Lectures on Evaluation Methods **Impact Evaluation Network**
Guido Imbens **October 2010, Miami**

Methods for Estimating Treatment Effects I:
Unconfounded Treatment Assignment

1. Introduction

There is a large literature these days on methods for estimating treatment effects. For recent reviews in the econometrics literature see Angrist and Pischke (2009), Caliendo (2006), Heckman and Vytlacil (2007), Imbens (2004), Imbens and Wooldridge (2009), and Lee (2005). For reviews in other areas in applied statistics see Rosenbaum (), Morgan and Winship (), and (). The technical literature has largely focused on establishing first order large sample properties of point and interval estimators. These first order large sample properties are identical for a number of proposed methods, limiting their usefulness for choosing between these methods. In addition, comparisons between the various methods based on Monte Carlo evidence are hampered by the dependence of many of the proposed methods on tuning parameters (e.g., bandwidth choices in kernel regression, or number of terms in series methods) for which rarely specific, data-dependent values are recommended. As a result, this literature leaves the empirical researcher with a bewildering choice of methods, with limited, and sometimes conflicting guidance on what to use in practice. Possibly in response, researchers have often avoided nonparametric estimators. Instead they continue to use simple estimators, including linear regression with an indicator for the treatment, or weighting or matching on the estimated propensity score, with the propensity score based on a specification that simply includes all available covariates linearly. These methods do not always work well. If the covariate distributions differ substantially by treatment status, simple regression methods rely heavily on extrapolation, and as a result can be sensitive to small changes in the specification. Matching or weighting on the propensity score may be sensitive to the precise specification of the propensity score, with linearity of the log odds ratio not always an accurate approximation.

Here I present some alternative estimators that capture more of the nonparametric spirit of the recently proposed estimators. In contrast to those estimators, however, the proposals in this paper are fully specified, with all tuning parameters chosen in a data dependent manner. I do not establish formal large sample optimality properties for these methods, and in fact they do not have such properties. Nevertheless, these methods may be useful as a benchmark or starting point for more sophisticated analyses, and at the same time be more robust against small changes in the specification than the simple estimators that currently dominate empirical work in economics.

Ultimately what I do here is in the current paper is to specify a function that takes as its argument the data, and gives an estimate of the average effect of the treatment, or the average effect of the treatment for the treated. The contribution is not a new estimator, or the derivation of formal properties of any of the existing estimators. Rather, it is a fully specified estimator that can be applied generally in problems of estimating average treatment effects under unconfoundedness. Thus, it leads to a function $\tau(\mathbf{Y}, \mathbf{W}, \mathbf{X})$ that, as a function of the vector of outcomes $\mathbf{Y}$, the vector of assignments $\mathbf{W}$, and the matrix of covariates $\mathbf{X}$, gives an estimate of the average treatment effect. The focus in this paper is on clarifying the details required for implementing the estimators. Technical conditions for formal properties will not be discussed. I illustrate the methods on two data sets, one of them the Lalonde data widely used in this literature, and argue that they give reasonable and robust answers in those cases. Software for implementing these methods is available on my website.

In specific situations, one may augment the implicit choices made regarding the specification of the regression function and the propensity score 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 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.

1 Set Up

The set up is the standard one in this literature, using the potential outcome framework that originates with Rubin (1974). See Imbens and Wooldridge (2009) for a recent survey and a discussion of the history of this framework.

The analysis is based on a sample of N units, indexed by $i = 1, \ldots, N$, drawn randomly from a large population. For each unit there are two potential outcomes, $Y_i(0)$ and $Y_i(1)$, the potential outcomes without and given treatment. Each unit is either assigned or not assigned to the treatment. In the first case, the treatment indicator is $W_i = 1$, otherwise $W_i = 0$, with $N_c = \sum_{i=1}^N (1 - W_i)$ the number of control units, and $N_t = \sum_{i=1}^N W_i$ the number of treated units. The observed outcome is

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

In addition a K-dimensional vector of covariates X_i, with $X_i \in \mathbb{X} \subset \mathbb{R}^K$ is observed.

Define

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

to be the population average effect of the treatment conditional on $X_i = x$. I focus on estimation of the average treatment effect conditional on the covariates,

$$\tau = \frac{1}{N} \sum_{i=1}^{N} \tau(X_i), \qquad \text{or the average for the treated,} \;\; \tau_{\text{treat}} = \frac{1}{N_t} \sum_{i:W_i=1} \tau(X_i).$$

Throughout the paper we assume unconfoundedness,

$$W_i \perp \left(Y_i(0), Y_i(1)\right) \;\Big|\; X_i.$$

This implies that we can estimate the average effects by adjusting for difference in covariates between treated and control units:

