Stat
The ISI's Journal for the Rapid Dissemination of Statistics Research
(wileyonlinelibrary.com) DOI: 10.100X/sta.0000

A doubly robust weighting estimator of the average treatment effect on the treated

Erica EM Moodie[a]*, Olli Saarela[b], David A Stephens[c]

Received 00 Month 2012; Accepted 00 Month 2012

We introduce an importance sampling derivation of the average treatment effect on the treated, and extend this to incorporate an augmentation term to allow doubly robust estimation of the average treatment effect on the treated. Unlike the matching estimator of the average treatment effect on the treated, the augmented inverse weighted estimator that results from the importance sampling approach has regular asymptotic properties and does not result in any datapoints being excluded from the estimation. Following simulations, we apply the doubly robust, augmented weighted estimator to a U.S. national survey of health to examine the impact of smoking on sleep, and use techniques developed for other doubly robust estimators to demonstrate model validity. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: Biostatistics; Causal inference; Medical statistics; Statistical modelling.

1. Introduction

For reasons of convenience, cost, or ethical considerations, observational data are typically used to draw inferences about the effect of a treatment or an exposure on an outcome. Causal inference reasoning provides a framework for specifying estimands of interest, and constructing estimators for them that are not biased by confounding or other distortions in the data. While in a randomized trial involving a particular treatment the population of interest would be those individuals who are eligible for that treatment (that is, with no contraindications, etc.), it is often the case that an observational dataset contains both those individuals who are treated, those who could have been treated but were not, and those who might never have been treated. Indeed, it is the imbalance between the distributions of confounding variables in the two treatment groups among those

[a]Department of Epidemiology & Biostatistics, McGill University, 1020 Pine Ave W, Montreal, QC Canada H3A 1A2
[†]erica.moodie@mcgill.ca
[b]Dalla Lana School of Public Health, University of Toronto, 155 College St, Toronto, ON Canada M5T 3M7
[†]olli.saarela@utoronto.ca
[c]Department of Mathematics & Statistics, McGill University, 805 Sherbrooke St W, Montreal, QC Canada H3A 2K6
[†]david.stephens@mcgill.ca
***Email: erica.moodie@mcgill.ca**

Prepared using staauth.cls [Version: 2012/05/12 v1.00]

eligible for treatment that may lead to spurious associations or biased estimates. Adjusting for these imbalances can be achieved either by performing an analysis that ensures that the confounder distribution in both treatment groups is the same as the overall distribution in the sample, or can be done so as to create balanced distributions that match one of the two treatment groups, typically the treated individuals. The first of these leads to an estimate of the *average treatment effect* (ATE), while latter yields the *average treatment effect on the treated* (ATT).

The ATT may often be a more relevant estimand when the effect of treatment is only of interest in those *actually exposed* to a particular treatment in the population. Consider, for example, a recent study examining the impact of the use of antidepressants during pregnancy and the risk of autism spectrum disorder in children (Boukhris et al., 2016). It could very reasonably be argued that the effect of interest is the impact of taking (versus not taking) antidepressants during pregnancy among those women who actually have taken antidepressants; this subpopulation likely includes women with symptoms of depression, anxiety, or other psychiatric disorders for which these medications are indicated. That is, we wish to understand what would be the effect (on average) of giving (versus withholding) antidepressants *among all women who are prescribed such medications*, rather than average effect of giving, versus withholding, antidepressants *among all women*. Similarly, it is likely more relevant to examine the effect of beta-blockers on adults diagnosed with hypertension (or who have another reasonable indication for the drugs) and taking beta-blockers, than in the population of all adults. When there exists no heterogeneity of the treatment effect by covariates (i.e. no interactions or effect modification), the ATE and ATT are identical if effects are additive (they need not coincide in logistic or multiplicative models even in the absence of effect modification). However, when treatment heterogeneity exists, the ATE and ATT differ.

In the biostatistical literature, there has been considerable attention devoted to estimation of the ATT via matching (Abadie & Imbens, 2006; Leacy & Stuart, 2014), and it is this parameter, rather than the ATE, that is the default target of estimation in a popular `R` package (Sekhon, 2011). Matching is appealing in that it is an intuitive approach to designing an analytic sample: if the treatment assignment mechanism is unconfounded, or the propensity score (Rosenbaum & Rubin, 1983) can be estimated correctly with respect to confounding variables, then matching can be used to create a sample in which measured confounders are balanced between treated and untreated subjects. The procedure has been implemented in several commonly used software packages such as `R`, SAS, and Stata. However, matching estimators also have several unattractive statistical properties (Abadie & Imbens, 2006): they are not guaranteed to be $n^{1/2}$ consistent due to a bias term, are not efficient even in settings where consistency is achieved, and variability cannot be reliably assessed via standard bootstrap procedures. While a bias correction has been proposed (Abadie & Imbens, 2011), it is has not seen much use in applications.

