IRP Lectures **Madison, WI, August 2008**

Lecture 2, Monday, Aug 4th, 10.00-11.00am

Estimation of Average Treatment Effects Under Unconfoundedness, Part II

1. Introduction

In this lecture we discuss assessing the two key assumptions, unconfoundedness and overlap in covariate distributions. We then illustrate the issues discussed in this and the previous lecture using data from a labor market program originally analyzed by Lalonde (1986).

2. Assessing Unconfoundedness

The unconfoundedness assumption used throughout this discussion is not directly testable. It states that the conditional distribution of the outcome under the control treatment, $Y_i(0)$, given receipt of the active treatment and given covariates, is identical to the distribution of the control outcome given receipt of the control treatment and given covariates. The same is assumed for the distribution of the active treatment outcome, $Y_i(1)$. Yet since the data are completely uninformative about the distribution of $Y_i(0)$ for those who received the active treatment and of $Y_i(1)$ for those receiving the control, the data cannot directly reject the unconfoundedness assumption. Nevertheless, there are often indirect ways of assessing the this, a number of which are developed in Heckman and Hotz (1989) and Rosenbaum (1987). These methods typically rely on estimating a causal effect that is known to equal zero. If based on the test we reject the null hypothesis that this causal effect varies from zero, the unconfoundedness assumption is considered less plausible. These tests can be divided into two broad groups.

The first set of tests focuses on estimating the causal effect of a treatment that is known not to have an effect, relying on the presence of multiple control groups (Rosenbaum, 1987). Suppose one has two potential control groups, for example eligible nonparticipants and ineligibles, as in Heckman, Ichimura and Todd (1997). One interpretation of the test is

to compare average treatment effects estimated using each of the control groups. This can also be interpreted as estimating an "average treatment effect" using only the two control groups, with the treatment indicator now a dummy for being a member of the first group. In that case the treatment effect is known to be zero, and statistical evidence of a non-zero effect implies that at least one of the control groups is invalid. Again, not rejecting the test does not imply the unconfoundedness assumption is valid (as both control groups could suffer the same bias), but non-rejection in the case where the two control groups are likely to have different biases makes it more plausible that the unconfoundness assumption holds. The key for the power of this test is to have available control groups that are likely to have different biases, if at all. Comparing ineligibles and eligible nonparticipants is a particularly attractive comparison. Alternatively one may use different geographic controls, for example from areas bordering on different sides of the treatment group.

One can formalize this test by postulating a three-valued indicator $T_i \in \{-0, 1, 1\}$ for the groups (e.g., ineligibles, eligible nonnonparticipants and participants), with the treatment indicator equal to $W_i = 1\{T_i = 1\}$, so that

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

If one extends the unconfoundedness assumption to independence of the potential outcomes and the three-valued group indicator given covariates,

$$Y_i(0), Y_i(1) \perp\!\!\!\perp T_i \,\Big|\, X_i,$$

then a testable implication is

$$Y_i(0) \perp\!\!\!\perp 1\{T_i = 0\} \,\Big|\, X_i, T_i \in \{-1, 0\},$$

and thus

$$Y_i \perp\!\!\!\perp 1\{T_i = 0\} \,\Big|\, X_i, T_i \in \{-1, 0\}.$$

An implication of this independence condition is being tested by the tests discussed above. Whether this test has much bearing on the unconfoundedness assumption depends on whether the extension of the assumption is plausible given unconfoundedness itself.

The second set of tests of unconfoundedness focuses 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.

3. Assessing Overlap

The second of the key assumptions in estimating average treatment effects requires that the propensity score is strictly between zero and one. Although in principle this is testable, as it restricts the joint distribution of observables, formal tests are not the main concern. In practice, this assumption raises a number of issues. The first question is how to detect a lack of overlap in the covariate distributions. A second is how to deal with it, given that such a lack exists.

3.1 Propensity Score Distributions

The first method to detect lack of overlap is to plot distributions of covariates by treatment groups. In the case with one or two covariates one can do this directly. In high dimensional cases, however, this becomes more difficult. One can inspect pairs of marginal distributions by treatment status, but these are not necessarily informative about lack of overlap. It is possible that for each covariate the distribution for the treatment and control groups are identical, even though there are areas where the propensity score is zero or one.

