HHS Public Access

Author manuscript

Stat Methods Med Res. Author manuscript; available in PMC 2016 August 24.

Published in final edited form as:

Stat Methods Med Res. 2017 June ; 26(3): 1199–1215. doi:10.1177/0962280215570722.

Estimation of causal effects of binary treatments in unconfounded studies with one continuous covariate

R Gutman[1] and **DB Rubin**[2]

[1]Department of Biostatistics, Brown University, Providence, RI, USA

[2]Department of Statistics, Harvard University, Cambridge, MA, USA

Abstract

The estimation of causal effects in nonrandomized studies should comprise two distinct phases: design, with no outcome data available; and analysis of the outcome data according to a specified protocol. Here, we review and compare point and interval estimates of common statistical procedures for estimating causal effects (i.e. matching, subclassification, weighting, and model-based adjustment) with a scalar continuous covariate and a scalar continuous outcome. We show, using an extensive simulation, that some highly advocated methods have poor operating characteristics. In many conditions, matching for the point estimate combined with within-group matching for sampling variance estimation, with or without covariance adjustment, appears to be the most efficient valid method of those evaluated. These results provide new conclusions and advice regarding the merits of currently used procedures.

Keywords

Causal inference; matching; regression adjustment; Rubin causal model; spline

1 Introduction

The causal effect of a binary treatment W on outcome Y for unit i ($i = 1, \ldots, N$) is the comparison of two "potential" outcomes, $Y_i(1)$ and $Y_i(0)$, corresponding to the two possible levels of W: $W_i = 1$ indicates the receipt of the active level of the treatment, and $W_i = 0$ indicates the receipt of the control level. The adjective "potential" is used because only one value of Y can be realized and observed: the potential outcome corresponding to the action actually taken at that time for that unit. The other potential outcome cannot be observed because its corresponding action was not taken.[1,2] We assume SUTVA (the stable unit

Corresponding author: R Gutman, Brown University, 121 S. main St, Box G-S121-7, Providence, RI 02912, USA. Roee_Gutman@brown.edu.

Conflict of interest

None declared.

Supplemental material

The online supplemental (available at: http://smm.sagepub.com/) includes additional tables and figures for nonmonotone response surfaces, as well as additional description on the implementation of the different methods.

treatment value assumption[3,4]), so that this notation is functionally well defined. The observable outcome for unit i can be written as

$$Y_i^{obs}=W_iY_i(1)+(1-W_i)Y_i(0) \quad (1)$$

This perspective is commonly referred to as the "Rubin Causal Model"[5] for work done in the 1970s (e.g. refer the literature[1–3,6,7]), which generalized Neyman's[8] use of potential outcomes in randomized experiments with randomization-based inference to other situations and other forms of inference.

We cannot directly observe the causal effect for unit i, so instead we observe multiple units, some exposed to the active level of the treatment and others exposed to the control level. For drawing causal inferences, there are other variables that are unaffected by W_i =1 versus W_i =0: covariates, $\mathbf{X}_i = (X_{i1}, \ldots, X_{iL})$. A crucial piece of information that is needed for causal inference is the reason each unit received the treatment it actually received or the assignment mechanism (AM)

$$P(W_i=1|\mathbf{X}_i,\ Y_i(0),\ Y_i(1), \phi) \quad (2)$$

where ϕ is a vector parameter governing this distribution; throughout we assume the standard mathematical statistical situation with independent modeling across units, implied by the notation in equation (2).

Ideally, given the covariates $\mathbf{X}_i$, the AM does not depend on the potential outcomes, so that it is unconfounded[4]

$$P(W_i=1|\mathbf{X}_i, Y_i(0), Y_i(1), \phi)=P(W_i=1|\mathbf{X}_i, \phi) \equiv e(\mathbf{X}_i) \quad (3)$$

where $e(\mathbf{X}_i)$ is the propensity score for unit i,[9] whose dependence on ϕ is notationally suppressed in $e(\mathbf{X}_i)$. We assume equation (3), thereby ensuring that comparing the observed outcomes for treated units and control units at $\mathbf{X}_i = x$ yields a valid estimate of the treatment effect at x. Here, we focus on the situation where both Y_i and X_i are scalar and continuous, so the propensity score is a scalar function of X_i, generally a many-to-one function.

Causal inference should be composed of two main phases: the design phase and the analysis phase. The design phase includes contemplating, collecting, organizing, and analyzing data without seeing any outcome data.[1] In randomized experiments, the design phase includes defining the randomization scheme, for example considering blocking on covariates that may influence Y, as well as specifying a protocol for the analysis of outcome data. In observational studies without randomization, methods such as subclassification[10] and matching on covariates or functions of them[9,11–13] have been proposed to help approximate hypothetical randomized experiments, and thus these activities are part of the design phase for observational studies. The design phase in observational studies has received more

attention recently than earlier (e.g. see literature[9,12]) but not dramatically so until recently.[2,14,15]

The statistical literature proposes many procedures for estimating treatment effects in unconfounded studies. However, published comparisons between these procedures have been limited to a small number of them at a time, limited simulation settings, or reliance on meta-analysis with mixed conclusions. More specifically, Rubin[12] and Rubin[16] compared matching, simple linear regression adjustment, and their combination and showed that in terms of bias reduction, the combination generally works best. On the other hand, Shah et al.[17] used meta-analysis to conclude that propensity score analysis generally lead to similar results as simple linear regression. Lunceford and Davidian[18] compared subclassification and inverse probability weighting (IPW), as well as their combination with regression adjustments, and concluded that weighting methods offer approximately unbiased inference for practical sample sizes, and that combining them with regression adjustments resulted in greater precision. In addition, they noticed that, in some cases, subclassification with regression adjustment resulted in more precision than a combination of weighting with regression adjustments. However, because subclassification with regression adjustment does not enjoy the "doubly robust" (DR) property,[19] Lunceford and Davidian[18] concluded that weighting with regression adjustment is preferable. Austin[20] compared subclassification, matching, IPW, and regression adjustment on the propensity score and concluded that matching and weighting had larger reductions in bias than subclassification with regression adjustment. Waernbaum[21] conducted a simulation and concluded that when the propensity score model and the outcome model are both misspecified, weighting with regression adjustment has larger bias and mean square error (MSE) than matching alone. Recently, Austin[22] examined different matching procedures to estimate the average effect of the treatment on the treated (ATT) and concluded that nearest neighbor caliper matching without replacement is the most optimal method for forming pairs of treated and untreated units. These inconsistent conclusions do not generate cogent advice.

This paper attempts, first, to identify promising methods with a scalar covariate that should be investigated in future studies with multiple covariates, and second, to identify methods that should not be considered further. Although, methods that perform poorly with a single covariate will generally perform poorly with multiple covariates, the best performing method with a scalar covariate can have worse performance with multiple covariates than methods that are marginally worse with a scalar covariate. Thus, it is important to identify a group of methods that perform well with a single covariate, rather than select a single best performing method.