Although estimators of the ATT by weighting have been proposed previously (Hirano et al., 2003, e.g.), the approach has not gained popularity in statistical analyses of health data in the way that it has for estimation of population-level treatment effects both cross-sectionally and longitudinally as in marginal structural models (Hernán et al., 2000). For instance, in a recent paper (Pirracchio et al., 2016) comparing the ATE and ATT, matching was used to estimate the ATT while weighting was employed for estimating the ATE. Furthermore, while a matching estimator of the ATT can be made doubly robust by performing regression adjustment after matching (Sekhon, 2011) or via non-parametric approaches (Hubbard et al., 2011), the use of doubly robust matching estimators appears to be extremely rare though exceptions exist (e.g. Wunsch & Lechner, 2008). This apparent aversion to weighting estimators of the ATT is not present in all disciplines. In econometrics, weighting-based estimators of the ATT have been considered in various forms. For example, Hahn (1998) proposes using “non-parametric imputation", estimating the counterfactual outcomes under both treatment conditions (present and absent) via inverse probability of treatment weighting, and then marginalizing these differences among the treated individuals. Hainmueller (2012) used a more conventional approach where weighting is used only to estimate the

mean among the treated individuals under the no treatment condition, where the weights were constrained to several balancing constraints and applied in a singly-robust fashion.

In this paper, we derive the ATT from the perspective of importance sampling (that is, change of measure) to provide a new and intuitive understanding of the usual weighted estimator, then extend this to include an augmentation term which yields a doubly robust estimator of the average treatment effect on the treated. We demonstrate its unbiasedness and efficiency relative to singly robust and matching estimators via simulation, and apply the estimator to examine the impact of current smoking on the number of hours slept in the 2011-2012 wave of the U.S. National Health and Nutrition Examination Study (www.cdc.gov/nchs/nhanes.htm).

2. The Average Treatment Effect on the Treated: Derivation and Estimation

2.1. A New Perspective on the Singly Robust Estimator

Let Y be the outcome, which we shall take to be continuous though the development below applied naturally to any variable for which a mean is well-defined (censored data requiring additional modifications to the estimator). Exposure will be denoted by $Z \in \{0, 1\}$ and X will denote a confounding variable, which could be vector-valued. Lower case will be used to denote realized values, while random variables are indicated with the use of upper-case letters. We shall work in with potential outcomes, where $Y(z)$ denotes the outcome that would be observed if treatment z is given; the counterfactuals are linked to the observed data by consistency: $Y = Y(0)(1 - Z) + Y(1)Z$. We assume that there are no unmeasured confounders; that is, X contains all the information needed to ensure that $Y(0), Y(1) \perp Z \mid X$.

In this section, we will review the form of the singly robust inverse probability of treatment weighted estimator of the average treatment effect on the treated, and show that it can be derived from the perspective of importance sampling. This approach to the derivation of the estimator leads to new insights that facilitate the extension of the estimator to one which is doubly robust.

The average treatment effect on the treated is $E(Y(1) - Y(0) \mid Z = 1)$, where, for $z = 0, 1$, we have:

$$E(Y(z) \mid Z = 1) = \int y f_{Y|X,Z}(y \mid x, z) f_{X|Z}(x \mid 1)\, dy\, dx.$$

When $z = 1$, we have $E(Y(1) \mid Z = 1) = \int y f_{Y|X,Z}(y \mid x, 1) f_{X|Z}(x \mid 1)\, dy\, dx$, which may be written

$$E(Y(1) \mid Z = 1) = \frac{\int \mathbb{1}_1(z) y f_{Y|X,Z}(y \mid x, z) f_{X|Z}(x \mid z) f_Z(z)\, dy\, dx\, dz}{\int \mathbb{1}_1(z) f_Z(z)\, dz}.$$

It is perhaps helpful to view the frequentist estimation of this quantity from data using moment-based methods as a form of Monte Carlo procedure where the Monte Carlo samples are the original sample data. As the integral in the numerator is taken with respect to the conditional joint density $f_{Y|X,Z}(y \mid x, 1) f_{X|Z}(x \mid 1)$, from which a sample is available from the original data as those samples for which Z is observed to take the value one, the 'Monte

Carlo' (non-parametric moment-based) estimator of the average treatment effect on the treated is

$$\widehat{E}(Y(1) \mid Z=1) = \frac{\sum_{i=1}^{n} \mathbb{1}_1(Z_i) Y_i}{\sum_{i=1}^{n} \mathbb{1}_1(Z_i)} = \frac{\sum_{i=1}^{n} Z_i Y_i}{\sum_{i=1}^{n} Z_i},$$

that is, it is simply the sample mean of the treated individuals.

