New Evidence on Linear Regression and Treatment Effect Heterogeneity

Słoczyński, Tymon

Warsaw School of Economics

27 June 2012

New Evidence on Linear Regression and Treatment Effect Heterogeneity*

Tymon Słoczyński†

June 27, 2012

In this paper I provide new evidence on the implications of treatment effect heterogeneity for least squares estimation when the effects are inappropriately assumed to be homogenous. I prove that under a set of benchmark assumptions linear regression provides a consistent estimator of the population average treatment effect on the treated times the population proportion of the nontreated individuals plus the population average treatment effect on the nontreated times the population proportion of the treated individuals. Consequently, in many empirical applications the linear regression estimates might not be close to any of the standard average treatment effects of interest.

*I am grateful to Joshua Angrist, Guido Imbens, Nicholas Longford, Łukasz Marć, and seminar and workshop participants at CEPS/INSTEAD and the Warsaw School of Economics for their comments and suggestions at different stages of this project. I also gratefully acknowledge the financial support from the "Weź stypendium – dla rozwoju" scholarship programme, funded by the European Social Fund and administered by the Warsaw School of Economics. All errors remain my own.

†Warsaw School of Economics, Department of Economics I, tymon.sloczynski@gmail.com.

1 Introduction

In his Nobel Lecture, James J. Heckman asserted that "[t]he most important discovery [of microeconometrics] was the evidence on the pervasiveness of heterogeneity and diversity in economic life" (Heckman 2001, p. 674). A large part of the literature on programme evaluation seeks, therefore, to explore heterogeneity in the response to treatment. For example, Heckman, Smith, and Clements (1997) and Djebbari and Smith (2008) develop a framework to study treatment effect heterogeneity using experimental data. Crump, Hotz, Imbens, and Mitnik (2008) propose and apply two nonparametric tests of treatment effect heterogeneity under unconfoundedness. Various estimators of quantile treatment effects (QTEs) have also been proposed (e.g., Abadie, Angrist, and Imbens 2002; Chernozhukov and Hansen 2005; Firpo 2007; Frölich and Melly 2008) and applied (e.g., Bitler, Gelbach, and Hoynes 2006, 2008) in recent papers. The empirical importance of treatment effect heterogeneity has invariably been confirmed.

At the same time, the homogeneous linear regression model is often believed to provide a good benchmark to study treatment effects, i.e. partial effects for a binary explanatory variable. A convincing explanation is given in Angrist and Pischke (2009), while many influential studies (e.g., Neal and Johnson 1996; Fryer and Levitt 2004) explicitly rely on linear regression to capture the possibly heterogeneous effects for a binary variable. In this paper my goal is to provide new evidence on the limitations of such an approach in light of "the pervasiveness of heterogeneity". In particular, what is the appropriate interpretation of the least squares estimand in the homogeneous linear model if treatment effects are actually heterogeneous? In this paper I provide a new answer to this question by exploiting the link between linear regression and the Oaxaca–Blinder decomposition (Blinder 1973; Oaxaca 1973) as well as utilizing a recent theoretical result in Elder, Goddeeris, and Haider (2010). I prove that under a set of benchmark assumptions linear regression provides a consistent estimator of the population average treatment effect on the treated times the population proportion of the nontreated individuals plus the population average treatment effect on the nontreated times the population proportion of the treated individuals. In other words, under the assumptions of (i) a single control variable

(ii) whose variance is equal in both subpopulations the linear regression estimand is a weighted average of both subpopulation-specific average treatment effects; while weights are equal to the population proportions of both groups, they are inappropriately interchanged between them. Consequently, the ability of linear regression to provide a good benchmark to study treatment effects is heavily data-dependent. Least squares estimation can be preferred on efficiency grounds if there is little heterogeneity in treatment effects or both subsamples are of approximately equal size; in the latter case both weights are more or less equal anyway. However, in other cases linear regression will provide biased estimates of all the standard average treatment effects of interest, even asymptotically.

Similar research on linear regression and treatment effect heterogeneity has been done in the past, although very little compared to the growing literature on impact heterogeneity. The key result is given in Angrist (1998). It is shown that in a saturated model the weights underlying linear regression are proportional to the variance of treatment at each combination of covariate values. This analysis has recently been extended in Humphreys (2009) by proving that the linear regression estimand is bounded by both subpopulation-specific average treatment effects whenever treatment assignment probabilities are monotonic in covariate-specific treatment effects.[1] My analysis distinguishes itself by completely relaxing the saturated model restriction. Such models are utilised in few applied studies (e.g., Angrist 1998; Black, Smith, Berger, and Noel 2003), while being inapplicable if any of the control variables are continuous. It is also unclear whether any theoretical analyses of saturated models can be generalised to the standard case of nonsaturated linear regression.

The remainder of the paper is organised as follows. Section 2 provides the main theoretical contribution of this study. Section 3 illustrates this proposition with a simple Monte Carlo experiment, while showing how the OLS weights on subpopulation-specific average treatment effects are inversely proportional to the sample proportions of both groups. Section 4 provides a further illustration through a reanalysis of the National

[1]Other related contributions include Yitzhaki (1996) and Angrist and Krueger (1999) who analyse the implicit weights on the estimated partial effects for a continuous explanatory variable in least squares estimation. Recently, Løken, Mogstad, and Wiswall (2012) have analysed the weighting of the partial effects in OLS, IV, and FE estimation when the estimated model is inappropriately assumed to be linear.

Supported Work (NSW) data (see, e.g., LaLonde 1986; Dehejia and Wahba 1999; Smith and Todd 2005). It is shown that good performance of linear regression in replicating the NSW experimental benchmark in Angrist and Pischke (2009) is a consequence of the large sample proportion of the nontreated individuals. Section 5 summarises and concludes.

2 Theory

In this section I use the standard potential outcomes framework and standard notation. Since this framework is now widespread in econometric literature, I do not provide a detailed description here.[2] Consider therefore a population of N individuals, indexed by $i = 1, \ldots, N$. The potential outcomes are denoted by y_{1i} (the treated outcome) and y_{0i} (the nontreated outcome), while the realised outcome is denoted by y_i. Also denote group membership (treatment) by d_i; consequently, $d_i = 1$ for the treated individuals and $d_i = 0$ for the nontreated individuals. We also observe X_i, a row vector of covariates (x_i if scalar).

The literature on programme evaluation seeks to identify and estimate *treatment effects* for various subpopulations. The individual-specific treatment effect is defined as $\tau_i = y_{1i} - y_{0i}$, i.e. the difference between the potential outcomes of a given individual. These individual-specific treatment effects are averaged and various average treatment effects are estimated. The population average treatment effect (PATE) is defined as:

$$\tau_{PATE} = \mathrm{E}[\tau_i] = \mathrm{E}[y_{1i} - y_{0i}]. \tag{1}$$

Similarly, one can define the population average treatment effect on the treated (PATT) and the population average treatment effect on the nontreated (PATN) as:

$$\tau_{PATT} = \mathrm{E}[\tau_i \mid d_i = 1], \tag{2}$$

$$\tau_{PATN} = \mathrm{E}[\tau_i \mid d_i = 0]. \tag{3}$$

[2]Recent reviews are given in Angrist and Pischke (2009), Blundell and Costa Dias (2009), Imbens and Wooldridge (2009), and Wooldridge (2010).

