Interpreting OLS Estimands When Treatment Effects Are Heterogeneous: Smaller Groups Get Larger Weights

Tymon Słoczyński

Applied work often studies the effect of a binary variable ("treatment") using linear models with additive effects. I study the interpretation of the OLS estimands in such models when treatment effects are heterogeneous. I show that the treatment coefficient is a convex combination of two parameters, which under certain conditions can be interpreted as the average treatment effects on the treated and untreated. The weights on these parameters are inversely related to the proportion of observations in each group. Reliance on these implicit weights can have serious consequences for applied work, as I illustrate with two well-known applications. I develop simple diagnostic tools that empirical researchers can use to avoid potential biases. Software for implementing these methods is available in R and Stata. In an important special case, my diagnostics only require the knowledge of the proportion of treated units.

Tymon Słoczyński is an Assistant Professor at the Department of Economics and International Business School, Brandeis University. E-mail: tslocz@brandeis.edu.

This version: May 18, 2020. This paper is based on portions of my previous working paper, Słoczyński (2018). I thank the editor and two anonymous referees for their helpful comments. I am very grateful to Alberto Abadie, Josh Goodman, Max Kasy, Pedro Sant'Anna, and Jeff Wooldridge for many comments and discussions. I also thank Arun Advani, Isaiah Andrews, Josh Angrist, Orley Ashenfelter, Richard Blundell, Stéphane Bonhomme, Carol Caetano, Marco Caliendo, Matias Cattaneo, Gary Chamberlain, Todd Elder, Alfonso Flores-Lagunes, Brigham Frandsen, Florian Gunsilius, Andreas Hagemann, James Heckman, Kei Hirano, Peter Hull, Macartan Humphreys, Guido Imbens, Krzysztof Karbownik, Shakeeb Khan, Toru Kitagawa, Pat Kline, Paweł Królikowski, Nicholas Longford, James MacKinnon, Łukasz Marć, Doug Miller, Michał Myck, Mateusz Myśliwski, Gary Solon, Jann Spiess, Michela Tincani, Alex Torgovitsky, Joanna Tyrowicz, Takuya Ura, Rudolf Winter-Ebmer, seminar participants at BC, Brandeis, Harvard–MIT, Holy Cross, IHS Vienna, Lehigh, MSU, Potsdam, SDU Odense, SGH, Temple, UCL, Upjohn, and WZB Berlin, and many conference participants for useful feedback. I thank Mark McAvoy for his excellent assistance in developing the R package hettreatreg that implements the results in this paper. I also thank David Card, Jochen Kluve, and Andrea Weber for providing me with supplementary data on the articles surveyed in Card et al. (2018). I acknowledge financial support from the National Science Centre (grant DEC-2012/05/N/HS4/00395), the Foundation for Polish Science (a "Start" scholarship), the "Weź stypendium—dla rozwoju" scholarship program, and the Theodore and Jane Norman Fund.

I Introduction

Many applied researchers study the effect of a binary variable ("treatment") on the expected value of an outcome of interest, holding fixed a vector of control variables. As noted by Imbens (2015), despite the availability of a large number of semi- and nonparametric estimators for average treatment effects, applied researchers often continue to use conventional regression methods. In particular, numerous studies use ordinary least squares (OLS) to estimate

$$y = \alpha + \tau d + X\beta + u, \tag{1}$$

where y denotes the outcome, d denotes the treatment, and X denotes the row vector of control variables, $(x_1, \ldots, x_K)$. Usually, τ is interpreted as the average treatment effect (ATE). This estimation strategy is used in many influential papers in economics (e.g., Voigtländer and Voth, 2012; Alesina et al., 2013; Aizer et al., 2016), as well as in other disciplines.

The great appeal of the model in (1) comes from its simplicity (Angrist and Pischke, 2009). At the same time, however, a large body of evidence demonstrates the importance of heterogeneity in effects (see, e.g., Heckman, 2001; Bitler et al., 2006), which is explicitly ruled out by this same model. In this paper I contribute to the recent literature on interpreting τ, the OLS estimand, when treatment effects are heterogeneous (Angrist, 1998; Humphreys, 2009; Aronow and Samii, 2016). I demonstrate that τ is a convex combination of two parameters, which under certain conditions can be interpreted as the average treatment effects on the treated (ATT) and untreated (ATU). Surprisingly, the weight that is placed by OLS on the average effect for each group is inversely related to the proportion of observations in this group. The more units are treated, the less weight is placed on ATT. One interpretation of this result is that OLS estimation of the model in (1) is generally inappropriate when treatment effects are heterogeneous.

It is also possible, however, to present a more pragmatic view of my main result. I derive a number of corollaries of this result which suggest several diagnostic methods that I recommend to applied researchers. These diagnostics are applicable whenever the researcher is: *(i)* studying the effects of a binary treatment, *(ii)* using OLS, and *(iii)* unwilling to maintain that ATT is exactly

equal to ATU. Typically, such a homogeneity assumption would be undesirably strong, because those choosing or chosen for treatment may have unusually high or low returns from that treatment, which would directly contradict the equality of ATT and ATU.

In deriving my diagnostics, I assume that the researcher is ultimately interested in ATE, ATT, or both, and that she wishes to estimate the model in (1) using OLS but is concerned about treatment effect heterogeneity. In this case, my diagnostics are able to detect deviations of the OLS weights from the pattern that would be necessary to consistently estimate a given parameter. These diagnostics are easy to implement and interpret; they are bounded between zero and one in absolute value and they give the proportion of the difference between ATU and ATT (or between ATT and ATU) that contributes to bias. Thus, if a given diagnostic is close to zero, OLS is likely a reasonable choice; but if a diagnostic is far from zero, other methods should be used.

In an important special case, these diagnostics become particularly simple and immediate to report. If we wish to estimate ATT, this "rule of thumb" variant of my diagnostic is equal to the proportion of treated units, $\mathrm{P}(d = 1)$; if our goal is to estimate ATE, the diagnostic is equal to $2 \cdot \mathrm{P}(d = 1) - 1$, twice the deviation of $\mathrm{P}(d = 1)$ from 50%. In short, OLS is expected to provide a reasonable approximation to ATE if both groups, treated and untreated, are of similar size. If we wish to estimate ATT, it is necessary that the proportion of treated units is very small.

It follows that OLS might often be substantially biased for ATE, ATT, or both. How common are these biases in practice? In a subset of 37 estimates from Card et al. (2018), a recent survey of evaluations of active labor market programs, the mean proportion of treated units is 17.7%.[1] Using the "rule of thumb" variants of my diagnostics, I establish that on average the difference between the OLS estimand and ATE is expected to correspond to 64.6% of the difference between ATT and ATU. Similarly, the expected difference between OLS and ATT is on average equal to 17.7% of the difference between ATU and ATT. In other words, these biases might often be large.