However, when $z=0$, we have

$$E(Y(0) \mid Z=1) = \int y f_{Y(0)|X}(y \mid x) f_{X|Z}(x \mid 1)\, dy\, dx = \int y f_{Y|X,Z}(y \mid x, 0) f_{X|Z}(x \mid 1)\, dy\, dx,$$

which cannot be computed directly using a similar strategy due to the 'incompatible' combination of conditional densities with different conditioning sets $f_{Y|X,Z}(y \mid x, 0) f_{X|Z}(x \mid 1)$. However, the expectation may be written using the importance sampling change of measure as

$$E(Y(0) \mid Z=1) = \int y f_{Y|X,Z}(y \mid x, 0) \frac{f_{X|Z}(x \mid 1)}{f_{X|Z}(x \mid 0)} f_{X|Z}(x \mid 0)\, dy\, dx$$

provided $f_{X|Z}(x \mid 0)$ is non-zero whenever $f_{X|Z}(x \mid 1)$ is non-zero to yield

$$E(Y(0) \mid Z=1) \equiv E\left(Y(0) \frac{f_{X|Z}(X \mid 1)}{f_{X|Z}(X \mid 0)} \,\middle|\, Z=1 \right).$$

where the right-hand side is an expectation with respect to the 'compatible' joint density $f_{Y|X,Z}(y \mid x, 0) f_{X|Z}(x \mid 0)$. Now,

$$\frac{f_{X|Z}(x \mid 1)}{f_{X|Z}(x \mid 0)} = \frac{f_{Z|X}(1 \mid x)}{f_{Z|X}(0 \mid x)} \frac{f_Z(0)}{f_Z(1)}$$

and so

$$\begin{aligned} E(Y(0) \mid Z=1) &= \frac{f_Z(0)}{f_Z(1)} \int y \frac{f_{Z|X}(1 \mid x)}{f_{Z|X}(0 \mid x)} f_{Y|X,Z}(y \mid x, 0) f_{X|Z}(x \mid 0)\, dy\, dx \\ &= \frac{1}{f_Z(1)} \int \mathbb{1}_0(z) y w(x) f_{Y|X,Z}(y \mid x, z) f_{X|Z}(x \mid z)\, dy\, dx\, dz \end{aligned}$$

say, where $w(x) = \pi(x)/(1-\pi(x))$ is a function of the propensity score, $\pi(x) = f_{Z|X}(1 \mid x)$. The Monte Carlo estimator is therefore

$$\widehat{E}(Y(0) \mid Z=1) = \frac{\sum_{i=1}^{n} \mathbb{1}_0(Z_i) w(X_i) Y_i}{\sum_{i=1}^{n} \mathbb{1}_1(Z_i)} = \frac{\sum_{i=1}^{n} (1-Z_i) w(X_i) Y_i}{\sum_{i=1}^{n} Z_i}.$$

That is, the estimator is a weighted sum of contributions from the untreated individuals: $\{(1-Z_i) w(X_i) Y_i\} \left(\sum_{i=1}^{n} Z_i \right)^{-1}$. Thus the average treatment effect on the treated estimator is

$$\frac{\sum_{i=1}^{n} (Z_i - (1-Z_i) w(X_i)) Y_i}{\sum_{i=1}^{n} Z_i}.$$

Under the standard assumptions listed above, this estimator is consistent for the average treatment effect on the treated and asymptotically normally distributed if $\pi(x)$ is correctly specified; that is, it is singly robust.

2.2. Augmenting to Achieve Double Robustness

To achieve double robustness, we proceed in the usual fashion and augment the estimand as follows:

$$E(Y(0) \mid Z=1) = E(Y(0) - \mu(0,X) \mid Z=1) + E(\mu(0,X) \mid Z=1)$$

where $\mu(Z,X) = E(Y \mid Z,X)$ is the modelled conditional mean for Y. To estimate the first term, we use the weighted 'Monte Carlo' estimator

$$\widehat{E}(Y(0) - \mu(0,X) \mid Z=1) = \frac{\sum_{i=1}^{n}(1-Z_i)w(X_i)(Y_i - \mu(0,X_i))}{\sum_{i=1}^{n} Z_i}.$$