A major strand in the treatment effects literature, typically referred to as selection on observables, is based on the so-called unconfoundedness assumption, i.e. it assumes that treatment is orthogonal to the potential outcomes, conditional on X_i. Under the unconfoundedness assumption the standard average treatment effects of interest are typically estimated using regression methods (see, e.g., Angrist and Pischke 2009), matching on covariates (see, e.g., Abadie and Imbens 2006, 2011), and methods based on the propensity score (see, e.g., Rosenbaum and Rubin 1983; Dehejia and Wahba 1999; Hirano, Imbens, and Ridder 2003). When the model for outcomes is linear ($y_i = X_i\beta_1 + \upsilon_{1i}$ if $d_i = 1$; $y_i = X_i\beta_0 + \upsilon_{0i}$ if $d_i = 0$), the PATT and the PATN can also be estimated using the two original versions of the Oaxaca–Blinder decomposition (see, e.g., Barsky, Bound, Charles, and Lupton 2002; Melly 2006; Fortin, Lemieux, and Firpo 2011; Kline 2011). Precisely:

$$\tau_{PATT} = \mathrm{E}[X_i \mid d_i = 1] \cdot (\beta_1 - \beta_0), \tag{4}$$

$$\tau_{PATN} = \mathrm{E}[X_i \mid d_i = 0] \cdot (\beta_1 - \beta_0). \tag{5}$$

Now, we can proceed to the main proposition of this paper which will provide a reinterpretation of the least squares estimand in the homogeneous linear model when treatment effects are actually heterogeneous. Its proof will utilise the link between linear regression and the Oaxaca–Blinder decomposition as well as the ability of this decomposition method to provide consistent estimates of various average treatment effects of interest.

Proposition. *Under the assumption of unconfoundedness the coefficient on a binary treatment variable in linear least squares regression is a consistent estimator of* $Pr[d_i = 0] \cdot \tau_{PATT} + Pr[d_i = 1] \cdot \tau_{PATN}$*, i.e. the population average treatment effect on the treated times the population proportion of the nontreated individuals plus the population average treatment effect on the nontreated times the population proportion of the treated individuals, provided that there is a single control variable whose variance is equal in both subpopulations.*

Proof. See Elder, Goddeeris, and Haider (2010, Appendix A) for a proof that the coeffi-

cient on a binary variable in linear least squares regression is computationally equivalent to the unexplained component from the extension of the Oaxaca–Blinder decomposition proposed by Cotton (1988), provided that there is a single control variable whose variance is equal in both subpopulations (also find a simplified version of this proof in the Appendix). Next, consider the following lemma which is an original contribution of the present paper.

LEMMA. *Under the assumption of unconfoundedness the unexplained component from the extension of the Oaxaca–Blinder decomposition proposed by Cotton (1988) is a consistent estimator of the population average treatment effect on the treated times the population proportion of the nontreated individuals plus the population average treatment effect on the nontreated times the population proportion of the treated individuals.*

PROOF. The generalised Oaxaca–Blinder decomposition (with the extension proposed by Cotton 1988 accommodated as a special case) decomposes the difference between average outcomes in both groups of interest in the following way:

$$\begin{aligned}
\mathrm{E}[y_i \mid d_i = 1] - \mathrm{E}[y_i \mid d_i = 0] &= \mathrm{E}[X_i \mid d_i = 1]\beta_1 - \mathrm{E}[X_i \mid d_i = 0]\beta_0 \\
&= \mathrm{E}[X_i \mid d_i = 1]\beta_1 - \mathrm{E}[X_i \mid d_i = 0]\beta_0 \\
&\quad + \mathrm{E}[X_i \mid d_i = 1]\beta^* - \mathrm{E}[X_i \mid d_i = 1]\beta^* \\
&\quad + \mathrm{E}[X_i \mid d_i = 0]\beta^* - \mathrm{E}[X_i \mid d_i = 0]\beta^* \\
&= \mathrm{E}[X_i \mid d_i = 1] \cdot (\beta_1 - \beta^*) + \mathrm{E}[X_i \mid d_i = 0] \cdot (\beta^* - \beta_0) \\
&\quad + (\mathrm{E}[X_i \mid d_i = 1] - \mathrm{E}[X_i \mid d_i = 0])\beta^*, \qquad (6)
\end{aligned}$$

where the unexplained component is equal to:

$$\begin{aligned}
\tau^* &= (\mathrm{E}[y_i \mid d_i = 1] - \mathrm{E}[y_i \mid d_i = 0]) - (\mathrm{E}[X_i \mid d_i = 1] - \mathrm{E}[X_i \mid d_i = 0])\beta^* \\
&= \mathrm{E}[X_i \mid d_i = 1] \cdot (\beta_1 - \beta^*) + \mathrm{E}[X_i \mid d_i = 0] \cdot (\beta^* - \beta_0). \qquad (7)
\end{aligned}$$

The extension of the Oaxaca–Blinder decomposition proposed by Cotton (1988) uses

$\beta^* = \beta^C = \Pr[d_i = 1] \cdot \beta_1 + \Pr[d_i = 0] \cdot \beta_0$. The difference between average outcomes in both groups of interest can thus be written as:

$$\begin{aligned}
\mathrm{E}[y_i \mid d_i = 1] - \mathrm{E}[y_i \mid d_i = 0] &= \mathrm{E}[X_i \mid d_i = 1] \cdot (\beta_1 - \beta^C) + \mathrm{E}[X_i \mid d_i = 0] \cdot (\beta^C - \beta_0) \\
&\quad + (\mathrm{E}[X_i \mid d_i = 1] - \mathrm{E}[X_i \mid d_i = 0])\beta^C \\
&= \mathrm{E}[X_i \mid d_i = 1] \cdot (\beta_1 - (\Pr[d_i = 1] \cdot \beta_1 + \Pr[d_i = 0] \cdot \beta_0)) \\
&\quad + \mathrm{E}[X_i \mid d_i = 0] \cdot ((\Pr[d_i = 1] \cdot \beta_1 + \Pr[d_i = 0] \cdot \beta_0) - \beta_0) \\
&\quad + (\mathrm{E}[X_i \mid d_i = 1] - \mathrm{E}[X_i \mid d_i = 0])\beta^C \\
&= \Pr[d_i = 0] \cdot \mathrm{E}[X_i \mid d_i = 1] \cdot (\beta_1 - \beta_0) \\
&\quad + \Pr[d_i = 1] \cdot \mathrm{E}[X_i \mid d_i = 0] \cdot (\beta_1 - \beta_0) \\
&\quad + (\mathrm{E}[X_i \mid d_i = 1] - \mathrm{E}[X_i \mid d_i = 0])\beta^C \\
&= \Pr[d_i = 0] \cdot \tau_{PATT} + \Pr[d_i = 1] \cdot \tau_{PATN} \\
&\quad + (\mathrm{E}[X_i \mid d_i = 1] - \mathrm{E}[X_i \mid d_i = 0])\beta^C, \qquad (8)
\end{aligned}$$

where the unexplained component is equal to:

$$\begin{aligned}
\tau^C &= (\mathrm{E}[y_i \mid d_i = 1] - \mathrm{E}[y_i \mid d_i = 0]) - (\mathrm{E}[X_i \mid d_i = 1] - \mathrm{E}[X_i \mid d_i = 0])\beta^C \\
&= \Pr[d_i = 0] \cdot \tau_{PATT} + \Pr[d_i = 1] \cdot \tau_{PATN}. \qquad (9)
\end{aligned}$$

Hence, the lemma is proven. A combination of this lemma with the above-mentioned result in Elder, Goddeeris, and Haider (2010, Appendix A) proves the proposition of the present paper.