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

2 The Strategy

In this section I lay out the strategy for estimating the average treatment effect of interest, either the average over the entire sample, or the average over the subsample of treated units. This will involve two distinct phases. In the first phase, which Rubin () calls the design phase, only the data on treatment indicators and covariates will be analyzed. Thus, at this stage the outcome data are not used, and in fact need not be available. In this phase overlap in covariate distributions is assessed, and if found lacking, a subsample is created with more overlap. In addition, analyses may be carried out to assess the plausibility of unconfoundedness. In practice this stage may involve various iterations between developing models for the data and assessing balance in the covariates. Because this doet not involve the outcome data, these steps do not affect inference conditional on covariates and treatment indicators. The result of this first stage is a sample, $(\mathbf{Y}_S, \mathbf{W}_S, \mathbf{X}_S) = f(\mathbf{Y}, \mathbf{W}, \mathbf{X})$, with the sample size in the selected sample, N_S less than or equal to N.

In the second stage the outcome data are used and estimates of the average treatment effect of interest are calculated, $\hat{\tau} = \hat{\tau}(\mathbf{Y}_S, \mathbf{W}_S, \mathbf{X}_S)$. This step is more delicate with concerns about pre-testing bias, and I propose simple methods without relying on specification checks.

In both phases estimators for the propensity score, and estimators for average treatment effects are used. I will discuss specific methods for doing this. For the propensity score this means specifying a function

$$\hat{e}(x) = \hat{e}(x|\mathbf{X}, \mathbf{W}). \tag{1}$$

Estimating the propensity score involves step-wise logistic regression where the number of (functions of) the explanatory variables entering into the logistic regression model is chosen sequentially. We apply this function first to the full sample $(\mathbf{Y}, \mathbf{W}, \mathbf{X})$ to construct the selected sample $(\mathbf{Y}_S, \mathbf{W}_S, \mathbf{X}_S)$, and possibly again to the selected sample.

For the average treatment effect I specify two pairs of functions, one for the average effect, and one for the average effect on the treated. In both cases I discuss two options, matching and blocking, leading to a total of four functions

$$\hat{\tau} = \hat{\tau}(\mathbf{Y}, \mathbf{W}, \mathbf{X}). \tag{2}$$

These estimators will be more or less nonparametric, but with specific, data-dependent, values recommended for the tuning parameters.

To be more specific, let me list the five different steps.

1. Assess the balance in covariates.
2. Estimate the propensity score $e(x)$ on the full sample $(\mathbf{Y}, \mathbf{W}, \mathbf{X})$.
3. Construct a subsample with sufficient overlap, $(\mathbf{Y}_S, \mathbf{W}_S, \mathbf{X}_S)$.
4. If possible assess the uconfoundedness assumption by estimating average effect on pseudo outcome.
5. Estimate the average effect τ_S on the subsample, $\hat{\tau}_S = \tau(\mathbf{Y}_S, \mathbf{W}_S, \mathbf{X}_S)$, and the standard error.

3 Estimating the Propensity Score

In this section I propose 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.

The unknown parameter γ is estimated by maximum likelihood:

$$\hat{\gamma}_{\mathrm{ml}} = \arg\max_{\lambda} L(\gamma),$$

where $L(\gamma)$ is the log likelihood function based on the logistic distribution:

$$L(\gamma) = \sum_{i=1}^{N} W_i \cdot h(X_i)'\gamma - \ln\left(1 + \exp\left(h(X_i)'\gamma\right)\right).$$

The estimated propensity score is then

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

The sole remaining issue is the choice of the vector of functions $h(\cdot)$. This vector of functions always includes a constant: $h_1(x) = 1$. With x a K-vector of covariates (not counting the intercept), I restrict the remaining elements of $h(x)$ to be either equal to an element of x, or to be equal to the product of two elements of x. In this sense the estimator is not completely 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 $(K+1)\cdot(K+2)/2 - 1$ possible 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. 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 $1 \leq K_B \leq K+1$. This vector always includes the intercept, but need not include any covariates beyond this, 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. This value will be denoted by C_{lin}. The value used in the applications is $C_{\text{lin}} = 1$. Finally, a threshold value for inclusion of second order terms has to be specified. This value will be denoted by C_{qua}. The value used in the applications is $C_{\text{qua}} = 2.71$.

Given these 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 $K + 1 - K_B$ logistic regression models where each model includes one 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 go back 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 $C_{\mathrm{lin}}/2$ (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. 6. 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 lead 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, the linear specification would come out of this algorithm 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).

4 Two Estimators for Average Treatment Effects

In this section I discuss two estimators for average treatment effects. There are many more estimators proposed in the literature. I choose these two because in my experience they tend to have attractive properties in realistic settings. The estimators can be written as functions of the data, $(\mathbf{Y}, \mathbf{W}, \mathbf{X})$. The first estimator is a blocking estimator, denoted by $\tau_{\text{block}}(\mathbf{Y}, \mathbf{W}, \mathbf{X})$. The second is a matching estimator, denoted by $\tau_{\text{match}}(\mathbf{Y}, \mathbf{W}, \mathbf{X})$.

