NBER WORKING PAPER SERIES

MATCHING METHODS IN PRACTICE: THREE EXAMPLES

Guido W. Imbens

Working Paper 19959
http://www.nber.org/papers/w19959

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2014

Financial support for this research was generously provided through NSF grants SES 0452590 and 0820361. This paper was prepared for the Journal of Human Resources. I am grateful for comments by three anonyous referees and the editor Sandra Black which greatly improved the presentation. This work builds on previous work coauthored with Alberto Abadie, Joshua Angrist, Susan Athey, Keisuke Hirano, Geert Ridder, Donald Rubin, and Jeffrey Wooldridge. I have learned much about these issues from them, and their influence is apparent throughout the paper, but they are not responsible for any errors or any of the views expressed here. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Matching Methods in Practice: Three Examples
Guido W. Imbens
NBER Working Paper No. 19959
March 2014
JEL No. C1

ABSTRACT

There is a large theoretical literature on methods for estimating causal effects under unconfoundedness, exogeneity, or selection--on--observables type assumptions using matching or propensity score methods. Much of this literature is highly technical and has not made inroads into empirical practice where many researchers continue to use simple methods such as ordinary least squares regression even in settings where those methods do not have attractive properties. In this paper I discuss some of the lessons for practice from the theoretical literature, and provide detailed recommendations on what to do. I illustrate the recommendations with three detailed applications.

Guido W. Imbens
Graduate School of Business
Stanford University
655 Knight Way
Stanford, CA 94305
and NBER
Imbens@stanford.edu

1 Introduction

There is a large literature on methods for estimating average treatment effects under the assumption of unconfoundedness (also referred to as selection-on-observables, exogeneity, ignorability, or simply the conditional independence assumption). Under this assumption the comparison of units with different treatments but identical pretreatment variables can be given a causal interpretation. Much of the econometric literature has focused on establishing asymptotic properties for a variety of estimators, without a firm conclusion on the relative merits of these estimators. As a result, the theoretical literature leaves the empirical researcher with a bewildering choice of methods, with limited, and often conflicting guidance on what to use in practice. Possibly in response to that, and possibly also because of lack of compelling evidence that simple methods (*e.g.,* ordinary least squares regression) are not adequate, few of the estimators proposed in the recent literature have been widely adopted. Instead researchers continue to use ordinary linear regression methods with an indicator for the treatment and other covariates entering additively and linearly. However, such simple methods do not always work well. If the covariate distributions differ substantially by treatment status, conventional regression methods can be sensitive to minor changes in the specification because of their heavy reliance on extrapolation.

In this paper I provide some advice for researchers estimating average treatment effects in practice, based on my personal view of this literature. The issues raised in this paper build on my earlier reviews (Imbens, 2004; Imbens and Wooldridge, 2009), and on my forthcoming book with Rubin (Imbens and Rubin, 2014). I focus on four specific issues. First, I discuss when and why the simple ordinary least squares estimator, which ultimately relies on the same fundamental unconfoundedness assumption as matching and propensity score estimators, but which combines that assumption with strong functional form restrictions, is likely to be inadequate for estimating average treatment effects. Second, I discuss in detail two specific, fully specified, estimators that in my view are attractive alternatives to least squares estimators because of their robustness

properties with respect to a variety of data configurations. Although I describe these two methods in some detail, the message is emphatically not that these are the only two estimators that are reasonably. Rather, the point is that unless a particular estimator is robust to modest changes in the implementation, any results should be viewed with suspicion. Third, I discuss trimming and other preprocessing methods for creating more balanced samples as an important first step in the analysis to ensure that the final results are more robust with any estimators, including least squares, weighting, or matching. Fourth, I discuss some supplementary analyses for assessing the plausibility of the unconfoundedness assumption. Although unconfoundedness is not testable, one should assess, whenever possible, whether unconfoundedness is plausible in specific applications, and in many cases one can in fact do so.

I illustrate the methods discussed in this paper using three different data sets. The first of these is a subset of the lottery data collected by Imbens, Rubin and Sacerdote (2001). The second data set is the experimental part of the Dehejia and Wahba (1999) version of the Lalonde (1986) data set containing information on participants in the Nationally Supported Work (NSW) program and members of the randomized control group. The third data set contains information on participants in the NSW program and one of Lalonde's non-experimental comparison groups. Both subsets of the Dehejia-Wahba version of the Lalonde data are available on Dehejia's website, http://users.nber.org/ rdehejia/. Software for implementing these methods will be available on my website.

2 Set Up and Notation

The set up used in the current paper is by now the standard one in this literature, using the potential outcome framework for causal inference that builds on Fisher (1925) and Neyman (1923) and was extended to observational studies by Rubin (1974). See Imbens and Wooldridge (2009) for a recent survey and historical context. Following Holland (1986), I will refer to this set up as the Rubin Causal Model, RCM for short.[1]

[1]There is some controversy around this terminology. In a personal communication Holland told me that he felt that Rubin's generalizations of the earlier work (*e.g.,* Neyman, 1923) and the central role

The analysis is based on a sample of N units, indexed by $i = 1, \ldots, N$, viewed as a random sample from an infinitely large population.[2] For unit i there are two potential outcomes, denoted by $Y_i(0)$ and $Y_i(1)$, the potential outcomes without and given the treatment respectively. Each unit in the sample is observed to receive or not receive a binary treatment, with the treatment indicator denoted by W_i. If unit i receives the treatment then $W_i = 1$, otherwise $W_i = 0$. There are $N_c = \sum_{i=1}^{N}(1 - W_i)$ control units and $N_t = \sum_{i=1}^{N} W_i$ treated units, so that $N = N_c + N_t$ is the total number of units. The realized (and observed) outcome for unit i is

$$Y_i^{\text{obs}} = Y_i(W_i) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1. \end{cases}$$

I use the superscript "obs" here to distinguish between the potential outcomes which are not always observed, and the observed outcome. In addition there is for each unit i a K-component covariate X_i, with $X_i \in \mathbb{X} \subset \mathbb{R}^K$. The key characteristic of these covariates is that they are known not to be affected by the treatment. We observe, for all units in the sample, the triple $(Y_i^{\text{obs}}, W_i, X_i)$. We use $\mathbf{Y}$, $\mathbf{W}$, and $\mathbf{X}$ to denote the vectors with tyical elements Y_i^{obs} and W_i, and the matrix with i-th row equal to X_i'.

Define the population average treatment effect conditional on the covariates,

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x],$$

and the average effect in the population, and in the subpopulation of treated units,

$$\tau = \mathbb{E}\left[Y_i(1) - Y_i(0)\right], \qquad \text{and} \quad \tau_{\text{treat}} = \mathbb{E}\left[Y_i(1) - Y_i(0)\middle| W_i = 1\right]. \tag{2.1}$$

There is a somewhat subtle issue that we can also focus on the average treatment effects conditional on the covariates in the sample,

$$\tau_{\text{sample}} = \frac{1}{N}\sum_{i=1}^{N} \tau(X_i), \qquad \text{and} \quad \tau_{\text{treat,sample}} = \frac{1}{N_t}\sum_{i:W_i=1} \tau(X_i). \tag{2.2}$$

the potential outcomes play in his (Rubin's) approach merited the label.

[2]The assumption that the sample can be viewed as a random sample from a large population is largely for convenience. Abadie, Athey, Imbens and Wooldridge (2014) discuss the implications of this assumption, as well as alternatives.

This matters for inference, although not for estimation. The asymptotic variances for the same estimators, viewed as estimators of τ_{sample} and $\tau_{\text{treat,sample}}$ are generally smaller than when we view τ and τ_{treat} as the target, unless there is no variation in $\tau(x)$ over the population distribution of X_i. See Imbens and Wooldridge (2009) for more details on this distinction.

The first key assumptions we make is *unconfoundedness* (Rubin, 1990),

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \;\Big|\; X_i, \tag{2.3}$$

where, using the Dawid (1979) conditional independence notation, $A \perp\!\!\!\perp B|C$ denotes that A and B are independent conditional on C. The second key assumption is *overlap*,

$$0 < e(x) < 1, \tag{2.4}$$

for all x in the support of the covariates, where

$$e(x) = \mathbb{E}[W_i|X_i = x] = \Pr(W_i = 1|X_i = x), \tag{2.5}$$

is the *propensity score* (Rosenbaum and Rubin, 1983). The combination of these two assumptions is referred to as *strong ignorability* by Rosenbaum and Rubin (1983). The combination implies that we can estimate the average effects by adjusting for difference in covariates between treated and control units. For example, given unconfoundedness, the population average treatment effect τ can be expressed in terms of the joint distribution of $(Y_i^{\text{obs}}, W_i, X_i)$ as

$$\tau = \mathbb{E}\Big[\mathbb{E}\left[Y_i^{\text{obs}}\big|\, W_i = 1, X_i\right] - \mathbb{E}\left[Y_i^{\text{obs}}\big|\, W_i = 0, X_i\right]\Big]. \tag{2.6}$$

To see that (2.6) holds, note that by definition,

$$\begin{aligned}
&\mathbb{E}\Big[\mathbb{E}\left[Y_i^{\text{obs}}\big|\, W_i = 1, X_i\right] - \mathbb{E}\left[Y_i^{\text{obs}}\big|\, W_i = 0, X_i\right]\Big] \\
&\qquad = \mathbb{E}\Big[\mathbb{E}\left[Y_i(1)|\, W_i = 1, X_i\right] - \mathbb{E}\left[Y_i(0)|\, W_i = 0, X_i\right]\Big].
\end{aligned}$$

By unconfoundedness we can drop the conditioning on W_i in these two terms, and

$$\mathbb{E}\Big[\mathbb{E}\left[Y_i(1)|\, W_i = 1, X_i\right] - \mathbb{E}\left[Y_i(0)|\, W_i = 0, X_i\right]\Big]$$

$$= \mathbb{E}\Big[\mathbb{E}\left[Y_i(1)|\, X_i\right] - \mathbb{E}\left[Y_i(0)|\, X_i\right]\Big]$$
$$= \mathbb{E}\Big[\mathbb{E}\left[Y_i(1) - Y_i(0)|\, X_i\right]\Big] = \tau,$$

thus proving (2.6).

To implement estimators based on this equality we also require that the conditional expectations $\mathbb{E}[Y_i(0)|X_i = x]$, $\mathbb{E}[Y_i(1)|X_i = x]$ and $e(x)$ are sufficiently smooth, that is, sufficiently many times differentiable. See the technical papers in this literature, referenced in Imbens and Wooldridge (2009) for details on the smoothness requirements.

The statistical challenge is now how to estimate objects such as

$$\mathbb{E}\Big[\mathbb{E}\left[Y_i^{\rm obs}\big|\, W_i = 1, X_i\right] - \mathbb{E}\left[Y_i^{\rm obs}\big|\, W_i = 0, X_i\right]\Big]. \tag{2.7}$$

We would like to estimate (2.7) without relying on strong functional form assumptions on the conditional distributions or conditional expectations. We would also like the estimators to be robust to minor changes in the implementation of the estimator.

Whether these two assumptions are reasonable is often controversial. Violations of the overlap assumptions are often easily detected. They motivate some of the design issues, and often imply that the researcher should change the estimand from the overall average treatment effect to an average for some subpopulation. See Crump, Hotz, Imbens and Mitnik (2009) for more discussion. Potential violations of the unconfoundedness assumption are more difficult to address. First, there are some methods for assessing the plausibility of the assumption, see the discussion in Section 4.2. To make the assumption plausible, it is important to have detailed information about the units on characteristics that are associated both with the potential outcomes and the treatment indicator. It is often particularly helpful to have lagged measures of the outcome, *e.g.,* detailed labor market histories in evaluations of labor market programs. At the same time, having detailed information on the background of the units makes the statistical challenge of adjusting for all these differences more challenging.

This set up has been studied extensively in the statistics and econometric literatures. For reviews and general discussions in the econometrics literature see Angrist and

Krueger (2000), Angrist and Pischke (2009), Caliendo (2006), Heckman and Vytlacil (2007), Imbens (2004), Imbens and Wooldridge (2009), Caliendo and Kopeinig (2011), Frölich (2004), Wooldridge (2002), and Lee (2005). For reviews in other areas in applied statistics see Rosenbaum (1995, 2010), Morgan and Winship (2007), Guo and Fraser (2010), and Murnane and Willett (2010). Much of the econometric literature is technical in nature and has primarily focused on establishing first order large sample properties of point and interval estimators using matching (Abadie and Imbens, 2006), non-parametric regression methods (Frolich, 2000; Chen, Hong and Tarozzi, 2008; Hahn, 1998; Hahn and Ridder, 2013), or weighting (Hirano, Imbens, and Ridder, 2003; Frölich, 2002). These first order asymptotic properties are identical for a number of proposed methods, limiting the usefulness of these results for choosing between these methods. In addition, comparisons between the various methods based on Monte Carlo evidence (e.g., Frölich, 2004; Busso, DiNardo and McCrary, 2008; Lechner, 2012) are hampered by the dependence of many of the proposed methods on tuning parameters (*e.g.*, the bandwidth choices in kernel regression, or on the number of terms or basis functions in sieve or series methods) for which rarely specific, data-dependent values are recommended, as well as by the difficulty in agreeing on realistic designs in simulation studies.

In this paper I will discuss my recommendations for empirical practice based on my reading of this literature. These recommendations are somewhat personal. In places they are also somewhat vague. I will emphatically not argue that one particular estimator should always be used. Part of the message is that there are no, and will not be, general results implying that in general some estimators are superior to all others. The recommendations are more specific in other places, for example suggesting supplementary analyses that indicate for some data sets that particular estimators will be sensitive to minor changes in implementation, thus making them unattractive in those settings. The hope is that these comments will be useful for empirical researchers. The recommendations will consist of a combination of particular estimators, supplementary analyses, and changes in estimands when (2.7) is particularly difficult to estimate.

3 Least Squares Estimation: When and Why Does it Not Work?

The most widely used method for estimating causal effects remains ordinary least squares (ols) estimation. Ordinary least squares estimation relies on unconfoundedness in combination with functional form assumptions. Typically researchers make these functional form assumptions for convenience, viewing them at best as approximations to the underlying functions. In this section I discuss merits and concerns with this method for estimating causal effects as opposed to using it for prediction. The main point is that the least squares functional form assumptions matter, and sometimes matter substantially. Moreover, one can assess from the data whether this will be the case.

To illustrate these ideas I will use a subset of the data collected by Lalonde. The treatment is participation in a labor market training program the National Supported Work (NSW) program. Later I will look at other parts of the data set, but here I focus on the non-experimental comparison between the trainees and a non-experimental comparison group from the Panel Study of Income Dynamics (PSID). The outcome of interest is an earnings variable, `earn'78`. I focus on a subset of the data with a single covariate, `earn'75`. There are $N_t = 185$ men in the treatment group, and $N_c = 2490$ men in the control group. From the experimental data it can be inferred that the average causal effect is approximately \$2000. Figures 1 and 2 present histograms of `earn'75` in the control and treatment group respectively. The average values of this variable in the two groups are 19.1 and 1.5, with standard deviations 13.6 and 3.2 respectively. Clearly the two distributions of `earn'75` in the trainee and comparison groups are far apart.