The remainder of the paper is organized as follows. Section II presents a leading example and the main theoretical results. Section III discusses two empirical applications. In a study of the

[1] This sample is restricted to studies that Card et al. (2018) coded as "selection on observables" and "regression."

effects of a training program (LaLonde, 1986), OLS estimates are very similar to $\widehat{\text{ATT}}$. On the other hand, in a study of the effects of cash transfers (Aizer et al., 2016), OLS estimates are similar to $\widehat{\text{ATU}}$. Section IV concludes. Proofs and several extensions are provided in the online appendices. The main results are implemented in newly developed R and Stata packages, `hettreatreg`.

II A Weighted Average Interpretation of OLS

A Leading Example

To illustrate the problem with OLS weights, consider the classic example of the National Supported Work (NSW) program. Because this program originally involved a social experiment, the difference in mean outcomes between the treated and control units provides an unbiased estimate of the effect of treatment. LaLonde (1986) studies the performance of various estimators at reproducing this experimental benchmark when the experimental controls are replaced by an artificial comparison group from the Current Population Survey (CPS) or the Panel Study of Income Dynamics (PSID). Angrist and Pischke (2009) reanalyze the NSW–CPS data and conclude that OLS estimates of the effect of NSW program on earnings in 1978 are similar to the experimental benchmark of \$1,794.[2] In particular, their richest specification delivers an estimate of \$794. As I will show, this conclusion is driven by the small proportion of treated units in these data.

In this example, ATT and ATU are likely to be substantially different. This is because the treated group, unlike the CPS comparison (untreated) group, was highly economically disadvantaged. It is plausible that ATU might be zero or, due to the opportunity cost of program participation, even negative. Also, only 1.1% of the sample was treated, so ATE and ATU will be similar.

To demonstrate this, I modify the model in (1) to include all interactions between d and X. Estimation of this expanded model, again using OLS, allows us to separately compute $\widehat{\text{ATE}}$, $\widehat{\text{ATT}}$, and $\widehat{\text{ATU}}$. This method is usually referred to as "regression adjustment" (Wooldridge, 2010) or "Oaxaca–Blinder" (Kline, 2011; Graham and Pinto, 2018). Using the control variables that deliver

[2] Subsequently to LaLonde (1986), these data were studied by Dehejia and Wahba (1999), Smith and Todd (2005), and many others. Angrist and Pischke (2009) analyze the subsample of the experimental treated units constructed by Dehejia and Wahba (1999), combined with "CPS-1" or "CPS-3," i.e. two of the nonexperimental comparison groups from CPS, constructed by LaLonde (1986). In this replication, I focus on "CPS-1."

the estimate of \$794, we obtain $\widehat{\text{ATE}} = -\$4{,}930$, $\widehat{\text{ATT}} = \$796$, and $\widehat{\text{ATU}} = -\$4{,}996$. It turns out that, since $\widehat{\text{ATE}}$ and $\widehat{\text{ATU}}$ are indeed negative, the OLS estimate and $\widehat{\text{ATE}}$ have different signs. Moreover, if we represent the OLS estimate as a weighted average of $\widehat{\text{ATT}}$ and $\widehat{\text{ATU}}$ with weights that sum to unity, we can write $\$794 = \hat{w}_{ATT} \cdot \$796 + (1 - \hat{w}_{ATT}) \cdot (-\$4{,}996)$, where $\hat{w}_{ATT}$ is the weight on $\widehat{\text{ATT}}$. Solving for $\hat{w}_{ATT}$ yields $\hat{w}_{ATT} = 99.96\%$. In other words, the hypothetical OLS weight on the effect on the treated is similar to the proportion of untreated units, 98.9%.

This "weight reversal" is not a coincidence. As I demonstrate below, the intuition from this example holds more generally, even though the OLS estimand is not necessarily a convex combination of two parameters from a procedure that controls for the full vector X.

B Main Result

This section presents my main result, which focuses on the algebra of OLS and "descriptive" estimands that I define below. A causal interpretation of OLS also requires introducing the notion of potential outcomes as well as certain conditions that I discuss in section IIC, including an ignorability assumption. However, this is not needed for my main result.

If $\mathrm{L}(\cdot \mid \cdot)$ denotes the linear projection, we are interested in the interpretation of τ in the linear projection of y on d and X,

$$\mathrm{L}(y \mid 1, d, X) = \alpha + \tau d + X\beta, \tag{2}$$

when this linear projection does not correspond to the (structural) conditional mean. Let

$$\rho = \mathrm{P}(d = 1) \tag{3}$$

be the unconditional probability of treatment and let

$$p(X) = \mathrm{L}(d \mid 1, X) = \alpha_p + X\beta_p \tag{4}$$

be the "propensity score" from the linear probability model or, equivalently, the best linear approximation to the true propensity score. Generally, the specification in (2) and (4) can be arbitrarily flexible, so this approximation can be made very accurate; in fact, we can think of equation (2) as partially linear, where we may include powers and cross-products of original control variables.

After defining $p(X)$, it is helpful to introduce two linear projections of y on $p(X)$, separately for $d = 1$ and $d = 0$, namely

$$\mathrm{L}\left[y \mid 1, p(X), d = 1\right] = \alpha_1 + \gamma_1 \cdot p(X) \tag{5}$$

and also

$$\mathrm{L}\left[y \mid 1, p(X), d = 0\right] = \alpha_0 + \gamma_0 \cdot p(X). \tag{6}$$

Note that equations (4), (5), and (6) are definitional. It is sufficient for my main result that the linear projections introduced so far exist and are unique.

Assumption 1. *(i)* $\mathrm{E}(y^2)$ *and* $\mathrm{E}(\|X\|^2)$ *are finite. (ii) The covariance matrix of* (d, X) *is nonsingular.*

Assumption 2. $\mathrm{V}\left[p(X) \mid d = 1\right]$ *and* $\mathrm{V}\left[p(X) \mid d = 0\right]$ *are nonzero, where* $\mathrm{V}(\cdot \mid \cdot)$ *denotes the conditional variance (with respect to* $\mathrm{E}\left[p(X) \mid d = j\right]$*,* $j = 0, 1$*).*

Assumption 1 guarantees the existence and uniqueness of the linear projections in (2) and (4). Similarly, Assumption 2 ensures that the linear projections in (5) and (6) exist and are unique.[3]

The next step is to use the linear projections in (5) and (6) to define the average partial linear effect of d as

$$\tau_{APLE} = (\alpha_1 - \alpha_0) + (\gamma_1 - \gamma_0) \cdot \mathrm{E}\left[p(X)\right] \tag{7}$$

as well as the average partial linear effect of d on group j ($j = 0, 1$) as

$$\tau_{APLE,j} = (\alpha_1 - \alpha_0) + (\gamma_1 - \gamma_0) \cdot \mathrm{E}\left[p(X) \mid d = j\right]. \tag{8}$$

These estimands are well defined under Assumptions 1 and 2, and have a causal interpretation under additional assumptions, as discussed in section IIC below.[4] When the linear projections in equations (5) and (6) represent the conditional mean of y, the average partial linear effects of d overlap with its average partial effects. It should be stressed, however, that Theorem 1, the main result of this paper, is more general and only requires Assumptions 1 and 2.