Q.E.D.

3 Monte Carlo Evidence

In the previous section I proved that under a set of benchmark assumptions linear regression might provide biased estimates of all the standard average treatment effects of

interest, since it inappropriately interchanges the implicit weights on both subpopulation-specific average treatment effects. However, linear regression is widely used in empirical research in many different disciplines and such a pessimistic result might seem counter-intuitive at first. Therefore, this section is an attempt to provide a simple Monte Carlo illustration of this proposition.

The data generating process I consider is designed in such a way that the assumptions of the proposition in Section 2 hold:

$$y_i = \alpha + \beta x_i + \tau_i d_i + \upsilon_i, \tag{10}$$

where $\alpha = 25$, $\beta = 20$, and $\upsilon_i \sim \mathcal{N}[0, 50]$. Although the true model in Equation 10 allows for treatment effect heterogeneity, a homogeneous linear regression model is estimated and its estimand is denoted as τ_{LR}. Clearly, there is a single control variable (x_i), but we also require its variance to be equal in both subpopulations. The joint distribution of x_i, d_i, and τ_i is thus presented in Table 1.

The conditional variance of x_i as well as all the standard average treatment effects of interest can be easily calculated using information in Table 1. Precisely, while $\mathrm{V}[x_i \mid d_i = 1] = \mathrm{V}[x_i \mid d_i = 0] = 1.5$, $\mathrm{E}[x_i \mid d_i = 1] = 2 \neq 4 = \mathrm{E}[x_i \mid d_i = 0]$. Also $\tau_{PATT} = \mathrm{E}[\tau_i \mid d_i = 1] = 110$ and $\tau_{PATN} = \mathrm{E}[\tau_i \mid d_i = 0] = 200$. Since $\Pr[d_i = 1] = 1/3$ and $\Pr[d_i = 0] = 2/3$, $\tau_{PATE} = \mathrm{E}[\tau_i] = 170$. Although intuition might suggest that $\tau_{LR} \approx \tau_{PATE}$, the proposition in Section 2 provides an assertion that linear least squares regression is actually based on an inappropriate weighting scheme, i.e. $\tau_{LR} = \tau_{PATT} \cdot \Pr[d_i = 0] + \tau_{PATN} \cdot \Pr[d_i = 1] = 140$. Such a claim is now illustrated with 10,000 replications of this data generating process. The results are presented in Figure 1.

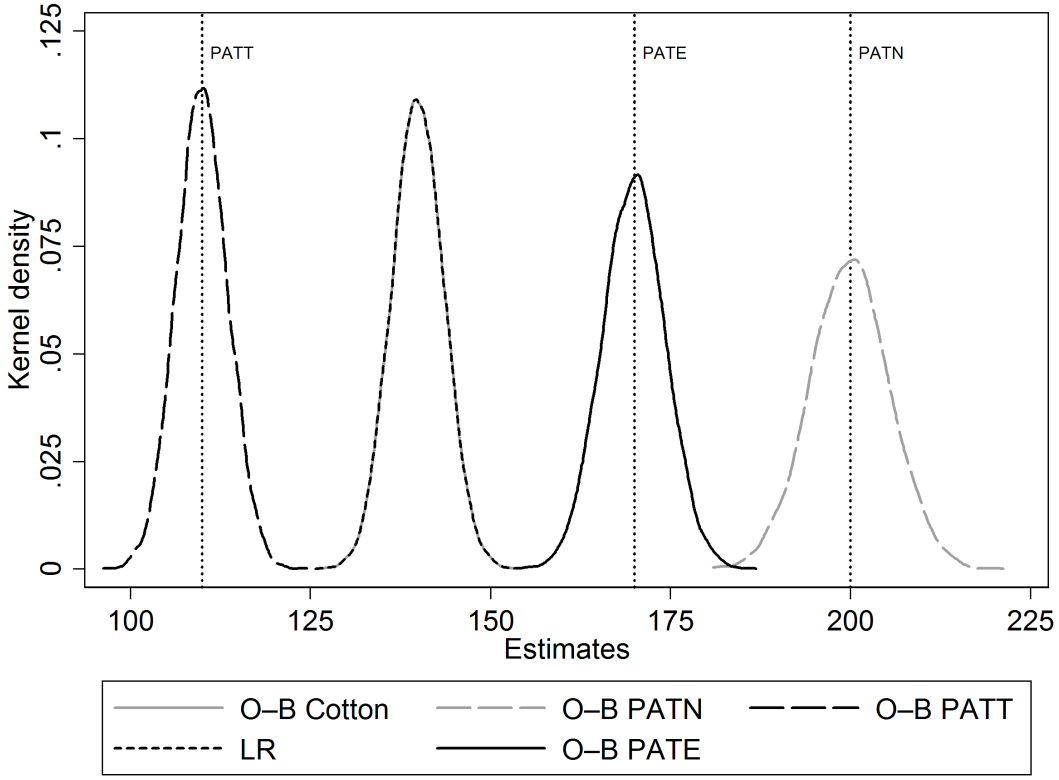
Figure 1 displays the empirical distribution of several linear estimators of various average treatment effects.[3] Four versions of the Oaxaca–Blinder decomposition are used in the simulation. The two original versions of this decomposition method (Equations 4 and 5) are used to estimate the PATT and the PATN, respectively. These estimators are denoted by $\hat{\tau}_{PATT}$ and $\hat{\tau}_{PATN}$. Moreover, Oaxaca–Blinder estimates of the PATE

[3] All the applications of the O–B decomposition use the *oaxaca* command in Stata (Jann 2008).

Table 1: The Joint Distribution of the Control Variable, the Treatment Variable, and the Treatment Effect

	Number of observations		
x_i	$d_i = 1$	$d_i = 0$	τ_i
1	500	150	$\mathcal{N}[65, 75]$
2	200	175	$\mathcal{N}[110, 75]$
3	150	75	$\mathcal{N}[155, 75]$
4	100	725	$\mathcal{N}[200, 75]$
5	50	875	$\mathcal{N}[245, 75]$
	1000	2000	

Figure 1: Monte Carlo Simulations of Linear Regression and Oaxaca–Blinder Estimators of Various Average Treatment Effects



are calculated as $\hat{\tau}_{PATE} = \hat{\tau}_{PATT} \cdot \hat{\Pr}[d_i = 1] + \hat{\tau}_{PATN} \cdot \hat{\Pr}[d_i = 0]$, while the extension of the Oaxaca–Blinder decomposition proposed by Cotton (1988) provides the following estimator: $\hat{\tau}_C = \hat{\tau}_{PATT} \cdot \hat{\Pr}[d_i = 0] + \hat{\tau}_{PATN} \cdot \hat{\Pr}[d_i = 1]$ (see Equation 9). As evident in Figure 1, the empirical distributions of $\hat{\tau}_{PATE}$, $\hat{\tau}_{PATT}$, and $\hat{\tau}_{PATN}$ are centred around the true values of the corresponding parameters. Similarly, $\hat{\tau}_C$ is centred around the PATT times the population proportion of the nontreated individuals plus the PATN times the population proportion of the treated individuals. At the same time, the theoretical result in Elder, Goddeeris, and Haider (2010) guarantees that in this setting $\hat{\tau}_C = \hat{\tau}_{LR}$ by construction. Indeed, this is evident in Figure 1 as well. Consequently, although $\tau_{PATE} = 170$, $\tau_{PATT} = 110$, and $\tau_{PATN} = 200$, the empirical distribution of $\hat{\tau}_{LR}$ is centred around 140.