3.1 Ordinary Least Squares Estimation

Initially I focus on the average effect for the treated, τ_{treat}, defined in (2.1), although the same issues arise for the average effect τ. Focusing on τ_{treat} simplifies some of the analytic calculations. Under unconfoundedness τ_{treat} can be written as a function of the

joint distribution of $(Y_i^{\rm obs}, W_i, X_i)$ as

$$\tau_{\rm treat} = \mathbb{E}\left[Y_i^{\rm obs}\middle|\, W_i = 1\right] - \mathbb{E}\Big[\mathbb{E}\left[Y_i^{\rm obs}\middle|\, W_i = 0, X_i\right]\Big|W_i = 1\Big]. \tag{3.8}$$

Define

$$\overline{Y}_c^{\rm obs} = \frac{1}{N_c}\sum_{i:W_i=0} Y_i^{\rm obs}, \qquad \overline{Y}_t^{\rm obs} = \frac{1}{N_t}\sum_{i:W_i=1} Y_i^{\rm obs}$$

$$\overline{X}_c = \frac{1}{N_c}\sum_{i:W_i=0} X_i, \qquad \text{and} \quad \overline{X}_t = \frac{1}{N_t}\sum_{i:W_i=1} X_i.$$

The first term in (3.8), $\mathbb{E}\left[Y_i^{\rm obs}\middle|\, W_i = 1\right]$, is directly estimable from the data as $\overline{Y}_t$. It is the second term, $\mathbb{E}\Big[\mathbb{E}\left[Y_i^{\rm obs}\middle|\, W_i = 0, X_i\right]\Big|W_i = 1\Big]$, that is more difficult to estimate. Suppose we specify a linear model for $Y_i(0)$ given X_i:

$$\mathbb{E}[Y_i(0)|X_i = x] = \alpha_c + \beta_c \cdot x.$$

The ols estimator for β_c equal to

$$\hat{\beta}_c = \frac{\sum_{i:W_i=0}(X_i - \overline{X}_c)\cdot(Y_i^{\rm obs} - \overline{Y}_c^{\rm obs})}{\sum_{i:W_i=0}(X_i - \overline{X}_c)^2}, \qquad \text{and} \quad \hat{\alpha}_c = \overline{Y}_c^{\rm obs} - \hat{\beta}_c \cdot \overline{X}_c.$$

For the Lalonde data we have

$$\hat{\beta}_c = 0.84 \ \ \text{(s.e. 0.03)}, \qquad \hat{\alpha}_c = 5.60 \ \ \text{(s.e. 0.51)}.$$

Then we can write the ols estimator of the average of the potential control outcomes for the treated, $\mathbb{E}\left[Y_i(0)\middle|\, W_i = 1\right]$, as

$$\widehat{\mathbb{E}\left[Y_i(0)\middle|\, W_i = 1\right]} = \overline{Y}_c + \hat{\beta}_c \cdot \left(\overline{X}_t - \overline{X}_c\right), \tag{3.9}$$

which for the Lalonde data leads to

$$\widehat{\mathbb{E}\left[Y_i(0)\middle|\, W_i = 1\right]} = 6.88 \ \ \text{(s.e. 0.48)}.$$

The question is how credible this is as an estimate of $\mathbb{E}\left[Y_i(0)\middle|\, W_i = 1\right]$. We focus on two aspects of this credibility.

3.2 Extrapolation and Misspecification

First we need some additional notation. Let us denote the true conditional expectation and variance of $Y_i(w)$ given $X_i = x$ by

$$\mu_w(x) = \mathbb{E}[Y_i(w)|X_i = x], \qquad \text{and} \quad \sigma_w^2(x) = \mathbb{V}(Y_i(w)|X_i = x).$$

Given the true conditional expectation the pseudo-true values for β_c, the probability limit of the ols estimator, is

$$\beta_c^* = \text{plim}(\hat{\beta}_c) = \frac{\mathbb{E}[\mu_c(X_i) \cdot (X_i - \mathbb{E}[X_i])|W_i = 0]}{\mathbb{E}[(X_i - \mathbb{E}[X_i])^2|W_i = 0]},$$

and the estimator for the predicted value for $\mathbb{E}[Y_i(0)|W_i = 1]$ will converge to

$$\begin{aligned}\text{plim}(\mathbb{E}\left[\widehat{Y_i(0)|W_i} = 1\right]) &= \mathbb{E}[Y_i(0)|W_i = 0] + \beta_c^* \cdot (\mathbb{E}[X_i|W_i = 1] - \mathbb{E}[X_i|W_i = 0]) \,. \\ &= \mathbb{E}[Y_i(0)|W_i = 1] - \Big(\mathbb{E}[\mu_t(X_i)|W_i = 1] - \mathbb{E}[\mu_t(X_i)|W_i = 0]\Big) \\ &\quad + \beta_c^* \cdot \Big(\mathbb{E}[X_i|W_i = 1] - \mathbb{E}[X_i|W_i = 0]\Big).\end{aligned}$$

The difference between the $\text{plim}(\mathbb{E}\left[\widehat{Y_i(0)|W_i} = 1\right])$ and the true average $\mathbb{E}\left[Y_i(0)|\,W_i = 1\right]$ depends on the difference between the average value of the regression function for the treated and the controls and the approximation of this difference of the average conditional expectations by the difference in the average values of the best linear predictors,

$$\begin{aligned}&\text{plim}(\mathbb{E}\left[\widehat{Y_i(0)|W_i} = 1\right]) - \mathbb{E}\left[Y_i(0)|\,W_i = 1\right] \\ &\quad = \beta_c^* \cdot \Big(\mathbb{E}[X_i|W_i = 1] - \mathbb{E}[X_i|W_i = 0]\Big) - \Big(\mathbb{E}[\mu_t(X_i)|W_i = 1] - \mathbb{E}[\mu_t(X_i)|W_i = 0]\Big).\end{aligned}$$

If the conditional expectations are truly linear, the difference is zero. In addition, if the average covariate values in the two subpopulations have the same distribution, $f_X(x|W_i = 0) = f_X(x|W_i = 1)$ for all x, the difference is zero. However, if both the conditional expectations are nonlinear, and the covariate distributions differ, there will in general be a bias. Establishing whether the conditional expectations are nonlinear can be difficult to do. However, it is straightforward to assess whether the distributions of the covariates in the two treatment arms are similar or not.

Now let us look at the sensitivity to different specifications. I consider two specification for the regression function. First, a simple linear regression:

$$\texttt{earn}'\texttt{78} = \alpha_c + \beta_c \cdot \texttt{earn}'\texttt{75} + \varepsilon_i \qquad \text{(linear)}.$$

Second, a linear regression after transforming the regressor by adding one unit (one thousand dollars) followed by taking logarithms:

$$\texttt{earn}'\texttt{78} = \alpha_c + \beta_c \cdot \ln\left(1 + \texttt{earn}'\texttt{75}\right) + \varepsilon_i \qquad \text{(log linear)}.$$

The predictions for $\mathbb{E}\left[Y_i(0) \middle| W_i = 1\right]$ based on the two models are quite different, given that the average causal effect is on the order of two (thousand) dollars:

$$\mathbb{E}\left[\widehat{Y_i(0)} \middle| W_i = 1\right]_{\text{linear}} = 6.88 \quad \text{(s.e. 0.48)}$$

$$\mathbb{E}\left[\widehat{Y_i(0)} \middle| W_i = 1\right]_{\text{log linear}} = 2.81 \quad \text{(s.e. 0.66)}.$$

The reason for the big difference (4.07) between the two predicted values can be seen in Figure 3 where I plot the two estimated regression functions. The dashed vertical lines present the 0.25 and 0.75 quartile of the distribution of `earn'75` for the PSID sample (9.84 and 24.50 respectively). This indicates the range of values where the regression function is most likely to approximate the conditional expectation accurately. The solid vertical lines present the 0.25 and 0.75 quantiles for the distribution of `earn'75` for the trainees (0 and 1.82 respectively). One can see in this figure that for the range where the trainees are and where we therefore need to predict $Y_i(0)$, with `earn'75` between 0 and 1.82, the predictions of the linear and the log-linear model are quite different.

The estimates here were based on regression outcomes on earnings in 1975 for the control units only. In practice an even more common approach is to simply estimate the regression

$$\texttt{earn}'\texttt{78} = \alpha + \tau \cdot W_i + \beta \cdot \texttt{earn}'\texttt{75} + \varepsilon_i \qquad \text{(linear)},$$

or

$$\texttt{earn}'\texttt{78} = \alpha + \tau \cdot W_i + \beta_c \cdot \ln\left(1 + \texttt{earn}'\texttt{75}\right) + \varepsilon_i \qquad \text{(log linear)},$$

on the full sample, including both treated and control units. The reason for focusing on the regression in the control sample here is mainly expositional: it simplifies the expressions for the estimated average of the potential outcomes. In practice it makes relatively little difference if we do the regressions with both treated and control units included. For these data, with many more control units (2490) than treated units (1985), the estimates for the average of the potential outcome $Y_i(0)$ for the treated are, under the linear and log linear model, largely driving by the values for the control units. Estimating the regression function as a single regression on all units leads to:

$$\mathbb{E}\left[\widehat{Y_i(0)|\,W_i}=1\right]_{\text{linear, full sample}}=6.93 \quad \text{(s.e. 0.47)}$$

$$\mathbb{E}\left[\widehat{Y_i(0)|\,W_i}=1\right]_{\text{log linear, full sample}}=3.47 \quad \text{(s.e. 0.38)},$$

with the difference in the average predicted value for $Y_i(0)$ for the treated units between the linear and the log linear specification much bigger than between the regression on the control sample versus the regression on the full sample.

3.3 Weights

Here I present a different way of looking at the sensitivity of the least squares estimator in settings with substantial differences in the covariates distributions. One can write the ols estimator of $\mathbb{E}\left[Y_i(0)|\,W_i=1\right]$ in a different way, as

$$\mathbb{E}\left[\widehat{Y_i(0)|\,W_i}=1\right]=\frac{1}{N_c}\sum_{i:W_i=0}\omega_i\cdot Y_i^{\text{obs}}, \tag{3.10}$$

where, for control unit i the weight ω_i is

$$\omega_i=1-\left(X_i-\overline{X}_c\right)\cdot\frac{\overline{X}_c-\overline{X}_t}{\overline{X^2}-\overline{X}\cdot\overline{X}}. \tag{3.11}$$

It is interesting to explore the properties of these weights ω_i. First, the weights average to one over the control sample, irrespective of the data. Thus, $\mathbb{E}\left[\widehat{Y_i(0)|\,W_i}=1\right]$ is a weighted average of the outcomes for the control units, with the weights adding up to one. Second, the normalized weights ω_i are all equal to 1 in the special case where the average of the covariates in the two groups are indentical. In that case $\mathbb{E}\left[\widehat{Y_i(0)|\,W_i}=1\right]=\overline{Y}_c^{\text{obs}}$. In a

randomized experiment $\overline{X}_c$ is equal to $\overline{X}_t$ in expectation, so the weights should be close to 1 in that case. The variation in the weights, and the possibility of relatively extreme weights, increases with the difference between $\overline{X}_c$ and $\overline{X}_t$. Finally, note that given a particular data set, one can inspect these weights directly, and assess if some units have excessively large weights.

Let us first look at the normalized weights ω_i for the Lalonde data. For these data the weights take the form

$$\omega_i = 2.8091 - 0.0949 \cdot X_i.$$

The average of the weights is, by construction, equal to 1. The standard deviation of the weights is 1.29. The largest weight is 2.08, for control units with `earn'75` equal to zero. It makes sense for control individuals with `earn'75` equal to zero to have substantial weights, as there are many individuals in the treatment group with identical or similar values for `earn'75`. The most negative weight is for the control unit with `earn'75` equal to 156.7, with a weight of -12.05. This is where the problem with ordinary least squares regression shows itself most clearly. The range of values of `earn'75` for trainees is $[0, 25.14]$. I would argue that whether an individual with `earn'75` equal to 156.7 thousand dollars makes \$100,000 or \$300,000 in 1978 should make no difference for a prediction for the average value of $Y_i(0)$ for the trainees in the program, $\mathbb{E}\left[Y_i(0) \middle| W_i = 1\right]$, given that the maximum value for `earn'75` for trainees is 25.14. Conditional on the specification of the linear model, however, that difference between \$100,000 and \$300,000 is very important for the prediction of earnings in 1978 for trainees. Given the weight of -12.05 for the control individual with `earn'75` equal to 156.7, the prediction for $\mathbb{E}\left[Y_i(0) \middle| W_i = 1\right]$ would be 6.829 if `earn'78` $= 100$ for this individual, and 5.861 if `earn'78` $= 300$ for this individual, with the difference equal to \$968, a substantial amount. The problem is that the least squares estimates take the linearity in the linear regression model very seriously, and given linearity, observations on units with extreme values for `earn'75` are in fact very informative. Note that one can carry out these calculations without even using the outcome data. It does not matter what the actual value is for `earn'78` for this unit, because

the weights is so large, the value of `earn'78` matters substantially for our prediction for $\mathbb{E}\left[Y_i(0)\middle| W_i = 1\right]$ where arguably it should have close to zero weight.

3.4 How to Think about Regression Adjustment

Of course the example in the previous subsection is fairly extreme. The two distributions are very different, with, for example, the fraction of the PSID control sample with `earn'75` exceeding the maximum value found in the trainee sample (which is 25.14) equal to 0.27. It would appear obvious that units with `earn'75` as large as 156.7 should be given little or no weight in the analyses. An obvious remedy would be to discard all PSID units with `earn'75` exceeding 25.14, and this would go a considerable way towards making the estimators more robust. However, it is partly because there is only a single covariate in this overly simplistic example that there are such simple remedies.[3] With multiple covariates it is difficult to see what trimming would need to be done. Moreover, a simple trimming rule such as the one describe above is very sensitive to outliers. If there were a single trainee with `earn'75` equal to 150, that would change the estimates substantially. The issue is that regression methods are fundamentally not robust to the substantial differences between treatment and control groups. I will be discussing alternatives to the simple regression methods used here that do systematically, and automatically, downweight the influence of such outliers, both in the case with scalar covariates and in the case with multivariate covariates. If, for example, instead of using regression methods, one matched all the treated units, the control units with `earn'75` exceeding 25.07 would receive zero weight in the estimation, and in fact few control units with `earn'75` between 18 and 25 would receive positive weights. Matching is therefore by design robust to the presence of such units.

In practice it may be useful to compare the least squares estimates to estimates based on more sophisticated methods, both as a check that calculations were carried out correctly, and to ensure that one understands what is driving any difference between the

[3]Note that Lalonde in his original paper does consider trimming the sample by dropping all individuals with earnings in 1975 over some fixed threshold. He considers difference values for this threshold.

estimates.

4 The Strategy

In this section I lay out the overall strategy for flexibly and robustly estimating the average effect of the treatment. The strategy will consist of two, sometimes three, distinct stages. In the first stage, which I will refer to, following Rubin (2005), as the *design* stage, the full sample will be trimmed by discarding some units to improve overlap in covariate distributions. In the second stage, the *supplementary analysis* stage, the unconfoundedness assumption is assessed. In the third stage, the *analysis* stage, the estimator for the average effect will be applied to the trimmed data set. Let me briefly discuss the three stages in more detail.

4.1 Stage I: Design

In this stage we do not use the outcome data and focus solely on the treatment indicators and covariates, $(\mathbf{X}, \mathbf{W})$. The first step is to assess overlap in covariate distributions. If this is found to be lacking, a subsample is created with more overlap by discarding some units from the original sample. The assessment of the degree of overlap is based on an inspection of some summary statistics, what I refer to as normalized differences in covariates, denoted by Δ_X. These are differences in average covariate values by treatment status, scaled by a measure of the standard deviation of the covariates. These normalized differences contrast with t-statistics for testing the null of no differences in means between treated and controls. The normalized differences provide a scale and sample size free way of assessing overlap.

