Estimating Average Treatment Effects: Supplementary Analyses and Remaining Challenges

By Susan Athey, Guido Imbens, Thai Pham, and Stefan Wager*

I. Introduction

There is a large literature in econometrics and statistics on semiparametric estimation of average treatment effects under the assumption of unconfounded treatment assignment. Recently this literature has focused on the setting with many covariates, where regularization of some kind is required. In this article we discuss some of the lessons from the earlier literature and their relevance for the many covariate setting.

II. The Set Up

We are interested in estimating an average treatment effect in a setting with a binary treatment. We use the potential outcome or Rubin Causal Model set up (Rubin [1974], Holland [1986], Imbens and Rubin [2015]). Each unit in a large population is characterized by a pair of potential outcomes $(Y_i(0), Y_i(1))$, with the estimand equal to the average causal effect:

$$\tau = \mathbf{E}[Y_i(1) - Y_i(0)],$$

or the average effect for the treated, $\tau_t = \mathbf{E}[Y_i(1) - Y_i(0)|W_i = 1]$. The treatment assignment for unit i is $W_i \in \{0, 1\}$. For each unit in a random sample from the population we observe the treatment received and the realized outcome,

$$Y_i^{\text{obs}} = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1, \end{cases}$$

and pretreatment variables or features X_i. To identify τ we assume unconfoundedness (Rosenbaum and Rubin [1983])

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \;\Big|\; X_i,$$

and overlap of the covariate distributions,

$$e(x) \in (0, 1),$$

where the propensity score (Rosenbaum and Rubin [1983]) is $e(x) = \text{pr}(W_i = 1|X_i = x)$. Define $p = \mathbf{E}[W_i]$, $\mu(w, x) = \mathbf{E}[Y_i(w)|X_i = x]$, $\mu_w = \mathbf{E}[Y_i(w)]$, and $\sigma^2(w, x) = \mathbf{V}(Y_i(w)|X_i = x)$. The efficient score for τ, which plays a key role in the discussion, is

$$\phi(y, w, x; \tau, \mu(\cdot, \cdot), e(\cdot)) = w\frac{y - \mu(1, x)}{e(x)} -$$

$$(1 - w)\frac{y - \mu(0, x)}{1 - e(x)} + \mu(1, x) - \mu(0, x) - \tau,$$

(Hahn [1998]) and the implied semiparametric variance bound is

$$\mathbf{AV} = \mathbf{E}\left[\phi(Y_i^{\text{obs}}, W_i, X_i; \tau, \mu(\cdot, \cdot), e(\cdot))^2\right].$$

For the average effect for the treated, τ_t, the efficient score function is

$$\phi'(y, w, x; \tau_t, \mu(\cdot, \cdot), e(\cdot)) = \frac{w}{p}(y - \mu(0, x) - \tau)$$

$$+\frac{(1 - w)e(x)}{p(1 - e(x))}(y - \mu(0, x)).$$

A wide range estimators for τ have been proposed in this setting, (see for a review Imbens and Wooldridge [2009]). Some of the proposed estimators rely on matching (Abadie and Imbens [2006]). Others rely on different characterizations of the average treatment effect, using the propensity score,

* Athey: Graduate School of Business, Stanford University, athey@stanford.edu. Imbens: Graduate School of Business, Stanford University, imbens@stanford.edu. Pham: Graduate School of Business, Stanford University, thaipham@stanford.edu. Wager Graduate School of Business, Stanford University, swager@stanford.edu.
We are grateful for discussions with Jasjeet Sekhon

(Hirano et al. [2001]),

$$\tau = \mathbf{E}\left[\frac{Y_i^{\text{obs}} \cdot W_i}{e(X_i)} - \frac{Y_i^{\text{obs}} \cdot (1 - W_i)}{1 - e(X_i)}\right],$$

the conditional expectation of the outcome,

$$\tau = \mathbf{E}\Big[\mu(1, X_i) - \mu(0, X_i)\Big],$$

(Hahn [1998]), or the efficient score representation

$$\tau = \mathbf{E}\Bigg[W_i \frac{Y_i^{\text{obs}} - \mu(1, X_i)}{e(X_i)} +$$

$$(1-W_i)\frac{Y_i^{\text{obs}} - \mu(0, X_i)}{1 - e(X_i)} + \mu(1, X_i) - \mu(0, X_i)\Bigg]$$

(van der Vaart [2000], Van Der Laan and Rubin [2006], Chernozhukov et al. [2016]), and similar estimators for the average effect for the treated.