For the second term, we simply have $\widehat{E}(\mu(0,X) \mid Z=1) = \left\{\sum_{i=1}^{n} Z_i \mu(0,X_i)\right\}\left(\sum_{i=1}^{n} Z_i\right)^{-1}$ so

$$\widehat{E}(Y(0) \mid Z=1) = \left\{\sum_{i=1}^{n}(1-Z_i)w(X_i)(Y_i - \mu(0,X_i)) + Z_i\mu(0,X_i)\right\}\left(\sum_{i=1}^{n} Z_i\right)^{-1}$$

which yields the augmented average treatment effect on the treated estimator

$$\frac{\sum_{i=1}^{n}(Z_i - (1-Z_i)w(X_i))(Y_i - \mu(0,X_i))}{\sum_{i=1}^{n} Z_i}. \tag{1}$$

Following standard asymptotic theory for semiparametric estimators (Tsiatis, 2006), the augmented estimator is asymptotically normally distributed and locally efficient under the standard conditions listed at the start of Section 2.1. The standard error of the augmented estimator is no greater than that of the singly robust weighted estimator under correct specification of the outcome and treatment models. Variance estimators are easily derived (see Appendix A); however because the estimator is smooth, bootstrapping may also be used.

3. Simulations

We briefly demonstrate the double robustness of the augmented inverse weighted estimator empirically in a simulation study, before turning to an analysis of the average effect of smoking on sleep duration among smokers. We consider three scenarios: in the first, both the treatment (propensity score) model and the outcome model are correctly specified; in the second, the treatment model is correctly specified but the outcome model is not; in the final, the treatment model is incorrect while the outcome model is correctly specified.

Data are generated as follows: for varying sample sizes, we generate a confounding variable X ~Uniform(0,10), a treatment Z ~Bernoulli(expit(−1.75+0.3X)), and an outcome Y ~Normal(μ,1) where $\mu = 2Z + 0.6ZX + X - 0.2X^2$ in the first and third scenarios, and $\mu = 2Z + 0.6ZX + X - 0.5X \times \log(X) + 5X/(1+X)$ in the second scenario. Using a Monte Carlo approach, the true average treatment effect on the treated is found to be 5.749209 based on a population of 50,000,000 observations. In all instances, the outcome regression model is fit as a function of X and X^2. For the first two scenarios, the treatment model is fit as a logistic regression of Z on X

whereas in the third, X is omitted from the model. In each case, results are aggregated over 1000 simulated datasets.

Results are presented in Table 1, and are as would be anticipated. When the treatment model is correctly specified, all estimators converge to the true value as sample size increases, while the doubly robust weighted estimator dominates in terms of variability and root mean squared error. When the outcome model is correctly specified but the treatment model is not, the singly robust approaches are biased. Both the doubly robust weighting and doubly robust matching procedures are unbiased and nearly equivalent with respect to efficiency. However, in the setting where the treatment model was incorrectly specified, the matching estimators sometimes failed to yield an estimate using the `Matching` package in `R`: this occurred up to 15% of the time when n=50; bias, standard error, and root mean squared error are reported only for those simulated datasets for which an estimate was found. The weighted estimators did not, in any scenario, fail to return an estimate.

In settings where correct model specification assumptions are met, bias is low and coverage is near the nominal level. Estimating the doubly robust weighting estimator via a nonparametric bootstrap yielded estimated standard errors that were very close to the Monte Carlo standard errors; the Abadie-Imbens estimator of the matching estimator standard errors deviated from the Monte Carlo standard errors by upto 20%.

We would expect greater efficiency with the weighted estimator, which does not discard observations, relative to the matched estimator, which does. In our simulations, approximately 44% of observations were treated, so that relatively few observations were discarded through matching. By changing the data generating mechanism so that only 25% were treated by setting $Z \sim \text{Bernoulli}(\text{expit}(-2.75 + 0.3X))$, little change in the relative efficiency was noted, however the frequency with which the matching failed to return an estimate increased somewhat to 18.1% for $n = 50$. It is perhaps unsurprising that matching does not perform as well as a smooth weighting, as matching can be viewed as a particular (extreme) form of weighting where weights are constrained to be binary.

4. Data Analysis

Poor sleep quality and duration has been associated with a number of negative health consequences including increased risk of mortality and cardiovascular outcomes (Cappuccio et al., 2010, 2011). There is some evidence to suggest that smoking worsens sleep quality (Phillips & Danner, 1995; Bellatorre et al., 2017). We investigate the impact of current smoking on sleep duration using the 2011-2012 wave of the U.S. National Health and Nutrition Examination Study (NHANES).