[3]Both assumptions are generally innocuous, although Assumption 2 rules out a small number of interesting applications, such as regression adjustments in Bernoulli trials and completely randomized experiments. In these cases, however, OLS is consistent for the average treatment effect under general conditions (Imbens and Rubin, 2015).

[4]Moreover, τ_{APLE} is similar to the "average regression coefficient" or "average slope coefficient" in Graham and Pinto (2018), which is also a descriptive estimand in the sense of Abadie et al. (2020).

Theorem 1 (Weighted Average Interpretation of OLS)**.** *Under Assumptions 1 and 2,*

$$\tau = w_1 \cdot \tau_{APLE,1} + w_0 \cdot \tau_{APLE,0},$$

where $w_1 = \frac{(1-\rho)\cdot \mathrm{V}\left[p(X)|d=0\right]}{\rho\cdot \mathrm{V}\left[p(X)|d=1\right]+(1-\rho)\cdot \mathrm{V}\left[p(X)|d=0\right]}$ *and* $w_0 = 1 - w_1 = \frac{\rho\cdot \mathrm{V}\left[p(X)|d=1\right]}{\rho\cdot \mathrm{V}\left[p(X)|d=1\right]+(1-\rho)\cdot \mathrm{V}\left[p(X)|d=0\right]}$.

Proof. See online appendix A. □

Theorem 1 shows that τ, the OLS estimand, is a convex combination of $\tau_{APLE,1}$ and $\tau_{APLE,0}$. The definition of $\tau_{APLE,j}$ makes it clear that τ is equivalent to the outcome of a particular three-step procedure. In the first step, we obtain $p(X)$, i.e. the "propensity score." Next, in the second step, we obtain $\tau_{APLE,1}$ and $\tau_{APLE,0}$, as in (8), from two linear projections of y on $p(X)$, separately for $d = 1$ and $d = 0$. This is analogous to the "regression adjustment" procedure in section IIA, although now we control for $p(X)$ rather than the full vector X. Finally, in the third step, we calculate a weighted average of $\tau_{APLE,1}$ and $\tau_{APLE,0}$. The weight on $\tau_{APLE,1}$, w_1, is decreasing in $\frac{\mathrm{V}\left[p(X)|d=1\right]}{\mathrm{V}\left[p(X)|d=0\right]}$ and ρ and the weight on $\tau_{APLE,0}$, w_0, is increasing in $\frac{\mathrm{V}\left[p(X)|d=1\right]}{\mathrm{V}\left[p(X)|d=0\right]}$ and ρ.[5] This is clearly undesirable, since $\tau_{APLE} = \rho \cdot \tau_{APLE,1} + (1-\rho) \cdot \tau_{APLE,0}$.

This weighting scheme is also surprising: the more units belong to group j, the less weight is placed on $\tau_{APLE,j}$, i.e. the effect *for this group*. There are several ways to provide intuition for this result. One is provided in the next section. Another intuition follows from an alternative proof of Theorem 1, which is provided with discussion in online appendix B2. It parallels the intuition in Angrist (1998) and Angrist and Pischke (2009) that OLS gives more weight to treatment effects that are better estimated in finite samples.[6]

C Causal Interpretation

The fact that Theorem 1 only requires the existence and uniqueness of several linear projections makes this result very general. On the other hand, one concern about this result might be that $\tau_{APLE,1}$ and $\tau_{APLE,0}$ do not necessarily correspond to the usual (causal) objects of interest. To define

[5] A formal proof that the relationship between ρ and w_1 (w_0) is indeed always negative (positive) is provided in online appendix B1. This proof additionally assumes that the conditional mean of d is linear in X.

[6] This proof uses a result from Deaton (1997) and Solon et al. (2015) as a lemma. The main proof of Theorem 1 uses a result on decomposition methods from Elder et al. (2010). See online appendix A for more details.

these objects, we need two potential outcomes, $y(1)$ and $y(0)$, only one of which is observed for each unit, $y = y(d) = y(1) \cdot d + y(0) \cdot (1 - d)$. The parameters of interest, ATE, ATT, and ATU, are defined as $\tau_{ATE} = \mathrm{E}\left[y(1) - y(0)\right]$, $\tau_{ATT} = \mathrm{E}\left[y(1) - y(0) \mid d = 1\right]$, and $\tau_{ATU} = \mathrm{E}\left[y(1) - y(0) \mid d = 0\right]$. A causal interpretation of OLS also entails the following assumptions.

Assumption 3 (Ignorability in Mean)**.** *(i)* $\mathrm{E}\left[y(1) \mid X, d\right] = \mathrm{E}\left[y(1) \mid X\right]$*; and (ii)* $\mathrm{E}\left[y(0) \mid X, d\right] = \mathrm{E}\left[y(0) \mid X\right]$.

Assumption 4. *(i)* $\mathrm{E}\left[y(1) \mid X\right] = \alpha_1 + \gamma_1 \cdot p\left(X\right)$*; and (ii)* $\mathrm{E}\left[y(0) \mid X\right] = \alpha_0 + \gamma_0 \cdot p\left(X\right)$.

Assumptions 3 and 4 ensure that τ admits a causal interpretation. Assumption 3 is standard in the program evaluation literature (Wooldridge, 2010). Assumption 4 is not commonly used. Sufficient for this assumption, but not necessary, is that the conditional mean of d is linear in X and the conditional means of $y(1)$ and $y(0)$ are linear in the true propensity score, which is now equal to $p\left(X\right)$. Linearity of $\mathrm{E}\left(d \mid X\right)$ is assumed in Aronow and Samii (2016) and Abadie et al. (2020). This assumption is not necessarily strong, since X might include powers and cross-products of original control variables. It is also satisfied automatically in saturated models, as in Angrist (1998) and Humphreys (2009). The linearity assumption for $\mathrm{E}\left[y(1) \mid p\left(X\right)\right]$ and $\mathrm{E}\left[y(0) \mid p\left(X\right)\right]$ dates back to Rosenbaum and Rubin (1983) but is restrictive. See also Imbens and Wooldridge (2009) and Wooldridge (2010) for a discussion.