The simulation-based comparisons use Neyman's framework of frequentist operating characteristics. An α-level interval estimate is "valid" if, under repeated sampling from the population (finite or super), the interval covers the estimand in at least α percent of the samples. Among valid procedures, one is more powerful (efficient) than another if it produces shorter intervals.[23] In addition to validity and efficiency of procedures, we compare the biases and MSEs of point estimators. Of all methods considered, and across most of the distributional conditions examined, when the distributions of the covariate in the control and treatment groups have reasonable overlap, matching with replacement for point

estimation, with or without covariance adjustment, combined with within-group matching for sampling variance estimation,[3] generally appear to be superior. Moreover, commonly used and accepted procedures such as model-based adjustments, subclassification, and matching with standard sampling variance estimation result in very poor operating characteristics when the treatment effect is nonnull. When the groups are far apart, all methods rely on unassailable assumptions and so all are ineffective in practice.

2 Common procedures for estimating treatment effects

We begin by reviewing previously suggested procedures that attempt to estimate the average difference between $Y_i(1)$ and $Y_i(0)$, also known as the super-population average treatment effect (ATE), Δ, which is the estimand of most commonly used procedures.

2.1 Linear regression

A common approach for adjusting for covariate imbalances in two treatment groups uses linear regression

$$E(Y_i(W)|X_i,\beta_0,\beta_X,\gamma_W)=\beta_0+X_i\beta_X+\gamma_W W \quad (4)$$

where the coefficient of W, γ_W, is regarded as the super-population treatment effect (to list only a few in the literature[24–26]); this is also known as "covariance adjustment" and originates with Fisher[27] (Section 49.1). Rubin[16] showed that for scalar X and Y, using equation (4) can be badly biased for Δ if the distributions of X in the treatment and control groups differ and the two response surfaces for Y given X are monotone but not linear (also see Cochran and Rubin[28]). Estimator (4) is approximately unbiased essentially only when the two response surfaces of Y given X are nearly linear and parallel, which is unknowable in practice, especially at the design stage. Moreover, usually the associated interval estimates are invalid due to their being badly miscentered. Because this method has been known for decades to be generally inapposite, it will not be investigated further here.

2.2 Polynomial regression, spline, and penalized spline

A procedure that attempts to address the likely misspecification of the linear model in equation (4) is regression adjustment with a nonlinear function of X_i,[29,30] with $\beta_0+X_i\beta_X$ in equation (4) replaced with a nonlinear function

$$E(Y_i(W)|X_i,\beta,\gamma_W)=h(X_i|\beta)+\gamma_W W \quad (5)$$

When X is scalar, h could be a Pth order polynomial function $h(X|\beta)=\beta_0+\beta_1X+\cdots+\beta_PX^P$. When P is small, the regression can be too inflexible to capture important features of the true $h(X|\beta)$, and when P is large, the model fitting can fail due to high multicollinearity.[31] A possible compromise is the regression spline,[32] where the polynomial terms are replaced by polynomial pieces, commonly forced to join smoothly at a sequence of "knots." A popular choice is the piecewise cubic spline that is constrained to be continuous and twice differentiable. Three options need to be specified for the spline: the basis of the

polynomials, the number of knots, and the location of knots. Due to its numerical stability, a commonly used basis is the "B-Spline."[33] Stone[34] found that more than five knots are seldom required in practice to approximate one response surface.

A simple procedure for defining the location of the knots is to set them at the quantiles of X.[35] However, the number of knots and their locations can have major influences on results[36] (Section 9.3). One standard way to address the knot-placement problem is to use smoothing splines, penalized splines, or "thin-plate splines."[37,38] With these alternatives, a relatively large number of knots are used, but excessively nonmonotone fitted models are avoided by applying "wiggliness" penalties.

The implicit assumption for the procedures described in this subsection is that the treatment effect is the same at every value of X (parallel response surfaces), which can lead to badly biased estimation of Δ when there are different distributions of X in the treatment and control groups, as documented since the early 1970s in Cochran and Rubin[30] and Rubin.[16]

2.3 Design-based matching or subclassification

Matching methods attempt to reduce the bias arising from the different X distributions by "balancing" the distributions of X in the treatment and control groups. Unlike the methods discussed in Sections 2.1 and 2.2, these are distinctly design-phase methods because they do not involve outcome data. Rubin[11] displayed the effectiveness of mean matching and "nearest neighbor" pair matching methods for bias reduction with scalar continuous X and Y, when estimating the ATT. An extension to pair matching is k:1 nearest neighbor matching,[15] where k control units are matched to each treatment unit; this extension discards fewer control units than pair matching, and thus can result in increased statistical efficiency, but this increase can be minimal, and there is also a chance of increased bias due to poor matches.[39,40]

After the matching is performed, the treatment effect can be estimated from each matched pair (or group) and averaged over the entire matched sample.[11,12,41] The standard error of the estimate of the ATE, $\hat{\Delta}$, is commonly obtained using the randomization-based sampling variance estimate (e.g. Austin[42]), but Abadie and Imbens[3] showed that this is typically an underestimate, because it ignores the variability induced by the random sampling from the super-population as well as the variability in the matching procedure. Abadie and Imbens[3] developed a sampling variance estimate that is consistent under certain conditions and matches units with similar X within each treatment group to estimate the variability of the unit level effects. This variance estimator has been implemented in statistical software[43,44] and has been extended to other matching-based point estimands.[45]

Subclassification methods partition all $n \leq N$ sampled units into subclasses with "similar" values of X_i. Cochran[10] showed that using only six subclasses can typically result in more than 90% reduction of the initial bias in scalar continuous X. The population ATE is obtained by estimating the average effect in each subclass and averaging across subclasses.[28,41]

A more sophisticated form of subclassification is "full matching,"[46] which constructs subclasses so that each subclass has exactly one observation from either the treatment group or the control group, and at least one from the other group. Let S_k be the set of units that are in subclass $k \in \{1,\ldots, K\}$; Hansen[47] assumed that the unit-level super-population expectations and variances are

$$E(Y_i(W)|\beta_k, \beta_{Wk}, i \in S_k) = \beta_k + \beta_{Wk} \times W$$
$$Var(Y_i(W)|i \in S_k) = \sigma^2 < \infty \quad (6)$$

Based on model (6), Δ can be estimated by $\sum_{k=1}^{K} \frac{|S_k|}{n} \hat{\beta}_{Wk}$, where $|S_k|$ is the cardinality of subclass k, and $\hat{\beta}_{Wk}$ is the estimate of β_{Wk}. Letting $\hat{V}_k$ be the estimated sampling variance of $\hat{\beta}_{Wk}$, the sampling variance of $\hat{\Delta}$ can be estimated by $\sum_{k=1}^{K} \left(\frac{|S_k|}{n}\right)^2 \hat{V}_k$.

2.4 Weighting

The covariates X can also be used to generate weights that are based on the inverse propensity score.[18,48] Under unconfoundedness

$$E\left(\frac{WY_i(W)}{e(X_i)}\right) = E\left\{E\left(\frac{WY_i(W)}{e(X_i)}|X_i, Y_i(W)\right)\right\} = E\left\{\frac{Y_i(W)}{e(X_i)}E(W|X_i, Y_i(W))\right\} = E(Y_i(W)) \quad (7)$$

So, an unbiased estimate of Δ is

$$\frac{1}{n}\sum_{i=1}^{n}\frac{W_i Y_i^{obs}}{e(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-W_i)Y_i^{obs}}{1-e(X_i)}$$

In practice, $e(X_i)$ is generally unknown and it is estimated from the data. Replacing $e(X_i)$ by its estimated value, $\hat{e}(X_i)$ divided by $\sum_i \frac{W_i}{\hat{e}(X_i)}$ can be more efficient than using $e(x_i)$.[49,50] A potential drawback of weighting occurs when some $e(X_i)$ (or $\hat{e}(X_i)$) are close to 0 or 1, so that a few observations dominate the estimated treatment effect resulting in large sampling variance. Moreover, when the estimated probabilities are misspecified, this method can suffer from large bias as well as from large true and estimated sampling variances[21,41,51]; Lunceford and Davidian[18] referred to equation (7) based on $\hat{e}(X_i)$ as IPW_1 and proposed additional weighting estimators IPW_2 and IPW_3 for such cases. They also derived the standard errors for all of these estimators.