4.1 Blocking with Regression

The first estimator relies on an initial estimate of the propensity score. For this I use the estimator discussed in the previous section. The estimator then requires the partition 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 $b_J = 1$. Within the blocks the average treatment effect is estimated using linear regression with the full set of covariates, and including an indicator for the treatment:

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

This leads to J estimates $\hat{\tau}_j$, for for each stratum or block. These are then averaged, using either the proportion of units in each block, $(N_{0j} + N_{1j})/N$, or the proportion of treated units in each block, N_{1j}/N as the weights:

$$\tau_{\text{block}}(\mathbf{Y},\mathbf{W},\mathbf{X}) = \sum_{j=1}^{J} \frac{N_{0j}+N_{1j}}{N} \cdot \hat{\tau}_j. \tag{3}$$

The only choice the researcher has to make in order to implement this estimator is the number and boundary values for the blocks. Following a result by Cochran (), researchers have often used five blocks, with an equal number of units in each block. Here I propose 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. It relies on comparing average values of the log odds ratios by treatment status, where

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

1. Start with a single block. Check whether the current stratification is adequate. This check is based on the comparison of three statistics.

   (a) 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

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

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

   and

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

(b) 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

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

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

The current block will be viewed as adequate if either the t-statistic is sufficiently small (less than $t_{\text{block}}^{\max}$, 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,

$$t_{\text{block}}^{\max} = 1.96,$$

or

$$\min\left(N_{-,j,0}, N_{-,j,1}, N_{+,j,0}, N_{+,j,1}\right) \leq 3,$$

or

$$\min\left(N_{-,j,0} + N_{-,j,1}, N_{+,j,0} + N_{+,j,1}\right) \leq K + 2.$$

2. 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.

4.2 Matching with Bias-Adjustment

This estimator consists of two steps. First all units are matched, both treated and controls. The matching, following Abadie and Imbens (2006), is with replacement, so the order 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.

Let the distance be based on the Mahalabonis metric:

$$d(x,z) = (x-z)'\hat{\Omega}_X^{-1}(x-z),$$

where

$$\hat{\Omega}_X = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)', \qquad \text{with } \overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i.$$

Now for each i, for $i = 1, \ldots, N$, let $\mathcal{J}(i)$ be the set of indices for the closest match, defined as

$$\mathcal{J}(i) = \{1, \ldots, N | W_i \neq W_j, d(X_i, X_j) = \min_{m:W_m \neq W_i} d(X_i, X_m)\}.$$

Unless there are ties, each set $\mathcal{J}(i)$ will consist of singletons. Given the sets $\mathcal{J}(i)$ define

$$\hat{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ \frac{1}{\#\mathcal{J}(i)}\sum_{j\in\mathcal{J}(i)} Y_j, & \text{if } W_i = 1, \end{cases} \qquad \hat{Y}_i(1) = \begin{cases} \frac{1}{\#\mathcal{J}(i)}\sum_{j\in\mathcal{J}(i)} Y_j, & \text{if } W_i = 0, \\ Y_i & \text{if } W_i = 1. \end{cases}$$

Define also

$$\hat{X}_i(0) = \begin{cases} X_i & \text{if } W_i = 0, \\ \frac{1}{\#\mathcal{J}(i)}\sum_{j\in\mathcal{J}(i)} X_j, & \text{if } W_i = 1, \end{cases} \qquad \hat{X}_i(1) = \begin{cases} \frac{1}{\#\mathcal{J}(i)}\sum_{j\in\mathcal{J}(i)} X_j, & \text{if } W_i = 0, \\ X_i & \text{if } W_i = 1. \end{cases}$$

This leads to a matched sample, with N pairs. For each pair I use the quintuple

$$\left(\hat{Y}_i(0), \hat{Y}_i(1), \hat{X}_i(0), \hat{X}_i(1), W_i\right).$$

The simple matching estimator proposed by Abadie and Imbens (2006) is

$$\hat{\tau}_{\rm sm} = \frac{1}{N} \sum_{i=1}^{N} \Big(\hat{Y}_i(1) - \hat{Y}_i(0) \Big).$$

Abadie and Imbens show that the bias of this estimator is $O_p(N^{-1/K})$, where K is the dimension of the covariates. They suggest improving the bias properties by using linear regression to remove biases associated with differences between $\hat{X}_i(0)$ and $\hat{X}_i(1)$. First run the two least squares regressions