Corollary 1 (Causal Interpretation of OLS)**.** *Under Assumptions 1, 2, 3, and 4,*

$$\tau \;=\; w_1 \cdot \tau_{ATT} + w_0 \cdot \tau_{ATU}.$$

Proof. Assumption 3 implies that $\mathrm{E}\left[y(1) - y(0) \mid X\right] = \mathrm{E}\left(y \mid X,\ d = 1\right) - \mathrm{E}\left(y \mid X,\ d = 0\right)$. Then, Assumption 4 implies that $\mathrm{E}\left[y(1) - y(0) \mid X\right] = (\alpha_1 - \alpha_0) + (\gamma_1 - \gamma_0) \cdot p\left(X\right)$, which in turn implies that $\tau_{ATT} = \tau_{APLE,1}$ and $\tau_{ATU} = \tau_{APLE,0}$. This, together with Theorem 1, completes the proof. □

Corollary 1 states that, under Assumptions 1, 2, 3, and 4, the OLS weights from Theorem 1 apply to the causal objects of interest, τ_{ATT} and τ_{ATU}. Hence, τ has a causal interpretation. The greater

the proportion of treated units, the smaller is the OLS weight on τ_{ATT}. Again, this is undesirable, since $\tau_{ATE} = \rho \cdot \tau_{ATT} + (1 - \rho) \cdot \tau_{ATU}$.

To aid intuition for this surprising result, recall that an important motivation for using the model in (1) and OLS is that the linear projection of y on d and X provides the best linear predictor of y given d and X (Angrist and Pischke, 2009). However, if our goal is to conduct causal inference, then this is not, in fact, a good reason to use this method. Ordinary least squares is "best" in predicting actual outcomes but causal inference is about predicting missing outcomes, defined as $y_m = y(1) \cdot (1 - d) + y(0) \cdot d$. In other words, the OLS weights are optimal for predicting "what is." Instead, we are interested in predicting "what would be" if treatment were assigned differently.

Intuition suggests that if our goal were to predict "what is" and, without loss of generality, group one were substantially larger than group zero, we would like to place a large weight on the linear projection coefficients of group one (α_1 and γ_1), because these coefficients can be used to predict actual outcomes of this group. As noted by Deaton (1997) and Solon et al. (2015), the OLS weights are consistent with this idea. Indeed, Theorem 1 also implies that

$$\tau = \left[\mathrm{E}\left(y \mid d = 1\right) - \mathrm{E}\left(y \mid d = 0\right)\right] - \left(w_0\gamma_1 + w_1\gamma_0\right) \cdot \left\{\mathrm{E}\left[p\left(X\right) \mid d = 1\right] - \mathrm{E}\left[p\left(X\right) \mid d = 0\right]\right\}. \quad (9)$$

Namely, the OLS estimand is equal to the simple difference in means of y plus an adjustment term that depends on the difference in means of $p\left(X\right)$ and a weighted average of γ_1 and γ_0. When group one is "large," w_0, the weight on γ_1, is large as well.

Conversely, if group one is "large" but our goal is to predict missing outcomes, we need to place a large weight on α_0 and γ_0, because these coefficients can be used to predict counterfactual outcomes of group one. To see this point, note that it follows from the discussion in Imbens and Wooldridge (2009) that when the conditional means of $y(1)$ and $y(0)$ are linear in X, we can write

$$\tau_{ATE} = \left[\mathrm{E}\left(y \mid d = 1\right) - \mathrm{E}\left(y \mid d = 0\right)\right] - \left[\left(1 - \rho\right)\beta_1 + \rho\beta_0\right] \cdot \left[\mathrm{E}\left(X \mid d = 1\right) - \mathrm{E}\left(X \mid d = 0\right)\right], \quad (10)$$

where β_1 and β_0 are the coefficients on X in the conditional means of $y(1)$ and $y(0)$, respectively. Equations (9) and (10) reiterate the point of Corollary 1 that τ and τ_{ATE} have a very similar structure but they differ substantially in how they assign weights. Indeed, in the case of τ_{ATE}, when group

one is "large," the weight on β_1 is small, the opposite of what we have seen for OLS.[7]

D Implications of Theorem 1

There are several practical implications of my main result. Throughout this section, I assume that the researcher is interested in estimating τ_{ATE}, τ_{ATT}, or both, and that she wishes to use OLS to estimate the model in (1) but is concerned about the implications of Theorem 1 and Corollary 1. In Corollaries 2 and 3, I show how to decompose the difference between τ and τ_{ATE} or τ and τ_{ATT} into components attributable to *(i)* the difference between $\tau_{APLE,1}$ and τ_{ATT}, *(ii)* the difference between $\tau_{APLE,0}$ and τ_{ATU} (jointly referred to as "bias from nonlinearity"), and *(iii)* the OLS weights on τ_{ATT} and τ_{ATU} ("bias from heterogeneity").[8] Because this paper generally focuses on what I now term "bias from heterogeneity," my discussion below is restricted to this source of bias, which is equivalent to implicitly making Assumptions 3 and 4.

Corollary 2. *Under Assumptions 1 and 2,*

$$\tau - \tau_{ATE} = \underbrace{w_0 \cdot (\tau_{APLE,0} - \tau_{ATU}) + w_1 \cdot (\tau_{APLE,1} - \tau_{ATT})}_{\textit{bias from nonlinearity}} + \underbrace{\delta \cdot (\tau_{ATU} - \tau_{ATT})}_{\textit{bias from heterogeneity}},$$

where $\delta = \rho - w_1 = \frac{\rho^2 \cdot \mathrm{V}\left[p(X)|d=1\right] - (1-\rho)^2 \cdot \mathrm{V}\left[p(X)|d=0\right]}{\rho \cdot \mathrm{V}\left[p(X)|d=1\right] + (1-\rho) \cdot \mathrm{V}\left[p(X)|d=0\right]}$. *Also, under Assumptions 1, 2, 3, and 4,*

$$\tau - \tau_{ATE} = \delta \cdot (\tau_{ATU} - \tau_{ATT}).$$

Corollary 3. *Under Assumptions 1 and 2,*

$$\tau - \tau_{ATT} = \underbrace{w_0 \cdot (\tau_{APLE,0} - \tau_{ATU}) + w_1 \cdot (\tau_{APLE,1} - \tau_{ATT})}_{\textit{bias from nonlinearity}} + \underbrace{w_0 \cdot (\tau_{ATU} - \tau_{ATT})}_{\textit{bias from heterogeneity}}.$$

Also, under Assumptions 1, 2, 3, and 4,

$$\tau - \tau_{ATT} = w_0 \cdot (\tau_{ATU} - \tau_{ATT}).$$

The proofs of Corollaries 2 and 3 follow from simple algebra and are omitted. These results show that, regardless of whether we focus on τ_{ATE} or τ_{ATT}, the bias from heterogeneity is equal to the

[7] Note that the (infeasible) linear projection of the missing outcome, y_m, on d and X would solve our problem of "weight reversal." The weights on τ_{ATT} and τ_{ATU} would still be different than ρ and $1-\rho$ if $\mathrm{V}\left[p(X) \mid d=1\right]$ and $\mathrm{V}\left[p(X) \mid d=0\right]$ were different; but, at least, the weight on τ_{ATT} (τ_{ATU}) would be increasing (decreasing) in ρ.