If the covariates are far apart in terms of this normalized difference metric, one may wish to change the target. Instead of focusing on the overall average treatment effect, one may wish to drop units with values of the covariates such that they have no counterparts in the other treatment group. The reason is that in general no estimation procedure will be give robust estimates of the average treatment effect in that case. To be specific,

following Crump, Hotz, Imbens and Mitnik (2008a, CHIM from hereon), I propose a rule for dropping units with extreme (that is, close to zero or one) values for the estimated propensity score. This step will require estimating the propensity score $\hat{e}(x : \mathbf{W}, \mathbf{X})$. One can capture the result of this first stage in terms of a function $\mathbf{I} = \mathbf{I}(\mathbf{W}, \mathbf{X})$, that determines which units are dropped as a function of the vector of treatment indicators $\mathbf{W}$ and the matrix of pretreatment variables $\mathbf{X}$. Here $\mathbf{I}$ denotes an N vector of binary indicators, with i-th element I_i equal to one if unit i is included in the analysis, and I_i equal to zero if unit i is dropped. Dropping the units with $I_i = 0$ leads to a trimmed sample $(\mathbf{Y}^T, \mathbf{W}^T, \mathbf{X}^T)$ with $N_s = \sum_{i=1}^{N} I_i$ units.

4.2 Stage II: Supplementary Analysis: Assessing Unconfoundedness

In the second stage, analyses are carried out to assess the plausibility of unconfoundedness. Again it should be stressed that these analyses can be carried out without access to the outcome data $\mathbf{Y}$, using solely $(\mathbf{W}, \mathbf{X})$. Here the set of pretreatment variables or covariates is partitioned into two parts, the first a vector of pseudo outcomes, $\mathbf{X}_{\rm p}$, and the second a matrix containing the remaining covariates, $\mathbf{X}_{\rm r}$. We then take the pseudo sample $(\mathbf{X}_{\rm p}, \mathbf{W}, \mathbf{X}_{\rm r})$ and estimate the average effect of the treatment on this pseudo outcome, adjusting only for differences in the remaining covariates $\mathbf{X}_{\rm r}$. This will involve first trimming this pseudo sample using the same procedures to construct a trimmed sample $(\mathbf{X}_{\rm p}^T, \mathbf{W}^T, \mathbf{X}_{\rm r}^T)$, and then estimating a "pseudo" average treatment effect on the pseudo outcome for this trimmed sample.

Specifically, I calculate

$$\hat{\tau}_X = \tau(\mathbf{X}_{\rm p}^T, \mathbf{W}^T, \mathbf{X}_{\rm r}^T),$$

for specific estimators $\tau(\cdot)$ described below, where the adjustment is only for the subset of covariates included in $\mathbf{X}_{\rm r}^T$. This estimates the "pseudo" causal effect, the causal effect of the treatment on the pretreatment variable $X_{\rm p}$, which is *a priori* known to be zero. If we find the estimate $\hat{\tau}_X$ is substantially and statistically close to zero, after adjusting for $\mathbf{X}_{\rm r}^T$,

this will be interpreted as evidence supportive of the assumption of unconfoundedness. If the estimate $\hat{\tau}_X$ differs from zero, either substantially or statistically, this will be viewed as casting doubt on the (ultimately untestable) unconfoundedness assumption, with the degree of evidence dependent on the magnitude of the deviations. The result of this stage will therefore be an assessment of the credibility of any estimates of the average effect of the treatment on the outcome of interest, without using the actual outcome data.

4.3 Stage III: Analysis

In the third stage the outcome data $\mathbf{Y}$ are used and estimates of the average treatment effect of interest are calculated. Using one of the recommended estimators, blocking (subclassification) or matching, applied to the trimmed sample, we obtain a point estimate:

$$\hat{\tau} = \tau(\mathbf{Y}^T, \mathbf{W}^T, \mathbf{X}^T).$$

This step is more delicate, with concerns about pre-testing bias, and I therefore propose no specification searches at this stage. The first estimator I discuss in detail is based on blocking or subclassification on the estimated propensity score in combination with regression adjustments within the blocks. The second estimator is based on direct one-to-one (or one-to-k) covariate matching with replacement in combination with regression adjustments within the matched pairs. Both are in my view attractive estimators because of their robustness.

5 Tools

In this section I will describe the calculation of the normalized differences $\Delta_{X,k}$ and the four functions, the propensity score, $e(x; \mathbf{W}, \mathbf{X})$, the trimming function, $I(\mathbf{W}, \mathbf{X})$, the blocking estimator, $\tau_{\text{block}}(\mathbf{Y}, \mathbf{W}, \mathbf{X})$ and the matching estimator, $\tau_{\text{match}}(\mathbf{Y}, \mathbf{W}, \mathbf{X})$, and the choices that go into these function.

I should note that in specific situations, one may augment the implicit choices made regarding the specification of the regression function and the propensity score in the proposed algorithm with substantive knowledge. Such knowledge is likely to improve the

performance of the methods. Such knowledge notwitstanding, there is often a part of the specification about which the researcher is agnostic. For example, the researcher may not have strong *a priori* views on whether a pretreatment variable such as age should enter linearly or quadratically in the propensity score in the context of the evaluation of a labor market program. The methods described here are intended to assist the researcher in such decisions by providing a benchmark estimator for the average effect of interest.

5.1 Assessing Overlap: Normalized Differences in Average Covariates

For element $X_{i,k}$ of the covariate vector X_i the normalized difference is defined as

$$\Delta_{X,k} = \frac{\overline{X}_{\mathrm{t},k} - \overline{X}_{\mathrm{c},k}}{\sqrt{(S^2_{X,\mathrm{t},k} + S^2_{X,\mathrm{c},k})/2}}, \tag{5.12}$$

where

$$\overline{X}_{\mathrm{c},k} = \frac{1}{N_\mathrm{c}} \sum_{i:W_i=0} X_{i,k}, \qquad \overline{X}_{\mathrm{t},k} = \frac{1}{N_\mathrm{t}} \sum_{i:W_i=1} X_{i,k},$$

$$S^2_{X,\mathrm{c},k} = \frac{1}{N_\mathrm{c}-1} \sum_{i:W_i=0} \left(X_{i,k} - \overline{X}_{\mathrm{c},k}\right)^2, \qquad \text{and} \quad S^2_{X,\mathrm{t},k} = \frac{1}{N_\mathrm{t}-1} \sum_{i:W_i=1} \left(X_{i,k} - \overline{X}_{\mathrm{t},k}\right)^2.$$

A key point is that the normalized difference is more useful for assessing the magnitude of the statistical challenge in adjusting for covariate differences between treated and control units than the t-statistic for testing the null hypothesis that the two differences are zero, or

$$t_{X,k} = \frac{\overline{X}_{\mathrm{t},k} - \overline{X}_{\mathrm{c},k}}{\sqrt{S^2_{X,\mathrm{t},k}/N_\mathrm{t} + S^2_{X,\mathrm{c},k}/N_\mathrm{c}}}. \tag{5.13}$$

The t-statistic $t_{X,k}$ may be large in absolute value simply because the sample is large and, as a result, small differences between the two sample means are statistically significant even if they are substantively small. Large values for the normalized differences, in contrast, indicate that the average covariate values in the two groups are substantially different. Such differences imply that simple methods such as regression analysis will be sensitive to specification choices and outliers. In that case there may in fact be no estimation method that leads to robust estimates of the average treatment effect.

5.2 Estimating the Propensity Score

In this section I discuss an estimator for the propensity score $e(x)$. The estimator is based on a logistic regression model, estimated by maximum likelihood. Given a vector of functions $h : \mathbb{X} \mapsto \mathbb{R}^M$, the propensity score is specified as

$$e(x) = \frac{\exp\left(h(x)'\gamma\right)}{1+\exp\left(h(x)'\gamma\right)},$$

with an M-dimensional parameter vector γ. There is no particular reason for the choice of a logistic model, as opposed to, say a probit model where $e(x) = \Phi(h(x)'\gamma)$. This is one of the places where it is important that the eventual estimates to be robust to this choice. In particular the trimming of the sample will improve the robustness to this choice, removing units with estimated propensity score values close to zero or one where the choice between logit and probit models may matter.

Given a choice for the function $h(x)$, the unknown parameter γ is estimated by maximum likelihood:

$$\hat{\gamma}_{\text{ml}}(\mathbf{W},\mathbf{X}) = \arg\max_{\gamma} L(\gamma|\mathbf{W},\mathbf{X}) = \arg\max_{\lambda} \sum_{i=1}^{N} \Big\{ W_i \cdot h(X_i)'\gamma - \ln\Big(1+\exp\left(h(X_i)'\gamma\right)\Big)\Big\}.$$

The estimated propensity score is then

$$\hat{e}(x|\mathbf{W},\mathbf{X}) = \frac{\exp\left(h(x)'\hat{\gamma}_{\text{ml}}(\mathbf{W},\mathbf{X})\right)}{1+\exp\left(h(x)'\hat{\gamma}_{\text{ml}}(\mathbf{W},\mathbf{X})\right)}.$$

The key issue is the choice of the vector of functions $h(\cdot)$. First note that the propensity score plays a mechanical role in balancing the covariates. It is not given a structural or causal interpretation in this analysis. In choosing a specification there is therefore little role for theoretical substantive arguments: we are mainly looking for a specification that leads to an accurate approximation to the conditional expectation. At most theory may help in judging which of the covariate are likely to be important here.

A second point is that there is little harm in specification searches at this stage. The inference for the estimators for the average treatment effect we consider is conditional on covariates, and is not affected by specification searches at this stage which do not involve the data on the outcome variables.

A third point is that although the penalty for including irrelevant terms in the propensity score is generally small, eventually the precision will go down if too many terms are included in the specification.

A simple, and the most common choice for the specification of the propensity score is to take the basic covariates themselves, $h(x) = x$. This need not be adequate for two reasons. The main reason is that this specification may not be sufficiently flexible. In many cases there are important interactions between the covariates that need to be adjusted for in order to get an adequate approximation to the propensity score. Second, it may be that the linear specification includes pretreatment variables that have very little association with the treatment indicator and therefore need not be included in the specification. Especially in settings with many pretreatment variables this may be wasteful and lead to low precision in the final estimates. I therefore discuss a more flexible specification for the propensity score based on stepwise regression, proposed in Imbens and Rubin (2014). This provides a data driven way to select a subset of all linear and second order terms. Details are provided in the Appendix. There are many, arguably more sophisticated, alternatives to choose a specification for the propensity score, which, however, have not been tried out much in practice. Other discussions include Millimet and Tchernis (2009). Recently lasso methods have been proposed for similar problems (Tibshirani, 1996). See Belloni, Chernozhukov and Hansen (2012) for applications to, and extensions for, the current treatment-effect setting. However, the point is again not to find a single method for estimating the propensity score that will out-perform all others. Rather, the goal is to find a reasonable method for estimating the propensity score that will, in combination with the subsequent adjustment methods, lead to estimates for the treatment effects of interest that are similar to those based on other reasonable strategies for estimating the propensity score. For the estimators here a finding that the final results for say the average treatment effect are sensitive to using the stepwise selection method compared to, say, the lasso or other shrinkage methods, would raise concerns that none of the estimates are credible.

This vector of functions always includes a constant: $h_1(x) = 1$. Let us consider

the case where x is a a K-component vector of covariates (not counting the intercept), I restrict the remaining elements of $h(x)$ to be either equal to an component of x, or to be equal to the product of two components of x. In this sense the estimator is not fully nonparametric: although one can generally approximate any function by a polynomial, here I limit the approximating functions to second order polynomials. In practice, however, for the purposes I will use the estimated propensity score for, this need not be a severe limitation.

The problem now is how to choose among the $(K+1)\times(K+2)/2-1$ first and second order terms. I select a subset of these terms in a stepwise fashion. Three choices are to be made by the researcher. First, there may be a subset of the covariates that will be included in the linear part of the specification, irrespective of their association with the outcome and the treatment indicator. Note that biases from omitting variables comes from both the correlation between the omitted variables and the treatment indicator and the correlation between the omitted variable and the outcome, just as in the conventional omitted variable formula for regression. In applications to job training programs these might include lagged outcomes, or other covariates that are *a priori* expected to be substantially correlated with the outcomes of interest. Let us denote this subvector by X_B, with dimension $1 \leq K_B \leq K+1$ (X_B always includes the intercept). This vector need not include any covariates beyond the intercept, if the researcher has no strong views regarding the relative importance of any of the covariates. Second, a threshold value for inclusion of linear terms has to be specified. The decision to include an additional linear term is based on a likelihood ratio test statistic for the null hypothesis that the coefficient on the additional covariate is equal to zero. The threshold value will be denoted by $C_{\rm lin}$, with the value used in the applications is $C_{\rm lin}=1$. Finally, a threshold value for inclusion of second order terms has to be specified. Again the decision is based on the likelihood ratio statistic for the test of the null hypothesis that the additional second order term has a coefficient equal to zero. The threshold value will be denoted by $C_{\rm qua}$, and the value used in the applications in the current paper is $C_{\rm qua}=2.71$. The values for $C_{\rm lin}$ and $C_{\rm qua}$ have been tried out in simulations, but there are no formal results demonstrating

that these are superior to other values. More research regarding these choices would be useful.

5.3 Blocking with Regression

The first estimator relies on an initial estimate of the propensity score and uses subclassification (Rosenbaum and Rubin, 1983, 1984), combined with regression within the subclasses. Conceptually there are a couple of advantages of this estimator over simple weighting estimators. First, the subclassification approximately averages the propensity score within the subclasses, and so smoothes over the extreme values of the propensity score. Second, the regression within the subclasses adds a lot of flexibility compared to a single weighted regression.

I use the estimator for the propensity score discussed in the previous section. The estimator then requires partitioning of the range of the propensity score, the interval $[0,1]$ into J intervals of the form $[b_{j-1}, b_j)$, for $j = 1, \ldots, J$, where $b_0 = 0$ and $b_J = 1$. Let $B_i(j) \in \{0,1\}$ be a binary indicator for the event that the estimated propensity score for unit i, $\hat{e}(X_i)$, satisfies $b_{j-1} < \hat{e}(X_i) \leq b_j$. Within each block the average treatment effect is estimated using linear regression with some of the covariates, and including an indicator for the treatment:

$$\left(\hat{\alpha}_j, \hat{\tau}_j, \hat{\beta}_j\right) = \arg\min_{\alpha,\tau,\beta} \sum_{i=1}^{N} B_i(j) \cdot \left(Y_i - \alpha - \tau \cdot W_i - \beta' X_i\right)^2.$$

The covariates included beyond the intercept may consist of all, or a subset of the covariates viewed as most important. Because the covariates are approximately balanced within the blocks, the regression does not rely heavily on extrapolation as it might do if applied to the full sample. This leads to J estimates $\hat{\tau}_j$, one for each stratum or block. These J within-block estimates $\hat{\tau}_j$ are then averaged over the J blocks, using either the proportion of units in each block, $(N_{\rm c}j + N_{\rm t}j)/N$, or the proportion of treated units in each block, N_{tj}/N_t as the weights:

$$\tau_{\rm block}(\mathbf{Y}, \mathbf{W}, \mathbf{X}) = \sum_{j=1}^{J} \frac{N_{cj} + N_{tj}}{N} \cdot \hat{\tau}_j, \qquad \text{and} \quad \tau_{\rm block,treat}(\mathbf{Y}, \mathbf{W}, \mathbf{X}) = \sum_{j=1}^{J} \frac{N_{\rm t}j}{N_{\rm t}} \cdot \hat{\tau}_j. \quad (5.14)$$

The only decision the researcher has to make in order to implement this estimator is the number of blocks, J, and boundary values for the blocks, b_j, for $j = 1, \ldots, J-1$. Following analytic result by Cochran (1968) for the case with a single normally distributed covariate, researchers have often used five blocks, with an equal number of units in each block. Here I use a data-dependent procedure for selecting both the number of blocks and their boundaries, that leads to a number of blocks that increases with the sample size, developed in Imbens and Rubin (2014).