2.5 Combined methods

Most methods in Section 2.3 do not perform any adjustments beyond grouping units that are similar in terms of the covariate. Rubin[16] observed that, for continuous scalar X and Y, combining linear regression (as in Section 2.1) and matching (as in Section 2.3) achieved, under most conditions studied, larger reductions in bias than either method alone. Furthermore, it was shown analytically that this combination asymptotically adjusts for possible bias in the commonly used ATE estimators.[52] Four procedures that are generated from the combination of either with covariance (C) adjustments or without (no adjustment-N) for point estimation, and either of these with standard (s) or the within-treatment-group matching (m) for sampling variance estimator, will be examined here. We label each of the four matching procedures as M-C-s, M-N-s, M-C-m, and M-N-m where the first letter represents across treatment group matching for point estimation, the second letter represents whether or not covariance adjustment was applied to create that point estimate, and the third letter represents whether or not within-group matching was performed for sampling variance estimation.

Regression can also be combined with subclassification to create a three-step procedure. First, partition units into K subclasses based on scalar X as in Section 2.3. Second, in each subclass, regress Y on a constant, X, and W as in Section 2.1. Define the estimated treatment effect in subclass k, γ_k, to be the estimated coefficient for W, $\hat{\gamma}_k$, and estimate its sampling variance, V_k. For example, when the estimand of interest is the population ATE, the treatment effect in the kth subclass, γ_k, is implicitly defined by

$$E(Y_i|X_i,\beta_{0k},\beta_{Xk},\gamma_k,k=1,\ldots,K)=\beta_{0k}+\gamma_k W_i+X_i\beta_{Xk} \quad (8)$$

The estimated sampling variance of $\hat{\gamma}_k$ can be obtained by the standard asymptotic approximation. The third step in the procedure combines across subclasses the estimated treatment effects and their estimated sampling variances, to estimate Δ, and its sampling variance, V

$$\hat{\Delta}\equiv\sum_{k=1}^{K}\frac{|S_k|}{n}\hat{\gamma}_k,\quad \hat{V}=\sum_{k=1}^{K}\left(\frac{|S_k|}{n}\right)^2\hat{V}_k$$

The intuition behind this procedure is that, within a subclass, the distributions of the covariates in the treatment and control groups are similar, and the regression estimate is not used to extrapolate out of the subclass.[41,53] This procedure is referred to as regression adjusted subclassification (RAS).

Although RAS is a useful method in practice, due to its simplicity and familiarity, the method has at least two limitations: (1) each subclass is modeled independently, which does not take into account that observations in adjacent subclasses are related, and using this fact could improve the final estimator; (2) it assumes a constant treatment effect in each subclass, which can lead to biased estimates when the response surfaces are not parallel within subclasses.[53]

Robins et al.[19] suggested combining weighting with regression adjustments to obtain DR estimators. The DR estimator is a modification of the IPW estimator planned to reduce the sensitivity to model misspecification and improve precision. The potential attractive feature of a DR estimator is that it is consistent and asymptotically normal when either the propensity score model or the regression model is correctly specified. In addition, these estimators achieve the lower bound for the sampling variance of semiparametric estimators.[18,19] Specifically, the DR estimate for Δ is

$$\hat{\Delta}_{DR} \equiv \frac{1}{n}\sum_{i=1}^{n}\frac{W_i Y_i^{obs} - (W_i - e(X_i))m_1(X_i, \hat{\beta}_1)}{e(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-W_i)Y_i^{obs} + (W_i - e(X_i))m_0(X_i, \hat{\beta}_0)}{1-e(X_i)} \quad (9)$$

where $m_w(X, \beta_w) = E(Y(W) | W = w, X) = \beta_{w0} + \beta_{w1}X$ is the regression of the response on X in treatment group W, depending on $\beta = \{\beta_{w0}, \beta_{w1}\}$, and $\hat{\beta}_W$ is an estimate for β based on subjects in group $W = w$. The standard error of $\hat{\Delta}_{DR}$ is equal to $\frac{1}{n^2}\sum_{i=1}^{n} I_i^2$, where

$$I_i = \frac{W_i Y_i^{obs} - (W_i - e(X_i))m_1(X_i, \hat{\beta}_1)}{e(X_i)} - \frac{(1-W_i)Y_i^{obs} + (W_i - e(X_i))m_0(X_i, \hat{\beta}_0)}{1-e(X_i)} - \hat{\Delta}_{DR} \quad (10)$$

As with weighting estimators, $e(X_i)$ is generally unknown and is replaced by an estimate of it, $\hat{e}(X_i)$.

3 Simulation design

Table 1 summarizes the 14 procedures described in Section 2; their exact implementation is provided in the online supplementary material. In addition to the factor defined by the 14 procedures, the factors that are evaluated in this simulation comprise two types. The first type describes the scalar covariate (X_i) distributions and sample sizes, both of which are either known to the investigator, or can be easily estimated without examining any outcome data. The second type of factors involves the response surfaces, which are neither known to the investigator, nor be empirically estimated at the design stage.

3.1 Factors known or estimable in the design phase

The values of the covariate for the n treated and $r \times n$ control units are generated from two different, possibly skewed, normal distributions[54]

$$X_i|W_i{=}1 \overset{\text{i.i.d.}}{\sim} Skewed-Normal(\mu,\nu_1^2,\eta_1),\ \ i{=}1,\ldots,n$$
$$X_i|W_i{=}0 \overset{\text{i.i.d.}}{\sim} Skewed-Normal(0,1,\eta_0),\ \ i{=}n{+}1,\ldots,n{+}r\times n \quad (11)$$

We parameterized the distance between the treated group and the control group means in terms of the standardized bias, $SB{=}\frac{\mu}{\sqrt{\frac{1+\nu_1^2}{2}}}$ as in the literature.[11,12,16,28,55] Equation (11) implicitly defines the unconfounded AM, $P(W_i = 1|X_i, \phi)$, as a function of simulation conditions (r, n, SB, ν_1, η_1, η_0).

Table 2 describes the three levels of each of four of the factors SB, ν_1^2, η_0, and η_1. The factors r and n also vary in each of the 3^4 settings, yielding a $3^4 \times 2 \times 3$ factorial design. In addition, a nested computational structure was used in order to reduce variability across configuration comparisons, as in Rubin[12] and Cangul et al.[56] Although SB is commonly used to summarize the overlap between two distributions, it is insensitive to differences in variances or skewnesses between distributions. One measure that quantifies these differences is the Jensen–Shannon divergence (JSD)[57]

$$\lambda Q_{KL}(H_1|\lambda H_1{+}(1-\lambda)H_0){+}(1-\lambda)Q_{KL}(H_0|\lambda H_1{+}(1-\lambda)H_0) \quad (12)$$

where H_0 and H_1 are the distributions of X in the control and active treatment, respectively, Q_{KL} is the Kullback–Leibler divergence,[58] and $\lambda{=}\frac{1}{r+1}$. It is helpful to compare JSD values to SB values using two normal distributions with equal variance but different means. When $SB = \{0.25,0.5,1\}$, the JSDs are $\{0.008,0.03,0.11\}$, respectively. $SB = 1$ is considered large in practical applications.[28] Thus, throughout this simulation, we restrict the analysis to configurations for which the JSD is less than 0.3. This value was chosen because it is about three times larger than the value expected for $SB = 1$, and on average at least 75% of the units are within the range of the other treatment's distribution.

