American Economic Review: Papers & Proceedings 2017, 107(5): 278–281
https://doi.org/10.1257/aer.p20171042

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 on semiparametric estimation of average treatment effects under unconfounded treatment assignment. Here we discuss lessons from this literature for the many covariate setting, and propose some supplementary analyses to assess the credibility of the analyses. Using the potential outcome or Rubin Causal Model setup (Imbens and Rubin 2015), each unit is characterized by the potential outcomes $(Y_i(0), Y_i(1))$, with the interest in the average causal effect: $\tau = E[Y_i(1) - Y_i(0)]$, or the average effect for the treated. The treatment assignment is $W_i \in \{0, 1\}$. We observe W_i and the realized outcome, $Y_i^{obs} = Y_i(W_i)$ and pretreatment variables X_i. We assume unconfoundedness (Rosenbaum and Rubin1983):

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i,$$

and overlap of the covariate distributions, $e(x) \in (0, 1)$, where the propensity score (Rosenbaum and Rubin 1983) is $e(x) = \Pr(W_i = 1 | X_i = x)$. Define $p = E[W_i]$, $\mu(w, x) = E[Y_i(w) | X_i = x]$, $\mu_w = E[Y_i(w)]$, and $\sigma^2(w, x) = V(Y_i(w) | X_i = x)$. The efficient score for τ 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). A number of estimators for τ have been proposed in this setting, relying on different characterizations of τ:

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

$$\tau = E[\mu(1, X_i) - \mu(0, X_i)],$$

$$\tau = E\left[W_i\frac{Y_i^{obs} - \mu(1, X_i)}{e(X_i)} - 1(1 - W_i)\frac{Y_i^{obs} - \mu(0, X_i)}{1 - e(X_i)} + \mu(1, X_i) - \mu(0, X_i)\right].$$

Because unconfoundedness imposes no restrictions on the joint distribution of the observed variables, it follows by Newey (1994) that all three approaches, substituting suitable nonparametric estimators of $e(\cdot)$ and $\mu(\cdot, \cdot)$, reach the semiparametric efficiency bound.

II. Four Issues

A. *Double Robustness*

A finding from the literature is that the best estimators in practice involve both estimation of $\mu(w, x)$ and $e(x)$, making them "doubly robust" (Robins and Rotnitzky 1995). One such

*Athey: Graduate School of Business, Stanford University, (e-mail: athey@stanford.edu); Imbens: Graduate School of Business, Stanford University, (e-mail: imbens@stanford.edu); Pham: Graduate School of Business, Stanford University, (e-mail: thaipham@stanford.edu); Wager: Department of Statistics, Columbia University, and Graduate School of Business, Stanford University (e-mail: swager@stanford.edu). We are grateful for discussions with Jasjeet Sekhon. We provide additional discussion in the working paper version (Athey et al. 2017).

†Go to https://doi.org/10.1257/aer.p20171042 to visit the article page for additional materials and author disclosure statement(s).

estimator for τ sets the average of the efficient score equal to zero as a function of τ given estimators $\hat{\mu}(\cdot,\cdot)$ and $\hat{e}(\cdot)$. As long as either the estimator for either $\mu(w,x)$ or $e(x)$ is consistent, the resulting estimator for τ is consistent.

B. *Modifying the Estimand*

A practical concern is that the average treatment effect may be difficult to estimate precisely if the propensity score is close to zero for a substantial fraction of the population. Then one may wish to focus on a weighted average effect of the treatment

$$\tau_{\omega(\cdot)} = E\big[\omega(X_i)\big(Y_i(1) - Y_i(0)\big)\big]/E\big[\omega(X_i)\big],$$

for $\omega(\cdot)$ that de-emphasize extreme values of the propensity score. The semiparametric efficiency bound for $\tau_{\omega(\cdot)}$ can be substantially smaller than the asymptotic variance bound for τ itself (Crump et al. 2009).