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

in both cases on N units, to get the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. Now adjust the imputed potential outcomes as

$$\hat{Y}_i^{\rm adj}(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ \frac{1}{\#\mathcal{J}(i)} \sum_{j \in \mathcal{J}(i)} Y_j + \hat{\beta}_c' \left(\hat{X}_i(1) - \hat{X}_i(0) \right), & \text{if } W_i = 1, \end{cases}$$

and

$$\hat{Y}_i^{\rm adj}(1) = \begin{cases} \frac{1}{\#\mathcal{J}(i)} \sum_{j \in \mathcal{J}(i)} Y_j + \hat{\beta}_t' \left(\hat{X}_i(0) - \hat{X}_i(1) \right), & \text{if } W_i = 0, \\ Y_i & \text{if } W_i = 1, \end{cases}$$

Now the bias-adjusted matching estimator is

$$\tau_{\rm match}(\mathbf{Y}, \mathbf{W}, \mathbf{X}) = \frac{1}{N} \sum_{i=1}^{N} \Big(\hat{Y}_i^{\rm adj}(1) - \hat{Y}_i^{\rm adj}(0) \Big). \tag{4}$$

This bias adjustment does not eliminate all biases associated with differences in the covariates in large samples sufficiently fast to achieve root-N convergence: for that one needs to add higher order terms in the covariates at a sample-size dependent rate. In practice, however, the linear regression adjustment eliminates a large part of the bias that remains after the simple matching.

5 A General Variance Estimator

Here we focus on estimation of the variance of estimators for average treatment effects, conditional on the covariates $\mathbf{X}$ and the treatment indicators $\mathbf{W}$. Both estimators discussed here, and in fact all estimators used in practice, are linear combinations of the outcomes,

$$\hat{\tau} = \sum_{i=1}^{N} \lambda_i(\mathbf{X}, \mathbf{W}) \cdot Y_i,$$

with the $\lambda(\mathbf{X}, \mathbf{W})$ known functions of the covariates and treatment indicators. Hence the conditional variance is

$$\mathbb{V}\left(\hat{\tau}|\mathbf{X}, \mathbf{W}\right) = \sum_{i=1}^{N} \lambda_i(\mathbf{X}, \mathbf{W})^2 \cdot \sigma_i,$$

where

$$\sigma_i^2 = \begin{cases} \sigma_c^2(X_i & \text{if } W_i = 0, \\ \sigma_t^2(X_i & \text{if } W_i = 1. \end{cases}$$

The only unknown component of this variance is σ_i^2. Rather than estimating this through nonparametric regression, AI suggest using matching to estimate $\sigma_w^2(x)$. To estimate $\sigma_{W_i}^2(X_i)$ one uses the closest match within the set of units with the same treatment indicator. Let $\tilde{\mathcal{M}}(i)$ be the set of indices with the same treatment indicator as i, closest to X_i:

$$\mathcal{M}(i) = \left\{ j \in \{1, \ldots, N\} | W_i = W_j, d(X_i, X_j) = \min_{m: W_i = W_m} d(X_i, X_m) \right\}.$$

Without ties, the sets $\mathcal{M}(i)$ would all be singletons. The conditional variance of the outcome variable for unit i, σ_i^2, can then be estimated as:

$$\hat{\sigma}_i^2 = \frac{1}{2 \cdot \#\mathcal{M}(i)} \sum_{m \in \mathcal{M}(i)} \left(Y_m - Y_i\right)^2,$$

Note that this estimator is not consistent for the conditional variance for unit i. However this is not important, as we are interested not in the variances at specific points in the covariates distribution, but in the variance of the average treatment effect. Following the process introduce above, this is estimated as:

$$\hat{\mathbb{V}}\left(\hat{\tau}|\mathbf{X},\mathbf{W}\right) = \sum_{i=1}^{N} \lambda_i(\mathbf{X},\mathbf{W})^2 \cdot \hat{\sigma}_i^2.$$

6 Design: Ensuring Overlap

Prior to estimating the average effect of the treatment, or the average effect of the treatment for the treated, one should assess whether there is sufficient overlap in the covariate distributions to expect reasonable results from any estimator. If, based on this assessment, one suspects that estimators are likely to be sensitive to minor modifications in their implementation, one may wish to construct a subsample of the original data set that is more balanced in the covariates. In this section I discuss two methods for doing so. 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_S elements, with assignment vector $\mathbf{W}_S$, covariates $\mathbf{X}_S$, and outcomes $\mathbf{Y}_S$. The units corresponding to these indices will then be used to apply the estimators for average treatment effects discussed in Section 4.

The first 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 $N_S = 2 \cdot N_t$ distinct units, half of them treated and half of them control units. These are then used in the analyses of Section 4.

The second drops units with extreme values of the propensity score. For such units it is difficult to find comparable units with the opposite treatment, and they tend to make analyses sensitive to minor changes in the specification and lower precision of the resulting estimates. The threshold at which units are dropped is based on a variance criterion, following Crump, Hotz, Imbens and Mitnik (2009).

