Using Post-Treatment Variables to Establish Upper Bounds on Causal Effects*

Adam N. Glynn† Nahomi Ichino‡

May 9, 2013

Abstract

We propose an adjustment based on post-treatment variables for pair matching estimators of the average treatment effect on the treated. Under relatively weak conditions, this adjusted estimator will provide an upper bound for the effect. Additionally, this approach allows for unbounded outcome variables and multiple mechanisms by which the treatment has an effect on the outcome. We also demonstrate that this adjustment will reduce the estimated effect in a wide variety of circumstances, and therefore, when the assumptions for the adjusted estimator are preferable to the assumptions for the unadjusted estimator, the adjustment can be used as a robustness check.

*An earlier version was presented at the Frontiers in the Analysis of Causal Mechanisms Conference at the Institute for Quantitative Social Science at Harvard University, March 24, 2012. We thank Konstantin Kashin for his helpful comments.

†Department of Government, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138. aglynn@fas.harvard.edu

‡Department of Government, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138. nichino@gov.harvard.edu

1 Introduction

In program evaluation, the effect of interest is often average treatment effect on those assigned to treatment. If treatment is randomly assigned, and if there is no missingness in the treatment or outcome, then unbiased estimates of this effect can be obtained without considering the mechanisms by which the treatment might affect the outcome (or post-treatment variables that might provide information about the mechanisms). Even in observational studies, where the treatment has not been randomly assigned, the most popular techniques for estimating this effect do not consider the mechanisms by which the treatment might affect the outcome (see Imbens and Wooldridge [2009] for a recent survey). This runs counter to the advice of Fisher (quoted in Cochran [1965]), Rosenbaum [2002b], Collier and Brady [2004], George and Bennett [2005], Brady et al. [2006], Hedström [2008], Heckman and Urzua [2009], Deaton [2009], and others who have argued for the explicit consideration of mechanisms when estimating causal effects. Furthermore, recent work provides methods that utilize mechanistic information for both the point identification [Pearl, 1995, 2009, Tian and Pearl, 2002a,b, Morgan and Winship, 2007] and partial identification [Joffe, 2001, Kuroki and Cai, 2008, Kaufman et al., 2009, Glynn and Quinn, 2011] of average treatment effects.

In this paper, we consider an adjustment, based on post-treatment variables, to pair matching estimators for the average treatment effect on the treated. This adjusted estimator will typically be biased, however under relatively weak assumptions the sign of this bias can be determined. Because one of these assumptions is a weakened version of the standard selection-on-observables assumption, in some cases the assumptions of the adjusted estimator will be preferable to the assumptions of the matching estimator. Finally, in a wide variety of circumstances, the sign of the bias of the adjusted estimator will be the same as the sign of the difference between the standard estimator and the adjusted estimator. If this holds, then the selection-on-observables assumption will be falsified. For example, in this paper we focus on the conditions such that the adjusted estimator has positive bias, and we consider the possibility that the adjusted estimator is smaller than the standard estimator. If the assumptions of the adjusted estimator are preferable to the selection-on-observables assumption, the adjusted estimator can function as a robustness check.

We first present traditional estimates in Section 2 before discussing our method and the properties of the upper bound within the context of a non-randomized study with treatment noncompliance in Section 3. We also discuss the implications of this methodology for the design and evaluation of programs.

2 Traditional Estimates

Suppose we have a data set with a binary treatment T_i and an outcome Y_i for observations $i = 1, ..., N$, and we are interested in the effect of the treatment for the treated individuals. We assume a lack of interference between individuals such that the individual-level causal effects of

T on Y are defined by the difference between the potential outcomes under treatment ($Y_i(1)$, the outcome that would be observed if $T_i = 1$) and control ($Y_i(0)$, the outcome that would be observed if $T_i = 0$),

$$\tau_i = Y_i(1) - Y_i(0), \text{ for } i = 1, ..., N \tag{1}$$

We further assume that we observe one of these two potential outcomes for each individual, determined by the treatment received:

$$Y_i = T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0).$$

We assume there are N_1 individuals with $T = 1$ and N_0 individuals with $T = 0$, where $N_1 \leq N_0$ such that we can pair match the N_1 individuals with $T = 1$ to the closest N_1 individuals with $T = 0$ (from among the N_0 individuals with $T = 0$). In this paper, we focus on pair matching in order to simplify notation. Each pair grouped together is denoted with $g = 1, ..., N_1$, so that the pair matched data can be re-indexed as Y_{g1} and T_g (outcome and treatment for the first unit in pair g) and Y_{g2} and $1 - T_g$ (outcome and treatment for the second unit in pair g). We also re-index the individual level effects:

$$\begin{aligned}
\tau_{g1} &= Y_{g1}(1) - Y_{g1}(0), \text{ and} \\
\tau_{g2} &= Y_{g2}(1) - Y_{g2}(0), \text{ for } g = 1, ..., N_1
\end{aligned}$$