[8] Because "bias from nonlinearity" arises when Assumptions 3 and/or 4 are violated, it might be more accurate to refer to this component as "bias from endogeneity and nonlinearity." Yet, I use the former term for brevity.

product of a particular measure of heterogeneity, namely the difference between τ_{ATU} and τ_{ATT}, and an additional parameter that is easy to estimate, δ for τ_{ATE} and w_0 for τ_{ATT}. While w_0 is guaranteed to be positive under Assumptions 1 and 2, δ may be positive or negative. Both w_0 and δ, however, are bounded between zero and one in absolute value. Thus, w_0 and $|\delta|$ can be interpreted as the percentage of our measure of heterogeneity, $\tau_{ATU} - \tau_{ATT}$, which contributes to bias.[9] It might be useful to report estimates of w_0 and δ in studies that use OLS to estimate the model in (1).

As an example, consider the empirical application in section IIA. In this case, $\hat{w}_0 = 0.017$ and $\hat{\delta} = -0.971$. The interpretation of these estimates is as follows: if our goal is to estimate τ_{ATT}, using the model in (1) and OLS is expected to bias our estimates by only 1.7% of the difference between τ_{ATU} and τ_{ATT}. If instead we wanted to interpret τ as τ_{ATE}, our estimates would be biased by an estimated 97.1% of the difference between τ_{ATT} and τ_{ATU}. Thus, in this application, it might perhaps be acceptable to interpret τ as τ_{ATT} but clearly not as τ_{ATE}.

Assumption 5. $\mathrm{V}\left[p\left(X\right) \mid d = 1\right] = \mathrm{V}\left[p\left(X\right) \mid d = 0\right]$.

The calculation of δ and w_0 is further simplified under Assumption 5. If we use δ^* and w_0^* to denote the values of δ and w_0 in this special case, we can write $\delta^* = 2\rho - 1$ and $w_0^* = \rho$. In this setting, the knowledge of δ and w_0 only requires information on ρ, the proportion of units with $d = 1$. Of course, the special case where $\mathrm{V}\left[p\left(X\right) \mid d = 1\right] = \mathrm{V}\left[p\left(X\right) \mid d = 0\right]$ is hardly to be expected in practice. Still, $\delta^* = 2\rho - 1$ and $w_0^* = \rho$ can potentially serve as a rule of thumb.

The practical implications of Assumption 5 are particularly clear when ρ is close to 0%, 50%, or 100%. When few units are treated, $\tau \simeq \tau_{ATT}$. When most of the units are treated, $\tau \simeq \tau_{ATU}$. Finally, when both groups are of similar size, $\tau \simeq \tau_{ATE}$. This can also be seen from Corollary 4.

Corollary 4. *Under Assumptions 1, 2, and 5,*

$$\tau = (1-\rho) \cdot \tau_{APLE,1} + \rho \cdot \tau_{APLE,0}.$$

[9] To be precise, $|\delta|$ can be interpreted as the percentage of $\mathrm{sgn}(\delta) \cdot (\tau_{ATU} - \tau_{ATT})$ that contributes to bias when focusing on τ_{ATE}. Both δ and w_0 also have an intuitive interpretation as the difference between *(i)* the weight that we should place on τ_{ATT} when focusing on τ_{ATE} or τ_{ATT} and *(ii)* the weight that OLS actually places on this parameter. Indeed, δ is equal to the difference between ρ and w_1. Similarly, $w_0 = 1 - w_1$.

Also, under Assumptions 1, 2, 3, 4, and 5,

$$\tau \quad = \quad (1-\rho)\cdot\tau_{ATT} + \rho\cdot\tau_{ATU}.$$

The proof follows immediately from simple algebra. Corollary 4 provides conditions under which OLS reverses the "natural" weights on $\tau_{APLE,1}$ and $\tau_{APLE,0}$ (or τ_{ATT} and τ_{ATU}). Indeed, under Assumption 5, τ is a convex combination of group-specific average effects, with "reversed" weights attached to these parameters. Namely, the proportion of units with $d = 1$ is used to weight the average effect of d on group zero, and vice versa.

The results in this section allow empirical researchers to interpret the OLS estimand when treatment effects are heterogeneous. Alternatively, it might be sensible to use any of the standard estimators for average treatment effects under ignorability, such as regression adjustment (see section IIA), weighting, matching, and various combinations of these approaches.[10] It might also help to estimate a model with homogeneous effects using weighted least squares (WLS). Indeed, in online appendix B3, I demonstrate that when we regress y on d and $p(X)$, with weights of $\frac{1-\rho}{w_0}$ for units with $d = 1$ and $\frac{\rho}{w_1}$ for units with $d = 0$, the WLS estimand is equal to τ_{APLE}. In practice, of course, τ_{APLE} can also be obtained directly from equation (7).

E Related Work

This section discusses the relationship between my main result and those in Angrist (1998) and Humphreys (2009). These papers focus on saturated models with discrete covariates, in which the estimating equation includes an indicator for each combination of covariate values ("stratum"). In particular, Angrist (1998) provides a representation of τ_n in

$$\mathrm{L}(y \mid d, x_1, \ldots, x_S) = \tau_n d + \sum_{s=1}^{S} \beta_{n,s} x_s, \tag{11}$$

where $x_1, \ldots, x_S$ are stratum indicators. More precisely, Angrist (1998) demonstrates that

$$\tau_n = \sum_{s=1}^{S} \frac{\mathrm{P}(x_s = 1)\cdot\mathrm{P}(d = 1 \mid x_s = 1)\cdot\mathrm{P}(d = 0 \mid x_s = 1)}{\sum_{t=1}^{S}\mathrm{P}(x_t = 1)\cdot\mathrm{P}(d = 1 \mid x_t = 1)\cdot\mathrm{P}(d = 0 \mid x_t = 1)}\cdot\tau_s, \tag{12}$$

[10] For recent reviews, see Imbens and Wooldridge (2009), Wooldridge (2010), and Abadie and Cattaneo (2018).

where $\tau_s = \mathrm{E}(y \mid d = 1, x_s = 1) - \mathrm{E}(y \mid d = 0, x_s = 1)$. In online appendix B4, I demonstrate that this result follows from Corollary 1 when the model for y is saturated.[11] At the same time, the interpretation of OLS in Angrist (1998) is different from Theorem 1 and Corollary 1. On the one hand, unlike Corollary 1 and Humphreys (2009), Angrist (1998) does not restrict the relationship between τ_s and $\mathrm{P}(d = 1 \mid x_s = 1)$ in any way. On the other hand, Theorem 1 and Corollary 1 make it arguably easier to identify whether in a given application the OLS estimand will be close to any of the parameters of interest (cf. Corollaries 2 to 4). In particular, Angrist (1998) does not recover a pattern of "weight reversal," which is discussed in detail in this paper.