Using a propensity score model that adjusts for age (natural and log-transformed scale), race (Black, Hispanic, Mexican, White, Other), education (eighth grade, grades 9 – 11, completed high school, some college, college graduate), poverty (a continuous variable) and body-mass index, balance is greatly improved through matching and even more so by weighting (results omitted). As noted in the context of g-estimation of optimal dynamic treatment regimes (Wallace et al., 2016), double robustness offers a means of model checking and validation when implemented in contexts other than matching. By holding one model fixed, and varying the other, the stability of the resulting estimates across re-samples of the data may be assessed. If, say, the treatment model is correctly specified, then the resulting estimator will be consistent whatever the outcome model and so varying the outcome model gives a means of assessing the similarity of estimates that result from different outcome models. We performed this model check using three different specifications of each of the treatment and the outcome model, using *simple* models (intercept only), *main* models whose results are presented in Table 2, and more *complex* models that included two-way interactions between smoking and each of education, poverty, and age as well as interactions between education and each of race and poverty. As Figure 1 demonstrates, the model specifications

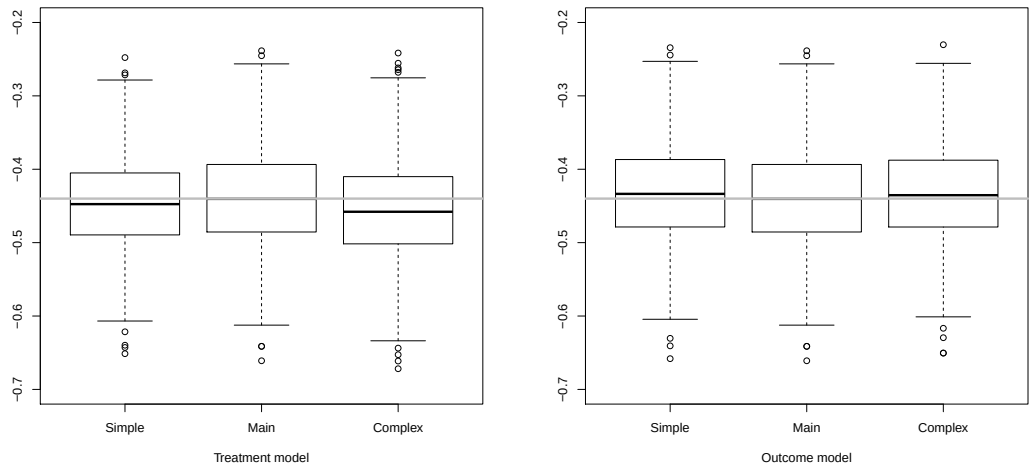

Figure 1. Estimates resulting from re-estimating the doubly robust weighted estimator under different specifications of the treatment model (left) whilst fixing the outcome model to the “main" specification, and the outcome model whilst fixing the treatment model to the “main" specification (right) under 1000 resamples of the NHANES (2011-2012) data. The thick grey line indicates the benchmark treatment effect.

used for the main analysis appear to be good, as the centre of the distribution of the estimates varies very little, even when an intercept-only model, which is almost surely incorrectly specified, is used for one of the two model specifications. The estimates are particularly stable when varying the treatment model, suggesting that the outcome model is correctly specified.

The doubly robust weighted estimator suggests that smokers lose approximately 26 minutes of sleep (-0.440$\times$60 minutes, 95% confidence interval: 18-35) due to smoking. The doubly robust matching estimate is quite similar (23 minutes), while the two singly robust estimates are larger at 31 minutes, although these values all lie within the confidence interval of the doubly robust weighting estimator. These results suggest a significant impact of smoking on sleep. Causal conclusions must be tempered by the fact that important confounders such as physical activity and caffeine intake were not available for this analysis.

While the NHANES analysis has revealed estimates in the expected direction, indicating a negative impact of smoking on sleep duration, the “true" value is unknown. In Appendix B, we provide a second data analysis using a benchmark dataset in which the true value of the association has been estimated from a randomized study.

5. Discussion

We have used an importance sampling derivation to arrive at an augmented weighted estimator for the average treatment effect on the treated that is doubly robust, regular, and locally efficient under the standard assumptions of semiparametric causal inference. We have demonstrated its use for continuous outcomes; the estimator extends naturally to binary or count outcomes to estimate a risk difference. The corresponding risk ratio estimator is $\sum_{i=1}^{n} Z_i Y_i \left\{ \sum_{i=1}^{n} (1 - Z_i) w(X_i)(Y_i - \mu(0, X_i)) \right\}^{-1}$. An augmented weighted estimator for standardized risk differences

and ratios for binary outcomes was independently proposed (Shinozaki & Matsuyama, 2015), albeit from an estimating equations viewpoint which lacks the intuitive foundation that our importance sampling perspective supplies. The weighted (importance sampling) estimator is straightforward to study theoretically, unlike matching estimators, and straightforward to compute, with the added advantage of having model checks afforded by the double robustness. In this paper we have examined only the use of conventional regression models for the propensity score, but more complex models and procedures (e.g. flexible parametric models, machine learning or ensemble approaches) may also be used

Acknowledgement

This work is supported by Discovery Grants from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Abadie, A & Imbens, GW (2006), 'Large sample properties of matching estimators for average treatment effects,' *Econometrica*, **74**(1), pp. 235–267.