6.1 Matching On the Propensity Score to Create a Balanced Sample

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 3. We then transform this to the log odds ratio,

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

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

$$\left|\hat{\ell}(X_{j(1)}) - \hat{\ell}(X_i)\right| = \min_{k=1,\ldots,N, W_k=0} \left|\hat{\ell}(X_k) - \hat{\ell}(X_i)\right|.$$

If there are ties, I take the first control unit (in terms of the original index) and use that as the match. Next, the second treated unit is matched to unit $j(2)$, where $W_{j(2)} = 0$, and

$$\left|\hat{\ell}(X_{j(2)}) - \hat{\ell}(X_i)\right| = \min_{k=1,\ldots,N, W_k=0, k\neq j(1)} \left|\hat{\ell}(X_k) - \hat{\ell}(X_i)\right|.$$

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.

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 selected 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.

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 Crump, Hotz, Imbens and Mitnik (2008). 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. Then define

$$\tau(\mathbb{A}) = \sum_{i=1}^{N} \mathbf{1}_{X_i \in \mathbb{A}} \cdot \tau(X_i) \Big/ \sum_{i=1}^{N} \mathbf{1}_{X_i \in \mathbb{A}}.$$

Crump et al calculate the efficiency bound for $\tau(\mathbb{A})$, assuming homoskedasticity, as

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

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 form

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

dropping observations with extreme values for the propensity score, with the cutoff value α determined by the equation

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

Crump et al then suggest estimating $\tau(\mathbb{A}^*)$. 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. The calculations in Crump *et al* suggest that $\alpha = 0.1$ provides a good approximation to the optimal $\mathbb{A}$.

7 Analyses with Outcome Data

Now the pieces are in place to describe the full algorithm for obtaining estimates of the average effect of the treatment, given a triple $(\mathbf{Y}, \mathbf{W}, \mathbf{X})$. First let me restate the decisions to be made by the researcher.

1. Choices for propensity score specification First a subset of the covariates, denoted by X_B, that will always be included in the propensity score, irrespective of correlations with the treatment indicator. This set can be empty, beyond the intercept. Second, thresholds C_{lin} and C_{qua} for including linear and quadratic terms in propensity score. The values used in the applications are $C_{\text{lin}} = 1$ and $C_{\text{qua}} = 2.71$.

2. Estimand Choose between estimating the average effect for full sample or for subsample of treated units.

3. Choices for blocking estimator Thresholds for t-statistic for splitting blocks, $t_{\text{max}} = 1.96$, and minimum block size, $K+2$, and minimum block size by treatment, 3. Within the blocks use linear regression with the full set of covariates.

4. Choices for matching estimator I recommend using the Mahalanobis distance metric, and bias reduction through regression on the full set of covariates.

Given these choices the algorithm for estimating the average effect of the treatment goes through the following steps.

1. Estimate the propensity score.

2. Trim the sample to ensure sufficient overlap in covariate distributions.

    (a) If the focus is on the average effect for the treated, match all the treated units based on the estimated propensity score, without replacement, to construct a sample of $N_S = 2 \cdot N_t$ distinct units.

   (b) If the focus is on the average effect for the full sample, drop units with extreme values of the propensity score, using threshold based on marginal distribution of propensity score.

3. For the trimmed subsample $(\mathbf{Y}_S, \mathbf{W}_S, \mathbf{X}_S)$, estimate the average effect using matching with bias-reduction or blocking.

   (a) Match on all covariates, with replacement, and reduce bias by regression. Match both the treated and the controls if the focus is on the average effect for the full sample, or match only the treated if the focus is on the effect for the treated. Given the matched sample use linear regression with all of the covariates.

   (b) Construct a number of strata. Within the strata use linear regression with all of the covariates.

4. Estimate the variance using the interpretation of the estimator as a linear combination of the outcomes.

8 Design: 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 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 labor market programs. If the estimated effect differs from zero, this implies that the treated observations are different from the controls in terms of this particular covariate given the others. If the treatment effect 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 $X_i = (Y_{i,-1}, \ldots, Y_{i,-T}, Z_i)$. 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:

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

and the second assumes stationarity and exchangeability:

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

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 tests. 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 but one appears plausible conditional on unconfoundedness given all lags, so the relevance of the test depends largely on the plausibility of the second assumption, stationarity and exchangeability.

9 Two Applications

In this section I will briefly introduce the two applications that I use to illustrate the methods discussed in this paper. The first data set was collected by Imbens, Rubin and Sacerdote (2001, IRS from hereon). They collected data 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 average difference in labor earnings for the first six years after winning the lottery.