An important negative consequence of the weighting scheme in linear regression (as proven in Section 2) is to attach the *greater* weight to $\hat{\tau}_{PATT}$ (i.e. the linear estimate of the effect on the treated), the *smaller* is the sample proportion of the treated individuals. This problematic property is illustrated in Figure 2 by manipulating the number of the nontreated individuals in simulated samples.

While the number of the nontreated individuals in simulated samples is manipulated in Figure 2, both the marginal distribution of x_i conditional on $d_i = 0$ and the number of the treated individuals are held constant. What follows, neither τ_{PATT} nor τ_{PATN} varies across sample compositions; on the other hand, τ_{PATE} does vary and the greater the sample proportion of the treated individuals, the smaller is τ_{PATE} (because $\tau_{PATN} = 200 > 110 = \tau_{PATT}$). As evident in Figure 2, however, the behaviour of $\hat{\tau}_{LR}$ is different. The *greater* the sample proportion of the treated individuals, the *greater* is $\hat{\tau}_{LR}$ and the *more distant* is its empirical distribution from τ_{PATT}, the population average treatment effect on the treated. Clearly, such a property of linear regression is highly undesirable.

4 An Application to the NSW Data

In the previous section I provided a simple illustration of the proposition in Section 2. The linear least squares estimator of average treatment effects, $\hat{\tau}_{LR}$, was shown to be the more distant from any of the group-specific average treatment effects, the larger is the sample proportion of the group in question. Yet the data generating process in Section 3 was designed in such a way that the assumptions of the proposition in Section 2 were satisfied. While these assumptions provide a useful benchmark, they are potentially quite restrictive. Therefore, in this section I examine empirically whether the proposition in Section 2 provides a good approximation to the behaviour of $\hat{\tau}_{LR}$ when the assumptions of this proposition do not hold.

In this study I use the well-known National Supported Work (NSW) data, analysed originally by LaLonde (1986) and subsequently by Heckman and Hotz (1989), Dehejia and Wahba (1999), Smith and Todd (2005), Angrist and Pischke (2009), Abadie and Imbens (2011), Kline (2011), and many others.[4] My starting point is a recent reanalysis by Angrist and Pischke (2009) who report that the linear regression estimates of the effect of the NSW training programme on subsequent earnings of programme participants are remarkably close to the experimental benchmark and other nonexperimental estimates.

4.1 Whose Effect of the NSW Training Programme Does Linear Regression Estimate?

In his seminal study of the NSW data, LaLonde (1986) discarded the control group from the original experimental evaluation of the NSW training programme and created six nonexperimental control groups using standard U.S. microeconomic datasets, the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID). These control groups have typically been referred to as CPS-1, CPS-2, CPS-3, PSID-1, PSID-2, and PSID-3. In a recent study of the NSW data, Angrist and Pischke (2009) restricted their attention to three control groups in total (the control group from the experiment,

[4]Since this dataset is described in many other papers (e.g., LaLonde 1986; Smith and Todd 2005), I do not provide a detailed description here.

CPS-1, and CPS-3) and used four different specifications for each control group. These variable selections together with all the available control groups and a wider array of linear estimators are used in the present paper. The estimates are presented in Table 2.

As evident in Table 2, although intuition might (incorrectly) suggest that $\tau_{LR} \approx \tau_{PATE}$, $\hat{\tau}_{LR}$ is generally far away from $\hat{\tau}_{PATE}$. Clearly, this is consistent with the proposition in Section 2, since neither there is little heterogeneity in the estimated effects of the NSW training programme (except for the comparison with the original control group and selected comparisons with PSID-3), nor both subsamples are of approximately equal size. At the same time, $\hat{\tau}_{LR}$ is generally quite close to $\hat{\tau}_{PATT}$ and $\hat{\tau}_{C}$. The latter remark is precisely the conclusion of the proposition in Elder, Goddeeris, and Haider (2010). However, please note that the assumptions of this proposition are false in the applications being considered. Such an observation would suggest that the propositions in Section 2 and in Elder, Goddeeris, and Haider (2010) might indeed provide a good approximation to the behaviour of $\hat{\tau}_{LR}$ even when the assumptions of these propositions do not hold. A further treatment of this observation is given in Figure 3.

Figure 3 displays a scatter plot of $\hat{\tau}_{LR}$ and $\hat{\tau}_{C}$, as presented in Table 2. As evident in Figure 3, the 45° line provides a good fit to these data, thus suggesting that even when the assumptions of the proposition in Section 2 do not hold, the linear least squares regression does indeed provide an estimator of the PATT times the population proportion of the nontreated individuals plus the PATN times the population proportion of the treated individuals. Consequently, since the sample proportion of the nontreated individuals is very large in most samples (especially CPS-1, CPS-2, and PSID-1), this estimated effect ($\hat{\tau}_{LR}$) is often very close to $\hat{\tau}_{PATT}$. At the same time, Oaxaca–Blinder estimates of the PATT replicate the experimental benchmark relatively well (see, e.g., Kline 2011). What follows, good performance of linear regression in replicating the NSW experimental benchmark in Angrist and Pischke (2009) is a direct consequence of the large sample proportion of the nontreated individuals. Were this proportion smaller, $\hat{\tau}_{LR}$ would deviate from $\hat{\tau}_{PATT}$. This corollary of the proposition in Section 2 is illustrated in the next subsection with a simple simulation-based study of the NSW data.