The average of the $2 \cdot N_1$ individual-level effects is defined as,

$$\begin{aligned}
\tau &= \frac{1}{N_1} \sum_{g=1}^{N_1} \frac{1}{2} [(Y_{g1}(1) - Y_{g1}(0)) + (Y_{g2}(1) - Y_{g2}(0))] \\
&= \frac{1}{N_1} \sum_{g=1}^{N_1} \frac{1}{2} [\tau_{g1} + \tau_{g2}]
\end{aligned}$$

In contrast, the average of the N_1 individual-level effects for the treated can be defined as the following,

$$\begin{aligned}
\tau_T &= \frac{1}{N_1} \sum_{g=1}^{N_1} [T_g(Y_{g1}(1) - Y_{g1}(0)) + (1 - T_g)(Y_{g2}(1) - Y_{g2}(0))] \\
&= \frac{1}{N_1} \sum_{g=1}^{N_1} [T_g \tau_{g1} + (1 - T_g)\tau_{g2}],
\end{aligned}$$

where τ_T depends on the assigned treatments within each pair and can therefore be considered a random variable. However, if the probability of receiving treatment is equal for both individuals

in all pairs, then $E[T_g] = 1/2$ for $g = 1, ..., N_1$, and it is straightforward to demonstrate that $E[\tau_T] = \tau$.

A typical summary statistic within this framework would be the average of the treatment-minus-control differences within each pair,

$$\begin{aligned}
\widehat{\tau}_T &= \frac{1}{N_1}\sum_{g=1}^{N_1}[T_g(Y_{g1} - Y_{g2}) + (1 - T_g)(Y_{g2} - Y_{g1})]) \\
&= \frac{1}{N_1}\sum_{g=1}^{N_1}[T_g(Y_{g1}(1) - Y_{g2}(0)) + (1 - T_g)(Y_{g2}(1) - Y_{g1}(0))]) \\
&= \frac{1}{N_1}\sum_{g=1}^{N_1}[T_g\tau_{g1} + (1 - T_g)\tau_{g2}] + [T_g - (1 - T_g)] \cdot [Y_{g1}(0) - Y_{g2}(0)] \\
&= \frac{1}{N_1}\sum_{g=1}^{N_1}\tau_T + [T_g - (1 - T_g)] \cdot [Y_{g1}(0) - Y_{g2}(0)],
\end{aligned}$$

where potential outcomes can be substituted into the second line according to which unit in the pair receives treatment, and the estimator can be re-written in terms of τ_T and an error term. When $E[T_g] = 1/2$ for all $g = 1, ..., N_1$, this estimator is unbiased for the random parameter τ_T in the sense that $E[\widehat{\tau}_T - \tau_T] = 0$.

3 Estimating Upper Bounds Using Post-Treatment Variables

When important pre-treatment matching variables are unmeasured or poorly measured, we may be unwilling to assume that treatment assignment is unconfounded within each pair. In particular, we may worry that $E[T_g] \neq 1/2$ for all $g = 1, ..., N_1$ pairs, and therefore we may find it useful to make alternative assumptions. In this section we demonstrate that some useful assumptions involve post-treatment variables.

Before considering post-treatment variables, we assume a direction of the bias due to confounding:

Assumption 1. $Y_{g1}(0) \geq Y_{g2}(0) \Rightarrow E[T_g] \geq 1/2$ *and* $Y_{g1}(0) \leq Y_{g2}(0) \Rightarrow E[T_g] \leq 1/2$

Note that when Assumption 1 holds, it is straightforward to demonstrate that $\widehat{\tau}_T$ will be positively biased as an estimator for τ_T ($E[\widehat{\tau}_T - \tau_T] \geq 0$).

With only Assumption 1, we could conduct a sensitivity analysis along the lines of Rosenbaum [2002b], but post-treatment variables may provide additional information. Suppose that we observe a post-treatment variable Z, such that we can define the potential outcomes $Z(1)$ and $Z(0)$. Because Z will often be an aggregation of the mechanisms (or side effects of mechanisms) by which T affects Y, we will utilize the standard terminology of treatment noncompliance: always takers

$(Z(1) = Z(0) = 1)$, never takers $(Z(1) = Z(0) = 0)$, compliers $(Z(1) - Z(0) = 1)$, and defiers $(Z(1) - Z(0) = -1)$. Given this terminology, we now discuss the key assumptions of the technique.

