Improving causal inference with a doubly robust estimator that combines propensity score stratification and weighting

Ariel Linden DrPH1,2

1 President, Linden Consulting Group, LLC Ann Arbor, MI, USA
2 Research Scientist, Division of General Medicine, Medical School, University of Michigan, Ann Arbor, MI, USA

Correspondence
Ariel Linden, DrPH, Linden Consulting Group, LLC, 1301 North Bay Drive, Ann Arbor, MI 48103, USA.
Email: alinden@lindenconsulting.org

Abstract
Rationale, aims and objectives When a randomized controlled trial is not feasible, health researchers typically use observational data and rely on statistical methods to adjust for confounding when estimating treatment effects. These methods generally fall into 3 categories: (1) estimators based on a model for the outcome using conventional regression adjustment; (2) weighted estimators based on the propensity score (ie, a model for the treatment assignment); and (3) “doubly robust” (DR) estimators that model both the outcome and propensity score within the same framework. In this paper, we introduce a new DR estimator that utilizes marginal mean weighting through stratification (MMWS) as the basis for weighted adjustment. This estimator may prove more accurate than treatment effect estimators because MMWS has been shown to be more accurate than other models when the propensity score is misspecified. We therefore compare the performance of this new estimator to other commonly used treatment effects estimators.

Method Monte Carlo simulation is used to compare the DR-MMWS estimator to regression adjustment, 2 weighted estimators based on the propensity score and 2 other DR methods. To assess performance under varied conditions, we vary the level of misspecification of the propensity score model as well as misspecify the outcome model.

Results Overall, DR estimators generally outperform methods that model one or the other components (eg, propensity score or outcome). The DR-MMWS estimator outperforms all other estimators when both the propensity score and outcome models are misspecified and performs equally as well as other DR estimators when only the propensity score is misspecified.

Conclusions Health researchers should consider using DR-MMWS as the principal evaluation strategy in observational studies, as this estimator appears to outperform other estimators in its class.

KEYWORDS causal inference, doubly robust, inverse probability of treatment weights, marginal mean weighting through stratification, propensity score, stratification, treatment effects

1 INTRODUCTION

When conducting a randomized controlled trial is not feasible, health researchers typically use observational data and rely on statistical methods to adjust for confounding when estimating treatment effects. Although conventional regression remains the most common adjustment approach, methods that explicitly model the treatment assignment—such as those using instrumental variables1,2 or based on the propensity score—that are now used more widely.

The propensity score is defined as the probability of assignment to the treatment group conditional on observed characteristics.3 Propensity scores are generally estimated via logistic regression, reducing each individual’s set of covariates into a single scalar. It has been demonstrated that, in large samples when treatment and control groups have similar distributions of the propensity score, the groups also usually have similar distributions of the underlying covariates used to create the propensity score. This implies that observed pretreatment covariates can be considered independent of treatment assignment.
Marginal mean weighting through stratification combines elements of both propensity score stratification and IPTW. Stratification (also known as subclassification) entails stratifying the analytic sample into quantiles of the propensity score, which reflects a coarser version of matching in which treated and nontreated individuals within each stratum are expected to be comparable on pretreatment characteristics. It has been shown that stratifying the propensity score into 5 quantiles can remove over 90% of the initial bias due to the covariates used to generate the propensity score. Next, a weight is generated for each individual based on their stratum and treatment assignment. The marginal mean weights are computed using the following formula:

\[
\frac{n_i \times \Pr(Z = 1)}{n_{2-5}},
\]

where \(n_i\) is the total number of individuals in a given stratum, \(s\), \(\Pr(Z = 1)\) is the estimated probability of assignment to treatment group 1, that is, the proportion of those actually receiving treatment \(z\) in the sample, and \(n_{2-5}\) is the total number of individuals in stratum \(s\) who were actually assigned to treatment \(z\). Thus, the weight is proportional to the ratio of the number of individuals in a given strata to the number of individuals within that strata actually receiving the treatment. Taken together, the stratification reduces bias in the observed covariates used to generate the propensity score, and the weighting standardizes each treatment group to the target population. The MMWS weights are then specified as sampling weights within the outcome regression model.