The second illustration is from a widely used data set, originally collected and analyzed by Lalonde (1986). Here I use the version of the data used by Dehejia and Wahba (1999), and which is available from Dehejia's website.[1] The data set contains information on participants in a job training program augmented with observations from the Current Population Survey (CPS). The focus is on estimating the average effect of the program for trainees using the comparison group from the CPS. Because of the availability of a randomly assigned control group we can assess whether the non-experimental estimators are accurate.

9.1 The Imbens-Rubin-Sacerdote Lottery Data

The first illustration is based on the IRS lottery data. We use a subset of 496 individuals with complete information on key variables. Of these 496 lottery players 237 won big one-time prizes, and 259 did not. We have information on them about 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 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.

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

Table 1 presents summary statistics for the covariates, including the normalized difference.

9.1.1 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 outcome, earnings after playing the lottery.

The algorithm leads to the inclusion of four additional terms, and ten second order terms. The parameter estimates for the final specification of the propensity score is given 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 prior to playing the lottery.

To assess the sensitivity of the propensity score estimates to the selection procedure, I consider two alternative specifications. In specification I, I do not selection any covariates for automatic inclusion. In Specification II I include all linear terms, but no second order terms. This corresponds to setting $C_{\text{lin}} = 0$ and $C_{\text{qua}} = \infty$. In Table 3 I report the correlations between the log odds ratios coming out of 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 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. Even though that specification has the same degrees of freedom, it las a value of the log likelihood function that is lower by

30.2.

9.1.2 Trimming for the Imbens-Rubin-Sacerdote Data

In the lottery application the focus is on the overall average effect. I therefore use the CHIM procedure 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 more than 0.9109. Table 4 presents details on the number of units dropped by treatment status and value of the propensity score.

This leaves 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 considerably 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 are presented in Table 6.

9.1.3 Assessing Unconfoundedness for the Imbens-Rubin-Sacerdote Data

Next I do some analyses to assess the plausibility of the unconfoundedness analyses. First I analyze the data using earnings from the the last pre-winning year as the outcome, and second I use average earnings for the last two pre-winning years as the outcome. In both cases I pre-select the same four covariates for automatic inclusion in the propensity score, with in each case the last pre-winning year of earnings given the new outcome. I re-do the entire analyses, including trimming and estimating the propensity score. Table 7 presents the results, for the blocking estimator.

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

Finally I estimate for the actual outcome of interest, average earnings in the six years after playing the lottery, the effect of winning a big prize. I report twelve estimates. One set of four uses no covariates in the regression part, one set uses the four pre-selected covariates in the regression part, and the final set of four uses all eighteen covariates in the regression part. In each set of four estimates there is one based on the full sample, and three based on the selected sample: one with no blocking, one with two blocks, and one with the optimal number of blocks. The results are reported in Table **??**. With five blocks the estimates range from -5.07 to -5.74, with the standard errors between 1.4 to 2.0. With two blocks the estimates are similar, with a little more variation. With no blocking the estimates vary considerably more, and even more in the full sample. Overall this suggests that the blocking estimates are robust, with a preferred estimate around -5.4, and a standard error of about 1.4.

9.2 The Lalonde Data (Dehejia-Wahba Sample)

Table 11 presents summary statistics for the covariates, including the normalized difference. Compared to the lottery data we see much larger differences between the two groups, with normalized differences as large as 2, and many larger than 1. This suggests that it will be important to carefully adjust for these differences, and that simple least squares adjustments will not lead to robust results.

9.2.1 Estimating the Propensity Score for the Lalonde Data

The first step is to estimate the propensity score on the full sample. I selected the two prior earnings measures, and 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 tmers, many of them involving the lagged earnings and employment measures.

9.2.2 Matching for the Lalonde Data

In the next step we matched the 185 treated individuals to the closest controls, without replacement. In Table 11 we report the normalized differences for the original sample (the same as before in 11), and the normalized differences in the matched sample. Now none on the normalized differences are larger than 0.28, with most smaller than 0.10. 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 are `married` and `nodegree`. In addition two second order terms are included.

9.2.3 Assessing Unconfoundedness for the Lalonde Data

Then we assess the plausibility of the unconfoundedness assumption by estimating treatment effects for the pseudo outcome, `E'75`. We report the results in Table 13. We find the effect is negative, fairly substantial and statistically highly significant. We also look at this outcome separately for those with positive and zero earnings in 1974, and separately at the probability of positive earnings and the mean earnings given that it is positive. Testing that all four differences are zero leads to a p-value less than 0.001. The conclusion is that we cannot be confident here that unconfoundedness holds based on these analyses.

9.2.4 Estimating Average Treatment Effects for the Lalonde Data

Finally, in Table 14we report the results for the actual outcome. We report estimates without regression adjustment, with regression adjustment for the four *a priori* selected covariates and with covariate adjustment for all ten covariates. We report these estimates for the full sample with 15,992 controls, for the matched sample, and 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 we 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.