Unlike Angrist (1998), Humphreys (2009) does not derive a new representation of τ_n, but instead presents further analysis of the result in equation (12). In particular, Humphreys (2009) notes that τ_n can take any value between $\min(\tau_s)$ and $\max(\tau_s)$. Then, he demonstrates that τ_n is also bounded by τ_{ATT} and τ_{ATU} if we restrict the relationship between τ_s and $\mathrm{P}(d = 1 \mid x_s = 1)$ to be monotonic. According to Corollary 1, τ is a convex combination of τ_{ATT} and τ_{ATU} if, among other things, both potential outcomes are linear in $p(X)$, which also implies a linear relationship between τ_s and $\mathrm{P}(d = 1 \mid x_s = 1)$ when the model for y is saturated. Of course, this linearity assumption is stronger than the monotonicity assumption in Humphreys (2009). However, in return, we are able to derive a closed-form expression for τ in terms of τ_{ATT} and τ_{ATU}, which is a major advantage over the earlier literature, such as Angrist (1998) and Humphreys (2009).[12]

III Empirical Applications

This section discusses two empirical illustrations of Theorem 1 and its corollaries.[13] In online appendices C and D, I discuss the implementation of these results in Stata and R. Throughout the current section τ_{APLE}, $\tau_{APLE,1}$, and $\tau_{APLE,0}$ are implicitly treated as equivalent to τ_{ATE}, τ_{ATT}, and

[11] Also, note that Aronow and Samii (2016) show that this result in Angrist (1998) is not specific to saturated models; instead, it is sufficient to assume that the model for d is linear in X. My analysis in online appendix B4 covers the results in both Angrist (1998) and Aronow and Samii (2016).

[12] Humphreys (2009) also provides a brief informal remark that the OLS estimand, as represented in Angrist (1998), is similar to τ_{ATT} (τ_{ATU}) if propensity scores are "small" ("large") in *every* stratum. This is a special case of the rule of thumb derived from Corollaries 3 and 4. My rule of thumb does not impose any such restrictions on the propensity score other than the requirement that the *unconditional* probability of treatment is close to zero or one.

[13] In a follow-up paper, I apply these results in the study of racial gaps in test scores and wages (Słoczyński, 2020).

τ_{ATU}, respectively. Although this might be restrictive, I also demonstrate that in both applications sample analogues of τ_{APLE}, $\tau_{APLE,1}$, and $\tau_{APLE,0}$, reported in the body of the paper, are similar to other estimates of τ_{ATE}, τ_{ATT}, and τ_{ATU}, reported in online appendix E.

A The Effects of a Training Program on Earnings

I first consider the example from section IIA in more detail. This replication of the study of the effects of NSW program in Angrist and Pischke (2009) constitutes an optimistic scenario for OLS. In this application, as I explained in section IIA, the effect for the treated group (ATT) is likely to be substantially larger than the effect for the CPS comparison group (ATU). Moreover, since the experimental benchmark of \$1,794 corresponds to $\widehat{\text{ATT}}$ and not to $\widehat{\text{ATU}}$, the researcher should also focus on ATT. It turns out that my diagnostic for estimating ATT, $\hat{w}_0$, indicates that this parameter should approximately be recovered by OLS, even if treatment effects are heterogeneous.[14]

The top and middle panels of Table 1 reproduce the estimates from Angrist and Pischke (2009) and report my diagnostics. The specification in column 4 was discussed in section IIA. It turns out that $\hat{w}_0$ is between 0.1% and 1.9% for all specifications; similarly, the "rule of thumb" value of this diagnostic, $\hat{w}_0^*$, is, as always, equal to the proportion of treated units (only 1.1% in this sample). These results are very simple to interpret. Namely, as in section IID, we estimate that the difference between the OLS estimand and ATT is less than 2% of the difference between ATU and ATT. In this case, it might indeed be sensible to rely on the OLS estimates of the effect of treatment.

The bottom panel of Table 1 provides an application of Corollary 1 to these results. In other words, the estimates from Angrist and Pischke (2009) are now decomposed into two components, $\widehat{\text{ATT}}$ and $\widehat{\text{ATU}}$. The difference between these estimates is substantial. In column 4, while the estimate of ATT is \$928, ATU is estimated to be –\$6,840. In other words, the OLS estimate of \$794, reported in Angrist and Pischke (2009) and discussed in section IIA, is actually a weighted average of these two estimates. The fact that it is close to \$928, and not to –\$6,840, is a consequence of the small proportion of treated units in this sample, 1.1%. The weight on \$928, $\hat{w}_1$, is 98.3% and the

[14] It is well known that, in the NSW–CPS data, there is limited overlap in terms of covariate values between the treated and untreated units (see, e.g., Dehejia and Wahba, 1999; Smith and Todd, 2005). Thus, it is important to note that my theoretical results in section II do not impose the overlap assumption.

Table 1: The Effects of a Training Program on Earnings

	(1)	(2)	(3)	(4)
	Original estimates			
OLS	–3,437***	–78	623	794
	(612)	(596)	(610)	(619)
	Diagnostics			
$\hat{w}_0$	0.019	0.001	0.017	0.017
$\hat{w}_0^* = \hat{\rho}$	0.011	0.011	0.011	0.011
$\hat{\delta}$	–0.970	–0.987	–0.971	–0.971
$\hat{\delta}^* = 2\hat{\rho} - 1$	–0.977	–0.977	–0.977	–0.977
	Decomposition			
$\widehat{\text{ATT}}$	–3,373***	–69	754	928
	(620)	(595)	(619)	(630)
$\hat{w}_1$	0.981	0.999	0.983	0.983
$\widehat{\text{ATU}}$	–6,753***	–6,289**	–6,841***	–6,840***
	(1,219)	(2,807)	(1,294)	(1,319)
$\hat{w}_0$	0.019	0.001	0.017	0.017
$\widehat{\text{ATE}}$	–6,714***	–6,218**	–6,754***	–6,751***
	(1,206)	(2,777)	(1,281)	(1,305)
Demographic controls	✓		✓	✓
Earnings in 1974				✓
Earnings in 1975		✓	✓	✓
$\hat{\rho} = \hat{\text{P}}(d = 1)$	0.011	0.011	0.011	0.011
Observations	16,177	16,177	16,177	16,177

Notes: The estimates in the top panel correspond to column 2 in Table 3.3.3 in Angrist and Pischke (2009, p. 89). The dependent variable is earnings in 1978. Demographic controls include age, age squared, years of schooling, and indicators for married, high school dropout, black, and Hispanic. For treated individuals, earnings in 1974 correspond to real earnings in months 13–24 prior to randomization, which overlaps with calendar year 1974 for a number of individuals. Formulas for w_0, w_1, and δ are given in Theorem 1 and Corollary 2. Following these results, OLS = $\hat{w}_1 \cdot \widehat{\text{ATT}} + \hat{w}_0 \cdot \widehat{\text{ATU}}$. Estimates of ATE, ATT, and ATU are sample analogues of τ_{APLE}, $\tau_{APLE,1}$, and $\tau_{APLE,0}$, respectively. Also, $\widehat{\text{ATE}} = \hat{\rho} \cdot \widehat{\text{ATT}} + (1 - \hat{\rho}) \cdot \widehat{\text{ATU}}$. Huber–White standard errors (OLS) and bootstrap standard errors ($\widehat{\text{ATE}}$, $\widehat{\text{ATT}}$, and $\widehat{\text{ATU}}$) are in parentheses. *Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

weight on –$6,840, $\hat{w}_0$, is only 1.7%.