Abadie, A & Imbens, GW (2011), 'Bias-corrected matching estimators for average treatment effects,' *Journal of Business & Economic Statistics*, **29**(1), pp. 1–11.

Bellatorre, A, Choi, K, Lewin, D, Haynie, D & Simons-Morton, B (2017), 'Relationships between smoking and sleep problems in black and white adolescents,' *Sleep*, **40**, p. zsw031.

Boukhris, T, Sheehy, O, Mottron, L & A, B (2016), 'Antidepressant use during pregnancy and the risk of autism spectrum disorder in children,' *JAMA Pediatrics*, **170**(2), pp. 117–124.

Cappuccio, FP, Cooper, D, D'Elia, L, Strazzullo, P & Miller, MA (2011), 'Sleep duration predicts cardiovascular outcomes: A systematic review and meta-analysis of prospective studies,' *European Heart Journal*, **32**, p. 1484 – 1492.

Cappuccio, FP, D'Elia, L, Strazzullo, P & Miller, MA (2010), 'Sleep duration and all-cause mortality: A systematic review and meta-analysis of prospective studies,' *Sleep*, **33**, p. 585 – 592.

Dehejia, RH & Wahba, S (1999), 'Causal effects in non-experimental studies: Reevaluating the evaluation of training programs,' *Journal of the American Statistical Association*, **94**, pp. 1053 – 1062.

Hahn, J (1998), 'On the role of the propensity score in efficient semiparametric estimation of average treatment effects,' *Econometrica*, **66**, pp. 315 – 331.

Hainmueller, J (2012), 'Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies,' *Political Analysis*, **20**, pp. 25 – 46.

Hernán, MA, Brumback, B & Robins, JM (2000), 'Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men,' *Epidemiology*, **11**, pp. 561 – 570.

Hirano, K, Imbens, GW & Ridder, G (2003), 'Efficient estimation of average treatment effects using the estimated propensity score,' *Econometrica*, **71**(4), pp. 1161–1189.

Hubbard, AE, Jewell, NP & van der Laan, MJ (2011), *Direct effects and effect among the treated*, Springer, New York.

Lalonde, R (1986), 'Evaluating the econometric evaluations of training programs,' *American Economic Review*, **76**, pp. 604–620.

Leacy, FP & Stuart, EA (2014), 'On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study,' *Statistics in Medicine*, **33**(20), pp. 3488–3508.

Phillips, BA & Danner, FJ (1995), 'Cigarette smoking and sleep disturbance,' *Archives of Internal Medicine*, **155**, pp. 734 – 737.

Pirracchio, R, Carone, M, Rigon, MR, Caruana, E, Mebazaa, A & Chevret, S (2016), 'Propensity score estimators for the average treatment effect and the average treatment effect on the treated may yield very different estimates,' *Statistical Methods in Medical Research*, **25**(5), pp. 1938–1954.

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

Sekhon, J (2011), 'Multivariate and propensity score matching software with automated balance optimization: The Matching package for R,' *Journal of Statistical Software*, **42**, pp. 1 – 52.

Shinozaki, T & Matsuyama, Y (2015), 'Doubly robust estimation of standardized risk difference and ratio in the exposed population,' *Epidemiology*, **26**, pp. 873 – 877.

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

Wallace, MP, Moodie, EEM & Stephens, DA (2016), 'Model assessment in dynamic treatment regimen estimation via double robustness,' *Biometrics*, **72**(3), pp. 855–864.

Wunsch, C & Lechner, M (2008), 'What did all the money do? on the general ineffectiveness of recent West German labour market programmes,' *Kyklos*, **61**(1), pp. 134–174.

Appendix A: Variance Derivation

For individual i, let

$$U_i(\beta, \alpha) = \{\omega_i(\alpha)(y_i - \mu_i(\beta)\}$$

denote the individual contribution to the (weighted) estimating function for the average treatment effect on the treated in (1), where $\mu_i(\beta)$ is a parametric function of the treatment, the covariates, and parameters β specifying the mean outcome among the treated, and $\omega_i(\alpha)$ is a weight based on a treatment model parameterized by α.

