Peter M. Aronow[1]

Data-Adaptive Causal Effects and Superefficiency

[1] Departments of Political Science and Biostatistics, Yale University, 77 Prospect St., New Haven, CT 06520, USA, E-mail: peter.aronow@yale.edu

Abstract: Recent approaches in causal inference have proposed estimating average causal effects that are local to some subpopulation, often for reasons of efficiency. These inferential targets are sometimes data-adaptive, in that they are dependent on the empirical distribution of the data. In this short note, we show that if researchers are willing to adapt the inferential target on the basis of efficiency, then extraordinary gains in precision can potentially be obtained. Specifically, when causal effects are heterogeneous, any asymptotically normal and root-n consistent estimator of the population average causal effect is superefficient for a data-adaptive local average causal effect.

Keywords: causal inference, superefficiency, data-adaptive target parameter, local average treatment effect

DOI: 10.1515/jci-2016-0007

1 Introduction

When causal effects are heterogeneous, then inferences depend on the population for which causal effects are estimated. Although population average causal effects have traditionally been the inferential targets, recent results have focused on estimating average causal effects that are *local* to some subpopulation for reasons of efficiency. These approaches include trimming observations based on the distribution of the propensity score [1], using regression adjustment to estimate reweighted causal effects [2–4], or implementing calipers for propensity-score matching [5, 6]. In some cases, the target parameter is dependent on the empirical distribution of the data, including cases where the researcher is explicitly conducting inference on, e.g., the average treatment effect among the treated conditional on the observed covariate distribution [7], or other causal sample functionals [8, 9], without revision to the estimator being used.

These approaches privilege efficiency in estimation over targeting population average causal effects, and often allow for the target to be defined on the basis of the observed data. We provide an example of how these approaches, taken to their extreme, can provide extraordinary gains in statistical certainty. We consider the case of a *data-adaptive* target parameter [10] that is allowed to vary with the data depending on which subpopulation's local average causal effect is best estimated. When treatment effects are heterogeneous, adaptively changing the target parameter on the basis of efficiency yields an unusual result: if the population average causal effect can be consistently estimated with a root-n consistent and asymptotically normal estimator $\hat{\theta}$, then the same estimator $\hat{\theta}$ is always superefficient (i.e., faster than root-n consistent) for a data-adaptive local average causal effect. Furthermore, with an additional regularity condition on mean square convergence, we show that the mean square error of $\hat{\theta}$ for a data-adaptive local average causal effect is of $o(n^{-1})$.

2 Results

Consider a full data probability distribution G with an associated causal effect distribution τ with finite expectation $\mathrm{E}_G[\tau]$, where $\mathrm{E}_G[.]$ denotes the expectation over the distribution G. Further denote the support of the distribution of τ as $\mathrm{Supp}_G[\tau]$. We impose a regularity condition on τ establishing non-degeneracy of τ.

Assumption 1: (Effect heterogeneity). $\min\left(\sup\left(\mathrm{Supp}_G[\tau]\right) - \mathrm{E}_G[\tau], \mathrm{E}_G[\tau] - \inf\left(\mathrm{Supp}_G[\tau]\right)\right) = c > 0.$

Assumption 1 is equivalent to assuming that causal effects are not constant across observations in the distribution G; i.e., causal effects are heterogeneous.

We do not observe the full data probability distribution G, but we observe an empirical distribution F_n. Suppose that, using F_n, we have a root-n consistent and asymptotically normal estimator of the average causal effect $\mathrm{E}_G[\tau]$, $\hat{\theta}$.

Peter M. Aronow is the corresponding author.

Definition 1: *An estimator $\hat{\theta}$ is root-n consistent and asymptotically normal for θ_0 if $\sqrt{n}(\hat{\theta} - \theta_0) = N(0, \sigma^2) + o_p(1)$, for some $0 < \sigma^2 < \infty$.*

We now define the target parameter, θ_{F_n}.

Definition 2: *Let the target parameter*

$$\theta_{F_n} = \begin{cases} \hat{\theta} & : |\hat{\theta} - \mathrm{E}_G[\tau]| \leq c \\ \mathrm{E}_G[\tau] + c & : \hat{\theta} - \mathrm{E}_G[\tau] > c \\ \mathrm{E}_G[\tau] - c & : \hat{\theta} - \mathrm{E}_G[\tau] < -c \end{cases},$$

where, as in Assumption 1, $c = \min\left(\sup\left(\mathrm{Supp}_G[\tau]\right) - \mathrm{E}_G[\tau], \mathrm{E}_G[\tau] - \inf\left(\mathrm{Supp}_G[\tau]\right)\right)$.

The target parameter adapts naturally to the closest value in an interval surrounding $\mathrm{E}_G[\tau]$, where the width of the interval is defined by the support of τ. We formalize how each θ_{F_n} is a local average treatment effect.