3.2 Factors empirically inestimable at the design stage

In each simulation replication, we randomly generate the continuous outcome data from

$$Y_i^{obs}{=}W_iG_1(X_i,\mathbf{B}_1){+}(1-W_i)G_0(X_i,\mathbf{B}_0) \quad (13)$$

where G_1 and G_0 are distributions unknown to the investigator, and $\mathbf{B}_1$ and $\mathbf{B}_0$ are parameters also unknown to the investigator. Specifically, G_1 and G_0 are Normal distribution with conditional mean $g_0(X)$ in the control group and conditional mean $\beta g_1(X) + \alpha$ in the treatment group, where β determines the correlation between X_i and $Y_i(1)$, and α is an additive effect. The values of the variances of the normal distributions, σ_W^2, are given in Table 2. Three g_W functions are used for monotone response surfaces: $\{\exp(X), \exp(-X), X\}$. These response surfaces were also used in Rubin[11,16] that examined the percent

reduction in bias for matching methods and matching methods combined with regression. These surfaces represent linear and moderately nonlinear surfaces, and depending on the overlap between the distributions of X in the active treatment and control groups, are either favorable or unfavorable for matching methods. For nonmonotone response surfaces, the three continuous, twice differentiable g_W functions are displayed in the online supplementary material; the primary differences between these functions are the slope at the point of inflection and that point's location. For nonmonotone response surfaces, we set $\beta = 1$ and $\alpha = 0$ for all configurations in order to limit simulation conditions.

Null treatment effects are generated by setting $g_1(X) = g_0(X)$; nonnull treatment effects are generated by allowing the response surfaces in the treatment group and the control group to differ. More specifically, for monotone response surfaces, we use the same previously described functions, but the potential outcomes, $Y_i(0)$, $Y_i(1)$ are generated using differing functions g_0, g_1, and in cases where $g_0 = g_1$, by setting $\alpha \neq 0$ or $\beta \neq 1$. These sets of configurations allow us to examine the performance of the different methods in settings with null effects, constant treatment effects, and heterogeneous treatment effects. For nonmonotone response surfaces, similar modifications are used, and different surfaces are simulated by different configurations of the shapes and points of inflection. There are 3^6 monotone response surface configurations and 3^5 nonmonotone response surface configurations. For monotone response surfaces, the median squared correlations (R^2) between Y and X over all configurations were 0.44 (range (0.03, 0.85)) and 0.42 (range (0.01, 0.96)), in the control and treated groups, respectively. For nonmonotone response surfaces, the median R^2s are 0.3 (range (0.00, 0.9)) in both the control and treatment groups. For each configuration of the factors, $N_{rep} = 100$ replications were produced.

All of the simulations were executed using R 2.15.0 software.[59] The full matching algorithm was implemented using the *optmatch* package,[60] and the different matching algorithms were implemented using the *Matching* package.[44]

4 Results

We compare the five classes of methods summarized in the rows of Table 1 for estimating Δ. Even with scalar X, weighting methods require estimation of the propensity score, which is done here using the algorithm described by Imai and Ratkovic.[61] For each procedure, at each factor's configuration, and at each of the 100 replications, we calculate the estimated treatment effect, the estimated sampling variance, the corresponding 95% interval width, and determine whether the interval covered or did not cover Δ. Then, we calculate for each procedure and each configuration, the mean coverage rate, the bias, the mean estimated sampling variance, and the mean interval length.

4.1 Results for 95% interval coverages

Table 3 displays the proportion of configurations in which the 14 different methods have intervals with over 90% coverage, as well as the median and interquartile range of the coverages when response surfaces are monotone. Overall, M-C-m and M-N-m have the highest coverages followed by IPW_1. The rest of the methods have fewer than 70% of the configurations with over 90% coverage.

When the treatment effect is null, all methods, except for IPW_3, M-C-s, and M-N-s, have median coverage similar to their nominal level. However, when the treatment effect is not null, only IPW_1, IPW_2, DR, full matching, M-N-m, and M-C-m have coverage close to or above 95%. Among these methods, the 25th percentile of coverage is the lowest for IPW_2 and the largest for M-N-m and M-C-m. M-N-m and M-C-m also have the interval coverages that are most concentrated around 95%. Similar results with slightly different numbers occur for nonmonotone response surfaces, where IPW_1 has the largest coverage with 25th coverage percentile of 1, which implies substantial overcoverage (see online supplementary material).

Because IPW_1, IPW_2, DR, full matching, M-N-m, and M-C-m have median coverages close to the nominal level across all configurations, we compared only them in subsequent evaluations. Figure 1 summarizes the median and the 25th percentile of the coverages as functions of the JSD for monotone response surfaces. The median coverages of M-N-m and M-C-m are above 95% for every value of the JSD. IPW_1, IPW_2, and DR have median coverages that are higher than 95% for lower JSD values, which decrease for larger JSD values. Full matching generally results in median coverages that are close to nominal. The median is a robust measure of central tendency that may require a large number of configurations to be lower than the nominal level to observe a significant difference, whereas the 25th percentile is a more sensitive measure for the lower part of the distribution. The 25th percentile plot shows that all methods have decreased coverage as the JSD increases, with M-N-m and M-C-m, having 25th percentiles that are close to 95%, even for large values of the JSD. IPW_1, IPW_2, and DR have the largest slope as JSD increases, with a significant drop for large JSD values. This drop occurs because some observations have weights that are very close to 0 or 1, resulting in biased point estimates and interval estimates that are too short. These phenomena are investigated further in the next subsection. Similar coverage trends are observed for nonmonotone response surfaces (see online supplementary material).

4.2 Results for biases, RMSEs, and intervals widths

Table 4 compares the median coverages, absolute biases, interval widths, and RMSEs for the generally statistically valid procedures identified in Section 4.1, for monotone response surfaces under the null and when the response surfaces are parallel. When JSD increases, as shown previously, IPW_1, IPW_2, and DR exhibit decreases in median coverages, increased coverages variability, and significant decreases in the percentage of configurations with above 90% coverage. M-N-m and M-C-m exhibit stable median coverages, stable coverage variability, and a moderate decrease in the percentage of configurations with above 90% coverage as JSD increases. Full matching exhibits stable median and IQR for coverages across JSD values, as well as stable percentages of configurations with above 90% coverage. When the distribution of X in the treatment and control groups differs substantially, none of the methods are statistically valid. To avoid these situations, it is important to examine the overlap of the distributions of X in the control and treated groups.