To implement the DR-MMWS estimator, we follow the framework proposed by Wooldridge, which applies IPTW together with RA (IPTW-RA), but we replace IPTW with MMWS. The DR-MMWS is operationalized in a multistep process. First, the propensity score model is estimated. Next, the sample is partitioned into strata of the propensity score (typically 5 quintiles are used, although an optimal stratification algorithm could be employed to determine if a different number should be used [Linden forthcoming]). Next, MMWS weights are computed for each individual in the sample. Next, using the MMWS as sampling weights, separate outcome models are fitted by a weighted regression for each treatment group, and treatment-specific predicted outcomes for each individual are obtained using the estimated coefficients from this weighted regression. Finally, the means of the treatment-specific predicted outcomes are computed. The contrasts between these averages provide the point estimates of the average treatment effects, and a bootstrapping procedure (which includes both the estimation of the propensity score and outcome models) is used to obtain valid SEs.

### 3 Monte Carlo Simulation Study

In this simulation study, we examine how well the DR-MMWS estimator compares to several other regression-based treatment effect methods.
estimators in reducing bias in treatment effects estimation. These models fall into 3 general categories: (1) estimators based on a model for the outcome variable using conventional RA; (2) estimators based on a model for the treatment assignment, using IPTW and MMWS; and (3) DR estimators that model both the treatment assignment and outcome variable within the same framework, using an augmented IPTW approach (A-IPTW). IPTW combined with RA (IPTW-RA), and the DR-MMWS estimator.

Our simulation design is a modified version of that described by Hong. The estimated propensity score is misspecified to varying degrees (4 scenarios), and the outcome model (which follows a non-linear normal distribution) is either correctly or incorrectly specified (2 scenarios). In each scenario, 10,000 replications are drawn from the data-generating process described below and repeated for sample sizes of 500 and 2000. For each replication, the treatment effect estimate and SE for each model are recorded. Bias (the difference between the simulated effect and the true effect of 1.0) and the root mean squared error (RMSE)—a measure that magnifies and severely penalizes large errors—are then calculated across all samples. Lower values for all measures indicate better bias reduction.

3.1 Data-generating process for the treatment model

As in Hong (Simulation II), the true propensity score assigns treatment according to a polynomial function of $X$:

$$\Pr = a_0 + a_1 X + a_2 X^2,$$

where $X$ is drawn from a standard normal distribution with a mean of 0 and a standard deviation of 1 and $a_0$, $a_1$, and $a_2$ are manipulated to induce varying degrees of nonlinearity as follows:

- Model 1: $a_0 = 1, a_1 = .2, a_2 = -.2$
- Model 2: $a_0 = 1, a_1 = .6, a_2 = -.2$
- Model 3: $a_0 = 1, a_1 = .2, a_2 = -.6$
- Model 4: $a_0 = 1, a_1 = .6, a_2 = -.6$

The treatment assignment indicator $Z$ is a Bernoulli random variable with the parameter of its distribution equal to the inverse logit of the true propensity score. A misspecified propensity score, which excludes the quadratic term $X^2$, is used in all simulation models.

3.2 Data-generating process for the outcome model

As in Hong, a nonlinear model for potential outcomes was generated for each set of simulations. The model generated 2 potential outcomes $Y(1)$ and $Y(0)$ corresponding to the experimental condition $Z = 1$ and the control condition $Z = 0$. Both $Y(1)$ and $Y(0)$ are polynomial functions of a standard normal covariate $X$:

$$Y(1) = 6 + 0.5X + 0.25X^2 - 0.125X^3 + \epsilon(1);$$
$$Y(0) = 5 + 0.5X + 0.25X^2 - 0.125X^3 + \epsilon(0);$$
$$\epsilon(1), \epsilon(0) \sim N(0, 0.25).$$

The misspecified outcome model excludes the polynomial functions $X^2$ and $X^3$. In all models, the true treatment effect equal to 1.