Because the unconfoundedness assumption imposes no restrictions on the joint distribution of the observed variables $(Y_i^{\text{obs}}, W_i, X_i)$, it follows by the general results for semiparametric estimators in Newey [1994] that all three approaches, using suitable nonparametric estimators of the propensity score and/ or the conditional expectations of the potential outcomes, reach the semiparametric efficiency bound.

III. Four Issues

First we wish to raise four issues that have come up in the fixed-number-of-covariate case, and which are even more relevant in the many covariate setting.

A. Double Robustness

A consistent finding from the observational study literature with a fixed number of pretreatment variables is that the best estimators in practice involve both estimation of the conditional expectations of the potential outcomes and estimation of the propensity score, rendering them less sensitive to estimation error in either, although this does not appear to be necessary in the case of a randomized experiment, where simply estimating the conditional expectation of the outcome is sufficient (Wager et al. [2016]). An important notion in the observational study literature is that of so called "doubly robust" estimators (Robins and Rotnitzky [1995], Robins et al. [1995], Scharfstein et al. [1999]) that rely for consistency only on consistent estimation of either the propensity score or the conditional outcome expectations, but not both. As a simple example to develop intuition for this, consider the standard omitted variable bias formule when estimating a regression function

$$Y_i^{\text{obs}} = W_i\tau + X_i'\beta + \varepsilon_i.$$

Omitting X_i from this regression leads to a bias if the included regressor W_i and the omitted regressor X_i are correlated, and the omitted regressor has a non-zero coefficient. In this setting weighting by the inverse of, or conditioning on, the propensity score removes the correlation between W_i and X_i. Therefore it eliminates the sensitivity to the parametric form in which X_i is included, without introducing bias if the weights are misspecified but the regression function is correct.

Here we view estimators as at least approximately doubly robust if they attempt to adjust directly for the association between the treatment indicator and the covariates, through balancing, weighting, or otherwise, and adjust directly for the association between the potential outcomes and the covariates. There are multiple ways of obtaining such estimators. One can do so by subclassification on the propensity score in combination with regression within the subclasses, or weighting in combination with regression. For example, suppose we parametrize the conditional means as $\mu(w, x) = w\tau + x'\beta$, and the propensity score as $e(x) = 1/(1 + \exp(x'\gamma))$, and estimate the regression by weighted linear regression with weights equal to $W_i/\sqrt{e(X_i;\hat{\gamma})} + (1 - W_i)/\sqrt{(1 - e(X_i;\hat{\gamma}))}$, then the estimator for τ is consistent if either the propensity score or the conditional expectations of the potential outcomes are

correctly specified. Similarly, using the efficient score, if we estimate the average treatment effect by solving

$$\frac{1}{N}\sum_{i=1}^{N}\phi\Big(Y_i^{\rm obs},W_i,X_i;\tau,\hat{\mu}(\cdot,\cdot),\hat{e}(\cdot)\Big)=0,$$

as a function of τ given estimators $\hat{\mu}(\cdot,\cdot)$ and $\hat{e}(\cdot)$, then as long as either the estimator for either $\mu(w,x)$ or $e(x)$ is consistent, the resulting estimator for τ is consistent.

If we use general nonparametric estimators for $\mu(\cdot),\cdot)$ and $e(\cdot)$, this last estimator also has the property that the estimator for the finite dimensional component τ is asymptotically uncorrelated with the estimator for the nonparametric components $\mu(w,x)$ and $e(x)$. This orthogonality property (Neyman and Scott [1948], Lancaster [2000], Chernozhukov et al. [2016]) is an important feature of the targeted maximum likelihood approach in Van Der Laan and Rubin [2006] and Van der Laan and Rose [2011], and, as noted in Chernozhukov et al. [2016], follows directly from the representation of the estimator in terms of the efficient score. Note that the properties are distinct: not all estimators that have the orthogonality property are doubly robust.