Proposition 1: *There exists a nonnegative weighting associated with each empirical distribution F_n, w_{F_n}, such that across all F_n, $\theta_{F_n} = \frac{\mathrm{E}_G[w_{F_n}\tau]}{\mathrm{E}_G[w_{F_n}]}$.*

A proof of Proposition 1 follows directly from the fact that a weighted mean can obtain any value in the interval defined by the infimum and supremum of its distribution's support. Proposition 1 asserts that across all realizations, the target parameter θ_{F_n} corresponds to an average causal effect for at least one subpopulation. (There in fact may be infinitely many subpopulations to which θ_{F_n} corresponds.) The composition of the subpopulation(s) associated with each θ_{F_n} is not directly knowable by the researcher and may vary across realizations of the data.

However, mirroring results on other data-adaptive parameters under random sampling, including the sample average causal effect, the target parameter θ_{F_n} will converge to the average causal effect $\mathrm{E}_G[\tau]$ at root-n rate. Proposition 2 proves that the data-adaptive local average causal effect is asymptotically equivalent to the average causal effect, and establishes its rate of convergence.

Proposition 2: *Suppose that $\hat{\theta}$ is a root-n consistent and asymptotically normal estimator of $\mathrm{E}_G[\tau]$. Then $\sqrt{n}(\theta_{F_n} - \mathrm{E}_G[\tau]) = O_p(1)$.*

A proof of Proposition 2 follows by noting that $\sqrt{n}(\hat{\theta} - \mathrm{E}_G[\tau]) = O_p(1)$ and that across every realization, $|\theta_{F_n} - \mathrm{E}_G[\tau]| \leq |\hat{\theta} - \mathrm{E}_G[\tau]|$.

We now turn to our primary result, proving the superefficiency of $\hat{\theta}$ in estimating θ_{F_n}.

Proposition 3: *Suppose that Assumption 1 holds and that $\hat{\theta}$ is a root-n consistent and asymptotically normal estimator of $\mathrm{E}_G[\tau]$. Then $\sqrt{n}(\hat{\theta} - \theta_{F_n}) = o_p(1)$.*

Proof: Decompose $\hat{\theta}$ into $\tilde{\theta} = N(\mathrm{E}_G[\tau], \sigma^2/n)$ and $u = o_p(n^{-1/2})$, so that $\hat{\theta} = \tilde{\theta} + u$. Since $(\tilde{\theta} - \theta_{F_n})$ is $o_p(a_n)$ for any positive sequence (a_n), the rate of convergence of $\hat{\theta}$ is at worst governed by the bound ensured by u's $o_p(n^{-1/2})$ convergence. To prove the claim, note that for any positive ε, $\Pr\left(\frac{|\tilde{\theta} - \theta_{F_n}|}{a_n} \geq \varepsilon\right) \leq \Pr\left(\tilde{\theta} - \theta_{F_n} \neq 0\right) = 2\Phi(-c\sqrt{n}/\sigma)$, where $\Phi(.)$ denotes the standard Normal CDF. Since $\lim_{n\to\infty} 2\Phi(-c\sqrt{n}/\sigma) = 0$, $(\tilde{\theta} - \theta_{F_n})$ is $o_p(a_n)$. Thus $\hat{\theta} - \theta_{F_n} = o_p(a_n) + o_p(n^{-1/2}) = o_p(n^{-1/2})$, yielding the result.

In short, Proposition 3 demonstrates that the probability that $\hat{\theta}$ falls inside the support of the effect distribution converges to one quickly; conditional on this event, then estimation error is zero (as the target parameter takes on the value as the estimator with probability one). To illustrate this result, we can consider a case where an interval defined by the support of the effect distribution encompasses the sampling distribution of the estimator.

Corollary 1: *Suppose that $c \geq \max\left(\sup\left(\mathrm{Supp}\left[\hat{\theta}\right]\right) - \mathrm{E}_G[\tau], \mathrm{E}_G[\tau] - \inf\left(\mathrm{Supp}\left[\hat{\theta}\right]\right)\right)$. Then $\Pr(\hat{\theta} = \theta_{F_n}) = 1$.*

A proof of Corollary 1 follows by noting that $\Pr(|\hat{\theta} - \mathrm{E}_G[\tau]| \leq c) = 0$, and applying Definition 2. In other words, if the support of the estimator being used lies entirely within the interval $[\mathrm{E}_G[\tau] - c, \mathrm{E}_G[\tau] + c]$, then estimation error is always zero. This condition necessarily holds if $\mathrm{Supp}_G[\tau] = \mathbb{R}$, then the value that any estimator $\hat{\theta}$ takes must coincide with a local average causal effect. But note that Corollary 1 would not hold if $\mathrm{Supp}_G[\tau] = \mathbb{R}^+$ and $\Pr(\hat{\theta} < 0) > 0$.

Our results can be generalized to stronger claims straightforwardly. When a regularity condition is imposed on the rate of convergence of $\hat{\theta}$ to normality, a stronger result can be obtained about the rate of mean square convergence.