3.3 Model estimation

In this section, we describe the estimation and inference procedures for each model and repetition over the simulation scenarios. All simulations and analyses reported in this paper were conducted using Stata version 14.2 (StataCorp, College Station, Texas).

For each scenario, 6 different models were used to estimate the potential outcome mean for each of the 3 treatment levels. (1) Regression adjustment was implemented by regressing the outcome $Y$ on all covariates (correctly specified model) or by regressing $Y$ on $X$ (misspecified model). (2) IPTW estimates were derived by, first, computing the IPTW weights as described earlier and then specifying the weights as sampling weights (pweights) in the outcome model where the outcome $Y$ was regressed on an indicator variable representing the 2 treatment levels of $Z$. (3) MMWS estimates were derived by, first, dividing the sample equally into 6 strata based on the estimated propensity score (in keeping with Hong), then by computing the MMWS weights by implementing a user-written command for Stata MMWS, and finally by regressing the outcome $Y$ on an indicator variable representing the treatment levels of $Z$, with the MMWS weights used as sample weights. (4) The A-IPTW estimator was implemented using the teffects ipw command. (5) The IPTW-RA estimator was implemented using the teffects ipwra command. (6) The DR-MMWS estimator was implemented as described in Section 2. All analyses were conducted with observations restricted to be within the region of common support (ie, all individuals have a corresponding counterfactual).

3.4 Monte Carlo simulation results

Table 1 presents the simulation results for sample sizes of 500 and 2000, when the outcome model is correctly specified. As expected with a correctly specified outcome model, the RA estimator had zero bias and low RMSE. Of the 2 estimators based on a model for the treatment assignment (IPTW and MMWS), MMWS consistently produces substantially lower bias and RMSE than IPTW, and that par increases as the amount of nonlinearity in the propensity score increases. All 3 DR models (IPTW-RA, A-IPTW, and DR-MMWS) perform best and produce unbiased estimates.

Table 2 presents the simulation results for sample sizes of 500 and 2000, when the outcome model is misspecified. The RA estimate is now biased because of the misspecification. The values for IPTW and MMWS are identical to those in Table 1 because these estimators are unaffected by misspecification of the outcome model. Inverse probability of treatment weights–RA outperformed A-IPTW, deriving estimates very close to those of IPTW, while A-IPTW appears to obtain results that split the difference between RA and IPTW. Doubly
TABLE 1 Monte Carlo results for estimators when the outcome is correctly specified and the propensity score is misspecified to varying degrees

<table>
<thead>
<tr>
<th>Propensity score parameters</th>
<th>N = 500</th>
<th>N = 2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>a0 a1 a2</td>
<td>Bias SE RMSE</td>
<td>Bias SE RMSE</td>
</tr>
<tr>
<td>RA</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td>IPTW</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td>MMWS</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td>IPTW-Ra</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td>A-IPTW</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td>Dr-MMWS</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Abbreviations: A-IPTW, augmented inverse probability of treatment weighting; DR-MMWS, doubly robust marginal mean weighting through stratification; IPTW, inverse probability of treatment weights; IPTW-Ra, inverse probability of treatment-weighted regression adjustment; MMWS, marginal mean weighting through stratification; RA, regression adjustment; RMSE, root mean squared error; SE, standard error.

robust-MMWS outperformed all the other estimators (save for MMWS) eliciting bias and RMSE estimates that are roughly half that of the other 2 DR estimators and RA.