B. *Modifying the Estimand*

A second issue is the choice of estimand. Much of the literature has focused on the average treatment effect $\mathbf{E}[Y_i(1)-Y_i(0)]$, or the average effect for the treated. A practical concern is that these estimands may be difficult to estimate precisely if the propensity score is close to zero for a substantial fraction of the population. This is a particular concern in settings with many covariates because regularization based on prediction criteria may downplay biases that are present in estimation of $\mu(w,x)$ in parts of the (w,x) space with few observations, even if those values are important for the estimation of the average treatment effect. In that case one may wish to focus on a weighted average effect of the treatment. One can do so by trimming or weighting. Crump et al. [2006, 2009] and Li et al. [2014] suggest estimating

$$\tau_{\omega(\cdot)}=\frac{\mathbf{E}\left[\omega(X_i)\cdot(Y_i(1)-Y_i(0))\right]}{\mathbf{E}\left[\omega(X_i)\right]},$$

for $\omega(x)=e(x)(1-e(x))$ or $\omega(x)=\mathbf{1}_{\alpha<e(x)<1-\alpha}$. The semiparametric efficiency bound for $\tau_{\omega(\cdot)}$ is (Hirano et al. [2001])

$$\mathbf{AV}=\frac{1}{\mathbf{E}[\omega(X_i)^2]}\mathbf{E}\Bigg[\frac{\omega(X_i)^2\sigma^2(1,X_i)}{e(X_i)}+\frac{\omega(X_i)^2\sigma^2(0,X_i)}{1-e(X_i)}+\omega(X_i)^2\left(\mu(1,X_i)-\mu(0,X_i)-\tau_{\omega(\cdot)}\right)^2\Bigg],$$

which can be an order of magnitude smaller than the asymptotic variance bound for τ itself.

In settings with limited or no heterogeneity in the treatment effects as a function of the covariates, these weights are particularly helpful and the weights $\omega(x)=e(x)(1-e(x))$ lead to efficient estimators for τ in that case.

C. *Weighting versus Balancing*

Although weighting by the inverse of the treatment assignment balances pretreatment variables in expectation, it does not do so in finite samples. Recently there have been a number of estimators proposed that focus directly on balancing the pretreatment variables, bypassing estimation of the propensity score (Hainmueller [2012], Zubizarreta [2015], Graham et al. [2012, 2016], Athey et al. [2016]). Specifically, given a set of pretreatment variables X_i, one can look for a set of weights λ_i such that

$$\frac{1}{N_t}\sum_{i=1}^{N}\lambda_i\cdot W_i\cdot X_i\approx\frac{1}{N_c}\sum_{i=1}^{N}\lambda_i\cdot(1-W_i)\cdot X_i,$$

where N_c and N_t are the number of control and treated units respectively. The advantage of such weights is that they eliminate any biases associated with linear and additive effects in the pretreatment variables in

the estimator

$$\hat{\tau} = \frac{\sum_{i=1}^N \lambda_i W_i Y_i^{\rm obs}}{\sum_{i=1}^N \lambda_i W_i} - \frac{\sum_{i=1}^N \lambda_i (1-W_i) Y_i^{\rm obs}}{\sum_{i=1}^N \lambda_i (1-W_i)},$$

whereas using the propensity score weights $\lambda_i = W_i/e(X_i) + (1-W_i)/(1-e(X_i))$ does so only in expectation.

D. Sensitivity

Consider the simple difference in average outcomes by treatment status as an estimator for the average treatment effect. The bias in this estimator arises from the presence of pretreatment variables that are associated with both the treatment and the potential outcomes. Pretreatment variables that are associated solely with the treatment, or solely with the potential outcomes may make it difficult to estimate the propensity score or the conditional expectations of the potential outcomes, but such variables do not compromise the estimates of the average treatment effects. As a result it is not so much sparsity of the propensity score or sparsity of the conditional expectations, but sparsity of the product of the respective coefficients that matter. A summary measure of this association is the characterization of the bias as an expected value,