Assumption 2. *For the never-takers and defiers with $T = 1$, T cannot have a positive effect on Y: $T_g = 1$ and $Z_{g1}(1) = 0 \Rightarrow \tau_{g1} \leq 0$, and $1 - T_g = 1$ and $Z_{g2}(1) = 0 \Rightarrow \tau_{g2} \leq 0$, for $g = 1, ..., N_1$*

Assumption 2 will be satisfied in a number of situations. For example, if we assume as in Angrist et al. [1996] that treatment assignment has a monotonic effect on Z (i.e., no defiers) and that an exclusion restriction holds (i.e., no direct effect), then Assumption 2 will hold. We have relaxed the no direct effect assumption to allow negative direct effects because this allows Z to represent an aggregation of only positive mechanisms. We have also relaxed the no defiers assumption because we may only be able to measure the side effects of mechanisms. For example, in the noncompliance setting we may only be able to obtain self-reports on treatment compliance.

Note that Assumption 2 implies an upper bound on the individual-level effects and the parameter τ_T,

$$\begin{aligned}
\tau_T &= \frac{1}{N_1}\sum_{g=1}^{N_1}[T_g\tau_{g1} + (1 - T_g)\tau_{g2}] \\
&\leq \frac{1}{N_1}\sum_{g=1}^{N_1}[T_g Z_{g1}(1)\tau_{g1} + (1 - T_g)Z_{g2}(1)\tau_{g2}] \qquad (2) \\
&\equiv \tau_T^{ub} \qquad (3)
\end{aligned}$$

The upper bound in (3) suggests an alternative estimator,

$$\begin{aligned}
\widehat{\tau}_T^{post} &= \frac{1}{N_1}\sum_{g=1}^{N_1}[T_g Z_{g1}(Y_{g1} - Y_{g2}) + (1 - T_g)Z_{g2}(Y_{g2} - Y_{g1})]) \\
&= \frac{1}{N_1}\sum_{g=1}^{N_1}[T_g Z_{g1}(1)(Y_{g1}(1) - Y_{g2}(0)) + (1 - T_g)Z_{g2}(1)(Y_{g2}(1) - Y_{g1}(0))]) \\
&= \tau_T^{ub} + \frac{1}{N_1}\sum_{g=1}^{N_1}[T_g Z_{g1}(1) - (1 - T_g)Z_{g2}(1)] \cdot [Y_{g1}(0) - Y_{g2}(0)],
\end{aligned}$$

where we only use the post-treatment variable Z from the treated unit, and therefore the variables Z_{g1} and Z_{g2} can be written as $Z_{g1}(1)$ and $Z_{g2}(1)$. Furthermore, when Assumption 2 holds, $\widehat{\tau}_T^{post}$ provides an estimate for an upper bound on τ_T where the second term represents the estimation error.

Unfortunately, $\widehat{\tau}_T^{post}$ will generally be a biased estimator for τ_T^{ub}. However, under certain conditions we can describe the direction of this bias. In particular, we are interested in the conditions under which the bias will be positive. If we write $\pi_g \equiv E[T_g]$ for all $g = 1, ..., N_1$, then the bias can

be written as the following:

$$E[\widehat{\tau}_T^{post} - \tau_T^{ub}] = E\{\frac{1}{N_1}\sum_{g=1}^{N_1}[T_g Z_{g1}(1) - (1-T_g)Z_{g2}(1)] \cdot [Y_{g1}(0) - Y_{g2}(0)]\}$$
$$= \frac{1}{N_1}\sum_{g=1}^{N_1}[\pi_g Z_{g1}(1) - (1-\pi_g)Z_{g2}(1)] \cdot [Y_{g1}(0) - Y_{g2}(0)]$$

If we assume, without loss of generality, that the individuals within each pair have been ordered such that $Y_{g1}(0) - Y_{g2}(0) \geq 0$, then $E[\widehat{\tau}_T^{post} - \tau_T^{ub}]$ is increasing in π_g and Assumption 1 implies that $\pi_g \geq .5$ for all $g = 1, ..., N_1$. Therefore, in order to derive the conditions under which the bias will be positive, it will be sufficient to consider the case where $\pi_g = .5$ for all $g = 1, ..., N_1$. In this condition, the bias simplifies to the following:

$$min\{E[\widehat{\tau}_T^{post} - \tau_T^{ub}]\} = \frac{1}{N_1}\sum_{g=1}^{N_1}[\frac{1}{2}(Z_{g1}(1) - Z_{g2}(1))] \cdot [Y_{g1}(0) - Y_{g2}(0)]$$

There are three things to note about this minimum bias. First, pairs in which $Z_{g1}(1) = Z_{g2}(1)$ will not contribute to the bias. Second, because the individuals within each pair have been ordered such that $Y_{g1}(0) - Y_{g2}(0) \geq 0$, pairs with $Z_{g1}(1) - Z_{g2}(1) = 1$ will make positive contributions to the bias while pairs with $Z_{g1}(1) - Z_{g2}(1) = -1$ will make negative contributions to the bias. Third, when the contributions from the $Z_{g1}(1) - Z_{g2}(1) = 1$ pairs outweighs the contributions for the $Z_{g1}(1) - Z_{g2}(1) = -1$ pairs, the minimum bias will be positive. This lead to our final assumption.