Table 2: Linear Regression and Oaxaca–Blinder Estimates of the Effects of the NSW Training Programme

Control Group	Specification[1]	$\hat{\Pr}[d_i = 1]$	Oaxaca–Blinder				$\hat{\tau}_{LR}$
			$\hat{\tau}_{PATT}$	$\hat{\tau}_{PATN}$	$\hat{\tau}_{PATE}$	$\hat{\tau}_{C}$	
NSW	(1)	0.416	1,786	1,478	1,606	1,658	1,670[a]
	(2)	0.416	1,753	1,746	1,749	1,750	1,750[a]
	(3)	0.416	1,751	1,455	1,578	1,628	1,636[a]
	(4)	0.416	1,785	1,471	1,602	1,654	1,676[a]
CPS-1	(1)	0.011	-3,417	-6,163	-6,132	-3,449	-3,437[a]
	(2)	0.011	-69	-6,289	-6,218	-140	-78[a]
	(3)	0.011	623	-5,017	-4,952	558	623[a]
	(4)	0.011	796	-4,996	-4,930	730	794[a]
CPS-2	(1)	0.072	-1,670	-2,770	-2,690	-1,750	-1,697
	(2)	0.072	-232	-2,753	-2,571	-415	-263
	(3)	0.072	415	-2,186	-1,997	226	362
	(4)	0.072	927	-2,141	-1,919	705	813
CPS-3	(1)	0.301	928	22	295	655	771[a]
	(2)	0.301	63	-465	-306	-96	-91[a]
	(3)	0.301	1,280	84	444	920	1,010[a]
	(4)	0.301	1,701	177	636	1,242	1,369[a]
PSID-1	(1)	0.069	-5,125	-12,728	-12,202	-5,651	-5,613
	(2)	0.069	-534	-12,010	-11,216	-1,328	-582
	(3)	0.069	507	-11,080	-10,279	-294	456
	(4)	0.069	827	-11,057	-10,235	5	795
PSID-2	(1)	0.422	-682	-2,702	-1,849	-1,535	-1,614
	(2)	0.422	1,023	-2,547	-1,039	-485	721
	(3)	0.422	1,592	-2,141	-564	15	874
	(4)	0.422	2,066	-2,028	-299	337	1,360
PSID-3	(1)	0.591	676	1,278	923	1,032	475
	(2)	0.591	1,420	1,266	1,357	1,329	1,370
	(3)	0.591	832	1,383	1,057	1,158	595
	(4)	0.591	1,462	1,481	1,470	1,473	1,107

[1] Specification (1) includes demographic controls only, i.e. age, age squared, years of schooling, and dummies for black, Hispanic, high school dropout, and married. Specification (2) includes 1975 earnings only. Specification (3) includes demographic controls and 1975 earnings. Specification (4) includes demographic controls, 1974 earnings, and 1975 earnings.

[a] Also appears in Angrist and Pischke (2009, p. 89).

4.2 Further Simulation Evidence

An unbiased estimator of τ_{PATT} or τ_{PATN} has the same expectation irrespective of the relative size of both subsamples. However, if a given estimator is unbiased for τ_{PATE}, its expectation changes whenever there is variation in the relative size of the treated group and the nontreated group. If, say, the number of the nontreated individuals *increases* and the number of the treated individuals is held constant, then the expectation of such an estimator moves *toward* τ_{PATN}. The contrary is true for $\hat{\tau}_{LR}$, as was proven in Section 2 under a set of benchmark assumptions. When the number of the nontreated individuals *increases*, $\hat{\tau}_{LR}$ moves *away* from τ_{PATN} and toward τ_{PATT}.

This undesirable property of linear regression is illustrated in Figure 4 for a situation in which the benchmark assumptions of Section 2 do not hold. A simulation-based study of the NSW data is performed with 10,000 replications for each sample size. In each of these replications, a random sample of size n is drawn from CPS-1 without replacement. Then, the new sample is merged with the treated group from the experimental evaluation of the NSW training programme and $\hat{\tau}_{LR}$ is calculated using the merged dataset.

As evident in Figure 4, $\hat{\tau}_{LR}$ is neither consistently centred around $\hat{\tau}_{PATT}$ nor around $\hat{\tau}_{PATN}$. It also does not change in accordance with $\hat{\tau}_{PATE}$ whenever n changes. On the contrary, when n grows, and hence the relative size of the nontreated group increases, $\hat{\tau}_{LR}$ converges to $\hat{\tau}_{PATT}$. In other words, again, the *larger* the sample proportion of a given group (treated or nontreated), the *more distant* is $\hat{\tau}_{LR}$ from the population average treatment effect on this group. This is indeed explained by the proposition in Section 2, although it should be noted that the rate of convergence in Figure 4 is faster than expected. For example, while the proposition in Section 2 would suggest that $\hat{\tau}_{LR} \approx -4700$ for $n = 10$, the empirical distribution of $\hat{\tau}_{LR}$ is centred around approx. -3000.

5 Conclusion

In this paper I have provided new evidence on the implications of treatment effect heterogeneity for least squares estimation when the effects are inappropriately assumed to

be homogenous. Although similar research is available in Angrist (1998) and Humphreys (2009), my contribution is novel in its complete relaxation of the saturated model restriction imposed in previous studies. In this paper I have proven that under a set of benchmark assumptions linear regression provides a consistent estimator of the population average treatment effect on the treated times the population proportion of the nontreated individuals plus the population average treatment effect on the nontreated times the population proportion of the treated individuals. Consequently, linear regression possesses a highly undesirable property in that it attaches the *greater* weight to the linear estimate of the population average treatment effect on the treated (nontreated), the *smaller* is the sample proportion of the treated (nontreated) individuals.

A general lesson to be drawn from this paper is that the weighting scheme in linear regression may drive the results in applied studies whenever heterogeneity in the response to treatment is sufficiently large and both subpopulations of interest are not of approximately equal size. In such a case linear regression will provide inconsistent estimates of all the standard average treatment effects of interest. In other cases linear regression might be preferred on efficiency and convenience grounds. However, the empirical importance of treatment effect heterogeneity has been confirmed by many applied studies (see, e.g., Heckman 2001 for a discussion), thus suggesting that the weighting scheme in linear regression is indeed a problem of substantial practical interest.

Appendix

This is a simplified version of the proof that the coefficient on a binary variable in linear least squares regression is computationally equivalent to the unexplained component from the extension of the Oaxaca–Blinder decomposition proposed by Cotton (1988), provided that there is a single control variable whose variance is equal in both subpopulations. The proof is due to Elder, Goddeeris, and Haider (2010, Appendix A). The (true) data generating process can be specified as:

$$y_i = \lambda_0 + \lambda_d d_i + \lambda_x x_i + \lambda_{dx} d_i x_i + \epsilon_i, \tag{11}$$

for which:

$$\mathrm{V}[x_i \mid d_i = 1] = \mathrm{V}[x_i \mid d_i = 0] = \sigma_d^2. \tag{12}$$

In such a setting, the unexplained component from the extension of the Oaxaca–Blinder decomposition proposed by Cotton (1988) can be written as:

$$\begin{aligned}
\tau^C &= (\mathrm{E}[y_i \mid d_i = 1] - \mathrm{E}[y_i \mid d_i = 0]) - (\mathrm{E}[x_i \mid d_i = 1] - \mathrm{E}[x_i \mid d_i = 0])\beta^C \\
&= \frac{\mathrm{Cov}[d_i, y_i]}{\mathrm{V}[d_i]} - \frac{\mathrm{Cov}[d_i, x_i]}{\mathrm{V}[d_i]} \\
&\quad \cdot (\Pr[d_i = 1] \cdot \frac{\mathrm{Cov}[x_i, y_i \mid d_i = 1]}{\mathrm{V}[x_i \mid d_i = 1]} + \Pr[d_i = 0] \cdot \frac{\mathrm{Cov}[x_i, y_i \mid d_i = 0]}{\mathrm{V}[x_i \mid d_i = 0]}) \\
&= \frac{\mathrm{Cov}[d_i, y_i]}{\mathrm{V}[d_i]} - \frac{\mathrm{Cov}[d_i, x_i]}{\mathrm{V}[d_i]} \\
&\quad \cdot \frac{\Pr[d_i = 1] \cdot \mathrm{Cov}[x_i, y_i \mid d_i = 1] + \Pr[d_i = 0] \cdot \mathrm{Cov}[x_i, y_i \mid d_i = 0]}{\sigma_d^2}.
\end{aligned} \tag{13}$$