$$\begin{aligned} \mathrm{B} &= \big(\mathbf{E}[Y_i^{\rm obs}|W_i=1] - \mathbf{E}[Y_i^{\rm obs}|W_i=0]\big) - \tau \\ &= \frac{1}{p(1-p)})\mathbf{E}\left[b(X_i)\right], \end{aligned}$$

where the bias funcction $b(\cdot)$ is

$$b(x) = (e(x)-p)$$
$$\times(p(\mu(0,x)-\mu_0) + (1-p)(\mu(1,x)-\mu_1).$$

Hence the bias is proportional to the covariance of the propensity score and a weighted average of the conditional expectations of the potential outcomes,

$$\mathrm{Cov}(e(X_i), p\mu(0,X_i) + (1-p)\mu(1,X_i).$$

The bias function at x measures the contribution to the overall bias B, coming from units with $X_i = x$. It is flat in a randomized experiment, or in cases where the pretreatment variables are not associated with the outcome. Settings where the bias B is large relative to the difference in average outcomes by treatment effects, or $b(\cdot)$ is very variable, are particularly challenging for estimating τ. In our calculations below we report summary statistics of $\hat{b}(X_i)$, scaled by the standard deviation of the outcome.

IV. Three Estimators

Here we briefly discuss three of the most promising estimators that have been proposed for the case with many pretreatment variables. All three address biases from the association between pretreatment variables and potential outcomes and between pretreatment variables and treatment assignment. There are other estimators using machine learning methods that focus only on one of these associations, for example inverse propensity score weighting estimators that estimate the propensity score using machine learning methods (McCaffrey et al. [2004]), but we do not expect those to perform well. The first two estimators we discuss assume linearity of the conditional expectation of the potential outcomes in the, potentially many, covariates. How sensitive the results are in practice to this linearity assumption in settings with many covariates, where some of the covariates may be functions of underlying variables, remains to be seen.

A. The Double Selection Estimator (DSE)

Belloni et al. [2013] propose using LASSO (Tibshirani [1996]) as a covariate selection method. They do so first to select pretreatment variables that are important for explaining the outcome, and then to select pretreatment variables that are important for explaning the treatment assignment. They then combine the two sets of pretreatment variables and estimate a regression of the outcome on the treatment indicator and the union of the selected pretreatment variables.

B. The Approximate Residual Balancing Estimator (ARBE)

Athey et al. [2016] suggest using elastic net (Zou and Hastie [2005]) or LASSO (Tibshirani [1996]) to estimate the conditional outcome expectation, and then using an approximate balancing approach in the spirit of Zubizarreta [2015] to further remove bias arising from remaining imbalances in the pretreatment variables.

C. The Targeted Maximum Likelihood Estimator (TMLE) and the Double Machine Learning Estimator (DMLE)

In the general discussion of semiparametric estimation van der Vaart [2000] suggest estimating the finite dimensional component as the average of the influence function, with the infinite dimensional components estimated nonparametrically. In the specific context of estimation of average treatment effects Van Der Laan and Rubin [2006] propose this estimator as a special case of the targeted maximum likelihood approach suggesting various machine learning methods for estimation of the conditional outcome expectation and the propensity score. Chernozhukov et al. [2016], in the context of much more general estimation problems, propose a closely related estimator focusing on the orthogonality properties arising from the use of the efficient score. In the Chernozhukov et al. [2016] approach the sample is partitioned into K subsamples, with the nonparametric component estimated on one subsample, and the parameter of interest estimated as the average of the influence function over the remainder of the sample. This is repeated K times, and the estimators for the parameter of interest averaged to obtain the final estimator. We report both the simple version of the TMLE and the averaged version DMLE.

V. Outstanding Challenges and Practical Recommendations

Here we present some practical recommendations for researchers estimating treatment effects, and discuss some of the remaining challenges for the theoretical researchers.

A. Recommendations

The main recommendation is to report analyses beyond the point estimates and the associated standard errors. Supporting analyses should be presented to convey to the reader that the estimates are credible (Athey and Imbens [2016]). By credible we do not mean whether the unconfoundedness property holds, but whether the estimates effectively adjust for differences in the covariates. Here are four specific recommendations to do so.