Assumption 3. *If paired units are arranged so that* $Y_{g1}(0) - Y_{g2}(0) \geq 0$ *for all* $g = 1, ..., N_1$*, then* $\frac{1}{N_1}\sum_{g=1}^{N_1}[Z_{g1}(1) - Z_{g2}(1)] \cdot [Y_{g1}(0) - Y_{g2}(0)] \geq 0$

Roughly, this implies a non-negative correlation between $Z(1)$ and $Y(0)$ within the pairs. In the noncompliance context, Assumption 3 implies that after matching, always-takers and compliers have higher baseline outcomes then never-takers and defiers on average.

When Assumptions 1, 2, and 3 hold, $\widehat{\tau}_T^{post}$ has nonnegative bias for an upper bound on τ_T. Additionally, it is possible for $\widehat{\tau}_T$ to be larger than $\widehat{\tau}_T^{post}$ (when $\frac{1}{N_1}\sum_{g=1}^{N_1}[T_g(1-Z_{g1})(Y_{g1} - Y_{g2}) + (1 - T_g)(1-Z_{g2})(Y_{g2} - Y_{g1})]) > 0$). Therefore, there are conditions under which the standard estimator ($\widehat{\tau}_T$) will be larger than an estimator known to have non-negative bias ($\widehat{\tau}_T^{post}$). Finally, it is straightforward to demonstrate that the variance for $\widehat{\tau}_T$ is at least as large as the the variance of $\widehat{\tau}_T^{post}$. Therefore, when Assumptions 1, 2, and 3 are preferable to the assumption of unconfoundedness, and $\widehat{\tau}_T$ is larger than $\widehat{\tau}_T^{post}$, then $\widehat{\tau}_T^{post}$ should be preferred and conservative confidence intervals can be formed by using the standard error of $\widehat{\tau}_T$.

4 Conclusion

This paper demonstrates that post-treatment variables can be used to test the robustness of causal estimates. Moreover, this approach allows for unbounded outcome variables, negative direct effects, and multiple positive mechanisms, as long as these can be aggregated into a single binary function of pre- and post-treatment variables including side effects of mediating variables. The generality of this result has two implications for future evaluations of non-randomized programs. First, when post-treatment variables are available such that Assumptions 2 and 3 are reasonable, these variables should be utilized in the analysis. Second, when designing an observational study or implementing a program where randomization is not possible, the potential mechanisms should be considered so that post-treatment variables that will satisfy Assumptions 2 and 3 can be measured. Whenever possible, a non-randomized study should be able to demonstrate that estimated effect is consistent with the hypothesized mechanisms.

References

Henry E. Brady, David Collier, and Jason Seawright. Toward a pluralistic vision of methodology. *Political Analysis*, 14:353–368, 2006.

William G. Cochran. The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society, A*, 128:134–155, 1965.

David Collier and Henry E. Brady. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Rowman & Littlefield, Lanham, MD, 2004.

Angus Deaton. Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. NBER Working Paper #14690, 2009.

A.L. George and A. Bennett. *Case studies and theory development in the social sciences*. Mit Press, 2005.

A.N. Glynn and K.M. Quinn. Why Process Matters for Causal Inference. *Political Analysis*, 19(3), 2011.

James Heckman and Sergio Urzua. Comparing iv with structural models: What simple iv can and cannot identify. NBER Working Paper, #14706, 2009.

Peter Hedström. Studying mechanisms to strengthen causal inferences in quantitative research. In *The Oxford Handbook of Political Methodology*. Oxford University Press, Oxford, 2008.

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

M.M. Joffe. Using information on realized effects to determine prospective causal effects. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 759–774, 2001.

Sol Kaufman, Jay S. Kaufman, and Richard F. MacLehose. Analytic bounds on causal risk differences in directed acyclic graphs with three observed binary variables. *Journal of Statistical Planning and Inference*, 139:3473–87, 2009.

M. Kuroki and Z. Cai. The Evaluation of Causal Effects in Studies with an Unobserved Exposure/Outcome Variable: Bounds and Identification. In *Proceedings from the 2008 Conference of Uncertainty in Artificial Intelligence*, 2008.

Stephen L. Morgan and Christopher Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, New York, 2007.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82:669–710, 1995.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2 edition, 2009.

Paul R. Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17:286–304, 2002a.

Paul R. Rosenbaum. *Observational Studies*. Springer, New York, second edition, 2002b.

J. Tian and J. Pearl. A general identification condition for causal effects. In *Proceedings of the National Conference on Artificial Intelligence*, pages 567–573. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2002a.

J. Tian and J. Pearl. On the identification of causal effects. In *Proceedings of the American Association of Artificial Intelligence*, 2002b.