At the same time, the (incorrectly specified) model be can specified as:

$$y_i = \alpha + \beta_d d_i + \beta_x x_i + \upsilon_i. \tag{14}$$

Our goal is therefore to prove that $\beta_d = \tau^C$. We can use $\tilde{d}_i$ to denote the residual from a regression of d_i on x_i and proceed with the proof:

$$\begin{aligned}
\beta_d &= \frac{\mathrm{Cov}[\tilde{d}_i, y_i]}{\mathrm{V}[\tilde{d}_i]} \\
&= \frac{\mathrm{Cov}[d_i - x_i \cdot \mathrm{Cov}[d_i, x_i]/\mathrm{V}[x_i], y_i]}{\mathrm{V}[\tilde{d}_i]} \\
&= \frac{\mathrm{Cov}[d_i, y_i]}{\mathrm{V}[\tilde{d}_i]} - \frac{\mathrm{Cov}[d_i, x_i]}{\mathrm{V}[x_i]} \cdot \frac{\mathrm{Cov}[x_i, y_i]}{\mathrm{V}[\tilde{d}_i]} \\
&= \frac{1}{\mathrm{V}[\tilde{d}_i]} \cdot (\mathrm{Cov}[d_i, y_i] - \frac{\mathrm{Cov}[d_i, x_i] \cdot \mathrm{Cov}[x_i, y_i]}{\mathrm{V}[x_i]}) \\
&= \frac{\mathrm{V}[x_i]}{\mathrm{V}[d_i] \cdot \mathrm{V}[x_i] - \mathrm{Cov}[d_i, x_i]^2} \cdot (\mathrm{Cov}[d_i, y_i] - \frac{\mathrm{Cov}[d_i, x_i] \cdot \mathrm{Cov}[x_i, y_i]}{\mathrm{V}[x_i]})
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\mathrm{Cov}[d_i, y_i] \cdot \mathrm{V}[x_i] - \mathrm{Cov}[d_i, x_i] \cdot \mathrm{Cov}[x_i, y_i]}{\mathrm{V}[d_i] \cdot \mathrm{V}[x_i] - \mathrm{Cov}[d_i, x_i]^2} \\
&= \frac{\mathrm{Cov}[d_i, y_i] \cdot \mathrm{V}[x_i] - \mathrm{Cov}[d_i, x_i]^2 \cdot \mathrm{Cov}[d_i, y_i]/\mathrm{V}[d_i]}{\mathrm{V}[d_i] \cdot \mathrm{V}[x_i] - \mathrm{Cov}[d_i, x_i]^2} \\
&\quad - \frac{\mathrm{Cov}[d_i, x_i] \cdot \mathrm{Cov}[x_i, y_i] - \mathrm{Cov}[d_i, x_i]^2 \cdot \mathrm{Cov}[d_i, y_i]/\mathrm{V}[d_i]}{\mathrm{V}[d_i] \cdot \mathrm{V}[x_i] - \mathrm{Cov}[d_i, x_i]^2} \\
&= \frac{\mathrm{Cov}[d_i, y_i]}{\mathrm{V}[d_i]} - \frac{\mathrm{Cov}[d_i, x_i]}{\mathrm{V}[d_i]} \cdot \frac{\mathrm{Cov}[x_i, y_i] \cdot \mathrm{V}[d_i] - \mathrm{Cov}[d_i, x_i] \cdot \mathrm{Cov}[d_i, y_i]}{\mathrm{V}[d_i] \cdot \mathrm{V}[x_i] - \mathrm{Cov}[d_i, x_i]^2} \\
&= \frac{\mathrm{Cov}[d_i, y_i]}{\mathrm{V}[d_i]} - \frac{\mathrm{Cov}[d_i, x_i]}{\mathrm{V}[d_i]} \cdot \frac{\mathrm{Cov}[x_i, y_i] - \mathrm{Cov}[d_i, x_i] \cdot \mathrm{Cov}[d_i, y_i]/\mathrm{V}[d_i]}{\mathrm{V}[x_i] - \mathrm{Cov}[d_i, x_i]^2/\mathrm{V}[d_i]} \\
&= \frac{\mathrm{Cov}[d_i, y_i]}{\mathrm{V}[d_i]} - \frac{\mathrm{Cov}[d_i, x_i]}{\mathrm{V}[d_i]} \\
&\quad \cdot \frac{\mathrm{Pr}[d_i = 1] \cdot \mathrm{Cov}[x_i, y_i \mid d_i = 1] + \mathrm{Pr}[d_i = 0] \cdot \mathrm{Cov}[x_i, y_i \mid d_i = 0]}{\sigma_d^2} \\
&= \tau^C, \qquad (15)
\end{aligned}
$$

where the penultimate equality follows from the decomposition of variance and the decomposition of covariance:

$$
\begin{aligned}
\mathrm{V}[x_i] &= \mathrm{Pr}[d_i = 1] \cdot \mathrm{V}[x_i \mid d_i = 1] + \mathrm{Pr}[d_i = 0] \cdot \mathrm{V}[x_i \mid d_i = 0] \\
&\quad + \mathrm{Pr}[d_i = 1] \cdot (\mathrm{E}[x_i \mid d_i = 1] - \mathrm{E}[x_i])^2 + \mathrm{Pr}[d_i = 0] \cdot (\mathrm{E}[x_i \mid d_i = 0] - \mathrm{E}[x_i])^2 \\
&= \sigma_d^2 + \mathrm{Pr}[d_i = 1] \cdot (\mathrm{Pr}[d_i = 0] \cdot (\mathrm{E}[x_i \mid d_i = 1] - \mathrm{E}[x_i \mid d_i = 0]))^2 \\
&\quad + \mathrm{Pr}[d_i = 0] \cdot (\mathrm{Pr}[d_i = 1] \cdot (\mathrm{E}[x_i \mid d_i = 1] - \mathrm{E}[x_i \mid d_i = 0]))^2 \\
&= \sigma_d^2 + (\mathrm{E}[x_i \mid d_i = 1] - \mathrm{E}[x_i \mid d_i = 0])^2 \\
&\quad \cdot (\mathrm{Pr}[d_i = 1] \cdot \mathrm{Pr}[d_i = 0]^2 + \mathrm{Pr}[d_i = 0] \cdot \mathrm{Pr}[d_i = 1]^2) \\
&= \sigma_d^2 + (\mathrm{Cov}[d_i, x_i]/\mathrm{V}[d_i])^2 \cdot (\mathrm{V}[d_i] \cdot \mathrm{Pr}[d_i = 0] + \mathrm{V}[d_i] \cdot \mathrm{Pr}[d_i = 1]) \\
&= \sigma_d^2 + \mathrm{Cov}[d_i, x_i]^2/\mathrm{V}[d_i] \qquad (16)
\end{aligned}
$$