The algorithm relies on comparing average values of the log odds ratios by treatment status, where the estimated log odds ratio is

$$\hat{\ell}(x) = \ln\left(\frac{\hat{e}(x)}{1-\hat{e}(x)}\right).$$

Start with a single block, $J = 1$. Check whether the current stratification or blocking is adequate. This check is based on the comparison of three statistics, one regarding the correlation of the log odds ratio with the treatment indicator within the current blocks, and two concerning the block sizes if one were to split them further. Calculate the t-statistic for the test of the null hypothesis that the average value for the estimated propensity score for the treated units is the same as the average value for the estimated propensity score for the control units in the block. Specifically, if one looks at block j, with $B_{ij} = 1$ if unit i is in block j (or $b_{j-1} \leq \hat{e}(X_i) < b_j$), the t-statistic is

$$t = \frac{\overline{\hat{\ell}_{tj}} - \overline{\hat{\ell}_{cj}}}{\sqrt{S^2_{\ell,j,t}/N_{tj} + S^2_{\ell,j,c}/N_{cj}}},$$

where

$$\overline{\hat{\ell}_{cj}} = \frac{1}{N_{cj}}\sum_{i=1}^{N}(1-W_i)\cdot B_i(j)\cdot\hat{\ell}(X_i), \qquad \overline{\hat{\ell}_{tj}} = \frac{1}{N_{tj}}\sum_{i=1}^{N} W_i\cdot B_i(j)\cdot\hat{\ell}(X_i),$$

$$S^2_{\ell cj} = \frac{1}{N_{cj}-1}\sum_{i=1}^{N}(1-W_i)\cdot B_{ij}\cdot\left(\hat{\ell}(X_i) - \overline{\hat{\ell}_{cj}}\right)^2,$$

and

$$S^2_{\ell tj} = \frac{1}{N_{tj}-1}\sum_{i=1}^{N} W_i\cdot B_i(j)\cdot\left(\hat{\ell}(X_i) - \overline{\hat{\ell}_{tj}}\right)^2.$$

If the block were to be split, it would be split at the median of the values of the estimated propensity score within the block (or at the median value of the estimated propensity score among the treated if the focus is on the average value for the treated). For block j, denote this median by $b_{j-1,j}$, and define

$$N_{-,c,j} = \sum_{i:W_i=0} 1_{b_{j-1} \leq \hat{e}(X_i) < b_{j-1,j}}, \qquad N_{-,t,j} = \sum_{i:W_i=1} 1_{b_{j-1} \leq \hat{e}(X_i) < b_{j-1,j}},$$

$$N_{+,c,j} = \sum_{i:W_i=0} 1_{b_{j-1,j} \leq \hat{e}(X_i) < b_j} \qquad \text{and} \quad N_{+,t,j} = \sum_{i:W_i=1} 1_{b_{j-1,j} \leq \hat{e}(X_i) < b_j}.$$

The current block will be viewed as adequate if either the t-statistic is sufficiently small (less than $t^{\max}_{\text{block}}$), or if splitting the block would lead to too small a number of units in one of the treatment arms or in one of the new blocks. Formally, the current block will be viewed as adequate if either,

$$t \leq t^{\max}_{\text{block}} = 1.96,$$

or

$$\min\left(N_{-,c,j}, N_{-,t,j}, N_{+,c,j}, N_{+,t,j}\right) \leq 3,$$

or

$$\min\left(N_{-,c,j} + N_{-,t,j}, N_{+,c,j} + N_{+,t,j}\right) \leq K + 2,$$

where K is the number of covariates. If all the current blocks are deemed adequate the blocking algorithm is finished. If at least one of the blocks is viewed as not adequate, it is split by the median (or at the median value of the estimated propensity score among the treated if the focus is on the average value for the treated). For the new set of blocks the adequacy is calculated for each block, and this procedure continues until all blocks are viewed as adequate.

5.4 Matching with Replacement and Bias-Adjustment

In this section I discuss the second general estimation procedure, matching. For general discussions of matching methods see Rosenbaum (1989), Rubin (1973a, 1979), Rubin and Thomas (1992ab), Sekhon (2009), and Hainmueller (2012). Here I discuss a specific procedure developed by Abadie and Imbens (2006).[4] This estimator consists of two steps. First all units are matched, both treated and controls. The matching is with replacement, so the order in which the units are matched does not matter. After matching all units, or all treated units if the focus is on the average effect for the treated, some of the remaining bias is removed through regression on a subset of the covariates, with the subvector denoted by Z_i.

For ease of exposition we focus on the case with a single match. The methods generalize in a straightforward manner to the case with multiple matches. See Abadie and Imbens (2006) for details. Let the distance between two values of the covariate vector x and x' be based on the Mahalanobis metric: $\|x, x'\| = (x - x')'\hat{\Omega}_X^{-1}(x - x')$, where $\hat{\Omega}_X$ is the sample covariance matrix of the covariates. Now for each i, for $i = 1, \ldots, N$, let $\ell(i)$ be the index for the closest match, defined as

$$\ell(i) = \arg \min_{j: W_j \neq W_i} \|X_i - X_j\|,$$

where we ignore the possibility of ties. Given the $\ell(i)$ define

$$\hat{Y}_i(0) = \begin{cases} Y_i^{\text{obs}} & \text{if } W_i = 0, \\ Y_{\ell(i)}^{\text{obs}} & \text{if } W_i = 1, \end{cases} \qquad \hat{Y}_i(1) = \begin{cases} Y_{\ell(i)}, & \text{if } W_i = 0, \\ Y_i^{\text{obs}} & \text{if } W_i = 1. \end{cases}$$

Define also the matched values for the covariates:

$$\hat{X}_i(0) = \begin{cases} X_i & \text{if } W_i = 0, \\ X_{\ell(i)}, & \text{if } W_i = 1, \end{cases} \qquad \hat{X}_i(1) = \begin{cases} X_{\ell(i)}, & \text{if } W_i = 0, \\ X_i & \text{if } W_i = 1. \end{cases}$$

This leads to a matched sample, with N pairs. Note that the same units may be used as a match more than once, because we match with replacement. The simple matching estimator discussed in Abadie and Imbens (2006) is $\hat{\tau}_{\text{sm}} = \sum_{i=1}^{N} (\hat{Y}_i(1) - \hat{Y}_i(0))/N$. Abadie

[4] The estimator has been implemented in the official version of Stata. See also Abadie, Drukker, Herr, and Imbens (2003), and Becker and Ichino (2002).

and Imbens (2006, 2010) suggest improving the bias properties of this simple matching estimator by using linear regression to remove biases associated with differences between $\hat{X}_i(0)$ and $\hat{X}_i(1)$. See also Quade (1982) and Rubin (1973b). First run the two least squares regressions

$$\hat{Y}_i(0) = \alpha_c + \beta_c' \hat{X}_i(0) + \varepsilon_{ci}, \qquad \text{and} \quad \hat{Y}_i(1) = \alpha_t + \beta_t' \hat{X}_i(1) + \varepsilon_{ti},$$

in both cases on N units, to get the least squares estimates $\hat{\beta}_c$ and $\hat{\beta}_t$. Now adjust the imputed potential outcomes as

$$\hat{Y}_i^{\text{adj}}(0) = \begin{cases} Y_i^{\text{obs}} & \text{if } W_i = 0, \\ \hat{Y}_i(0) + \hat{\beta}_c' \left(\hat{X}_i(1) - \hat{X}_i(0) \right), & \text{if } W_i = 1, \end{cases}$$

and

$$\hat{Y}_i^{\text{adj}}(1) = \begin{cases} \hat{Y}_i(1) + \hat{\beta}_t' \left(\hat{X}_i(0) - \hat{X}_i(1) \right), & \text{if } W_i = 0, \\ Y_i^{\text{obs}} & \text{if } W_i = 1, \end{cases}$$

Now the bias-adjusted matching estimator is

$$\tau_{\text{match}} = \frac{1}{N} \sum_{i=1}^{N} \Big(\hat{Y}_i^{\text{adj}}(1) - \hat{Y}_i^{\text{adj}}(0) \Big), \tag{5.15}$$

and the bias-adjusted matching estimator for the average effect for the treated is

$$\tau_{\text{match}} = \frac{1}{N_t} \sum_{i:W_i=1} \Big(\hat{Y}_i^{\text{adj}}(1) - \hat{Y}_i^{\text{adj}}(0) \Big). \tag{5.16}$$

In practice the linear regression bias-adjustment eliminates a large part of the bias that remains after the simple matching. Note that the linear regression used here is very different from linear regression in the full sample. Because the matching ensures that the covariates are well-balanced in the matched sample, linear regression does not rely much on extrapolation the way it may in the full sample if the covariate distributions are substantially different.

5.5 A General Variance Estimator

In this section I will discuss an estimator for the variance of the two estimators for average treatment effects. Note that the bootstrap is not valid in general because matching estimators are not asymptotically linear. See Abadie and Imbens (2008) for detailed discussions.

5.5.1 The Weighted Average Outcome Representation of Estimators, and Asymptotic Linearity

The first key insight is that most estimators for average treatment effects share a common structure. This common structure is useful for understanding some of the commonalities of and differences between the estimators. These estimators, including the blocking and matching estimators discussed in Sections 5.3 and 5.4, can be written as a weighted average of observed outcomes,

$$\hat{\tau} = \frac{1}{N_t} \sum_{i:W_i=1} \lambda_i \cdot Y_i^{\text{obs}} - \frac{1}{N_c} \sum_{i:W_i=0} \lambda_i \cdot Y_i^{\text{obs}}, \tag{5.17}$$

with

$$\frac{1}{N_t} \sum_{i:W_i=1} \lambda_i = 1, \quad \text{and} \quad \frac{1}{N_c} \sum_{i:W_i=0} \lambda_i = 1.$$

Moreover, the weights λ_i do not depend on the outcomes $\mathbf{Y}^{\text{obs}}$, only on the covariates $\mathbf{X}$ and the treatment indicators $\mathbf{W}$. The specific functional form of the dependence of the weights λ_i on the covariates and treatment indicators depends on the particular estimator, whether linear regression, matching, weighting, blocking, or some combination thereof. Given the choice of the estimator, and given values for $\mathbf{W}$ and $\mathbf{X}$, the weights can be calculated. See Appendix B for the results for some common estimators.

5.5.2 The Conditional Variance

Here we focus on estimation of the variance of estimators for average treatment effects, conditional on the covariates $\mathbf{X}$ and the treatment indicators $\mathbf{W}$. We exploit the weighted linear average characterization of the estimators in (5.17) Hence the conditional variance is

$$\mathbb{V}\left(\hat{\tau}|\mathbf{X},\mathbf{W}\right) = \sum_{i=1}^{N} \left(\frac{W_i}{N_t^2} + \frac{1-W_i}{N_c^2} \right) \cdot \lambda_i^2 \cdot \sigma_{W_i}^2(X_i).$$

The only unknown components of this variance is $\sigma_{W_i}^2(X_i)$. Rather than estimating these conditional variances through nonparametric regression, following Abadie and Imbens (2006) I suggest using matching. Suppose unit i is a treated unit. Then find closest

match within the set of all other treated units in terms of the covariates. Ignoring ties, let $h(i)$ be the index of the unit with the same treatment indicator as i, closest to X_i:

$$h(i) = \arg \min_{j=1,\ldots,N, j\neq i, W_j = W_i} \|X_i - X_j\|.$$

Because $X_i \approx X_{h(i)}$, and thus $\mu_1(X_i) \approx \mu_1(X_{h(i)})$, it follows that we can approximate the difference $Y_i - Y_{h(i)}$ by

$$Y_i - Y_{h(i)} \approx \Big(Y_i(1) - \mu_1(X_i)\Big) + \Big(Y_{h(i)}(1) - \mu_1(X_{h(i)})\Big). \tag{5.18}$$

The right hand side of (5.18) has expectation zero and variance equal to $\sigma_1^2(X_i) + \sigma_1^2(X_{h(i)}) \approx 2\sigma_1^2(X_i)$. This motivates estimating $\sigma_{W_i}^2(X_i)$ by

$$\hat{\sigma}_{W_i}^2(X_i) = \frac{1}{2}\left(Y_i^{\text{obs}} - Y_{h(i)}^{\text{obs}}\right)^2.$$

Note that this estimator $\hat{\sigma}_{W_i}^2(X_i)$ is not a consistent estimator for $\sigma_{W_i}^2(X_i)$. However, this is not important, because we are interested not in the variances at specific points in the covariates distribution, but rather in the variance of the average treatment effect. Following the procedure introduce above, this variance is estimated as:

$$\hat{\mathbb{V}}\left(\hat{\tau}|\mathbf{X},\mathbf{W}\right) = \sum_{i=1}^{N}\left(\frac{W_i}{N_t^2} + \frac{1-W_i}{N_c^2}\right)\cdot\lambda_i^2\cdot\hat{\sigma}_{W_i}^2(X_i).$$

In principle one can generalize this variance estimator using the nearest L matches rather than just using a single match. In practice there is little evidence that this would make much of a difference. Hanson and Sunderam (2012) discuss extensions to clustered sampling.

5.6 Design: Ensuring Overlap

In this section I discuss two methods for constructing a subsample of the original data set that is more balanced in the covariates. Both take as input the vector of assignments $\mathbf{W}$ and the matrix of covariates $\mathbf{X}$, and select a set of units, a subset of the set of indices $\{1, 2, \ldots, N\}$, with N^T elements, leading to a trimmed sample with assignment vector $\mathbf{W}^T$, covariates $\mathbf{X}^T$, and outcomes $\mathbf{Y}^T$. The units corresponding to these indices will

then be used to apply the estimators for average treatment effects discussed in Sections 5.3 and 5.4.

The first method is aimed at settings with a large number of controls relative to the number of treated units, and the focus is on the average effect of the treated. This method constructs a matched sample where each treated unit is matched to a distinct control unit. This creates a sample of size $2 \cdot N_t$ distinct units, half of them treated and half of them control units. This sample can then be used in the analyses of Sections 5.3 and 5.4.

The second method for creating a sample with overlap drops units with extreme values of the propensity score. For such units it is difficult to find comparable units with the opposite treatment. Their presence makes analyses sensitive to minor changes in the specification to the presence of outliers in terms of outcome values. In addition their presence increases the variance of many of the estimators. The threshold at which units are dropped is based on a variance criterion, following CHIM.

5.6.1 Matching Without Replacement On the Propensity Score to Create a Balanced Sample

The first methods creates balance by matching on the propensity score. Here we match without replacement. Starting with the full sample with N units, N_t treated and $N_c > N_t$ controls, the first step is to estimate the propensity score using the methods from Section 5.2. We then transform this to the log odds ratio,

$$\hat{\ell}(x; \mathbf{W}, \mathbf{X}) = \ln\left(\frac{\hat{e}(x; \mathbf{W}, \mathbf{X})}{1 - \hat{e}(x; \mathbf{W}, \mathbf{X})}\right).$$

To simplify notation I will drop the dependence of the estimated log odds ratio on $\mathbf{X}$ and $\mathbf{W}$ and simply write $\hat{\ell}(x)$. Given the estimated log odds ratio, the N_t treated observations are ordered, with the treated unit with the highest value of the estimated log odds ratio score first. Then, the first treated unit (the one with the highest value of the estimated log odds ratio), is matched with the control unit with the closest value of the estimated log odds ratio.