For all methods, the median and IQR for absolute bias, interval width, and RMSE increase with increasing JSD. Compared to M-N-m, full matching has generally similar biases, but

larger interval widths and RMSEs. IPW_1 and IPW_2 have larger and more variable absolute biases in comparison to M-N-m, M-C-m, and full matching for all JSD values. Among all of the methods, M-N-m and M-C-m have the shortest interval width for small JSD, with M-C-m having a slight advantage. For larger JSD values, DR has the shortest interval and RMSE. These values are deceiving because DR also has the lowest coverages, with many of the configurations resulting in substantially less than nominal coverage.

Table 5 compares the median coverages, absolute biases, interval widths, and RMSEs for monotone response surfaces when the response surfaces are not parallel. Compared to the other methods, M-N-m and M-C-m are generally valid and they have point estimates with the smallest biases. As with the parallel surfaces case, IPW_1 and IPW_2 have the largest median absolute biases and largest IQR, followed by DR. These results also appear for nonmonotone response surfaces, which support the conclusions of Kang and Schafer[54] and Waernbaum.[24]

For small JSD values, full matching has the shortest median interval width, followed by DR, IPW_2, M-N-m, M-C-m, and IPW_1. For large JSD values, DR has the shortest median width followed by IPW_2, IPW_1, full matching, M-N-m, and M-C-m. These values are deceiving because full matching, DR, IPW_1, and IPW_2 have the lowest coverages, with many of the configurations resulting in substantially less than nominal coverage. Thus, all four methods have more concentrated intervals around biased estimates. Comparing full matching to M-N-m or M-C-m reveals that the median biases are similar for all three methods, and the gain in efficiency when using full matching results in a procedure that has less than nominal coverages. Similar trends are observed for the median RMSEs, where full matching and DR have the smallest RMSEs, but also lower than nominal coverages. Compared to IPW_1 and IPW_2, M-N-m and M-C-m have generally lower RMSE, as well as better coverage.

For both parallel and nonparallel response surfaces, the ratio of median absolute bias to median interval width is generally larger than 0.5 for IPW_1, IPW_2, and DR for $0.2 < \text{JSD} < 0.3$. These JSD values also yield significant decreases in coverages for these methods, which is because the biases in point estimates cannot be buried in their variances. M-N-m, M-C-m, and full matching have ratio of median absolute bias to median interval width that is always smaller than 0.3 for $0.2 < \text{JSD} < 0.3$, and in many cases significantly less than 0.3.

5 Discussion

This manuscript compares previously suggested procedures for estimating causal effects when there is a key covariate that is unbalanced between treatment groups. The simulations demonstrate that all regression adjustments relying on monotone and nonmonotone functions of X with a constant treatment effect generally result in statistically invalid methods. Similar results are observed for RAS, which assumes a constant treatment effect within each subclass. Weighting methods generally have larger and more variable absolute biases, compared to M-N-m, M-C-m, and full matching, resulting in below nominal coverage when the distributions of X in the control and treatment groups differ markedly. In addition, when the distributions of X are relatively similar, IPW_1 and IPW_2 have larger and more variable RMSEs, and DRs have more variable RMSEs than any of the valid

subclassification and matching methods. These results are exacerbated when the response surfaces are not parallel and are similar to these observed by Waernbaum.[21]

In comparison to M-N-m and M-C-m, full matching has the lowest and most variable coverages when the response surfaces are not parallel, mainly because the intervals are too short. One possible reason for the short intervals is that full matching does not account for possible variability in the matching procedure and treats the matching as known, which is apparent when considering matching as a method of single imputation without uncertainty. M-N-m has coverages that are closest to nominal among all of the methods that were examined. Combining matching for sampling variance estimate with covariance adjustment for point estimate (M-C-m) results in slightly lower bias and smaller RMSE, but it may also result in slightly lower coverage than M-N-m when the distributions of the covariate in treatment control groups are further apart. M-C-s and M-N-s result in coverages that are lower than nominal, reinforcing the conclusions of Abadie and Imbens.[1] For small JSD, the statistical validity of the confidence intervals obtained using M-N-m and M-C-m does not arise from significantly wider intervals, but primarily from smaller biases. For large JSD, M-N-m and M-C-m still enjoy the smallest bias among all of the methods, but they also have larger interval widths, reflecting the larger uncertainty about the values of the missing potential outcomes due to less overlap.

The results presented here have been obtained for a single covariate, but we believe that methods that do not perform well with a single covariate will perform even worse with multiple covariates due to the added complexity of the response surfaces and estimation of the propensity score. Gutman and Rubin[65] examine the characteristics of a new method for estimating treatment effects when the outcome is binary, based on imputation of the potential outcomes, which appears to be valid and relatively efficient. The sampling variance estimate of M-N-m has some similarities to this imputation method, because M-N-m uses units with the same W_i values and similar X_i values to obtain the variability of the unit-level effects. Thus, it seems that imputing the potential outcomes using some modeling can result in a generally valid and more efficient procedure.

When it is impossible to obtain good overlap of the covariate distributions between the treatment group and the control group, no method can provide generally valid statistical inferences. In such cases, the investigator should consider discarding observations that do not overlap, so that a valid inference on a restricted population can be obtained. The new estimand differs from the original estimand, but the latter is generally impossible to estimate well without making empirically unassailable assumptions.

In conclusion, this manuscript shows that M-N-m and M-C-m are generally valid standard procedures with scalar X. These procedures are also efficient in comparison to common methods used for causal effect estimation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974; 66:688.
2. Rubin DB. Bayesian inference for causal effects: the role of randomization. Ann Stat. 1978; 6:34–58.
3. Rubin DB. Comment on randomization analysis of experimental data: the fisher randomization test. J Am Stat Assoc. 1980; 75:591–593.
4. Rubin DB. Formal modes of statistical inference for causal effects. J Stat Plan Infer. 1990; 25:279–292.
5. Holland PW. Statistics and causal inference (with discussion). J Am Stat Assoc. 1986; 81:945–970.
6. Rubin, DB. Proceedings of the Social Statistics Section of the American Statistical Association. Alexandria, VA: American Statistical Association; 1975. Bayesian inference for causality: the importance of randomization; p. 233-239.
7. Rubin DB. Assignment to treatment group on the basis of a covariate. J Educ Behav Stat. 1977; 2:1–26.
8. Neyman J. Sur les applications de la thar des probabilities aux experiences agaricales: essay de principle. English translation of excerpts by Dabrowska D and Speed T (1990). Stat Sci. 1923; 5:465–472.
9. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70:41–55.
10. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics. 1968; 24:295–313. [PubMed: 5683871]
11. Rubin DB. Matching to remove bias in observational studies. Biometrics. 1973; 29:159–183.
12. Rubin DB. Using multivariate matched sampling and regression adjustment to control bias. J Am Stat Assoc. 1979; 74:318–328.
13. Rubin DB, Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates. J Am Stat Assoc. 2000; 95:573–585.
14. Rubin, DB. Matched sampling for causal effects. New York, NY: Cambridge University Press; 2006.
15. Rosenbaum, PR. Design of observational studies. New York, NY: Springer Verlag; 2009.
16. Rubin DB. The use of matched sampling and regression adjustment to remove bias in observational studies. Biometrics. 1973; 29:185–203.
17. Shah BR, Laupacis A, Hux JE, et al. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. J Clin Epidemiol. 2005; 58:550–559. [PubMed: 15878468]
18. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat Med. 2004; 23:2937–2960. [PubMed: 15351954]
19. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc. 1994; 89:846–866.
20. Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. Med Decis Making. 2009; 29:661–677. [PubMed: 19684288]
21. Waernbaum I. Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation. Stat Med. 2012; 31:1572–1581. [PubMed: 22359267]