We might expect that if the proportion of treated units was larger, the weight on $\widehat{\text{ATT}}$ would be smaller and the "performance" of OLS in replicating the experimental benchmark would deteriorate. I confirm this conjecture in online appendix E1 by quasi-discarding "random" subsamples

of untreated units over a range of sample sizes. In particular, I reestimate the model in (1) using WLS, with weights of 1 for treated and $\frac{1}{k}$ for untreated units. Figures E1.1 to E1.4 show that in this application WLS estimates become more negative as k increases. This is because larger values of k correspond to greater proportions of untreated units being "discarded," and hence *larger* weights on $\widehat{\text{ATU}}$, which is substantially more negative than $\widehat{\text{ATT}}$.

Additional extensions of my analysis are also presented in online appendix E1. For each specification in Table 1, I provide both a linear and a nonparametric estimate of the conditional mean of the outcome given $p(X)$, separately for treated and untreated units (Figures E1.5 to E1.8). A visual comparison of both estimates provides an informal test of Assumption 4, which is necessary for a causal interpretation of τ_{APLE}, $\tau_{APLE,1}$, and $\tau_{APLE,0}$. The linearity assumption appears to be approximately satisfied for the treated but usually not for the untreated units.

Thus, as a robustness check, I also report a number of alternative estimates of the effects of NSW program in Table E1.1. I consider regression adjustment, as in section IIA, as well as matching on $p(X)$ and on the logit propensity score.[15] In each case, I separately estimate ATE, ATT, and ATU. These estimates are consistent with the claim that the general pattern of results in Table 1 is driven by the OLS weights. The estimates of ATE and ATU are always negative and large in magnitude; the estimates of ATT are much closer to the experimental benchmark.

Finally, I repeat the following exercise from section IIA. When we match the OLS estimates in Table 1 with the corresponding estimates of ATT and ATU in Table E1.1, we can write $\hat{\tau} = \hat{w}_{ATT} \cdot \hat{\tau}_{ATT} + (1 - \hat{w}_{ATT}) \cdot \hat{\tau}_{ATU}$. Unless $\hat{\tau}_{ATT}$ and $\hat{\tau}_{ATU}$ are sample analogues of $\tau_{APLE,1}$ and $\tau_{APLE,0}$, $\hat{w}_{ATT}$ does not need to be bounded between zero and one. Yet, we can solve for $\hat{w}_{ATT}$ for each set of estimates. The mean of $\hat{w}_{ATT}$ across all sets of estimates in Table E1.1 is 98.3%, which is nearly identical to the sample proportion of untreated units, 98.9%. This is reassuring for my claims.

B The Effects of Cash Transfers on Longevity

In my second application, I replicate a recent paper by Aizer et al. (2016) and study the effects of cash transfers on longevity of the children of their beneficiaries, as measured by their log age

[15] In particular, the estimates discussed in section IIA are reported in column 4 of the bottom panel of Table E1.1.

at death. In particular, Aizer et al. (2016) analyze the administrative records of applicants to the Mothers' Pension (MP) program, which supported poor mothers with dependent children in pre-WWII United States. In this study, the untreated group consists only of children of mothers who applied for a transfer, were initially deemed eligible, but were ultimately rejected. This strategy is used to ensure that treated and untreated individuals are broadly comparable, and hence an ignorability assumption might be plausible. Nevertheless, rejected mothers were slightly older and came from slightly smaller and richer families than accepted mothers. Thus, as before, there is no reason to believe that ATT and ATU are equal, although it is perhaps less clear a priori which is larger. Unlike in section IIIA, it seems plausible that the researcher might be interested either in the average effect of cash transfers, ATE, or in their average effect for accepted applicants, ATT.

The top and middle panels of Table 2 reproduce the baseline estimates from Aizer et al. (2016) and report my diagnostics. While the OLS estimates are positive and statistically significant, my diagnostics indicate that these results should be approached with caution. Namely, treated units constitute the vast majority (or 87.5%) of the sample. It follows that OLS is expected to place a disproportionately large weight on $\widehat{\text{ATU}}$, in which case the OLS estimates might be very biased for both ATE and ATT (cf. Corollaries 2 and 3). Indeed, my estimates of δ suggest that the difference between the OLS estimand and ATE is equal to 65.9–74.5% of the difference between ATU and ATT. Also, the estimates of w_0 suggest that the difference between OLS and ATT corresponds to 78.4–87.0% of this measure of heterogeneity. The estimates of δ^* and w_0^* are similar. It turns out that in this application the OLS estimates might be substantially biased for both of our parameters of interest. This would be a pessimistic scenario for OLS.

The results in the bottom panel of Table 2 suggest that these biases are indeed substantial. In this panel, following Corollary 1, each OLS estimate from Aizer et al. (2016) is represented as a weighted average of estimates of two effects, on accepted (ATT) and rejected (ATU) applicants. The estimates of ATU are consistently larger than those of ATT. Thus, OLS overestimates both ATE (since $\hat{\delta} > 0$) and ATT. While the implicit OLS estimates of these parameters remain statistically significant in columns 1 and 2, this is no longer the case in columns 3 and 4, following the

Table 2: The Effects of Cash Transfers on Longevity

	(1)	(2)	(3)	(4)
	Original estimates			
OLS	0.0157***	0.0158***	0.0182***	0.0167***
	(0.0058)	(0.0059)	(0.0062)	(0.0061)
	Diagnostics			
$\hat{w}_0$	0.861	0.870	0.784	0.784
$\hat{w}_0^* = \hat{\rho}$	0.875	0.875	0.875	0.875
$\hat{\delta}$	0.736	0.745	0.659	0.659
$\hat{\delta}^* = 2\hat{\rho} - 1$	0.750	0.750	0.750	0.750
	Decomposition			
$\widehat{\text{ATT}}$	0.0129**	0.0149**	0.0097	0.0089
	(0.0064)	(0.0071)	(0.0078)	(0.0079)
$\hat{w}_1$	0.139	0.130	0.216	0.216
$\widehat{\text{ATU}}$	0.0162***	0.0160***	0.0206***	0.0188***
	(0.0057)	(0.0059)	(0.0063)	(0.0064)
$\hat{w}_0$	0.861	0.870	0.784	0.784
$\widehat{\text{ATE}}$	0.0133**	0.0150**	0.0110	0.0102
	(0.0063)	(0.0068)	(0.0073)	(0.0074)
State fixed effects	✓			
County fixed effects			✓	✓
Cohort fixed effects	✓	✓	✓	✓
State characteristics		✓	✓	✓
County characteristics		✓		
Individual characteristics		✓	✓	✓
$\hat{\rho} = \hat{\text{P}}(d = 1)$	0.875	0.875	0.875	0.875
Observations	7,860	7,859	7,859	7,857