Formally, if the treated units are indexed by $i = 1, \ldots, N_t$, with $\hat{\ell}(X_i) \geq \hat{\ell}(X_{i+1})$, the index of the matched control $j(1)$ satisfies $W_{j(1)} = 0$, and

$$j(1) = \arg \min_{i:W_i=0} \left| \hat{\ell}(X_i) - \hat{\ell}(X_1) \right|.$$

Next, the second treated unit is matched to unit $j(2)$, where $W_{j(2)} = 0$, and

$$j(2) = \arg \min_{i:W_i=0, i \neq j(1)} \left| \hat{\ell}(X_i) - \hat{\ell}(X_2) \right|.$$

Continuing this for all N_t treated units leads to a sample of $2 \cdot N_t$ distinct units, half of them treated and half of them controls. Although before the matching the average value of the propensity score among the treated units must be at least as large as the average value of the propensity score among the control units, this need not be the case for the matched samples.

I do not recommend simply estimating the average treatment effect for the treated by differencing average outcomes in the two treatment groups in this sample. Rather, this sample is used as a trimmed sample, with possibly still a fair amount of bias remaining, but one that is more balanced in the covariates than the original full sample, and as a result more likely to lead to credible and robust estimates. One may augment this procedure by dropping units for whom the match quality is particularly poor, say dropping units where the absolute value of the difference in the log odds ratio between the unit and its match is larger than some threshold.

5.6.2 Dropping Observations with Extreme Values of the Propensity Score

The second method for addressing lack of overlap we discuss is based on the work by CHIM. Their starting point is the definition of average treatment effects for subsets of the covariate space. Let $\mathbb{X}$ be the covariate space, and $\mathbb{A} \subset \mathbb{X}$ be some subset of the covariate space. Then define $\tau(\mathbb{A}) = \mathbb{E}[Y_i(1) - Y_i(0) | X_i \in \mathbb{A}]$. The idea in CHIM is to choose a set $\mathbb{A}$ such that there is substantial overlap between the covariate distributions for treated and control units within this subset of the covariate space. That is, we wish to exclude from the set $\mathbb{A}$, values for the covariates for which there are few treated units compared

the number of control units or *vice versa.* The question is how to operationalize this. CHIM suggest looking at the asymptotic efficiency bound for the efficient estimator for $\tau(\mathbb{A})$. The motivation for that criterion is as follows. If there is a value for the covariate such that there are few treated units relative to the number of control units, then for this value of the covariates the variance for an estimator for the average treatment effect will be large. Excluding units with such covariate values should therefore improve the asymptotic variance of the efficient estimator. It turns out that this is a feasible criterion that leads to a relatively simple rule for trimming the sample. The trimming also improves the robustness properties of the estimators. The trimmed units tend to be units with high leverage whose presence makes estimators sensitive to outliers in terms of outcome values.

CHIM calculate the efficiency bound for $\tau(\mathbb{A})$, building on the work by Hahn (1998), assuming homoskedasticity so that $\sigma^2 = \sigma_0^2(x) = \sigma_1^2(x)$ for all x, and a constant treatment effect, as

$$\frac{\sigma^2}{q(\mathbb{A})} \cdot \mathbb{E}\left[\frac{1}{e(X)} + \frac{1}{1-e(X)}\middle| X \in \mathbb{A}\right],$$

where $q(\mathbb{A}) = \Pr(X \in \mathbb{A})$. They derive the characterization for the set $\mathbb{A}$ that minimizes the asymptotic variance and show that it has the simple form

$$\mathbb{A}^* = \{x \in \mathbb{X} | \alpha \leq e(X) \leq 1-\alpha\}, \tag{5.19}$$

where α satisfies

$$\frac{1}{\alpha \cdot (1-\alpha)} = 2 \cdot \mathbb{E}\left[\frac{1}{e(X)\cdot(1-e(X))}\middle| \frac{1}{e(X)\cdot(1-e(X))} \leq \frac{1}{\alpha\cdot(1-\alpha)}\right].$$

Crump et al then suggest dropping units with $X_i \notin \mathbb{A}^*$. Note that this subsample is selected solely on the basis of the joint distribution of the treatment indicators and the covariates, and therefore does not introduce biases associated with selection based on the outcomes.

Implementing the CHIM suggestion is straightforward once one has estimated the propensity score. Define the function $g : \mathbb{X} \mapsto \mathbb{R}$, with

$$g(x) = \frac{1}{\hat{e}(x)\cdot(1-\hat{e}(x))},$$

and the function $h : \mathbb{R} \mapsto \mathbb{R}$, with

$$h(\lambda) = \frac{1}{\left(\sum_{i=1}^{N} \mathbf{1}_{\{g(X_i) \leq \lambda\}}\right)^2} \sum_{i=1}^{N} \mathbf{1}_{\{g(X_i) \leq \lambda\}} \cdot g(X_i),$$

and let $\hat{\lambda} = \arg\min_\lambda h(\lambda)$. The value of α in (5.19) that defines $\mathbb{A}^*$ is $\hat{\alpha} = 1/2 - \sqrt{1/4 - \hat{\lambda}/2}$. Finding $\hat{\lambda}$ is straightforward. The function $h(\lambda)$ is a step function so one can simply evaluate the function at $\lambda = g(X_i)$ for $i = 1, \ldots, N$ and select the one that minimizes $h(\lambda)$.

The simulations reported in CHIM suggest that in many settings in practice the choice $\alpha = 0.1$, leading to

$$\tilde{\mathbb{A}} = \{x \in \mathbb{X} | 0.1 \leq e(X) \leq 0.9\},$$

provides a good approximation to the optimal $\mathbb{A}^*$.

5.7 Assessing Unconfoundedness

Although the unconfoundedness assumption is not testable, the researcher can often do calculations to assess the plausibility of this critical assumption. These calculations focus on estimating the causal effect of the treatment on a pseudo outcome, a variable known to be unaffected by it, typically because its value is determined prior to the treatment itself. Such a variable can be time-invariant, but the most interesting case is in considering the treatment effect on a lagged outcome, commonly observed in evaluations of labor market programs. If the estimated effect differs from zero, it is less plausible that the unconfoundedness assumption holds, and if the treatment effect on the pseudo outcome is estimated to be close to zero, it is more plausible that the unconfoundedness assumption holds. Of course this does not directly test this assumption; in this setting, being able to reject the null of no effect does not directly reflect on the hypothesis of interest, unconfoundedness. Nevertheless, if the variables used in this proxy test are closely related to the outcome of interest, the test arguably has more power. For these tests it is clearly helpful to have a number of lagged outcomes.

To formalize this, let us suppose the covariates consist of a number of lagged outcomes $Y_{i,-1}, \ldots, Y_{i,-T}$ as well as time-invariant individual characteristics Z_i, so that the full set of covariates is $X_i = (Y_{i,-1}, \ldots, Y_{i,-T}, Z_i)$. By construction only units in the treatment group after period -1 receive the treatment; all other observed outcomes are control outcomes. Also suppose that the two potential outcomes $Y_i(0)$ and $Y_i(1)$ correspond to outcomes in period zero. Now consider the following two assumptions. The first is unconfoundedness given only $T-1$ lags of the outcome:

$$Y_i(1), Y_i(0) \perp\!\!\!\perp W_i \,\Big|\, Y_{i,-1}, \ldots, Y_{i,-(T-1)}, Z_i,$$

and the second assumes stationarity and exchangeability:

$$f_{Y_{i,s}(0)|Y_{i,s-1}(0),\ldots,Y_{i,s-(T-1)}(0),Z_i,W_i}(y_s|y_{s-1}, \ldots, y_{s-(T-1)}, z, w), \text{ does not depend on } i \text{ and } s.$$

Then it follows that

$$Y_{i,-1} \perp\!\!\!\perp W_i \,\Big|\, Y_{i,-2}, \ldots, Y_{i,-T}, Z_i,$$

which is testable. This hypothesis is what the procedure described above assesses, by analyzing the data with $X_{\rm p} = Y_{-1}$ and $X_{\rm r} = (Y_{-2}, \ldots, Y_{-T}, Z)$. Whether this test has much bearing on unconfoundedness depends on the link between the two assumptions and the original unconfoundedness assumption. With a sufficient number of lags unconfoundedness given all lags except the very last one would appear to be a plausible assumption if unconfoundedness given all lags holds. Therefore the relevance of the test depends largely on the plausibility of the second assumption, stationarity and exchangeability.

6 Three Applications

In this section I will apply the methods discussed in the previous sections to three data sets. The first data set, the "lottery data", collected by Imbens, Rubin and Sacerdote (2001), contains information on individuals who won large prizes in the Massachusetts lottery, as well as on individuals who won small one-time prizes. They study the effect of lottery winnings on labor market outcomes. Here I look at the effect of winning a

large prize on average difference in labor earnings for the first six years after winning the lottery. The second illustration is based a widely used data set, the "experimental Lalonde data", originally collected and analyzed by Lalonde (1986). Labor market programs are a major area of application for the evaluation methods discussed here in this manuscript, with important applications in Ashenfelter and Card (1985), Lalonde (1986), and Card and Sullivan (1988). Here I use the version of the data put together by Dehejia and Wahba (1999), and which is available from Dehejia's website.[5] The data set contains information on men participating in an experimental job training program. The focus is on estimating the average effect of the program on subsequent earnings. The third data set, the "non-experimental Lalonde data", replaces the individuals in the control group of the experimental Lalonde data set with observations on men from the Current Population Survey (CPS). It is also available on Dehejia's website. The individuals in the CPS are substantially different from those in the experiment, which creates severe challenges for comparisons between the two groups.

6.1 The Imbens-Rubin-Sacerdote Lottery Data

We use a subset of 496 lottery players with complete information on key variables. Of these 496 individuals 237 won big prizes (on average approximately $50,000 per year, for twenty years), and 259 did not. For clarity we refer to the former group as the "winners" and the latter as the "losers," although the latter did win small one-time prizes. We have information on them concerning their characteristics and economic circumstances at the time of playing the lottery, and social security earnings for six years before and after playing the lottery. Although obviously lottery numbers are drawn randomly to determine the winners, within the subset of individuals who responded to our survey (approximately 50%) there may be systematic differences between individuals who won big prizes and individuals who did not. Moreover, even if there was no non-response, differential ticket buying behavior implies that simple comparisons between winners and non-winners do not necessarily have a causal interpretation.

[5]The webpage is http://www.nber.org/~rdehejia/nswdata.html.

6.1.1 Summary Statistics for the Imbens-Rubin-Sacerdote Data

Table 1 presents summary statistics for the covariates, including the normalized difference. The normalized differences are generally modest, with only three out of eighteen normalized differences larger than 0.30 in absolute value. These three are `Tickets Bought` (the number of tickets bought per week), `Age`, and `Years of Schooling`. The latter two may be related to the willingness to respond to the survey. Note that the overall response rate in the survey was approximately 50%.

6.1.2 Estimating the Propensity Score for the Imbens-Rubin-Sacerdote Data

Four covariates are selected for automatic inclusion in the propensity score, `Tickets Bought`, `Years of Schooling`, `Working Then`, and `Earnings Year -1`. The reason for including `Tickets Bought` is that by the nature of the lottery, individuals buying more tickets are more likely to be in the big winner sample, and therefore it is *a priori* known that this variable should affect the propensity score. The other three covariates are included because on *a priori* grounds they are viewed as likely to be associated with the primary outcome, average yearly earnings for the six years after playing the lottery.

The algorithm discussed in detail in the Appendix leads to the inclusion of four additional linear terms, and ten second order terms. The parameter estimates for the final specification of the propensity score are presented in Table 2. The covariates are listed in the order they were selected for inclusion in the specification. Note that only one of the earnings measures was included beyond the (automatically selected) earnings in the year immediately prior to playing the lottery.

To assess the sensitivity of the propensity score estimates to the selection procedure, I consider three alternative specifications. In the second specification, I do not select any covariates for automatic inclusion. In the third specification I include all linear terms, but no second order terms. In the fourth specification I use lasso (Tibshirani, 1996) to select among all first and second order terms for inclusion in the propensity score. In Table 3 I report the correlations between the log odds ratios (the logarithm of the ratio of the propensity score and one minus the propensity score) based on these three specifications,

and the number of parameters estimated and the value of the log likelihood function. It turns out in this particular data set, the automatic inclusion of the four covariates does not affect the final specification of the propensity score. All four covariates are included by the algorithm for choosing the specification of the propensity score even if they had not be pre-selected. The fit of the propensity score model selected by the algorithm is substantially better than that based on the specification with all eighteen covariates included linearly with no second order terms or the lasso. Even though the linear specification has the same degrees of freedom, it las a value of the log likelihood function that is lower by 30.2. The lasso selects fewer terms, twelve in total, and also has a substantially lower value for the log likelihood function.

6.1.3 Trimming the Imbens-Rubin-Sacerdote Data

In the lottery application the focus is on the overall average effect. I therefore use the CHIM procedure discussed in Section 5.6.2 for trimming. The threshold for the propensity score coming out of this procedure is 0.0891. This leads to dropping 86 individuals with an estimated propensity score less than 0.0891, and 87 individuals with an estimated propensity score in excess of 0.9109. Table 4 presents details on the number of units dropped by treatment status and value of the propensity score.

The trimming leaves us with a sample consisting of 323 individuals, 151 big winners and 172 small winners. Table 5 presents summary statistics for this trimmed sample. One can see that the normalized differences between the two treatment groups are substantially smaller. For example, in the full sample, the normalized difference in the number of tickets bought was 0.90, and in the trimmed sample it is 0.51. I then re-estimate the propensity score on the trimmed sample. The same four variables as before were selected for automatic inclusion. Again four additional covariates were selected by the algorithm for inclusion in the propensity score. These were the same four as in the full sample, with the exception of `male`, which was replaced by `pos earn year -5`. For the trimmed sample only four second order terms were selected by the algorithm. The parameter estimates for the propensity score estimated on the trimmed sample are presented in Table

6.

6.1.4 Assessing Unconfoundedness for the Imbens-Rubin-Sacerdote Data

Next I do some analyses to assess the plausibility of the unconfoundedness analyses. There are eighteen covariates, six characteristics from the time of playing the lottery, six annual earnings measures, and six indicators for positive earnings for those six years. I use three different pseudo-outcomes. First I analyze the data using earnings from the the last pre-winning year as the pseudo outcome, and second I use average earnings for the last two pre-winning years as the pseudo outcome, and in the third case I use average earnings for the last three pre-winning years as the pseudo outcome. In all three cases I pre-select four covariates for automatic inclusion in the propensity score, the same three characteristics as before with in each case the last pre-winning year of earnings given the new outcome. I re-do the entire analyses, including estimating the propensity score, trimming the sample and re-estimating the propensity score. Table 7 presents the results for the blocking and matching estimators. In all three cases the both blocking and matching estimates are substantively and statistically close to zero. The estimated effects for the three pseudo outcomes range from -$1,160 to -$390, with none of the three statistically significantly different from zero at the 10% level. This suggests that unconfoundedness may be a reasonable assumption in this setting.

6.1.5 Estimating Average Treatment Effects for the Imbens-Rubin-Sacerdote Data

Now we turn to the primary outcome, average earnings in the six years after playing the lottery, the effect of winning a big prize. Note that up to now we had not used data on the outcome in any of the analyses. I report eighteen estimates. One set of six uses no covariates in the regression part, one set uses the four pre-selected covariates in the regression part, and the final set of six uses all eighteen covariates in the regression part. In each set of six estimates there are two based on the full sample (one blocking estimator with a single block, and one matching estimator), and four based on the trimmed sample: one with no blocking, one with two blocks, one with the optimal number of blocks, and one

matching estimator. The results are reported in Table 8. With five blocks the estimates range from -$5,070 to -$5,740, with the standard errors between $1,400 to $2,000. With two blocks the estimates are similar, with a little more variation. With no blocking the estimates vary considerably more, and even more so in the full sample.