Proposition 4: *Suppose that Assumption 1 holds and* $\hat{\theta}$ *obeys* $\sqrt{n}(\hat{\theta} - \mathrm{E}_G[\tau]) = N(0, \sigma^2) + \varepsilon$, *where* $\mathrm{E}[\varepsilon^2] = o(n^{-1/2})$. *Then* $\mathrm{E}[(\hat{\theta} - \theta_{F_n})^2] = o(n^{-1})$.

Proof: We will show that the mean square error of $(\tilde{\theta} - \theta_{F_n})$ converges to zero sufficiently quickly, implying that the rate of convergence of $\hat{\theta}$ is at worst governed by the mean square error bound ensured by ε's convergence rate. To obtain the rate of convergence of the mean square error of $\tilde{\theta}$, we integrate over its squared deviation from the target parameter. Within c of $\mathrm{E}_G[\tau]$, the squared deviation is zero, thus we need only integrate over the squared deviation over the tails of the normal distribution. To ease calculations, we obtain an upper bound by integrating over the squared deviation from $\mathrm{E}_G[\tau]$, rather than from θ_{F_n}:

$$\begin{aligned}
\mathrm{E}_G[(\tilde{\theta} - \theta_{F_n})^2] &\leq 2\int_c^\infty x^2 \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} e^{-\frac{x^2 n}{2\sigma^2}} \\
&= c\sigma\sqrt{\frac{2}{\pi}} \frac{e^{-\frac{c^2 n}{2\sigma^2}}}{\sqrt{n}} + 2\sigma^2 \frac{\Phi\left(\frac{-c\sqrt{n}}{\sigma}\right)}{n} \\
&= o(n^{-1}).
\end{aligned}$$

Since $\mathrm{E}[(\tilde{\theta} - \theta_{F_n})^2] = o(n^{-1})$ and $n^{-1/2}\mathrm{E}[\varepsilon^2] = o(n^{-1})$, the Cauchy-Schwarz inequality ensures that $\mathrm{E}[(\hat{\theta} - \theta_{F_n})^2] = o(n^{-1}) + o(n^{-1}) = o(n^{-1})$.

3 Discussion

Our results highlight the additional certainty obtained by data-adaptively choosing the population for which average causal effects are measured on the basis of efficiency. It is well known that efficiency gains may be obtained through data-adaptive inference. But the extent to which the researcher can benefit from such practice has been understated. Under treatment effect heterogeneity – a precondition for locality to be a concern – all root-n consistent and asymptotically normal estimators of the average treatment effect are superefficient for a local average treatment effect.

There is of course a cost to this superefficiency: the target parameter is likely not of intrinsic interest. This issue is not unique to our setting, and other methods that change the inferential target based on efficiency concerns may be subject to this critique. As Crump etal. ([1], p. 188) notes, "external validity may be lost by changing the focus to average treatment effects for a subset of the original sample." This is exacerbated in our setting by the researcher's lack of knowledge about the characteristics of the subpopulation under study. Our result represents an extreme case of privileging efficiency over targeting population average causal effects. However, our results provide insight into a potential pathology of data-adaptivity purely on efficiency concerns: the gains in statistical certainty may be essentially unbounded without further restrictions. We hope that future work in the domain of efficiency theory for data-adaptive parameters will consider classes of restrictions that would exclude the case considered here.

Acknowledgement

The author thanks Don Green, Cyrus Samii, Jas Sekhon, Mark van der Laan, and two anonymous reviewers for helpful comments. The author expresses particular gratitude to Jas Sekhon for suggesting a parsimonious proof strategy for Proposition 3 and to an anonymous reviewer for inspiring Corollary 1. All remaining errors are the author's responsibility.

References

1. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. Biometrika 2009.
2. Humphreys M. Bounds on least squares estimates of causal effects in the presence of heterogeneous assignment probabilities Columbia University, 2009;Manuscript.
3. Angrist JD, Pischke JS. Mostly harmless econometrics: An empiricist's companion. Princeton, NJ: Princeton University Press, 2009.
4. Aronow PM, Samii C. Does regression produce representative estimates of causal effects? Am J Pol Sci 2016;60(1):250–267.
5. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm Stat 2011;10(2):150–161.

6. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat 1985;39(1):33–38.
7. Abadie A, Imbens G. Simple and bias-corrected matching estimators for average treatment effects. NBER technical working paper no. 283 2002.
8. Aronow PM, Green DP, Lee DK. Sharp bounds on the variance in randomized experiments. Ann Stat 2014;42(3):850–871.
9. Balzer LB, Petersen ML, van der Laan MJ. Targeted estimation and inference for the sample average treatment effect Bepress, 2015 Berkeley, CA.
10. van der Laan MJ, Hubbard AE, Pajouh SK. Statistical inference for data adaptive target parameters. Princeton, NJ: Bepress, 2013.