22. Austin PC. A comparison of 12 algorithms for matching on the propensity score. Stat Med. 2014; 33:1057–1069. [PubMed: 24123228]
23. Lehmann, EL., Romano, JP. Testing statistical hypotheses. 3rd. New York, NY: Springer; 2005.
24. Chambers CD, Johnson KA, Dick LM, et al. Birth outcomes in pregnant women taking fluoxetine. N Engl J Med. 1996; 335:1010–1015. [PubMed: 8793924]
25. Gastil J, Pierre Deess E, Weiser P. Civic awakening in the jury room: a test of connection between jury deliberation and political participation. J Polit. 2002; 64:585–595.
26. O'Dea JA, Wilson R. Socio-cognitive and nutritional factors associated with body mass index in children and adolescents: possibilities for childhood obesity prevention. Health Educ Res. 2006; 21:796–805. [PubMed: 17095571]
27. Fisher, RA. Statistical methods for research workers. 4th. Edinburgh: Oliver & Boyed; 1932.
28. Cochran WG, Rubin DB. Controlling bias in observational studies: a review. Sankhya Indian J Stat Ser A. 1973; 35:417–466.
29. McCandless LC, Gustafson P, Austin PC. Bayesian propensity score analysis for observational data. Stat Med. 2009; 28:94–112. [PubMed: 19012268]
30. Myers JA, Louis TA. Comparing treatments via the propensity score: stratification or modeling? Health Serv Outcomes Res Methodol. 2012; 12:1–15.
31. Marsh, LC., Cormier, DR. Spline regression models. Thousand Oaks, CA: Sage University; 2002.
32. Eubank, RL. Nonparametric regression and spline smoothing. 2nd. New York, NY: Marcel Dekker; 1999.
33. Boor, C. A practical guide to splines. New York, NY: Springer; 2001.
34. Stone CJ. Generalized additive models: comment. Stat Sci. 1986; 1:312–314.
35. Harrell, FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York, NY: Springer-Verlag; 2001.
36. Hastie, T., Tibshirani, R. Generalized additive models. 1st. London, UK: Chapman and Hall; 1990.
37. Wahba, G. Spline models for observational data. 1st. Philadelphia, PA: SIAM; 1990.
38. Wood SN. On confidence intervals for generalized additive models based on penalized regression splines. Aust NZ J Stat. 2003; 48:445–464.
39. Rubin DB, Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates. J Am Stat Assoc. 2000; 95:573–585.
40. Stuart EA. Matching methods for causal inference. Stat Sci. 2010; 25:1–21. [PubMed: 20871802]
41. Imbens GW, Wooldridge JM. Recent developments in the econometrics of program evaluation. J Econ Literature. 2009; 47:5–86.
42. Austin PC. Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. Int J Biostat. 2009; 5(1):Article 13. [PubMed: 20949126]
43. Abadie A, Drukker D, Herr JL, et al. Implementing matching estimators for average treatment effects in stata. Stata J. 2004; 4:290–311.
44. Sekhon JS. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. J Stat Softw. 2011; 42:1–52.
45. Imbens, G., Rubin, DB. Causal inference in statistics, and in the social and biomedical sciences. New York, NY: Cambridge University Press; 2015.
46. Rosenbaum PR. A characterization of optimal designs for observational studies. J Roy Stat Soc Ser B (Methodological). 1991; 53:597–610.
47. Hansen BB. Full matching in an observational study of coaching for the sat. J Am Stat Assoc. 2004; 99:609–618.
48. Rosenbaum PR. Model-based direct adjustment. J Am Stat Assoc. 1987; 82:387–394.
49. Rubin DB, Thomas N. Characterizing the effect of matching using linear propensity score methods with normal distributions. Biometrika. 1992; 79:797–809.
50. Hirano K, Imbens G, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica. 2003; 71:1161–1189.
51. Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Stat Sci. 2007; 4:1–18.

52. Abadie A, Imbens GW. Bias-corrected matching estimators for average treatment effects. J Bus Econ Stat. 2011; 29:1–11.

53. Espindle, LP. Improving confidence coverage for the estimate of the treatment effect in a subclassification setting, Bachelor's thesis, Department of Statistics. Harvard University; Cambridge, MA: May. 2004

54. Azzalini A. A class of distributions which includes the normal ones. Scand J Stat. 1985; 12:171–178.

55. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat. 1985; 39:33–38.

56. Cangul MZ, Chretien YR, Gutman R, et al. Testing treatment effects in unconfounded studies under model misspecification: logistic regression, discretization, and their combination. Stat Med. 2009; 28:2531–2551. [PubMed: 19572258]

57. Lin J. Divergence measures based on Shannon entropy. IEEE Trans Inform Theory. 1991; 37:145–151.

58. Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat. 1951; 22:79–86.

59. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014. http://www.R-project.org (accessed August 2014)

60. Hansen BB, Klopfer SO. Optimal full matching and related designs via network flows. J Comput Graph Stat. 2006; 15:609–627.

61. Imai K, Ratkovic M. Covariate balancing propensity score. J Roy Stat Soc Ser B (Stat Methodol). 2014; 76(1):243–263.

62. Gutman R, Rubin DB. Robust estimation of causal effects of binary treatments in unconfounded studies with dichotomous outcomes. Stat Med. 2013; 32:1795–1814. [PubMed: 23019093]