6.1.6 Sensitivity to the Specification of the Propensity Score for the Imbens-Rubin-Sacerdote Data

Finally we assess the sensitivity of the estimates of the average treatment effects to the specification of the propensity score. We compare three specifications, one with $C_L = 1$ and $C_Q = 2.71$, which leads to the inclusion of some linear and some second order terms, and the second corresponding to the inclusion of all linear terms and no second order terms, which formally corresponds to $C_L = 0$ and $C_Q = \infty$, and in the third specification we use the lasso to selection among all first and second order terms. In Table 9 we present results for the same six estimators we considered before. The range of estimates for the preferred $(C_L = 1, C_Q = 2.71)$ specification is $[-5.66, -4.19]$, with width 1.47. The corresponding range for the linear propensity score estimator with $(C_L = 0, C_Q = \infty)$ specification is $[-5.21, -2.39]$, with width 3.82. The lasso estimates range from -3.91 to -2.97, somewhat lower than the other estimates.

6.1.7 Conclusions for the Imbens-Rubin-Sacerdote Data

Overall the analyses suggests that the blocking estimates are robust, with a preferred estimate of the effect of winning a large prize on average yearly earnings around -$5,400, and a standard error of about $1,400. Moreoever, the supporting analyses suggest that this is credible as an estimate of the causal effect of winning the lottery.

6.2 The Lalonde Experimental Data (Dehejia-Wahba Sample)

The second data set comes from an experimental evaluation of a labor market training program. It was originally analyzed by Lalonde (1986), and subsequently by many other researchers. Various versions of the Lalonde data have been widely used in this literature (Dehejia and Wahba, 1999; Heckman and Hotz, 1989; Smith and Todd, 2005; Crump,

Hotz, Imbens and Mitnik, 2008b). The data we use here were analyzed by Dehejia and Wahba (1999) and are available on Dehejia's website. The data set contains information on 185 men who were randomly selected to the training program, and 260 men assigned to the control group. We focus on estimating the average effect of the training.

6.2.1 Summary Statistics for the Experimental Lalonde Data

Table 10 presents summary statistics for the experimental Lalonde data. Not surprisingly the normalized differences are modest. The largest is 0.30, for `no degree`. The t-statistic for this covariate is 3.1. If we had all the data from the randomized experiment this would be extremely unlikely, but given attrition it is not uncommon even in a randomized experiment to find some covariates whose distributions differ in the two treatment groups.

6.2.2 Estimating the Propensity Score for the Experimental Lalonde Data

Next we estimate the propensity score. If we had data from a completely randomized experiment with no attrition, the true propensity score would be constant. Even in that case the algorithm might select some variables for inclusion in the propensity score and so the estimated propensity score need not be constant. Here we include the four earnings variables, earnings in 1974 and 1975, and indicators for these earnings being equal to zero for inclusion into the propensity score. The algorithm selects three additional terms, `nodegree`, `hispanic`, and `education`, for inclusion in the linar part of the propensity score. In addition, the algorithm selects three second order terms, the interaction of `nodegree` and `education`, `earn '74` and `nodegree`, and `unempl '75` and `education`, for a total of eleven covariates. The results are in Table 11.

6.2.3 Trimming the Experimental Lalonde Data

Based on the estimated propensity score we discard individuals with very low and very high values of the propensity score. The procedure described in Section 5.6.2 leads to a threshold of 0.1299 for low values (and 0.8701 for high values). This leads to dropping four control units with low values of the propensity score, and one treated unit with a

low value of the propensity score (0.12) and two treated units with a high value of the propensity score. It should not come as a surprise that in this case the procedure does not lead to a large number of units being dropped. The randomization ensures that most units are comparable, and propensity score values are clustered around the average value 0.42, with a standard deviation of 0.13. The minimum and maximum values for the estimated propensity score are 0.03 and 0.91. Table 12 presents the subsample sizes.

We re-estimated the propensity score on the trimmed sample, but because the sample is so similar to the full sample the algorithm selects the exact same specification and the parameter estimates are very similar. For that reason they are not reported here.

6.2.4 Assessing Unconfoundedness for the Experimental Lalonde Data

The next step is to assess the unconfoundedness assumption. I use two pseudo outcomes. The first is `earn '75`, earnings in 1975, and the second is the average of earnings in 1974 and 1975. In the first case I use the eight covariates (the ten original ones minus earnings in 1975 and the indicator for zero earnings in 1975), and in the second case I use six covariates (the original ten minus the four earnings related covariates). In both cases I do the full analysis, including trimming the sample and estimating the propensity score. Table 14 presents the results for these analyses. I look at the subclassification and matching estimators, using the all covariates to adjust for remaining bias. Over the four cases (the two estimators and the two pseudo outcomes) the estimated effect of the treatment on the pseudo outcome ranges from -$80 to $220. In none of the four cases can we reject the null hypothesis of no effect of the treatment on the pseudo outcome (all four t-statistics are less than or equal to 1 in absolute value). Thus, the data are supportive of the unconfoundedness assumption, not surprisingly given that the data came from a randomized experiment.

6.2.5 Subclasses for the Experimental Lalonde Data

To gain further understanding in the workings of these methods I report details on the blocking procedure. The algorithm discussed in Section 5.3 and the Appendix leads for

the experimental Lalonde data to three blocks. Table 13 presents the block boundaries and the number of control and treated units in each block. In the trimmed sample the average estimated propensity for the treated individuals is 0.45 and for the control individuals 0.39, with a difference of 0.06. Within each of the three blocks, however, the difference in average propensity score values for treated and controls is 0.02 or less.

6.2.6 Estimating Average Treatment Effects for the Experimental Lalonde Data

Finally we turn to the outcome data, earnings in 1978. In Table 15 I present results for the blocking and matching estimators, using no covariates, the four earnings variables, or all ten covariates for adjusting beyond the blocking or matching. The results range from \$1,460 to \$2,300, in all cases significantly different from zero at the 5% level.

6.2.7 Conclusions for the Experimental Lalonde Data

For the experimental Lalonde data we find that the program had a positive effect on earnings, with point estimates somewhere between \$1,500 and \$2,300. The variation in the estimates is probably due to the small sample size and the thick-tailed earnings distribution. The data suggest that unconfoundedness is plausible.

6.3 The Lalonde Non-experimental Data (Dehejia-Wahba Sample)

The final data set is the non-experimental Lalonde data. Here we take the men who were assigned to the training program and compare them to a comparison group drawn from the Current Population Survey. This is a data set that has been widely used as a testing ground for estimators for treatment effects.

6.3.1 Summary Statistics for the Non-experimental Lalonde Data

Table 16 presents summary statistics for the covariates for the non-experimental Lalonde data, including the normalized difference. Compared to the lottery data or the experimental Lalonde data we see substantially larger differences between the two groups, with

normalized differences as large as 2 (for the indicator for being African-American), and many larger than 1 (including the *a priori* likely to be important earnings variables). These summary statistics suggest that, irrespective of whether one believes the unconfoundedness assumption is a plausible one, it will be a severe statistical challenge to adjust for these covariate differences. Simple least squares regression adjustments are likely to be sensitive to the exact implementation.

6.3.2 Estimating the Propensity Score for the Non-experimental Lalonde Data

The first step is to estimate the propensity score on the full sample. I selected the two prior earnings measures, and the two indicators for these measures being positive for automatic inclusion in the propensity score. The selection algorithm selected five additional linear terms, leaving only `education` as a covariate that was not selected. The algorithm then selected five second order terms, many of them involving the lagged earnings and employment measures. Table 17 presents the parameter estimates for this specification.

Clearly the second order terms are important here. The value of the log likelihood function at the maximum likelihood estimates is -408.8. If instead we use a simple linear specification with all ten covariates, the value of the log likelihood function is -474.5. The correlation between the log odds ratio (the logarithm of the ratio of the propensity score over one minus the propensity score) for the specification involving second order terms versus only linear terms is only 0.70. It appears unlikely that the linear specification of the propensity score will lead to an accurate approximation to the true conditional expectation.

6.3.3 Matching the Non-experimental Lalonde Data to Improve Balance

In this case we are interested in the average effect on the treated. Because we have a large number of potential controls we use the matching from Section 5.6.1 to obtain a more balanced sample. Specifically, I match the 185 treated individuals to the closest controls, without replacement. The match quality is not always high for this data set. Out of

the 185 matches, there are 30 matches with difference in log odds ratio larger than 1.00. These are treated units with a propensity score between 0.74 and 0.89, matched with control units with propensity scores between 0.49 and 0.69. Although these differences are substantial, it is plausible that the further adjustment procedures suggested in this paper are effective in removing them. In Table 18 we report the normalized differences for the original sample (the same as before in 16), the normalized differences in the matched sample, and the ratio of the two. Compared to the normalized differences prior to the matching, the normalized differences after propensity score matching are substantially smaller. Whereas before there were six (out of ten) normalized differences in excess of 1.00, now none of the normalized differences are larger than 0.28, with most smaller than 0.10. The remaining normalized differences are still substantial, and they highlight that the simple matching has not removed all covariate differences between treated and control units, as also evidenced by the poor match quality for some of the matches. Nevertheless, because the differences are so much smaller, it will now be more reasonable to expect various estimators to be able to remove most of the remaining differences.

We then re-estimate the propensity score on the matched sample. This time only two covariates are selected for inclusion in the linear propensity score beyond the four automatically selected. These two covariates are `married` and `nodegree`. In addition two second order terms are included. Table 19 presents the parameter estimates for this specification.

6.3.4 Assessing Unconfoundedness for the Non-experimental Lalonde Data

The next step is to assess the plausibility of the unconfoundedness assumption by estimating treatment effects for two pseudo outcomes, `Earn '75`, and (`Earn '74`+`Earn '75`)/2. in Table 20 presents the results, using both the bias-adjusted matching estimator and the blocking estimator. All four estimates are substantively large and we can in all four cases soundly reject the null hypothesis that the effect of the treatment on the pseudo outcome is zero. The conclusion is that we cannot be confident here that unconfoundedness holds based on these analyses.

It is interesting though to note that the estimated effect for the second pseudo outcome, (`Earn '74`+`Earn '75`)/2, where we only adjust for the six non-earnings related covariates, is much larger in absolute value than the estimated effect for the first pseudo outcome, `Earn '75`. In the latter case we adjust for differences in earnings in 1974 for treated and controls. This combination of results does suggest that with additional earnings measures unconfoundedness might be more plausible.

6.3.5 Subclasses for the Non-experimental Lalonde Data

For comparison with the experimental Lalonde data it is useful to compare the subclasses constructed by the blocking method. Here the method leads to five blocks. Table 21 presents the block boundaries and the number of control and treated units in each block. In the matched sample, before the blocking, the average estimated propensity score for the 185 treated individuals is 0.53 and the average estimated propensity score for the 185 matched control individuals is 0.47, with a difference of 0.06. Within four of the five blocks the difference in average propensity score values for treated and controls is 0.01 or less, with only in the first block the difference in average values for the propensity score equal to 0.05.

6.3.6 Estimating Average Treatment Effects for the Non-experimental Lalonde Data

Finally, Table 22 presents the results for the actual outcome, earnings in 1978. I report estimates without regression adjustment, with regression adjustment for the four *a priori* selected covariates and with covariate adjustment for all ten covariates. I report these estimates for the full sample with 15,992 controls, for the matched sample without further blocking, for the matched sample with two blocks and finally, for the matched sample with the optimal number of blocks. The estimates are fairly robust once I use at least two blocks for the matched samples, or even in the single block case with regression adjustment for the full set of ten covariates. In addition, the estimates are fairly close to the estimates based on the experimental sample.

6.3.7 Conclusions for the Non-experimental Lalonde Data

The results for the non-experimental Lalonde data are interesting and quite different from those of the other data sets. The point estimates are fairly close to the estimates based on the experimental data, and the estimates are robust to changes in the specification. However, the estimates based on the pseudo outcomes suggest we would not have been able to ascertain that the unconfoundedness assumption was plausible based on the non-experimental data alone. We would have needed additional lagged measures of earnings to get a better assessment of the unconfoundedness assumption.

7 Conclusion

In this paper I have outlined a strategy for estimating average treatment effects in settings with unconfoundedness. I have described a specific algorithm to estimate the propensity score in a flexible manner and to implement subclassification and matching methods based on Imbens and Rubin (2013). I stressed the importance of dropping units with extreme values of the propensity score to obtain a more balance sample. I also described a method for assessing the plausibility of the unconfoundedness assumption. These methods are illustrated on three data sets, and are seen to lead to robust estimates for average treatment effects in all three cases. In two of the examples, with the lottery data and the experimental Lalonde data, the data pass the test that suggests that unconfoundedness is plausible. In the third case, with the non-experimental Lalonde data, we cannot be confident that unconfoundedness holds based on the data. Thus we would be less sure about the findings absent the experimental estimates as a benchmark, which in a normal study the researcher does not have.

Appendix

A. Estimating the Propensity Score

Given the choices, X_B, C_{lin}, and C_{qua}, the algorithm selects covariates for inclusion in the specification of the propensity score using the following eleven steps.

1. Estimate the logistic regression model, by maximum likelihood, with the basic covariates X_B.

2. Estimate $K+1-K_B$ additional logistic regression models where each model includes a single additional element of X not included in X_B. In each case calculate the likelihood ratio test statistic for the null hypothesis that the coefficient on this additional variable is equal to zero against the alternative hypothesis that the coefficient on this additional covariate differs from zero.

3. If the largest of the $K+1-K_B$ likelihood ratio test statistics is smaller than C_{lin}, go to step 6. If the the largest of the likelihood ratio test statistics is larger than or equal to C_{lin}, select the corresponding covariate for inclusion in the vector $h(\cdot)$, and go to step 4.

4. At this stage K_B+K_L linear terms have been selected for inclusion in the propensity score. Estimate $K+1-K_B-K_L$ logistic regressions, each with the already selected K_B+K_L covariates, plus one of the remaining covariates at a time. For each case calculate the likelihood ratio test statistic for the null hypothesis that the coefficient on this additional variable is equal to zero against the alternative hypothesis that it differs from zero.

5. If the largest largest of the likelihood ratio test statistics is smaller than C_{lin}, go to Step 6. If the the largest of the likelihood ratio test statistics is larger than or equal to C_{lin}, select the corresponding covariate for inclusion in the vector $h(\cdot)$, and return to Step 4.

6. At this stage K_B+K_L linear terms have been selected (including the intercept), and none of the remaining covariates would improve the log likelihood more than by $C_{\text{lin}}/2$ (given that the likelihood ratio statistic is twice the difference in log likelihood values). Now I will select a subset of the second order terms. I only consider second order terms for covariates that have been selected for inclusion in the linear part of the specification. Excluding the intercept that leaves K_B+K_L-1 linear terms, and thus $(K_B+K_L-1)\times(K_B+K_L)/2$ potential second order terms. I follow essentially the same algorithm as for the linear case for deciding which of these second order terms to include, but with the threshold for the likelihood ratio test statistic equal to C_{qua} instead of C_{lin}.

7. Estimate $(K_B+K_L-1)\times(K_B+K_L)/2$ logistic regression models, each including the K_B+K_L linear terms, and one second order term. Calculate the likelihood ratio test statistics for the null hypothesis that the coefficient on the second order term is equal to zero.