C. *Weighting versus Balancing*

Recently there have been a number of estimators proposed that focus directly on balancing the pretreatment variables (Hainmueller 2012; Zubizarreta 2015; Graham, Campos de Xavier Pinto, and Egel 2012). Specifically, given a set of pretreatment variables X_i, one can look for a set of weights λ_i such that the weighted average covariates are the same in the two treatment groups, which eliminates any biases associated with linear and additive effects in the pretreatment variables whereas using the propensity score weights does so only in expectation.

D. *Sensitivity*

The bias in the simple difference in average outcomes by treatment status as an estimator for the average treatment effect arises from the presence of pretreatment variables that are associated with both the treatment and the potential outcomes. As a result, it is the sparsity of the product of the respective coefficients that matter. Defining

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

we can characterize the bias as $E\big[b(X_i)\big]$, proportional to the covariance of the propensity score and the conditional expectations of the potential outcomes. Settings where $b(\cdot)$ is very variable, are particularly challenging for estimating τ.

III. Three Estimators

Here we briefly discuss three of the most promising estimators that have been proposed for the case with many pretreatment variables.

Belloni et al. (2013) propose the double selection estimator (DSE) using LASSO (Tibshirani 1996) as a covariate selection method. They first select pretreatment variables that are important for explaining the outcome, then select pretreatment variables that are important for explaining the treatment assignment and combine the two sets of pretreatment variables.

Athey, Imbens, and Wager (2016) propose the approximate residual balancing estimator (ARBE) using elastic net or LASSO 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.

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, leading to a doubly robust estimator (DRE) in the spirit of Robins and Rotnitzky (1995). In the specific context of estimation of average treatment effects van der Laan and Rubin (2006) propose a closely related estimator as a special case of the targeted maximum likelihood approach. Chernozhukov et al. (2016), in the context of much more general estimation problems, propose a closely related double machine learning estimator (DMLE) that also incorporates sample splitting to further improve the properties.

IV. Outstanding Challenges and Practical Recommendations

A. *Recommendations*

The main recommendation is to report analyses beyond the point estimates and the associated

Table 1—An Illustration Based on the Connors et al. (1996) Heart Catherization Data

	ATT	(SE)	Trimmed ATT	SBB	cov split Mean	cov split SD
Naïve	0.074	0.014	0.038	−0.002	0.073	0.009
OLS	0.064	0.014	0.056	0.704	0.063	0.004
DSE	0.062	0.014	0.057	−0.213	0.061	0.007
ARBE	0.061	0.015	0.050	−0.157	0.058	0.004
DRE	0.038	0.012	0.039	0.084	0.038	0.003
DMLE	0.037	0.014	0.036	0.341	0.044	0.005
				Quantiles		
	mean	0.025	0.250	0.500	0.750	0.975
$\hat{b}(X_i)/\text{std}(Y_i)$	0.07	−1.29	−0.054	0.25	0.58	1.31

standard errors. Supporting analyses (Athey and Imbens 2016) should be presented to convey to the reader that the estimates effectively adjust for differences in the covariates. Here are four specific recommendations to do so.

Robustness.—Do not rely on a single estimation method. If the substantive results are not robust to the specific choice of estimator relying on the same identification assumptions, it is unlikely that the results are credible.

Overlap.—Compare 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.

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 2015).

Scaled Bootstrap Bias.—Report estimates of the bias of the estimator, based on repeatedly half-sampling

B. *An Illustration*

We illustrate these recommendations with the Connors et al. (1996) heart catherization data, with 72 covariates, with additional illustrations in the working paper version (Athey et al. 2017). We report the simple difference in average outcomes by treatment status, the OLS estimator with all covariates, and the estimators DSE, ARBE, DRE, and DMLE. In addition we report simple bootstrap standard errors, the scaled bootstrap bias (SBB, based on half-sampling, scaled by the bootstrap standard error), the 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 and then average the estimates (Athey and Imbens 2015), and summary statistics of $\hat{b}(X_i)$.

