $\hat{\tau}_{\text{reg}}(x_i)$, on the other hand, have higher bootstrap variances than their respective unadjusted counterparts $\hat{\tau}_{\text{unadj}}(x_i^{\text{mim}})$ and $\hat{\tau}_{\text{unadj}}(x_i)$, illustrating the possible loss in precision by regression adjustment.

5. Further discussion on the role of the outcome model

Theorems 1 and 2 assume that the missingness model for the outcome is correctly specified. When this assumption fails, $\hat{\tau}_{\text{ps}}$ is inconsistent, while $\hat{\tau}_{\text{reg}}$ remains consistent if the linear outcome model is correct. The use of the outcome model ensures this double robustness property of $\hat{\tau}_{\text{reg}}$; see Proposition S2 in the Supplementary Material. Analogously, we can also augment $\hat{\tau}_{\text{ps}}$ with the outcome model:

$$\hat{\tau}_{\text{ps-reg}} = \frac{1}{N}\sum_{i=1}^N \left[\hat{m}_1(x_i) + \frac{R_i^Y}{\hat{p}_i}\frac{Z_i}{\hat{e}_i}\{Y_i - \hat{m}_1(x_i)\}\right] - \frac{1}{N}\sum_{i=1}^N \left[\hat{m}_0(x_i) + \frac{R_i^Y}{\hat{p}_i}\frac{1-Z_i}{1-\hat{e}_i}\{Y_i - \hat{m}_0(x_i)\}\right]$$

with $\hat{m}_z(x_i)$, $z = 0, 1$, the estimated outcome model. The augmented estimator $\hat{\tau}_{\text{ps-reg}}$ is doubly robust in that it is consistent if either the outcome model or the outcome missingness model is correct.

As a special case, we can construct $\hat{\tau}_{\text{ps-reg}}$ as the coefficient of Z_i from the weighted-least-squares fit of the interacted regression over $\{i\colon R_i^Y = 1\}$ with weight $\hat{p}_i^{-1}\hat{\pi}_i$ for unit i. The corresponding $\hat{m}_z(x_i)$, $z = 0, 1$, equals the estimated outcome model from the same weighted-least-squares fit; see Proposition S3 in the Supplementary Material. This integrates the regression adjustment and the propensity score weighting in the last two rows of Table 1.

Acknowledgement

Ding thanks the U.S. National Science Foundation. Li thanks the Patient-Centered Outcomes Research Institute. We thank Jerry Chang, Wang Rui, Sean O'Brien, Luke Miratrix and participants of Harvard Data Science Initiative causal seminars for stimulating discussions, the associate editor and two reviewers for constructive comments and the investigators of the Best Apnea Interventions for Research trial for providing the data.

Supplementary material

The Supplementary Material includes additional results and proofs.

References

Bakker, J. P., Wang, R., Weng, J., Aloia, M. S., Toth, C., Morrical, M. G., Gleason, K. J., Rueschman, M., Dorsey, C., Patel, S. R. et al. (2016). Motivational enhancement for increasing adherence to CPAP: a randomized controlled trial. *Chest* **150**, 337–45.

Chang, C.-R., Song, Y., Li, F. & Wang, R. (2023). Covariate adjustment in randomized clinical trials with missing covariate and outcome data. *Statist. Med.* **42**, 3919–35.

Fisher, R. A. (1935). *The Design of Experiments*, 1st ed. London: Oliver and Boyd.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66**, 315–31.

Hirano, K., Imbens, G. W. & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**, 1161–89.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique. *Ann. Appl. Statist.* **7**, 295–318.

Negi, A. (2024). Doubly weighted M-estimation for nonrandom assignment and missing outcomes. *J. Causal Infer.* **12**, 20230016.

Negi, A. & Wooldridge, J. M. (2021). Revisiting regression adjustment in experiments with heterogeneous treatment effects. *Economet. Rev.* **40**, 504–34.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* **89**, 846–66.

Robins, J., Sued, M., Lei-Gomez, Q. & Rotnitzky, A. (2007). Comment: performance of double-robust estimators when 'inverse probability' weights are highly variable. *Statist. Sci.* **22**, 544–59.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rosenbaum, P. R. & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Statist. Assoc.* **79**, 516–24.

Seaman, S. R. & White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statist. Meth. Med. Res.* **22**, 278–95.

Shen, C., Li, X. & Li, L. (2014). Inverse probability weighting for covariate adjustment in randomized studies. *Statist. Med.* **33**, 555–68.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

Tsiatis, A. A., Davidian, M., Zhang, M. & Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statist. Med.* **27**, 4658–77.

Williamson, E. J., Forbes, A. & White, I. R. (2014). Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statist. Med.* **33**, 721–37.

Zeng, S., Li, F., Wang, R. & Li, F. (2021). Propensity score weighting for covariate adjustment in randomized clinical trials. *Statist. Med.* **40**, 842–58.

Zhao, A. & Ding, P. (2024). To adjust or not to adjust? Estimating the average treatment effect in randomized experiments with missing covariates. *J. Am. Statist. Assoc.* **119**, 450–60.

[*Received on* 17 *November* 2023. *Editorial decision on* 27 *February* 2024]