8. If the largest largest of the likelihood ratio test statistics is smaller than C_{qua}, go to Step 11. If the largest of the likelihood ratio test statistics is larger than or equal to C_{lin}, select the corresponding second order term for inclusion in the vector $h(\cdot)$.

9. At this point there are $K_B + K_L$ linear terms selected, and K_Q second order terms. Estimate $(K_B + K_L - 1) \times (K_B + K_L)/2 - K_Q$ logistic regression models, each including the $K_B + K_L + K_Q$ terms already selected, and one of the remaining second order terms. Calculate the likelihood ratio test statistic for testing the null that the additional second order term has a zero coefficient.

10. If the largest largest of the likelihood ratio test statistics is smaller than C_{qua}, go to Step 11. If the largest of the likelihood ratio test statistics is larger than or equal to C_{qua}, select the corresponding second order term for inclusion in the vector $h(\cdot)$, and go to step 9.

11. The vector $h(\cdot)$ now consists of the K_B terms selected *a priori*, the K_L linear terms, and the K_Q second order terms. Estimate the propensity score by maximum likelihood using this specification.

This algorithm will always converge to some specification for the propensity score. It need not select all covariates that are important, and it may select some that are not important, but it can generally provide a reasonable starting point for the specification of the propensity score. It is likely that it will lead to a substantial improvement over simply including all linear terms and no second order terms. Incidentally, this algorithm would lead to the linear specification if one fixed $C_{\text{lin}} = 0$ (so that all linear terms would be included), and $C_{\text{qua}} = \infty$ (so that no second order terms would be included).

B. Weights for Various Estimators

Here I briefly show what the weights are for some widely used estimators.

B.1 Difference Estimator

For the difference in average outcomes for treated and control units, $\hat{\tau} = \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$ we have

$$\lambda_i = 1.$$

The weights for all units are equal to one.

B.2 Regression Estimator

Consider the least squares estimator in the regression with a scalar covariate X_i,

$$Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + \beta \cdot X_i + \varepsilon_i.$$

The weights are

$$\lambda_i = \frac{S_X^2(N_c(N-1)/N^2) + (N_t/N)(N_c/N)(\overline{X}_t - \overline{X}_c)(X_i - \overline{X})}{(N_t(N-1)/N^2)S_X^2 - (N_t/N)(N_c/N)(\overline{X}_t - \overline{X}_c)^2},$$

where $S_X^2 = \sum_{i=1}^N (X_i - \overline{X})^2/(N-1)$ is the sample variance of X_i. Here the weights are not necessarily all positive.

B.3 Subclassification Estimator

For the subclassification estimator let the number of units in subclass j be equal to N_j, and the number of control and treated units in this subclass be equal to $N_{c,j}$ and $N_{t,j}$ respectively,

and let $B_i(j)$ be an indicator for unit i falling in subclass j. Then

$$\lambda_i = \begin{cases} (N_c/N_{cj}) \cdot (N_j/N), & \text{if } W_i = 0, B_i(j) = 1, \\ (N_t/N_{tj}) \cdot (N_j/N), & \text{if } W_i = 1, B_i(j) = 1. \end{cases}$$

B.4 Matching Estimator

A matching estimator with 1 matches for each treated and control unit has the form,

$$\hat{\tau}^{\text{match}} = \frac{1}{N} \sum_{i=1}^{N} \Big(\hat{Y}_i(1) - \hat{Y}_i(0) \Big),$$

where

$$\hat{Y}_i(w) = \begin{cases} Y_i^{\text{obs}} & \text{if } W_i = w, \\ Y_{j(i)}^{\text{obs}} & \text{if } W_i \neq w, \end{cases}$$

ensuring that $\hat{Y}_i(w)$ is a linear combination of Y_j^{obs} with positive weights.

B.5 Weighting Estimator

The weighting estimator (Hirano, Imbens, and Ridder, 2003), has the form

$$\hat{\tau} = \sum_{i:W_i=1} \frac{Y_i}{e(X_i)} \Big/ \sum_{i:W_i=1} \frac{1}{e(X_i)} - \sum_{i:W_i=0} \frac{Y_i}{1-e(X_i)} \Big/ \sum_{i:W_i=0} \frac{1}{1-e(X_i)},$$

so that

$$\lambda_i = \begin{cases} \frac{N_c}{1-e(X_i)} \Big/ \sum_{j:W_j=0} \frac{1}{1-e(X_j)}, & \text{if } W_i = 0, \\ \frac{N_t}{e(X_i)} \Big/ \sum_{j:W_j=1} \frac{1}{e(X_j)}, & \text{if } W_i = 1. \end{cases}$$

Here the weights are all positive.

References

Abadie, A., S. Athey, G. Imbens, and J. Wooldridge, (2014), "Finite Population Standard Errors,"Unpublished Manuscript.

Abadie, A., and G. Imbens, (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74(1), 235-267.

Abadie, A., and G. Imbens, (2008), "On the Failure of the Bootstrap for Matching Estimators," *Econometrica*.

Abadie, A., D. Drukker, H. Herr, and G. Imbens, (2003), "Implementing Matching Estimators for Average Treatment Effects in STATA," *The Stata Journal*, 4(3), 290-311.

Angrist, J. D. and A. B. Krueger (2000), "Empirical Strategies in Labor Economics,"in A. Ashenfelter and D. Card eds. Handbook of Labor Economics, vol. 3. New York: Elsevier Science.

Angrist, J., and S. Pischke (2009), *Mostly Harmless Econometrics: An Empiricists' Companion*, Princeton University Press, Princeton, NJ.

Ashenfelter, O., and D. Card, (1985), "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs", *Review of Economics and Statistics*, 67, 648-660.

Becker, S., and A. Ichino, (2002), "Estimation of Average Treatment Effects Based on Propensity Scores," *The Stata Journal*, 2(4): 358-377.

Belloni, A., V. Chernozhukov, and C. Hansen (2012), "Inference on treatment effects after selection amongst high-dimensional controls,"CMAP working paper.

Busso, M., J. DiNardo, , and J. McCrary. (2008), "Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects,", Unpublished Manuscript, Department of Economics, University of Michigan.

Caliendo, M., (2006), *Microeconometric Evaluation of Labour Market Policies*, Springer Verlag, Berlin.

Caliendo, M., and S. Kopeinig, (2011), "Some practical guidance for the implementation of propensity score matching", *Journal of Economic Surveys*, Vol. 22(1): 31-72.

Card, D., and D. Sullivan, (1988), "Measuring the Effect of Subsidized Training Programs on Movements In and Out of Employment", *Econometrica*, vol. 56, no. 3, 497-530.

Chen, X., H. Hong, and . Tarozzi, (2008), "Semiparametric efficiency in GMM models with auxiliary data," *Annals of Statistics*.

Belloni, A., V. Chernozhukov, and C. Hansen, (2012), "Inference on treatment effects after selection amongst high-dimensional controls,"Working Paper, MIT.

Cochran, W., (1968) "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies", *Biometrics* 24, 295-314.

Crump, R., V. J. Hotz, V. J., G. Imbens, and O.Mitnik, (2008a), "Dealing with Limited Overlap in Estimation of Average Treatment Effects," *Biometrika*.

Crump, R., V. J. Hotz, V. J., G. Imbens, and O.Mitnik, (2008b), "Nonparametric Tests for Treatment Effect Heterogeneity," *Review of Economics and Statistics*.

Dawid, P. (1979), "Conditional Independence in Statistical Theory," *Journal of the Royal Statistical Society*, Series B, Vol. 41(1), 1-31.

Dehejia, R., and S. Wahba, (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs", *Journal of the American Statistical Association*, 94, 1053-1062.

Fisher, R. A., (1925),*The Design of Experiments*, 1st ed, Oliver and Boyd, London.

Frölich, M. (2000), "Treatment Evaluation: Matching versus Local Polynomial Regression,"Discussion paper 2000-17, Department of Economics, University of St. Gallen.

Frölich, M. (2004), "Finite Sample Properties of Propensity-Score Matching and Weighting Estimators ," *Review of Economics and Statistics*, Vol 86(1): 77-90.

Frölich, M. (2002), "A Note on the Role of the Propensity Score for Estimating Average Treatment Effects," *Econometric Reviews* Vol 23(2): 167-174.

Frölich, M. (2004), "Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators," *Review of Economics and Statistics*, Vol. 86(1), 77-90.

Guo, S., and M. Fraser, (2010), *Propensity Score Analysis*, Sage, Los Angeles.

Hahn, J., (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66 (2), 315-331.

Hahn, J., and G. Ridder, (2013), "Asymptotic Variance of Semiparametric Estimators With Generated Regressors " *Econometrica* 81(1): 315-340.

Hainmueller, J., (2012), "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies" *Political Analysis*.

Hanson, and A. Sunderam, (2012), "The Variance of Non-Parametric Treatment Effect Estimators in the Presence of Clustering," *The Review of Economics and Statistics*, Vol. 94, No. 4, Pages 1197-1201.

Heckman, J., and J. Hotz, (1989), "Alternative Methods for Evaluating the Impact of Training Programs", (with discussion), *Journal of the American Statistical Association.*, Vol. 84, No. 804, 862-874.

Heckman, J., and E. Vytlacil (2007), "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation,"in J. Heckman and E. Leamer eds. *Handbook of Econometrics*, vol. 6B, Chapter 70, 4779-4874. New York: Elsevier Science.

Hirano, K., G. Imbens, and G. Ridder, (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score,"*Econometrica*, 71(4): 1161-1189. July

Imbens, G., (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review,"*Review of Economics and Statistics*, 86(1): 1-29.

Imbens, G., and D. Rubin, (forthcoming), *Causal Inference*, Cambridge University Press.

Imbens, G., D. Rubin, and B. Sacerdote, (2000), "Estimating the Effect of Unearned Income,"*American Economic Review*.

Imbens, G, and J. Wooldridge, (2009), ""*Journal of Economic Literature*, (): .

Lalonde, R.J., (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data,"*American Economic Review*, 76, 604-620.

Lee, M.-J., (2005), *Micro-Econometrics for Policy, Program, and Treatment Effects* Oxford University Press, Oxford.

Millimet, D., and R. Tchernis, (2009), "On the Specification of Propensity Scores, With Applications to the Analysis of Trade Policies,"*Journal of Business and Economic Statistics*, 397-415.

Morgan, S. and C. Winship, (2007), *Counterfactuals and Causal Inference*, Cambridge University Press, Cambridge.

Murnane, R., and J. Willett, (2010), *Methods Matter*, Oxford University Press, Oxford.

Neyman, J., (1923), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9,"translated in *Statistical Science*, (with discussion), Vol 5, No 4, 465–480, 1990.

Quade, D., (1982), "Nonparametric Analysis of Covariance by Matching", *Biometrics*, 38, 597-611.

Rosenbaum, P., (1989), "Optimal Matching in Observational Studies", *Journal of the American Statistical Association*, 84, 1024-1032.

Rosenbaum, P., (1995), *Observational Studies*, Springer Verlag, New York.

Rosenbaum, P., (2010), *Design of Observational Studies*, Springer Verlag, New York.

Rosenbaum, P., and D. Rubin, (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70, 41-55.

Rosenbaum, P., and D. Rubin, (1984), "Reducing the Bias in Observational Studies Using Subclassification on the Propensity Score", *Journal of the American Statistical Association*, 79, 516-524.

Rubin, D., (1973a), "Matching to Remove Bias in Observational Studies", *Biometrics*, 29, 159-183.

Rubin, D., (1973b), "The Use of Matched Sampling and Regression Adjustments to Remove Bias in Observational Studies", *Biometrics*, 29, 185-203.

RUBIN, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology*, 66, 688-701.

RUBIN, D., (1979), "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association*, 74, 318-328.

RUBIN, D. B., (1990), "Formal Modes of Statistical Inference for Causal Effects," *J*ournal of Statistical Planning and Inference, 25, 279-292.

RUBIN, D., AND N. THOMAS, (1992a), "Affinely Invariant Matching Methods with Ellipsoidal Distributions," *Annals of Statistics* 20 (2) 1079-1093.

RUBIN, D., AND N. THOMAS, (1992b), "Characterizing the effect of matching using linear propensity score methods with normal distributions," *Biometrika* 79 797-809.

SEKHON, J., (2009), "Opiates for the Matches: Matching Methods for Causal Inference" *Annual Review of Political Science*.

TIBSHIRANI, R., (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Vol. 58(1): 267-288.

WOOLDRIDGE, J., (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.

Table 1: Summary Statistics Lottery Data

Covariate	Losers (N=259) mean	(s.d.)	Winners (N=237) mean	(s.d.)	t-stat	nor-dif
Year Won	6.38	1.04	6.06	1.29	-3.0	-0.27
# Tickets	2.19	1.77	4.57	3.28	9.9	0.90
Age	53.2	12.9	47.0	13.8	-5.2	-0.47
Male	0.67	0.47	0.58	0.49	-2.1	-0.19
Education	14.4	2.0	13.0	2.2	-7.8	-0.70
Working Then	0.77	0.42	0.80	0.40	0.9	0.08
Earn Y -6	15.6	14.5	12.0	11.8	-3.1	-0.27
Earn Y -5	16.0	15.0	12.1	12.0	-3.2	-0.28
Earn Y -4	16.2	15.4	12.0	12.1	-3.4	-0.30
Earn Y -3	16.6	16.3	12.8	12.7	-2.9	-0.26
Earn Y -2	17.6	16.9	13.5	13.0	-3.1	-0.27
Earn Y -1	18.0	17.2	14.5	13.6	-2.5	-0.23
Pos Earn Y-6	0.69	0.46	0.70	0.46	0.3	0.03
Pos Earn Y -5	0.68	0.47	0.74	0.44	1.6	0.14
Pos Earn Y -4	0.69	0.46	0.73	0.44	1.1	0.10
Pos Earn Y -3	0.68	0.47	0.73	0.44	1.4	0.13
Pos Earn Y -2	0.68	0.47	0.74	0.44	1.6	0.15
Pos Earn Y-1	0.69	0.46	0.74	0.44	1.2	0.10

Table 2: Estimated Parameters of Propensity Score for the Lottery Data

Variable	est	s.e.
intercept	30.24	0.13
preselected linear terms		
Tickets Bought	0.56	0.38
Education	0.87	0.62
Working Then	1.71	0.55
Earnings Year -1	-0.37	0.09
additional linear terms		
Age	-0.27	0.08
Year Won	-6.93	1.41
Pos Earnings Year -5	0.83	0.36
Male	-4.01	1.71
second order terms		
Year Won× Year Won	0.50	0.11
Earnings Year -1× Male	0.06	0.02
Tickets Bought× Tickets Bought	-0.05	0.02
Tickets Bought× Working Then	-0.33	0.13
Education× Education	-0.07	0.02
Education× Earnings Year -1	0.01	0.00
Tickets Bought× Education	0.05	0.02
Earnings Year -1× Age	0.002	0.001
Age × Age	0.002	0.001
Year Won× Male	0.44	0.25

Table 3: Alternative Specifications of the Propensity Score

	Baseline	No Pre-selected	$C_{\text{lin}} = 0, C_{\text{qua}} = \infty$	lasso
Degrees Of Freedom	18	18	18	12
Log Likelihood Function	-201.5	-201.5	-231.7	-229.1
Correlations of Log Odds Ratios				
Baseline	1.00	1.00	0.86	0.86
No Pre-selected	1.00	1.00	0.86	0.86
$C_{\text{lin}} = 0, C_{\text{qua}} = \infty$	0.86	0.86	1.00	0.98
lasso	0.86	0.86	0.98	1.99