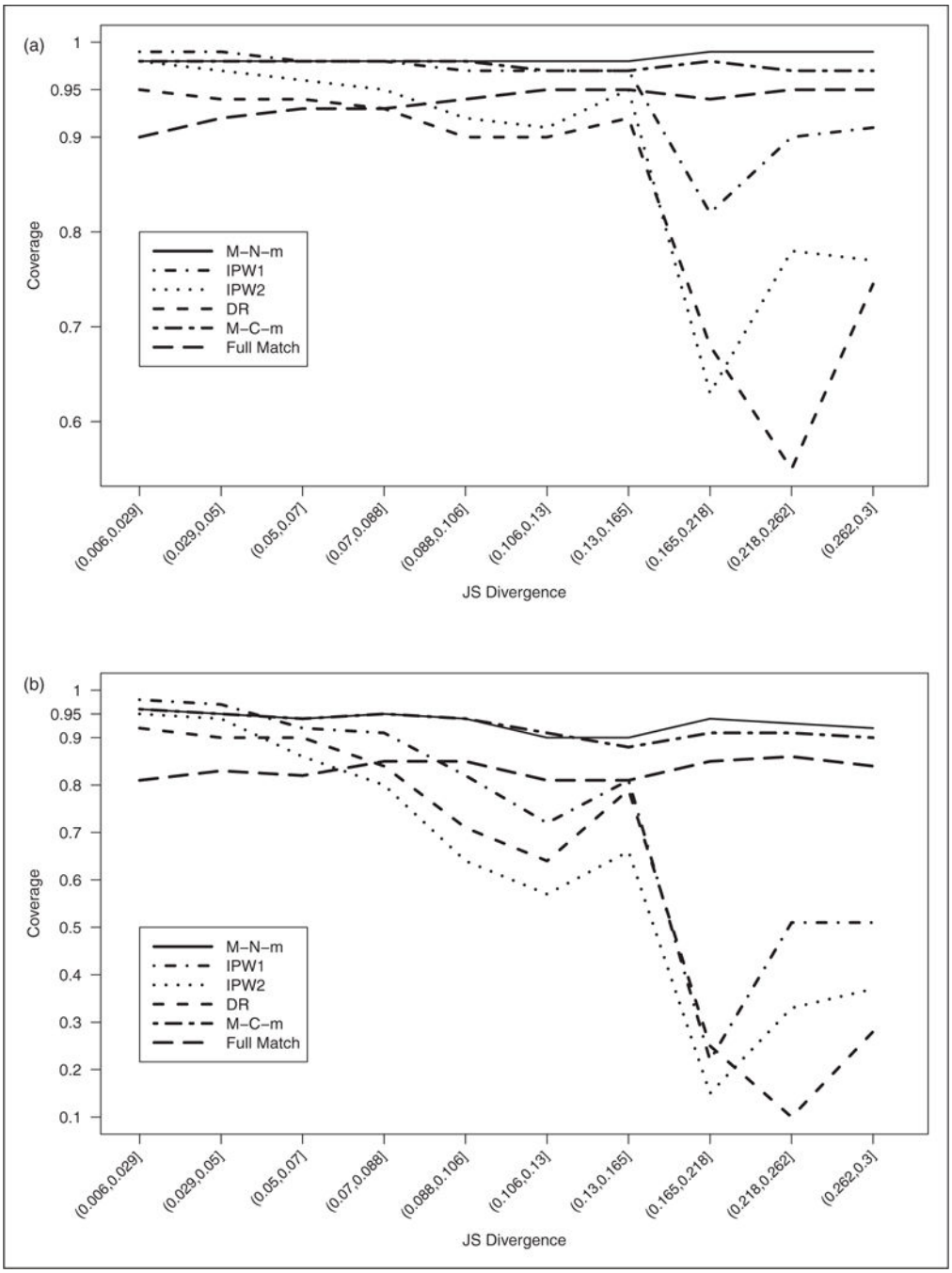

Figure 1.

Median and 25th percentile of 95% interval estimates coverage for the average treatment effect across simulated configurations by JSD for top performing procedures in terms of coverage (monotone response surface with scalar X). (a) Median Coverage and (b) 25th Percentile Coverage.

Table 1

List of the methods used in simulation analysis for scalar X.

Method Class	Method	Reference Section	References
Weighting	IPW_1	Section 2.4	Lunceford and Davidian[18]
	IPW_2	Section 2.4	Lunceford and Davidian[18]
	IPW_3	Section 2.4	Lunceford and Davidian[18]
Subclassification	Full matching	Section 2.3	Rosenbaum[46]
Matching	Within-treatment-group matching for sampling variance estimation (M-N-m)	Section 2.3	Abadie and Imbens[3]
	Standard sampling variance estimation (M-N-s)	Section 2.3	Austin[42]
Regression	Polynomial regression	Section 2.2	
	Spline 6 knots	Section 2.2	McCandless et al.[29]
	Spline 15 knots	Section 2.2	McCandless et al.[29]
	TPS	Section 2.2	Myers and Louis[30]
Combined	RAS	Section 2.5	Rubin[16]
	Covariance adjusted matching with standard sampling variance estimate (M-C-s)	Section 2.5	Abadie and Imbens[52]
	Covariance adjusted matching with matching within treatment group for sampling variance estimate (M-C-m)	Section 2.5	Abadie and Imbens[3,52]
	DR	Section 2.5	Robins et al.[19]

Table 2

Factors and corresponding levels used in the simulation analysis.

Factor	Levels	Description
r	{1, 2}	Ratio of sample sizes
n	{300, 600, 1200}	Treatment group population size
η_0	{−3.5, 0, 3.5}	Skewness of the covariate in the treatment group
η_1	{−3.5, 0, 3.5}	Skewness of the covariate in the control group
SB	$\left\{\frac{1}{4}, \frac{1}{2}, 1\right\}$	Standardized bias for the covariate
ν_1^2	$\left\{\frac{1}{2}, 1, 2\right\}$	Ratio of variances
β	{0.5, 1, 2}	Correlation between X_j and $Y_i(W)$
α	{0, 0.5, 1}	Additive constant effect
σ_W^2	$\left\{\frac{1}{2}, 1, 2\right\}$ or {0.005, 0.01, 0.02}	Deviation of outcome from mean response surface for monotone and nonmonotone response surfaces, respectively.
$g_0(x)$ and $g_1(x)$	Described in Section 3.2	Response surfaces

Table 3

Median, 25% percentile, and 75% percentile of the 95% coverage for the average treatment effect across 262,440 simulated configurations with JSD < 0.3 for monotone response surfaces.

		Overall	Null			Nonnull		
	Method	**% Configurations over 90% coverage**	**Median**	**25%**	**75%**	**Median**	**25%**	**75%**
Weighting	IPW_1	70	0.98	0.86	1	0.97	0.82	1
	IPW_2	58	0.95	0.72	0.99	0.93	0.66	0.98
	IPW_3	47	0.86	0.37	0.97	0.88	0.44	0.96
Subclassification	Full matching	64	0.96	0.93	0.99	0.93	0.83	0.98
Matching	M-N-m	84	0.94	0.90	0.97	0.98	0.94	1
	M-N-s	8	0.46	0.23	0.65	0.62	0.33	0.79
Regression	Polynomial regression	18	0.93	0.87	0.96	0.32	0	0.83
	Spline 6 knots	20	0.95	0.92	0.97	0.33	0	0.83
	Spline 15 knots	21	0.95	0.93	0.97	0.3	0	0.84
	TPS	20	0.95	0.93	0.97	29	0	0.83
Combined	M-C-m	84	0.93	0.89	0.96	0.98	0.94	1
	M-C-s	7	0.46	0.22	0.65	0.59	0.31	0.78
	RAS	27	0.94	0.89	0.97	0.73	0.28	0.9
	DR	55	0.92	0.70	0.96	0.91	0.65	0.96

The median, 25%, and 75% are the percentiles across $3^{10} \times 2$ simulation configurations.

Table 4

Medians across 29,160 configurations with JSD < 0.3 of absolute bias, interval width, RMSE, and coverage for top performing procedures, by JSD (monotone response surfaces) for parallel response surfaces.