The four main estimators range from 0.037 to 0.062. This range is substantial compared to the difference relative to the naïve estimator of 0.074, and relative to the standard error. Trimming does not reduce this range substantially. The scaled bootstrap bias is as large as 29 percent of the standard error, so coverage of confidence intervals may not be close to nominal. Splitting systematically on all covariates generates substantial variation in the estimates. The tentative conclusion is that under unconfoundedness the average effect is likely to be positive, but with a range substantially wider than suggested by the standard errors for any of the estimators.

C. *Challenges*

Choice of Regularization.—The regularization methods are based on optimal prediction rather than focusing on the ultimate object of interest, the average treatment effect, and do not take account of the fact that not all errors in estimating the unknown functions matter equally.

Choice of Prediction Methods.—The leading estimators allow for different prediction methods of the unknown functions, without guidance for practitioners on how to make this choice in practice.

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

REFERENCES

Athey, Susan, and Guido Imbens. 2015. "A Measure of Robustness to Misspecification." *American Economic Review* 105 (5): 476–80.

Athey, Susan, and Guido Imbens. 2016. "The State of Applied Econometrics—Causality and Policy Evaluation." https://arxiv.org/abs/1607.00699.

Athey, Susan, Guido W. Imbens, and Stefan Wager. 2016. "Efficient Inference of Average Treatment Effects in High Dimensions via Approximate Residual Balancing." Stanford Graduate School of Business Working Paper 3408.

Athey, Susan, Guido Imbens, Thai Pham, and Stefan Wager. 2017. "Estimating Average Treatment Effects: Supplementary Analyses and Remaining Challenges." https://arxiv.org/abs/1702.01250.

Belloni, Alexandre, Victor Chernozhukov, Ivan Fernandez-Val, and Christian Hansen. 2013. "Program Evaluation with High-Dimensional Data." https://arxiv.org/abs/1311.2645.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. 2016. "Double Machine Learning for Treatment and Causal Parameters." https://arxiv.org/abs/1608.00060.

Connors, Alfred F. Jr., Theodore Speroff, Neal V. Dawson, Charles Thomas, Frank E. Harrell, Douglas Wagner, Norman Desbiens, et al. 1996. "The Effectiveness of Right Heart Catheterization in the Initial Care of Critically Ill Patients." *Journal of the American Medical Association* 276 (11): 889–97.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. 2009. "Dealing with Limited Overlap in Estimation of Average Treatment Effects." *Biometrika* 96 (1): 187–99.

Graham, Bryan S., Cristine Campos de Xavier Pinto, and Daniel Egel. 2012. "Inverse Probability Tilting for Moment Condition Models with Missing Data." *Review of Economic Studies* 79 (3): 1053–79.

Hahn, Jinyong. 1998. "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica* 66 (2): 315–31.

Hainmueller, Jens. 2012. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis* 20 (1): 25–46.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge, UK: Cambridge University Press.

Newey, Whitney K. 1994. "The Asymptotic Variance of Semiparametric Estimators." *Econometrica* 62 (6): 1349–82.

Robins, James M., and Andrea Rotnitzky. 1995. "Semiparametric Efficiency in Multivariate Regression Models with Missing Data." *Journal of the American Statistical Association* 90 (429): 122–29.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B* (*Methodological*) 58 (1): 267–88.

van der Laan, Mark J., and Daniel Rubin. 2006. "Targeted Maximum Likelihood Learning." *International Journal of Biostatistics* 2 (1).

van der Vaart, A. W. 2000. *Asymptotic Statistics.* Cambridge, UK: Cambridge University Press.

Zubizarreta, Jose R. 2015. "Stable Weights that Balance Covariates for Estimation with Incomplete Outcome Data." *Journal of the American Statistical Association* 110 (511): 910–22.