Letting $\dot{\ell}_\alpha(\alpha)$ and $\ddot{\ell}_\alpha(\alpha)$ denote the score function of the treatment model and its derivative, respectively, and taking

$$U_{\text{adj}}(\beta, \alpha) \quad = \quad U(\beta, \alpha) - \mathbb{E}\left[\frac{\partial}{\partial \alpha} U(\beta, \alpha)\right] \left(\mathbb{E}\left[\ddot{\ell}_\alpha(\alpha)\right]\right)^{-1} \dot{\ell}_\alpha(\alpha)$$

yields a variance formula for estimators of parameters in the outcome regression model, $\widehat{\beta}$, of

$$\text{Var}[\widehat{\beta}] = \mathbb{E}\left[\left\{\left(\mathbb{E}\left[\frac{\partial}{\partial \beta} U_{\text{adj}}(\beta, \alpha)\right]\right)^{-1} U_{\text{adj}}(\beta, \alpha)\right\}^{\otimes 2}\right].$$

The weighted estimator is smooth and so bootstrapping may be used to estimate standard errors; performance of the bootstrap estimator is good even in small samples (see §3).

Appendix B: Benchmark Analysis

In 1986, Lalonde published a landmark paper examining the impact of the National Supported Work Demonstration labour training program on subsequent income. Thanks to the availability of a randomized component to the intervention, a 'true' effect of the program was known and thus evaluation of statistical methods applied to a composite dataset that includes both randomized and non-randomized individuals can be compared to this benchmark value. Lalonde concluded that many traditional approaches to estimation were not able to recover the benchmark estimate of $1,794 (Lalonde, 1986); subsequent work has shown that propensity score methods such as stratification and matching are successful in estimating the benchmark value (Dehejia & Wahba, 1999; Sekhon, 2011).

Using a propensity score model that adjusts for age, education (as a linear and quadratic term), and indicators for being black, being Hispanic, being married, and having no high school degree, balance is greatly improved through weighting (see Table 3, and also largely improved by matching except for the variable age, where some imbalance persists (results not shown). As seen in Table 4, the singly- and doubly-robust weighting estimates are remarkably close to the benchmark value, differing by less than $25. The matching estimators are not quite so close, differing by $188 and $165 for the singly robust and doubly robust versions, respectively. Efficiency gains through weighting are also evident: the standard error of the weighted estimators relative to their matched counterparts are 76.7% and 79.5% for the robust and doubly robust estimators, respectively. As in the NHANES data, we used double robustness to assess estimator consistency, and found that the model specification appear to be good (results not shown).

Appendix C: Sample code

We provide code to generate and analyze a single dataset according to the structure used in §3.

```
n<-1000
al<-c(-1.75,0.3)
theta<-c(2,0.6)

ORgood <- 1 # indicator of whether fitting a correctly-specified outcome model
PSgood <- 0 # indicator of whether fitting a correctly-specified treatment model

Nbig <- 50000000
x<-runif(Nbig,0,10)
pi.vec<-1/(1+exp(-cbind(1,x) %*% al))
Z<-rbinom(Nbig,1,pi.vec)
att <- theta[1] + theta[2]*mean(x[Z==1])
att
att <- 5.749209  # based on n=50000000 as per above


h.func<-function(xv){return(xv-0.5*xv*log(xv)+5*xv/(1+xv))}
if(ORgood==1) { h.func<-function(xv){return(xv-0.2*xv*xv)} }
```


Prepared using staauth.cls

```
# Generate data
x<-runif(n,0,10)
pi.vec<-1/(1+exp(-cbind(1,x) %*% al))
Z<-rbinom(n,1,pi.vec)
Y<-cbind(Z,Z*x)%*%theta+h.func(x)+rnorm(n)

## Estimate treatment model and compute IP weight
pi.hat<-fitted(glm(Z~1,family=binomial))
if(PSgood==1) { pi.hat<-fitted(glm(Z~x,family=binomial)) }
w.hat<-pi.hat/(1-pi.hat)

## Singly-robust estimation
sr.att <-sum((Z-(1-Z)*w.hat)*Y)/sum(Z)          #IPW

## Doubly-robust estimation
dat0 <- dat <- data.frame(cbind(Y,Z,x)); dat0$Z<-0
mu0 <- predict(lm(Y~Z*x+I(x^2),data=dat),dat0)
dr.att<-sum((Z-(1-Z)*w.hat)*(Y-mu0))/sum(Z)     #AIPW
```

Table 1. Absolute bias, root mean squared error (rMSE), the ratio of the estimated standard error to the Monte Carlo standard error (SE/MCSE) and coverage of 95% confidence intervals of the singly robust (SR) matching and weighting estimators, and the doubly robust (DR) matching and augmented weighting estimators when both the outcome and propensity score models are correctly specified, when only the propensity score is correctly specified, and when only the outcome mode is correctly specified.