10 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 and to implement subclassification and matching methods. I also described a method for assessing the plausibility of the unconfoundedness assumption. These methods are illustrated on two data sets, and are seen to lead to robust estimates for average treatment effects in both cases. In one of the cases the data pass the test that suggests that unconfoundedness is plausible. In the other case, with the Lalonde data, we cannot be confident that unconfoundedness holds based on the data, and as a result are less sure about the findings.

References

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, (2007), "On the Failure of the Bootstrap for Matching Estimators,"unpublished manuscript, department of Economics, UC Berkeley.

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.

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

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

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, (2005), "Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors, Missing Data and Treatment Effects,"unpublished working paper, Department of Economics, New York University.

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,"forthcoming *Biometrika*.

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

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.

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

Frölich, M. (2002), "What is the Value of knowing the propensity score for estimating average treatment effects", Department of Economics, University of St. Gallen.

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

Heckman, J., and R. Robb, (1985), "Alternative Methods for Evaluating the Impact of Interventions,"in Heckman and Singer (eds.), *Longitudinal Analysis of Labor Market Data*, Cambridge, Cambridge University Press.

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., W. Newey and G. Ridder, (2003), "Mean-squared-error Calculations for Average Treatment Effects,"unpublished manuscript, Department of Economics, UC Berkeley.

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.

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

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., and D. Rubin, (1983a), "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.

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

Table 1: Summary Statistics Lottery Data

	Losers	(N=259)	Winners	(N=237)		
Covariate	mean	(s.d.)	mean	(s.d.)	t-stat	nor-dif
Year Won	6.23	1.18	6.38	6.06	-3.0	-0.27
# Tickets	3.33	2.86	2.19	4.57	9.9	0.90
Age	50.2	13.7	53.2	47.0	-5.2	-0.47
Male	0.63	0.48	0.67	0.58	-2.1	-0.19
Education	13.73	2.20	14.43	12.97	-7.8	-0.70
Working Then	0.78	0.41	0.77	0.80	0.9	0.08
Earn Y -6	13.8	13.4	15.6	12.0	-3.1	-0.27
Earn Y -5	14.12	13.76	15.96	12.12	-3.2	-0.28
Earn Y -4	14.21	14.06	16.20	12.04	-3.4	-0.30
Earn Y -3	14.80	14.77	16.62	12.82	-2.9	-0.26
Earn Y -2	15.62	15.27	17.58	13.48	-3.1	-0.27
Earn Y -1	16.3	15.7	18.0	14.5	-2.5	-0.23
Pos Earn Y-6	0.69	0.46	0.69	0.70	0.3	0.03
Pos Earn Y -5	0.71	0.45	0.68	0.74	1.6	0.14
Pos Earn Y -4	0.71	0.45	0.69	0.73	1.1	0.10
Pos Earn Y -3	0.70	0.46	0.68	0.73	1.4	0.13
Pos Earn Y -2	0.71	0.46	0.68	0.74	1.6	0.15
Pos Earn Y-1	0.71	0.45	0.69	0.74	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
Years of Schooling× Years of Schooling	-0.07	0.02
Years of Schooling× Earnings Year -1	0.01	0.00
Tickets Bought× Years of Schooling	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	$C_{\text{lin}} = 0, C_{\text{qua}} = \infty$
Degrees Of Freedom	19	19	19
Log Likelihood Function	-201.5	-201.5	-231.7
Correlations of Log Odds Ratios			
Baseline	1.00	1.00	0.86
No Pre-selected	1.00	1.00	0.86
$C_{\text{lin}} = 0, C_{\text{qua}} = \infty$	0.86	0.86	1.00

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

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

Table 5: Summary Statistics Trimmed Lottery Data

	Losers	(N=172)	Winners	(N=151)		
Covariate	mean	(s.d.)	mean	(s.d.)	t-stat	nor-dif
Year Won	6.36	1.15	6.40	6.32	-0.55	-0.06
# Tickets	2.99	2.52	2.40	3.67	4.55	0.51
Age	51.00	13.29	51.48	50.44	-0.70	-0.08
Male	0.63	0.48	0.65	0.60	-1.02	-0.11
Education	13.55	2.13	14.01	13.03	-4.23	-0.47
Work Then	0.80	0.40	0.79	0.80	0.24	0.03
Earn Year -6	14.33	13.28	15.49	13.02	-1.68	-0.19
Earn Year -5	14.72	13.67	15.99	13.29	-1.79	-0.20
Earn Year -4	15.02	13.98	16.44	13.40	-1.98	-0.22
Earn Year -3	15.63	14.64	16.84	14.26	-1.60	-0.18
Earn Year -2	16.34	15.27	17.77	14.71	-1.82	-0.20
Earn Year -1	17.01	15.67	18.38	15.45	-1.70	-0.19
Pos Earn Year -6	0.71	0.45	0.71	0.71	-0.01	-0.00
Pos Earn Year -5	0.72	0.45	0.70	0.74	0.88	0.10
Pos Earn Year -4	0.72	0.45	0.71	0.74	0.52	0.06
Pos Earn Year -3	0.71	0.45	0.70	0.72	0.23	0.03
Pos Earn Year -2	0.71	0.45	0.70	0.72	0.48	0.05
Pos Earn Year -1	0.71	0.45	0.72	0.71	-0.13	-0.01