4 | DISCUSSION

In this paper, we used Monte Carlo simulations to compare the performance of the DR-MMWS estimator to several other adjustment techniques commonly used for estimating treatment effects in nonrandomized studies. Our overall simulation results can be briefly summarized as follows: (1) When the outcome model is correctly specified but the propensity score model is misspecified, RA and all DR estimators provide unbiased estimates, while methods based solely on modeling the propensity score (ie MMWS and IPTW) provide biased estimates. That said, MMWS provides substantially less biased estimates than IPTW. (2) When both the propensity score and outcome models are misspecified, MMWS and DR-MMWS substantially outperform all other estimators.

In these simulations, the advantage DR-MMWS holds over these other estimators—when both treatment and outcomes models are misspecified—is due to the better performance of MMWS over IPTW when the propensity score is misspecified. That is, the DR-MMWS estimator is much more influenced by the propensity score model (and thus MMWS) than RA. Similarly, IPTW-Ra is much more influenced by the propensity score model (and thus IPTW) than RA. On the other hand, the A-IPTW framework appears to split the difference between the results of the IPTW and RA models.

Why does the MMWS outperform IPTW when the propensity score model is misspecified? Hong17 suggests that given IPTW is computed as a direct function of the estimated propensity score; when the estimated propensity score is misspecified, the IPTW will systematically deviate from the true weight (leading to bias in the treatment effect estimates). Conversely, misspecification of the propensity score does not change propensity score stratum membership for units in either treatment group. Given that MMWS weights are estimated as a ratio of the sample sizes within each stratum, the computed weights will remain consistent even under misspecification, and therefore, estimated treatment effects will remain robust.17

Other empirical studies examining a similar array of adjustment methods have shown that DR methods provide unbiased estimates when either the propensity score or outcome model is misspecified.16,27,29–32 However, there currently appears to be no consensus as to which estimator is most appropriate if both models are misspecified.30,31,33 Thus, from a practical standpoint, investigators may be best served by analyzing their data—as we have here—using DR-MMWS along with other estimators as a sensitivity analysis.34 If all methods obtain similar treatment effect estimates, investigators will...
have greater confidence that the study results are unbiased. If, on the other hand, estimates differ substantially, a close examination of the results may clarify whether the inconsistencies are found between treatment model estimators based on the MMWS vs those using IPTW. If this appears to be where the discrepancy occurs, then investigators may either assume that the estimates of the DR-MMWS are more accurate (ie, less biased) than those derived from estimators using IPTW, or they should consider re-estimating the propensity score, perhaps using machine learning techniques, which have been shown to outperform logistic regression in estimating the propensity score (ie, predicting treatment assignment).35–40

The primary limitation of this simulation study is that the performance of the various estimators on treatment effects was considered in the context of a specific data generating process. Second, our simulation assumed strong ignorability, although observational data in health research are typically laden with confounding from unobservable variables such as unmeasured motivation to change health behaviors.41,42

Thus, future research should compare the performance of the DR-MMWS estimator to other methods in the context of more diverse data-generating processes (including additional variable types and distributions) and violations to assumptions of the causal model. Finally, while simulation is, in and of itself, a form of cross-validation, future comparisons using empirical data should be coupled with cross-validation techniques (ie, k-fold or leave-one-out cross-validation)63 to assess if DR-MMWS generalizes better than other estimators to individuals outside of the original estimation sample.44

In summary, the results of our simulation study suggest that the DR-MMWS estimator outperforms other regression-based treatment effect estimators when both the propensity score and outcome models are misspecified and perform equally as well as other DR estimators when only the propensity score is misspecified. Health researchers should consider using DR-MMWS as the principal evaluation strategy in observational studies, as it is unlikely that he or she will know which of the 2 models (or both) is misspecified.

ACKNOWLEDGMENT

I wish to thank Julia Adler-Milstein for reviewing the manuscript and providing many helpful comments.

REFERENCES


How to cite this article: Linden A. Improving causal inference with a doubly robust estimator that combines propensity score stratification and weighting. J Eval Clin Pract. 2017. doi: 10.1111/jep.12714