1) **(Robustness)** Do not rely on a single estimation method. Many of the methods have attractive properties under slightly different sets of regularity conditions but rely on the same fundamental set of identifying assumptions. These regularity conditions are difficult to assess in practice. Therefore, if the substantive results are not robust to the specific choice of estimator, it is unlikely that the results are credible.

2) **(Overlap)** Assess concerns with overlap by comparing the variance bound for τ and $\tau_{\omega(\cdot)}$ for a choice of $\omega(\cdot)$ that de-emphasizes parts of the covariate space with limited overlap. If there is a substantial efficiency difference between the τ and $\tau_{\omega(\cdot)}$, report results for both.

3) **(Bootstrap Bias)** Report bootstrap estimates of the bias of the estimator, calculated as the estimator minus the average of estimates based on bootstrap samples, created by randomly splitting the original sample into two equal-sized subsamples. Asymptotic results rely on bias components of the asymptotic distribution vanishing. Bootstrap estimates may shed light on the validity of such approximations. For example, it could reveal sensitivity to the choice of regularization parameter.

4) **(Specification Sensitivity)** Split the sample based on median values of each of the covariates in turn, estimate the parameter of interest on both subsamples and average the estimates to assess sensitivity to the model specification (e.g., Athey and Imbens [2015a]).

B. Some Illustrations

Here we illustrate these recommendations with three data sets widely used in the evaluation literature, the experimental Lalonde data, the non-experimental Lalonde data (LaLonde [1986], Dehejia and Wahba [1999]), both with ten covariates, and the Connors et al. [1996] heart catherization data, with 72 covariates.

Four each of the data sets we report six estimators, the simple difference in average outcomes by treatment status, the OLS estimator with all covariates, the DS estimator (Belloni et al. [2013]), the ARB estimator (Athey et al. [2016]), and the closely related TML and DML estimators (Van Der Laan and Rubin [2006], Chernozhukov et al. [2016]). In addition to the point estimates, we report simple bootstrap standard errors, a goodness of fit measure for the potential control outcome (the square root of the average squared error), the scaled bootstrap bias (SBB, calculated as the average difference between the estimates based on equal size sample splits and the overall estimate, scaled by the bootstrap standard error, and an estimate of the bias), $\hat{B}$, equal to the difference between the estimator and the naive estimator equal to the difference in average outcomes by treatment status. In addition we report average of the estimator based on sample splits., one for each covariate, where we split the sample by the median value of each covariate in turn. Given the splits we calculate the estimator for each of the two subsamples, and then average those. See Athey and Imbens [2015a] for details. We also report summary statistics of $\hat{b}(X_i)$, the average, the median and the 0.025 and 0.975 quantiles, based on random forest methods. We also present histograms of $\hat{b}(x)$ for the three data sets.

For the Connors et al. [1996] data the methods do vary substantially, with the four estimators (ignoring the naive difference in means and the ols estimator) ranging from 0.038 to 0.062. This range is substantial compare to the difference relative to the naive estimator of 0.074. Trimming does not reduce this range substantially. The bootstrap bias is as large as 34% of the standard error, so coverage of confidence intervals may not be close to nominal. Splitting systematically on the 70 covariates generates substantial variation in the estimates, with the standard deviation of the estimates of the same order of magnitude as the standard errors of the original estimates. The tentative conclusion is that under unconfoundedness the average effect is likely to be positive, but with a range substantially wider than that captured by the confidence intervals based on any of the estimators.

C. Challenges

There are now more credible methods available for estimating average treatment effects under unconfoundedness with many covariates than there used to be, but there remain challenges in making these methods useful to practitioners. Here are some of the challenges remaining.

1) **(Choice of Regularization)** The regularization methods used continue to be based on optimal prediction for the infinitely dimensional components of the influence function. Although in some cases this may be optimal in large samples, e.g., Wager et al. [2016], in many cases these methods do not focus on the ultimate object of interest, the average treatment effect, and the implication that not all errors in estimating the unknown functions matter equally. See for some discussion of this issue Athey and Imbens [2015b].