Table 4: Sample Sizes for Subsamples with the Propensity Score between α and $1-\alpha$ ($\alpha = 0.0891$) by Treatment Status.

	low $e(x) < \alpha$	middle $\alpha \leq e(X) \leq 1-\alpha$	high $1-\alpha < e(X)$	All
Losers	82	172	5	259
Winners	4	151	82	237
All	86	323	87	496

Table 5: Summary Statistics Trimmed Lottery Data

Covariate	Losers (N=172) mean	(s.d.)	Winners (N=151) mean	(s.d.)	t-stat	nor-dif
Year Won	6.40	1.12	6.32	1.18	-0.6	-0.06
# Tickets	2.40	1.88	3.67	2.95	4.6	0.51
Age	51.5	13.4	50.4	13.1	-0.7	-0.08
Male	0.65	0.48	0.60	0.49	-1.0	-0.11
Education	14.0	1.9	13.0	2.2	-4.2	-0.47
Work Then	0.79	0.41	0.80	0.40	0.2	0.03
Earn Year -6	15.5	14.0	13.0	12.4	-1.7	-0.19
Earn Year -5	16.0	14.4	13.3	12.7	-1.8	-0.20
Earn Year -4	16.4	14.9	13.4	12.7	-2.0	-0.22
Earn Year -3	16.8	15.6	14.3	13.3	-1.6	-0.18
Earn Year -2	17.8	16.4	14.7	13.8	-1.8	-0.20
Earn Year -1	18.4	16.6	15.4	14.4	-1.7	-0.19
Pos Earn Year -6	0.71	0.46	0.71	0.46	-0.0	-0.00
Pos Earn Year -5	0.70	0.46	0.74	0.44	0.9	0.10
Pos Earn Year -4	0.71	0.46	0.74	0.44	0.5	0.06
Pos Earn Year -3	0.70	0.46	0.72	0.45	0.2	0.03
Pos Earn Year -2	0.70	0.46	0.72	0.45	0.5	0.05
Pos Earn Year -1	0.72	0.45	0.71	0.46	-0.1	-0.01

Table 6: Estimated Parameters of Propensity Score for the Trimmed Lottery Data

Variable	est	s.e.
`intercept`	21.77	0.13
pre-selected linear terms		
`Tickets Bought`	-0.08	0.46
`Years of Schooling`	-0.45	0.08
`Working Then`	3.32	1.95
`Earnings Year -1`	-0.02	0.01
additional linear terms		
`Age`	-0.05	0.01
`Pos Earnings Year -5`	1.27	0.42
`Year Won`	-4.84	1.53
`Earnings Year -5`	-0.04	0.02
second order terms		
`Year Won` × `Year Won`	0.37	0.12
`Tickets Bought` × `Year Won`	0.14	0.06
`Tickets Bought` × `Tickets Bought`	-0.04	0.02
`Working Then` × `Year Won`	-0.49	0.30

Table 7: Assessing Unconfoundedness for the Lottery Data: Estimates of Average Treatment Effects for Pseudo Outcomes

Pseudo Outcome	Blocking		Matching	
	est	(s.e.)	est	(s.e.)
Y_{-1}	-0.53	(0.78)	-0.10	(0.95)
$\frac{Y_{-1}+Y_{-2}}{2}$	-1.16	(0.83)	-0.88	(0.94)
$\frac{Y_{-1}+Y_{-2}+Y_{-3}}{3}$	-0.39	(0.95)	-0.81	(0.98)

Table 8: Lottery Data: Estimates of Average Treatment Effects

	Full Sample		Trimmed			
Cov	1 Block	Match	1 Block	2 Blocks	5 Blocks	Match
No	-6.16	-4.03	-6.64	-6.05	-5.66	-4.53
	(1.34)	(1.32)	(1.66)	(1.87)	(1.99)	(1.36)
Few	-2.85	-4.29	-3.99	-5.57	-5.07	-4.19
	(0.99)	(1.31)	(1.16)	(1.30)	(1.46)	(1.36)
All	-5.08	-5.77	-5.34	-6.35	-5.74	-5.00
	(0.93)	(1.31)	(1.10)	(1.29)	(1.40)	(1.36)

Table 9: Lottery Data: Sensitivity of Estimates of Average Treatment Effcts to Propensity Score Specification

	Preferred Specification $C_L = 1$, $C_Q = 2.71$		Linear Specification $C_L = 0$, $C_Q = \infty$		lasso	
Cov	Blocked	Match	Blocked	Match	Blocked	Match
No	-5.66	-4.53	-4.61	-3.62	-3.34	-3.50
	(1.99)	(1.36)	(2.08)	(1.56)	1.95	1.24
Few	-5.07	-4.19	-4.52	-2.39	-3.91	-2.97
	(1.46)	(1.36)	(1.46)	(1.56)	1.35	1.24
All	-5.74	-5.00	-5.21	-3.71	-3.64	-3.78
	(1.40)	(1.36)	1.60	1.56	1.31	1.24

Table 10: Summary Statistics for Experimental Lalonde Data

	experimental controls (N_c=260)		trainees (N_t=185)			
Covariate	mean	(s.d.)	mean	(s.d.)	t-stat	nor-dif
`Black`	0.83	0.38	0.84	0.36	0.5	0.04
`Hisp`	0.11	0.31	0.06	0.24	-1.9	-0.17
`Age`	25.05	7.06	25.82	7.16	1.1	0.11
`Married`	0.15	0.36	0.19	0.39	1.0	0.09
`Nodegree`	0.83	0.37	0.71	0.46	-3.1	-0.30
`Education`	10.09	1.61	10.35	1.97	1.4	0.14
`E'74`	2.11	5.69	2.10	4.89	-0.0	-0.00
`U'74`	0.75	0.43	0.71	0.46	-1.0	-0.09
`E'75`	1.27	3.10	1.53	3.22	0.9	0.08
`U'75`	0.68	0.47	0.60	0.49	-1.8	-0.18

Table 11: Estimated Parameters of Propensity Score for the Lalonde Experimental Data

Variable	est	s.e.
intercept	-3.48	0.10
pre-selected linear terms		
`earn '74`	0.03	0.05
`unempl '74`	-0.24	0.39
`earn '75`	0.06	0.05
`unempl '75`	-3.48	1.65
additional linear terms		
`nodegree`	7.33	4.25
`hispanic`	-0.65	0.39
`education`	0.29	0.37
second order terms		
`nodegree` $\times$ `education`	-0.67	0.35
`earn '74` $\times$ `nodegree`	-0.13	0.06
`unempl '75` $\times$ `eduction`	0.30	0.16

Table 12: Sample Sizes for Subsamples with the Propensity Score between α and $1-\alpha$ ($\alpha = 0.1299$) by Treatment Status.

	low $e(x) < \alpha$	middle $\alpha \leq e(X) \leq 1-\alpha$	high $1-\alpha < e(X)$	All
Controls	4	256	0	260
Treated	1	182	2	185
All	5	438	2	445

Table 13: Subclasses

Subclass	Pscore Min	Pscore Max	# Controls	# Treated	ave pscore controls	ave pscore treated	average difference in pscore	t-stat
1	0.07	0.38	152	67	0.32	0.33	0.01	0.8
2	0.38	0.49	52	42	0.42	0.42	0.01	1.0
3	0.49	0.85	52	73	0.56	0.58	0.02	1.4

Table 14: Assessing Unconfoundedness for the Lottery Data: Estimates of Average Treatment Effects for Pseudo Outcomes

Pseudo Outcome	Blocking est	Blocking (s.e.)	Matching est	Matching (s.e.)
`earn '75`	0.22	(0.22)	0.03	(0.27)
(`earn '74`+`earn '75`)/2	0.03	(0.36)	-0.08	(0.41)

Table 15: Experimental Lalonde Data: Estimates of Average Treatment Effects

	Full Sample		Trimmed Sample			
Cov	1 Block	Match	1 Block	2 Blocks	3 Blocks	Match
No	1.79	2.21	1.69	1.49	1.48	2.30
	(0.67)	(0.82)	(0.66)	(0.68)	(0.68)	(0.81)
Few	1.74	2.15	1.60	1.54	1.52	2.26
	(0.67)	(0.82)	(0.66)	(0.66)	(0.68)	(0.81)
All	1.67	2.11	1.56	1.56	1.46	2.26
	(0.64)	(0.82)	(0.65)	(0.64)	(0.65)	(0.81)

Table 16: Summary Statistics for Non-experimental Lalonde Data

	CPS controls (N_c=15,992)		trainees (N_t=185)			
Covariate	mean	(s.d.)	mean	(s.d.)	t-stat	nor-dif
`Black`	0.07	0.26	0.84	0.36	28.6	2.43
`Hisp`	0.07	0.26	0.06	0.24	-0.7	-0.05
`Age`	33.23	11.05	25.82	7.16	-13.9	-0.80
`Married`	0.71	0.45	0.19	0.39	-18.0	-1.23
`Nodegree`	0.30	0.46	0.71	0.46	12.2	0.90
`Education`	12.03	2.87	10.35	2.01	-11.2	-0.68
`E'74`	14.02	9.57	2.10	4.89	-32.5	-1.57
`U'74`	0.12	0.32	0.71	0.46	17.5	1.49
`E'75`	13.65	9.27	1.53	3.22	-48.9	-1.75
`U'75`	0.11	0.31	0.60	0.49	13.6	1.19

Table 17: Estimated Parameters of Propensity Score for the Lalonde Non-experimental (CPS) Data

Variable	est	s.e.
intercept	-16.20	(0.69)
pre-selected linear terms		
`earn '74`	0.41	(0.11)
`unempl '74`	0.42	(0.41)
`earn '75`	-0.33	(0.06)
`unempl '75`	-2.44	(0.77)
additional linear terms		
`iblack`	4.00	(0.26)
`married`	-1.84	(0.30)
`nodegree`	1.60	(0.22)
`hispanic`	1.61	(0.41)
`age`	0.73	(0.09)
second order terms		
`age` × `age`	-0.007	(0.002)
`unempl '74` × `unempl '75`	3.41	(0.85)
`earn '74` × `age`	-0.013	(0.004)
`earn '75` × `married`	0.15	(0.06)
`unempl '74` × `earn '75`	0.22	(0.09)

Table 18: Normalized Differences Before and After Matching for Non-experimental Lalonde Data

	Full Sample nor-dif	Matched Sample nor-dif	ratio of nor-dif
Black	2.43	0.00	0.00
Hispanic	-0.05	0.00	-0.00
Age	-0.80	-0.15	0.19
Married	-1.23	-0.28	0.22
Nodegree	0.90	0.25	0.28
Education	-0.68	-0.18	0.26
E'74	-1.57	-0.03	0.02
U'74	1.49	0.02	0.02
E'75	-1.75	-0.07	0.04
U'75	1.19	0.02	0.02

Table 19: Estimated Parameters of Propensity Score for the Matched Lalonde Nonexperimental (CPS) Data

Variable	est	s.e.
intercept	-0.15	0.11
pre-selected linear terms		
earn '74	0.03	0.04
unempl '74	-0.00	0.42
earn '75	-0.06	0.05
unempl '75	0.26	0.36
additional linear terms		
married	-0.52	0.55
nodegree	0.26	0.26
second order terms		
unempl '75 × married	-1.24	0.55
married × nodegree	1.10	0.55

Table 20: Assessing Unconfoundedness for the Non-experimental Lalonde Data: Estimates of Average Treatment Effects for Pseudo Outcomes

Pseudo Outcome	Blocking		Matching	
	est	(s.e.)	est	(s.e.)
`earn '75`	-1.22	(0.25)	-1.24	(0.30)
(`earn '74`+`earn '75`)/2	-6.13	(0.49)	-6.37	(0.67)

Table 21: Subclasses

Subclass	Pscore		# Controls	# Treated	average pscore		difference in	
	Min	Max			controls	treated	average pscore	t-stat
1	0.00	0.37	31	7	0.20	0.25	0.05	1.75
2	0.37	0.43	5	7	0.39	0.40	0.00	0.39
3	0.43	0.46	26	22	0.44	0.44	0.00	0.18
4	0.46	0.53	36	36	0.50	0.50	0.00	0.51
5	0.53	1.00	87	113	0.57	0.58	0.01	1.14

Table 22: Non-experimental Lalonde Data: Estimates of Average Treatment Effects

	Full Sample		Trimmed Sample			
Cov	1 Block	Match	1 Block	2 Blocks	4 Blocks	Match
No	-8.50	1.72	1.72	1.81	1.79	1.98
	(0.58)	(0.90)	(0.74)	(0.75)	(0.76)	(0.85)
Few	0.69	1.73	1.81	1.80	2.10	1.98
	(0.59)	(0.90)	(0.73)	(0.73)	(0.75)	(0.85)
All	1.07	1.81	1.97	1.90	1.93	2.06
	(0.55)	(0.90)	(0.66)	(0.67)	(0.70)	(0.85)

Fig 1: Histogram of Earnings in 1975 for Control Group

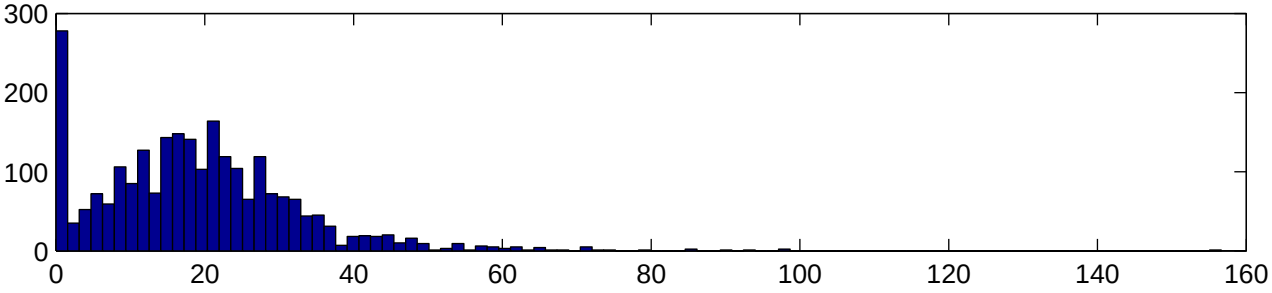

Fig 2: Histogram of Earnings in 1975 for Treatment Group

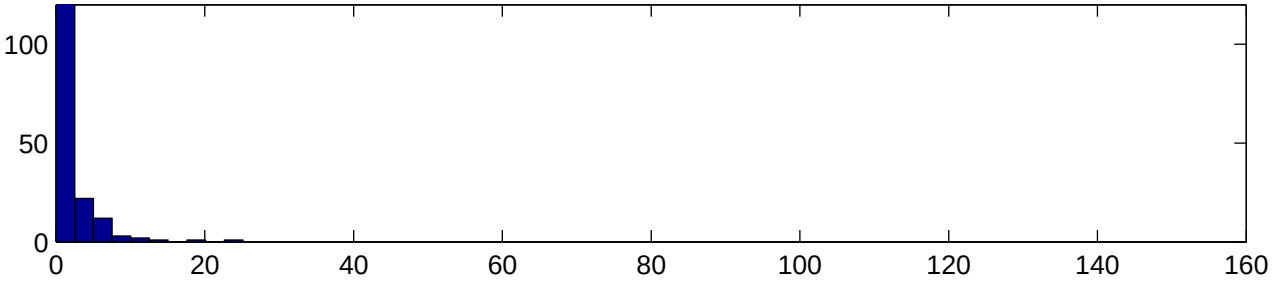

Fig 3: Linear and Log Linear Regression Functions, and Quartiles of Earnings Distributions