JSD		(0.006,0.05]	(0.05,0.088]	(0.088,0.13]	(0.13,0.218]	(0.218,0.3]
IPW1	Bias	0.01 (0.06)	0.06 (0.22)	0.10 (0.24)	0.17 (0.43)	0.22 (0.46)
	Width	0.25 (0.26)	0.26 (0.46)	0.30 (0.36)	0.36 (0.74)	0.44 (0.48)
	RMSE	0.13 (0.15)	0.16 (0.31)	0.24 (0.31)	0.35 (0.60)	0.38 (0.50)
	Coverage	1.00 (0.02)	0.99 (0.05)	0.98 (0.22)	0.98 (0.32)	0.93 (0.47)
	% Above 0.9	94	82	68	64	53
IPW2	Bias	0.01 (0.06)	0.06 (0.23)	0.11 (0.22)	0.22 (0.47)	0.29 (0.47)
	Width	0.22 (0.20)	0.23 (0.36)	0.26 (0.29)	0.30 (0.55)	0.35 (0.28)
	RMSE	0.12 (0.12)	0.15 (0.27)	0.22 (0.23)	0.31 (0.62)	0.36 (0.43)
	Coverage	0.99 (0.04)	0.97 (0.09)	0.93 (0.41)	0.91 (0.58)	0.79 (0.60)
	% Above 0.9	90	75	55	51	31
DR	Bias	0.01 (0.05)	0.03 (0.15)	0.04 (0.17)	0.09 (0.52)	0.22 (0.58)
	Width	0.17 (0.14)	0.18 (0.24)	0.21 (0.31)	0.33 (0.52)	0.30 (0.46)
	RMSE	0.09 (0.10)	0.11 (0.22)	0.16 (0.23)	0.29 (0.53)	0.29 (0.55)
	Coverage	0.95 (0.06)	0.94 (0.10)	0.92 (0.23)	0.88 (0.52)	0.78 (0.73)
	% Above 0.9	81	69	56	45	33
M-N-m	Bias	0.01 (0.03)	0.02 (0.05)	0.05 (0.13)	0.10 (0.22)	0.16 (0.24)
	Width	0.15 (0.09)	0.19 (0.12)	0.25 (0.15)	0.40 (0.36)	0.54 (0.33)
	RMSE	0.08 (0.05)	0.10 (0.07)	0.15 (0.12)	0.24 (0.26)	0.34 (0.26)
	Coverage	0.94 (0.04)	0.94 (0.05)	0.93 (0.09)	0.94 (0.09)	0.94 (0.08)
	% Above 0.9	86	76	63	66	74
M-C-m	Bias	0.01 (0.02)	0.02 (0.06)	0.03 (0.12)	0.10 (0.30)	0.07 (0.29)
	Width	0.15 (0.09)	0.19 (0.12)	0.25 (0.15)	0.40 (0.36)	0.54 (0.33)
	RMSE	0.08 (0.05)	0.11 (0.07)	0.15 (0.11)	0.24 (0.29)	0.30 (0.26)
	Coverage	0.95 (0.03)	0.94 (0.06)	0.93 (0.08)	0.92 (0.09)	0.93 (0.08)
	% Above 0.9	89	75	65	56	63
Full matching	Bias	0.01 (0.03)	0.02 (0.06)	0.06 (0.13)	0.10 (0.23)	0.17 (0.24)
	Width	0.17 (0.10)	0.22 (0.14)	0.30 (0.22)	0.44 (0.50)	0.60 (0.45)

JSD		(0.006,0.05]	(0.05,0.088]	(0.088,0.13]	(0.13,0.218]	(0.218,0.3]
	RMSE	0.09 (0.06)	0.12 (0.08)	0.16 (0.15)	0.25 (0.30)	0.35 (0.31)
	Coverage	0.96 (0.05)	0.96 (0.06)	0.96 (0.06)	0.96 (0.06)	0.96 (0.07)
	% Above 0.9	87	86	88	86	82

% Above 0.9 represents the percent of simulation configurations with 95% coverage interval that have coverage larger than 90%. Parentheses are the interquartile range.

Table 5

Medians across 233,280 configurations with JSD < 0.3 of absolute bias, interval width, RMSE, and coverage for top performing procedures, by JSD (monotone response surfaces) for nonparallel response surfaces.

JSD		(0.006,0.05]	(0.05,0.088]	(0.088,0.13]	(0.13,0.218]	(0.218,0.3]
IPW1	Bias	0.02 (0.06)	0.09 (0.26)	0.11 (0.26)	0.33 (0.65)	0.30 (0.64)
	Width	0.28 (0.25)	0.32 (0.45)	0.33 (0.35)	0.47 (0.83)	0.52 (0.61)
	RMSE	0.15 (0.14)	0.21 (0.33)	0.25 (0.31)	0.51 (0.78)	0.48 (0.68)
	Coverage	0.99 (0.03)	0.98 (0.09)	0.97 (0.23)	0.93 (0.51)	0.90 (0.48)
	% Above 0.9	95	76	67	54	50
IPW2	Bias	0.02 (0.06)	0.10 (0.26)	0.12 (0.26)	0.33 (0.65)	0.30 (0.65)
	Width	0.24 (0.21)	0.27 (0.38)	0.27 (0.29)	0.37 (0.72)	0.40 (0.38)
	RMSE	0.13 (0.12)	0.18 (0.30)	0.22 (0.27)	0.45 (0.78)	0.39 (0.64)
	Coverage	0.97 (0.04)	0.95 (0.17)	0.92 (0.39)	0.81 (0.61)	0.78 (0.59)
	% Above 0.9	89	65	52	42	34
DR	Bias	0.02 (0.06)	0.04 (0.14)	0.08 (0.18)	0.18 (0.52)	0.26 (0.51)
	Width	0.20 (0.18)	0.22 (0.33)	0.26 (0.35)	0.38 (0.63)	0.35 (0.48)
	RMSE	0.11 (0.11)	0.14 (0.25)	0.20 (0.23)	0.33 (0.68)	0.38 (0.58)
	Coverage	0.95 (0.06)	0.94 (0.08)	0.90 (0.28)	0.87 (0.49)	0.65 (0.70)
	% Above 0.9	79	69	47	40	25
M-N-m	Bias	0.01 (0.02)	0.02 (0.05)	0.04 (0.11)	0.09 (0.22)	0.12 (0.23)
	Width	0.24 (0.19)	0.30 (0.24)	0.38 (0.26)	0.63 (0.62)	0.79 (0.63)
	RMSE	0.13 (0.11)	0.17 (0.15)	0.22 (0.17)	0.40 (0.43)	0.49 (0.43)
	Coverage	0.98 (0.04)	0.99 (0.05)	0.99 (0.06)	0.99 (0.07)	0.99 (0.06)
	% Above 0.9	95	89	82	79	79
M-C-m	Bias	0.01 (0.02)	0.02 (0.07)	0.04 (0.10)	0.09 (0.31)	0.09 (0.27)
	Width	0.24 (0.19)	0.30 (0.24)	0.38 (0.26)	0.63 (0.62)	0.79 (0.63)
	RMSE	0.13 (0.10)	0.17 (0.15)	0.22 (0.17)	0.39 (0.46)	0.46 (0.41)
	Coverage	0.98 (0.04)	0.98 (0.05)	0.98 (0.06)	0.98 (0.08)	0.98 (0.09)
	% Above 0.9	96	89	83	77	76
Full Matching	Bias	0.01 (0.03)	0.02 (0.05)	0.04 (0.11)	0.09 (0.23)	0.13 (0.23)
	Width	0.17 (0.10)	0.21 (0.14)	0.29 (0.23)	0.44 (0.46)	0.58 (0.46)

JSD		(0.006,0.05]	(0.05,0.088]	(0.088,0.13]	(0.13,0.218]	(0.218,0.3]
	RMSE	0.09 (0.06)	0.13 (0.10)	0.19 (0.16)	0.31 (0.39)	0.39 (0.38)
	Coverage	0.90 (0.15)	0.92 (0.15)	0.93 (0.18)	0.94 (0.19)	0.95 (0.17)
	% Above 0.9	49	58	59	61	62

% Above 0.9 represents the percent of simulation configurations with 95% coverage interval that have coverage larger than 90%. Parentheses are the interquartile range.