2) **(Choice of Prediction Methods)** The leading estimators allow for the use of many different prediction methods of the infinitely dimensional components, without guidance for practioners how to choose among these

TABLE 1—THREE ILLUSTRATIONS

Heart Catherization Data

					Covariate Split	
	$\hat{\tau}$	(s.e.)	trimmed	SBB	mean	s.t.d.
$\overline{Y}_t - \overline{Y}_c$	0.074	(0.014)	0.038	0.00	0.069	(0.015)
OLS	0.064	(0.014)	0.056	0.01	0.063	(0.009)
DSE	0.062	(0.014)	0.058	-0.24	0.059	(0.009)
ARBE	0.061	(0.015)	0.050	-0.16	0.060	(0.011)
TMLE	0.038	(0.012)	0.039	-0.07	0.042	(0.010)
DMLE	0.045	(0.014)	0.036	-0.29	0.042	(0.010)

		Quantiles				
	mean	0.025	0.25	0.5	0.75	.975
$\hat{b}(X_i)/\text{std}(Y_i)$	0.07	-1.29	-0.54	0.25	0.58	1.29

methods in practice.

3) **(Supporting Analyses)** There is more work needed on supporting analyses that are intended to provide evidence that in a particular data analysis the answer is credible.

REFERENCES

Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

Susan Athey and Guido Imbens. A measure of robustness to misspecification. *The American Economic Review*, 105(5):476–480, 2015a.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *arXiv preprint arXiv:1504.01132*, 2015b.

Susan Athey and Guido Imbens. The state of applied econometrics-causality and policy evaluation. *arXiv preprint arXiv:1607.00699*, 2016.

Susan Athey, Guido Imbens, and Stefan Wager. Efficient inference of average treatment effects in high dimensions via approximate residual balancing. *arXiv preprint arXiv:1604.07125*, 2016.

Alexandre Belloni, Victor Chernozhukov, Ivan Fernández-Val, and Chris Hansen. Program evaluation with high-dimensional data. *arXiv preprint arXiv:1311.2645*, 2013.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, et al. Double machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*, 2016.

Alfred F Connors, Theodore Speroff, Neal V Dawson, Charles Thomas, Frank E Harrell, Douglas Wagner, Norman Desbiens, Lee Goldman, Albert W Wu, Robert M Califf, et al. The effectiveness of right heart catheterization in the initial care of critically iii patients. *Jama*, 276(11):889–897, 1996.

Richard Crump, V Joseph Hotz, Guido Imbens, and Oscar Mitnik. Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand, 2006.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, pages 187–199, 2009.

Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062, 1999.

Bryan Graham, Christine Pinto, and Daniel Egel. Inverse probability tilting for moment condition models with missing data. *Review of Economic Studies*, pages 1053–1079, 2012.

Bryan Graham, Christine Pinto, and Daniel Egel. Efficient estimation of data combination models by the method of auxiliary-to-study tilting (ast). *Journal of Business and Economic Statistics*, 34 (2):288–301, 2016.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

Keisuke Hirano, Guido Imbens, Geert Ridder, and Donald Rubin. Combining panels with attrition and refreshment samples. *Econometrica*, pages 1645–1659, 2001.

Paul W Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–970, 1986.

Guido Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, pages 1–29, 2004.

Guido Imbens and Jeffrey Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86, 2009.

Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

Tony Lancaster. The incidental parameter problem since 1948. *Journal of econometrics*, 95(2):391–413, 2000.

Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing covariates via propensity score weighting. *arXiv preprint arXiv:1404.1785*, 2014.

Daniel F McCaffrey, Greg Ridgeway, and Andrew R Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4):403, 2004.

Whitney K Newey. The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382, 1994.

Jerzy Neyman and Elizabeth L Scott. Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, pages 1–32, 1948.

James Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(1):122–129, 1995.

James Robins, Andrea Rotnitzky, and L.P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(1): 106–121, 1995.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94 (448):1096–1120, 1999.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data.* Springer Science & Business Media, 2011.

Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

Aad W. van der Vaart. *Asymptotic Statistics.* Number 3. Cambridge Univ Pr, 2000.

Stefan Wager, Wenfei Du, Jonathan Taylor, and Robert J Tibshirani. High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, 113 (45):12673–12678, 2016.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2): 301–320, 2005.

José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110 (511):910–922, 2015.