Notes: The estimates in the top panel correspond to columns 1 to 4 in panel A of Table 4 in Aizer et al. (2016, p. 952). The dependent variable is log age at death, as reported in the MP records (columns 1 to 3) or on the death certificate (column 4). State, county, and individual characteristics are listed in Table E2.1 in online appendix E2. Formulas for w_0, w_1, and δ are given in Theorem 1 and Corollary 2. Following these results, OLS $= \hat{w}_1 \cdot \widehat{\text{ATT}} + \hat{w}_0 \cdot \widehat{\text{ATU}}$. Estimates of ATE, ATT, and ATU are sample analogues of τ_{APLE}, $\tau_{APLE,1}$, and $\tau_{APLE,0}$, respectively. Also, $\widehat{\text{ATE}} = \hat{\rho} \cdot \widehat{\text{ATT}} + (1 - \hat{\rho}) \cdot \widehat{\text{ATU}}$. Huber–White standard errors (OLS) and bootstrap standard errors ($\widehat{\text{ATE}}$, $\widehat{\text{ATT}}$, and $\widehat{\text{ATU}}$) are in parentheses.
*Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

inclusion of county fixed effects. Perhaps more importantly, these estimates of ATT are half smaller than the corresponding OLS estimates. Clearly, this difference is economically quite meaningful.

To assess the robustness of these findings, I present several extensions of my analysis in online appendix E2. The informal test of Assumption 4, as discussed in section IIIA, appears to suggest that the conditional mean of the outcome given $p(X)$ is approximately linear for both the treated and untreated units (see Figures E2.5 to E2.8). I also report a number of alternative estimates of the effects of cash transfers in Table E2.1. These additional results support my conclusion. Only one in twelve estimates of ATT is statistically different from zero, and four of the insignificant estimates are negative. While it is possible that cash transfers increase longevity, the OLS estimates reported in Aizer et al. (2016) are almost certainly too large. Interestingly, this bias appears to be driven by the implicit OLS weights on ATT and ATU, which were the focus of this paper.[16]

IV Conclusion

This paper proposed a new interpretation of the OLS estimand for the effect of a binary treatment in the standard linear model with additive effects. According to the main result of this paper, the OLS estimand is a convex combination of two parameters, which under certain conditions are equivalent to the average treatment effects on the treated (ATT) and untreated (ATU). Surprisingly, the weights on these parameters are inversely related to the proportion of observations in each group, which can lead to substantial biases when interpreting the OLS estimand as ATE or ATT.

One lesson from this result is that it might be preferable, as suggested by a body of work in econometrics, to use any of the standard estimators of average treatment effects under ignorability, such as regression adjustment, weighting, matching, and various combinations of these approaches. Empirical researchers with a preference for OLS might instead want to use the diagnostic tools that this paper also provided. These diagnostics, which are implemented in the `hettreatreg` package in R and Stata, are applicable whenever the researcher is: *(i)* studying the effects of a binary treatment, *(ii)* using OLS, and *(iii)* unwilling to maintain that ATT is exactly equal to ATU. In an important special case, these diagnostics only require the knowledge of the proportion of treated units.

[16] I also repeat two further exercises from section IIIA. First, after I reestimate the model in (1) using WLS, with weights of 1 for treated and $\frac{1}{k}$ for untreated units, I demonstrate in Figures E2.1 to E2.4 that these estimates become more positive as k increases. As before, larger values of k translate into larger weights on $\widehat{\text{ATU}}$, which is now greater than $\widehat{\text{ATT}}$. Second, when I use the estimates of ATT and ATU in Table E2.1 to recover the hypothetical OLS weights, I obtain 22.8% as the mean of $\hat{w}_{ATT}$. This is reasonably similar to the proportion of untreated units, 12.5%.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, 88:265–296.

Abadie, A. and Cattaneo, M. D. (2018). Econometric methods for program evaluation. *Annual Review of Economics*, 10:465–503.

Aizer, A., Eli, S., Ferrie, J., and Lleras-Muney, A. (2016). The long-run impact of cash transfers to poor families. *American Economic Review*, 106:935–971.

Alesina, A., Giuliano, P., and Nunn, N. (2013). On the origins of gender roles: Women and the plough. *Quarterly Journal of Economics*, 128:469–530.

Angrist, J. D. (1998). Estimating the labor market impact of voluntary military service using Social Security data on military applicants. *Econometrica*, 66:249–288.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Aronow, P. M. and Samii, C. (2016). Does regression produce representative estimates of causal effects? *American Journal of Political Science*, 60:250–267.

Bitler, M. P., Gelbach, J. B., and Hoynes, H. W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review*, 96:988–1012.

Card, D., Kluve, J., and Weber, A. (2018). What works? A meta analysis of recent active labor market program evaluations. *Journal of the European Economic Association*, 16:894–931.

Deaton, A. (1997). *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Johns Hopkins University Press.

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

Elder, T. E., Goddeeris, J. H., and Haider, S. J. (2010). Unexplained gaps and Oaxaca–Blinder decompositions. *Labour Economics*, 17:284–290.

Graham, B. S. and Pinto, C. C. d. X. (2018). Semiparametrically efficient estimation of the average linear regression function. NBER Working Paper no. 25234.

Heckman, J. J. (2001). Micro data, heterogeneity, and the evaluation of public policy: Nobel Lecture. *Journal of Political Economy*, 109:673–748.

Humphreys, M. (2009). Bounds on least squares estimates of causal effects in the presence of heterogeneous assignment probabilities. Unpublished.

Imbens, G. W. (2015). Matching methods in practice: Three examples. *Journal of Human Resources*, 50:373–419.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

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

Kline, P. (2011). Oaxaca–Blinder as a reweighting estimator. *American Economic Review: Papers & Proceedings*, 101:532–537.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76:604–620.

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

Słoczyński, T. (2018). A general weighted average representation of the ordinary and two-stage least squares estimands. IZA Discussion Paper no. 11866.

Słoczyński, T. (2020). Average gaps and Oaxaca–Blinder decompositions: A cautionary tale about regression estimates of racial differences in labor market outcomes. *ILR Review*, 73:705–729.

Smith, J. A. and Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125:305–353.

Solon, G., Haider, S. J., and Wooldridge, J. M. (2015). What are we weighting for? *Journal of Human Resources*, 50:301–316.

Voigtländer, N. and Voth, H.-J. (2012). Persecution perpetuated: The medieval origins of anti-Semitic violence in Nazi Germany. *Quarterly Journal of Economics*, 127:1339–1392.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.