and

$$
\mathrm{Cov}[x_i, y_i] = \mathrm{Pr}[d_i = 1] \cdot \mathrm{Cov}[x_i, y_i \mid d_i = 1] + \mathrm{Pr}[d_i = 0] \cdot \mathrm{Cov}[x_i, y_i \mid d_i = 0]
$$

$$
\begin{aligned}
&+ \Pr[d_i = 1] \cdot (\mathrm{E}[x_i \mid d_i = 1] - \mathrm{E}[x_i]) \cdot (\mathrm{E}[y_i \mid d_i = 1] - \mathrm{E}[y_i]) \\
&+ \Pr[d_i = 0] \cdot (\mathrm{E}[x_i \mid d_i = 0] - \mathrm{E}[x_i]) \cdot (\mathrm{E}[y_i \mid d_i = 0] - \mathrm{E}[y_i]) \\
&= \Pr[d_i = 1] \cdot \mathrm{Cov}[x_i, y_i \mid d_i = 1] + \Pr[d_i = 0] \cdot \mathrm{Cov}[x_i, y_i \mid d_i = 0] \\
&+ \Pr[d_i = 1] \cdot \Pr[d_i = 0] \cdot (\mathrm{E}[x_i \mid d_i = 1] - \mathrm{E}[x_i \mid d_i = 0]) \\
&\cdot \Pr[d_i = 0] \cdot (\mathrm{E}[y_i \mid d_i = 1] - \mathrm{E}[y_i \mid d_i = 0]) \\
&+ \Pr[d_i = 0] \cdot \Pr[d_i = 1] \cdot (\mathrm{E}[x_i \mid d_i = 1] - \mathrm{E}[x_i \mid d_i = 0]) \\
&\cdot \Pr[d_i = 1] \cdot (\mathrm{E}[y_i \mid d_i = 1] - \mathrm{E}[y_i \mid d_i = 0]) \\
&= \Pr[d_i = 1] \cdot \mathrm{Cov}[x_i, y_i \mid d_i = 1] + \Pr[d_i = 0] \cdot \mathrm{Cov}[x_i, y_i \mid d_i = 0] \\
&+ (\mathrm{E}[x_i \mid d_i = 1] - \mathrm{E}[x_i \mid d_i = 0]) \cdot (\mathrm{E}[y_i \mid d_i = 1] - \mathrm{E}[y_i \mid d_i = 0]) \\
&\cdot (\Pr[d_i = 1] \cdot \Pr[d_i = 0]^2 + \Pr[d_i = 0] \cdot \Pr[d_i = 1]^2) \\
&= \Pr[d_i = 1] \cdot \mathrm{Cov}[x_i, y_i \mid d_i = 1] + \Pr[d_i = 0] \cdot \mathrm{Cov}[x_i, y_i \mid d_i = 0] \\
&+ (\mathrm{Cov}[d_i, x_i] \cdot \mathrm{Cov}[d_i, y_i] / \mathrm{V}[d_i]^2) \cdot (\mathrm{V}[d_i] \cdot \Pr[d_i = 0] + \mathrm{V}[d_i] \cdot \Pr[d_i = 1]) \\
&= \Pr[d_i = 1] \cdot \mathrm{Cov}[x_i, y_i \mid d_i = 1] + \Pr[d_i = 0] \cdot \mathrm{Cov}[x_i, y_i \mid d_i = 0] \\
&+ \mathrm{Cov}[d_i, x_i] \cdot \mathrm{Cov}[d_i, y_i] / \mathrm{V}[d_i]. \qquad (17)
\end{aligned}
$$

References

[1] Abadie, A., Angrist, J., and Imbens, G. (2002), "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91–117.

[2] Abadie, A., and Imbens, G. W. (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74, 235–267.

[3] Abadie, A., and Imbens, G. W. (2011), "Bias-Corrected Matching Estimators for Average Treatment Effects," *Journal of Business & Economic Statistics*, 29, 1–11.

[4] Angrist, J. D. (1998), "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica*, 66, 249–288.

[5] Angrist, J. D., and Krueger, A. B. (1999), "Empirical Strategies in Labor Economics," in *Handbook of Labor Economics* (Vol. 3), eds. O. Ashenfelter and D. Card, Amsterdam: Elsevier.

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

[7] Barsky, R., Bound, J., Charles, K. K., and Lupton, J. P. (2002), "Accounting for the Black-White Wealth Gap: A Nonparametric Approach," *Journal of the American Statistical Association*, 97, 663–673.

[8] 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.

[9] Bitler, M. P., Gelbach, J. B., and Hoynes, H. W. (2008), "Distributional Impacts of the Self-Sufficiency Project," *Journal of Public Economics*, 92, 748–765.

[10] Black, D. A., Smith, J. A., Berger, M. C., and Noel, B. J. (2003), "Is the Threat of Reemployment Services More Effective than the Services Themselves? Evidence

from Random Assignment in the UI System," *American Economic Review*, 93, 1313–1327.

[11] Blinder, A. S. (1973), "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, 8, 436–455.

[12] Blundell, R., and Costa Dias, M. (2009), "Alternative Approaches to Evaluation in Empirical Microeconomics," *Journal of Human Resources*, 44, 565–640.

[13] Chernozhukov, V., and Hansen, C. (2005), "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261.

[14] Cotton, J. (1988), "On the Decomposition of Wage Differentials," *Review of Economics and Statistics*, 70, 236–243.

[15] Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2008), "Nonparametric Tests for Treatment Effect Heterogeneity," *Review of Economics and Statistics*, 90, 389–405.

[16] 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.

[17] Djebbari, H., and Smith, J. (2008), "Heterogeneous Impacts in PROGRESA," *Journal of Econometrics*, 145, 64–80.

[18] Elder, T. E., Goddeeris, J. H., and Haider, S. J. (2010), "Unexplained Gaps and Oaxaca-Blinder Decompositions," *Labour Economics*, 17, 284–290.

[19] Firpo, S. (2007), "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 75, 259–276.

[20] Fortin, N., Lemieux, T., and Firpo, S. (2011), "Decomposition Methods in Economics," in *Handbook of Labor Economics* (Vol. 4), eds. O. Ashenfelter and D. Card, San Diego and Amsterdam: Elsevier.

[21] Frölich, M., and Melly, B. (2008), "Unconditional Quantile Treatment Effects under Endogeneity," IZA Discussion Paper 3288, Institute for the Study of Labor (IZA).

[22] Fryer, R. G., and Levitt, S. D. (2004), "Understanding the Black-White Test Score Gap in the First Two Years of School," *Review of Economics and Statistics*, 86, 447–464.

[23] Heckman, J. J. (2001), "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *Journal of Political Economy*, 109, 673–748.

[24] Heckman, J. J., and Hotz, V. J. (1989), "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association*, 84, 862–874.

[25] Heckman, J. J., Smith, J., and Clements, N. (1997), "Making the Most out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies*, 64, 487–535.

[26] Hirano, K., Imbens, G. W., and Ridder, G. (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189.

[27] Humphreys, M. (2009), "Bounds on Least Squares Estimates of Causal Effects in the Presence of Heterogeneous Assignment Probabilities," unpublished manuscript, Columbia University.

[28] Imbens, G. W., and Wooldridge, J. M. (2009), "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86.