		Both models correct			Only propensity score correct			Only outcome model correct		
Method	n	bias	rMSE	SE/MCSE	Cover	bias	rMSE	SE/MCSE	Cover	bias	rMSE	SE/MCSE	Cover
SR matching	50	0.15	0.67	0.94	91.40	0.02	0.55	1.09	95.80	2.33	2.44	0.96	8.93
SR weighting	50	0.02	1.24	1.15	92.30	0.00	0.57	1.27	96.80	2.32	2.43	1.01	12.00
DR matching	50	0.00	0.58	1.04	95.10	0.04	0.60	1.04	96.10	0.01	0.50	0.95	93.18
DR weighting	50	0.01	0.50	1.01	92.90	0.01	0.50	1.07	96.20	0.02	0.49	1.02	94.90
SR matching	100	0.03	0.41	1.10	96.00	0.02	0.40	1.13	96.60	2.30	2.36	0.93	0.34
SR weighting	100	0.02	0.81	1.03	94.40	0.00	0.35	1.09	96.10	2.30	2.36	0.97	0.60
DR matching	100	0.02	0.40	1.12	96.80	0.00	0.40	1.13	96.80	0.02	0.34	1.00	93.50
DR weighting	100	0.02	0.34	1.00	94.70	0.01	0.34	1.02	95.30	0.02	0.33	1.01	94.20
SR matching	250	0.01	0.25	1.12	97.30	0.01	0.24	1.19	97.00	2.30	2.32	0.98	0.00
SR weighting	250	0.01	0.47	1.01	95.60	0.00	0.21	1.07	96.30	2.30	2.32	1.02	0.00
DR matching	250	0.00	0.25	1.14	97.20	0.00	0.24	1.19	96.90	0.00	0.21	1.02	95.45
DR weighting	250	0.00	0.22	0.98	94.50	0.00	0.21	1.05	96.20	0.00	0.21	1.00	94.40
SR matching	500	0.01	0.17	1.18	97.80	0.01	0.17	1.18	97.50	2.31	2.32	0.97	0.00
SR weighting	500	0.00	0.34	0.97	94.30	0.01	0.15	1.00	94.70	2.31	2.32	1.00	0.00
DR matching	500	0.00	0.17	1.18	97.90	0.01	0.17	1.18	97.40	0.00	0.14	1.07	95.85
DR weighting	500	0.00	0.15	1.01	95.10	0.01	0.15	1.00	94.50	0.00	0.14	1.03	95.20
SR matching	1000	0.01	0.12	1.22	98.20	0.01	0.12	1.16	97.80	2.30	2.31	0.96	0.00
SR weighting	1000	0.00	0.23	1.01	95.60	0.00	0.11	0.97	93.30	2.30	2.31	0.99	0.00
DR matching	1000	0.01	0.12	1.22	98.40	0.01	0.12	1.16	97.80	0.01	0.11	1.04	94.93
DR weighting	1000	0.01	0.10	1.04	96.40	0.00	0.11	0.97	93.20	0.01	0.11	1.00	94.30

Table 2. Point estimates, standard errors (SE), and 95% confidence intervals (CI) in the study of the effect of smoking on sleep duration among smokers in the 2011-2012 NHANES data.

Estimator	Estimate	SE	95% CI
Singly robust matching	-0.517	0.094	(-0.701, -0.333)
Singly robust weighting	-0.387	0.075	(-0.533, -0.240)
Doubly robust matching	-0.521	0.094	(-0.706, -0.336)
Doubly robust weighting	-0.440	0.069	(-0.575, -0.305)

Table 3. Standardized mean difference between those who did and did not receive the labour training in the Lalonde data in the original sample, a matched sample, and an inverse probability weighted sample.

Variable	Original	Matched	Weighted
Age	0.176	0.053	0.002
Race	0.099	0.111	0.039
Education	0.586	0.173	0.018
Poverty	0.411	0.031	0.012
BMI	0.213	0.037	0.011

Table 4. Point estimates, standard errors (SE), and 95% confidence intervals (CI) for the average treatment effect on the treated in the Lalonde data.

Estimator	Estimate	SE	95% CI
Singly robust matching	1982.28	738.17	(535.46, 3429.09)
Singly robust weighting	1816.67	587.16	(665.83, 2967.51)
Doubly robust matching	1959.75	742.29	(504.87, 3414.63)
Doubly robust weighting	1771.07	568.98	(655.87, 2886.26)