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

Variable	est	s.e.
intercept	21.77	0.13
preselected 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 Outc	Covariates	Selected Cov	est	(s.e.)
Y_{-1}	$X, Y_{-6:-2}, Y_{-6:-2} > 0$	X_2, X_5, X_6, Y_{-2}	-0.53	(0.78)
$\frac{Y_{-1}+Y_{-2}}{2}$	$X, Y_{-6:-3}, Y_{-6:-3} > 0$	X_2, X_5, X_6, Y_{-3}	-1.16	(0.83)
$\frac{Y_{-1}+Y_{-2}+Y_{-3}}{3}$	$X, Y_{-6:-4}, Y_{-6:-4} > 0$	X_2, X_5, X_6, Y_{-4}	-0.39	(0.95)
$\frac{Y_{-1}+...+Y_{-4}}{4}$	$X, Y_{-6:-5}, Y_{-6:-5} > 0$	X_2, X_5, X_6, Y_{-5}	-0.56	(0.97)
$\frac{Y_{-1}+...+Y_{-5}}{5}$	$X, Y_{-6}, Y_{-6} > 0$	X_2, X_5, X_6, Y_{-6}	-0.41	(0.92)
$\frac{Y_{-1}+...+Y_{-6}}{6}$	X	X_2, X_5, X_6	-2.56	(2.17)
Actual Outcome				
Y^{obs}	$X, Y_{-1:-6}, Y_{-6:-1} > 0$	X_2, X_5, X_6, Y_{-1}	**-5.74**	**(1.40)**

Table 8: Lottery Data: Estimates of Average Treatment Effects

	Full Sample		Selected		Selected		Selected	
Cov	1 Block		1 Block		2 Blocks		5 Blocks	
No	-6.16	1.34	-6.64	1.66	-6.05	1.87	-5.66	1.99
Few	-2.85	0.99	-3.99	1.16	-5.57	1.30	-5.07	1.46
All	-5.08	0.93	-5.34	1.10	-6.35	1.29	-5.74	1.40

Table 9: Summary Statistics for Lalonde Data

	CPS controls (N=15,992)		trainees (N=185)			
Covariate	mean	(s.d.)	mean	(s.d.)	t-stat	nor-dif
Black	0.08	0.27	0.07	0.84	28.6	2.4
Hisp	0.07	0.26	0.07	0.06	-0.7	-0.1
Age	33.1	11.0	33.2	25.8	-13.9	-0.8
Married	0.71	0.46	0.71	0.19	-18.0	-1.2
Nodegree	0.30	0.46	0.30	0.71	12.2	0.9
Education	12.0	2.9	12.0	10.4	-11.2	-0.7
E’74	13.9	9.6	14.0	2.1	-32.5	-1.6
U’74	0.13	0.33	0.12	0.71	17.5	1.5
E’75	13.5	9.3	13.7	1.5	-48.9	-1.7
U’75	0.11	0.32	0.11	0.60	13.6	1.2

Table 10: Estimated Parameters of Propensity Score for the Lalonde Non-experimental (CPS) Data

Variable	est	s.e.
intercept	-16.20	0.69
preselected 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		
`black`	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`	-0.01	0.00
`unempl '74` × `unempl '75`	3.41	0.85
`earn '74`× `age`	-0.012	0.002
`earn '75` × `married`	0.15	0.06
`unempl '74` × `earn '75`	0.22	0.08

Table 11: Summary Statistics for Matched Lalonde Data

	All 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 12: Estimated Parameters of Propensity Score for the Lalonde Non-experimental (CPS) Data

Variable	est	s.e.
intercept	-0.15	0.11
preselected 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 13: Assessing Unconfoundedness for the Lalonde Data: Estimates of Average Treatment Effects for Pseudo Outcomes

			p-value
earnings 1975:	-0.90	(0.33)	0.006
chi-squared test	53.8	(dof=4)	< 0.001

Table 14: Lalonde Data: Estimates of Average Treatment Effects

	Full Sample		Matched		Matched		Matched	
Cov	1 Block		1 Block		2 Blocks		4 Blocks	
No	-8.50	0.58	1.72	0.74	1.81	0.75	1.86	0.76
Few	0.69	0.59	1.81	0.73	1.80	0.73	1.99	0.75
All	1.07	0.55	1.97	0.66	1.90	0.67	2.06	0.66