[29] Jann, B. (2008), "The Blinder-Oaxaca Decomposition for Linear Regression Models," *Stata Journal*, 8, 453–479.

[30] Kline, P. (2011), "Oaxaca-Blinder as a Reweighting Estimator," *American Economic Review*, 101, 532–537.

[31] LaLonde, R. J. (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76, 604–620.

[32] Løken, K. V., Mogstad, M., and Wiswall, M. (2012), "What Linear Estimators Miss: The Effects of Family Income on Child Outcomes," *American Economic Journal: Applied Economics*, 4, 1–35.

[33] Melly, B. (2006), "Applied Quantile Regression," unpublished Ph.D. dissertation, University of St. Gallen.

[34] Neal, D. A., and Johnson, W. R. (1996), "The Role of Premarket Factors in Black-White Wage Differences," *Journal of Political Economy*, 104, 869–895.

[35] Oaxaca, R. (1973), "Male-Female Wage Differentials in Urban Labor Markets," *International Economic Review*, 14, 693–709.

[36] 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.

[37] Smith, J. A., and Todd, P. E. (2005), "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?," *Journal of Econometrics*, 125, 305–353.

[38] Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data* (2nd ed.), Cambridge and London: MIT Press.

[39] Yitzhaki, S. (1996), "On Using Linear Regressions in Welfare Economics," *Journal of Business & Economic Statistics*, 14, 478–486.

Figure 2: Monte Carlo Simulations of the Linear Least Squares Regression Under Different Sample Compositions

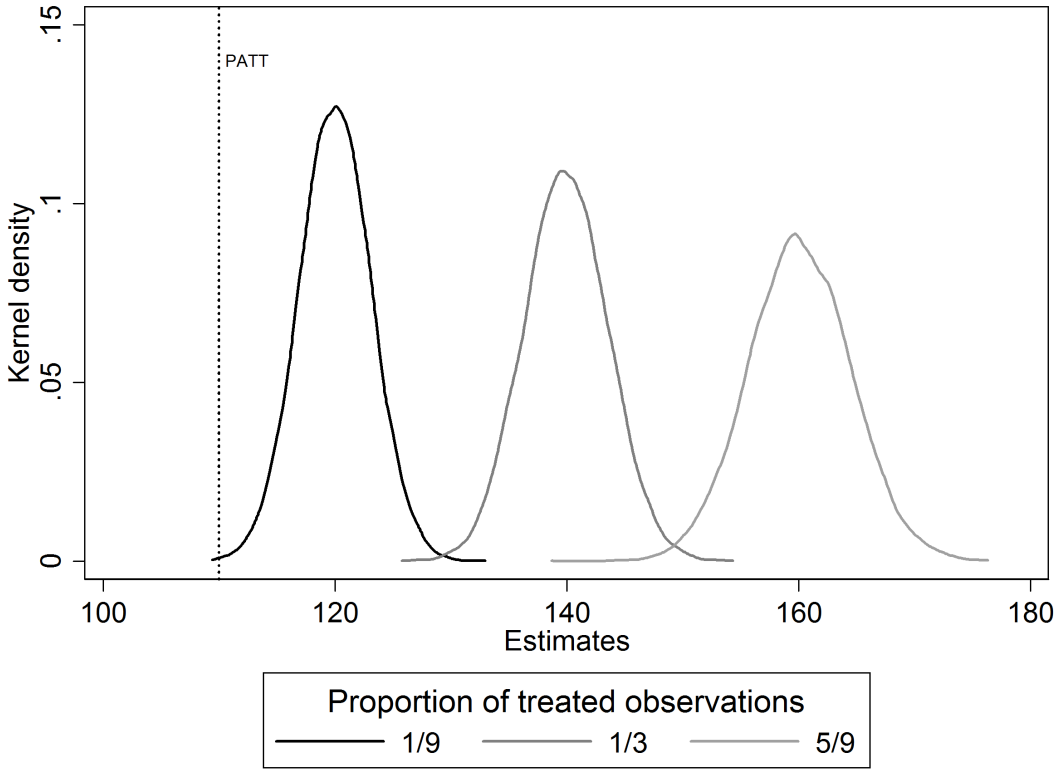


Figure 3: The Relationship Between Linear Regression and Cotton (1988) Estimates of the Effects of the NSW Training Programme

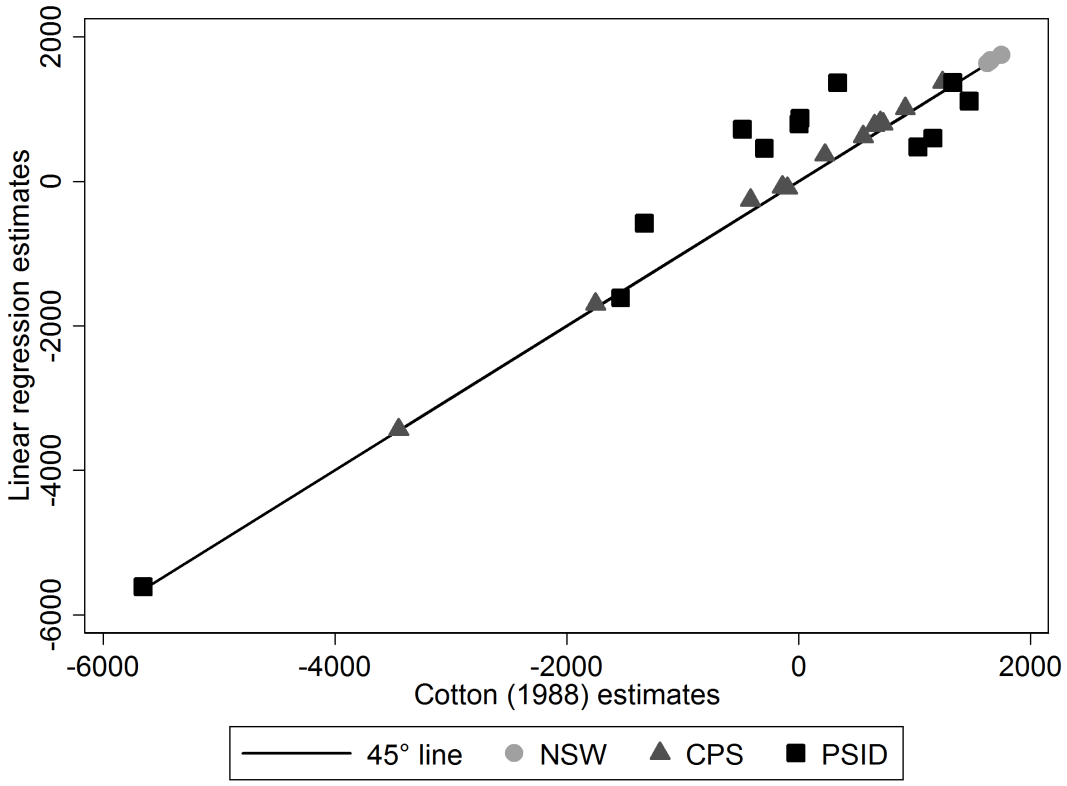


Figure 4: The Behaviour of the Linear Least Squares Regression in the Simulated NSW–CPS-1 Datasets Under Different Sample Compositions