A more direct method is to inspect the distribution of the propensity score in both treatment groups, which can reveal lack of overlap in the multivariate covariate distributions. Its implementation requires nonparametric estimation of the propensity score, however, and misspecification may lead to failure in detecting a lack of overlap, just as inspecting various

marginal distributions may be insufficient. In practice one may wish to undersmooth the estimation of the propensity score, either by choosing a bandwidth smaller than optimal for nonparametric estimation or by including higher order terms in a series expansion.

3.2 Selecting a Sample with Overlap

Once one determines that there is a lack of overlap one can either conclude that the average treatment effect of interest cannot be estimated with sufficient precision, and/or decide to focus on an average treatment effect that is estimable with greater accuracy. To do the latter it can be useful to discard some of the observations on the basis of their covariates. For example one may decide to discard control (treated) observations with propensity scores below (above) a cutoff level. To do this sytematically, we follow Crump, Hotz, Imbens and Mitnik (2006), who focus on sample averate treatment effects. 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

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

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

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

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 form

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

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

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

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. Calculations for Beta distributions for the propensity score suggest that $\alpha = 0.1$ approximates the optimal set well in practice.

4. The Lalonde Data

Here we look at application of the ideas discussed in these notes. We take the NSW job training data orinally collected by Lalonde (1986), and subsequently analyzed by Dehejia and Wahba (1999). The starting point is an experimental evaluation of this training program. Lalonde then constructed non-experimental comparison groups to investigate the ability of various econometric techniques to replicate the experimental results. In the current analysis we use three subsamples, the (experimental) trainees, the experimental controls, and a CPS comparison group.

In the next two subsections we do the design part of the analysis. Without using the outcome data we assess whether strong ignorability has some credibility.

4.1 Summary Statistics

First we give some summary statistics

Table 1: Summary Statistics for Experimental Sample

	Controls (N=260)		Trainees (N=185)			CPS (N=15,992)		
	mean	(s.d.)	mean	(s.d.)	diff / sd	mean	(s.d.)	diff / sd
Age	25.05	7.06	25.82	7.16	0.11	33.23	11.05	-0.67
Black	0.83	0.38	0.84	0.36	0.04	0.07	0.26	2.80
Education	10.09	1.61	10.35	2.01	0.14	12.03	2.87	-0.59
Hispanic	0.11	0.31	0.06	0.24	-0.17	0.07	0.26	-0.05
Married	0.15	0.36	0.19	0.39	0.09	0.71	0.45	-1.15
Earnings '74	2.11	5.69	2.10	4.89	-0.00	14.02	9.57	-1.24
Earnings '75	1.27	3.10	1.53	3.22	0.08	0.12	0.32	1.77
Unempl '74	0.75	0.43	0.71	0.46	-0.09	13.65	9.27	-1.30
Unempl. '75	0.68	0.47	0.60	0.49	-0.18	0.11	0.31	1.54

In this table we report averages and standard deviations for the three subsamples. In addition we report for the trainee/experimental-control and for the trainee/CPS-comparison-group pairs the difference in average covariate values by treatment status, normalized by the standard deviation of these covariates. So, in Table 1 we see that in the experimental data set the difference in average age between treated and controls is 0.11 standard deviations. In the nonexperimental comparison the difference in age is 0.67 standard deviations.

Note that we do not report the t-statistic for the difference. Essentially the t-statistic is equal to the normalized difference multiplied by the square root of the sample size. As such, the t-statistic partly reflects the sample size. Given a difference of 0.25 standard deviations between the two groups in terms of average covariate values, a larger t-statistic just indicates a larger sample size, and therefore in fact an easier problem in terms of finding credible estimators for average treatment effects. As this example illustrates, a larger t-statistic for the difference between average covariates by treatment group does not indicate that the problem of finding credible estimates of the treatment effect is more difficult. A larger normalized difference does unambiguously indicate a more severe overlap problem.

In general a difference in average means bigger than 0.25 standard deviations is substantial. In that case one may want to be suspicious of simple methods like linear regression with a dummy for the treatment variable. Recall that estimating the average effect essentially amounts to using the controls to estimate the conditional mean $\mu_0(x) = \mathbb{E}[Y_i|W_i = 1, X_i = x]$ and using this estimated regression function to predict the (missing) control outcomes for the treated units. With such a large difference between the two groups in covariate distributions, linear regression is going to rely heavily on extrapolation, and thus will be sensitive to the exact functional form.

Right away we can see that the experimental data set is well balanced. The difference in averages between treatment and control group is never more than 0.18 standard deviations. In contrast, with the CPS comparison group the differences between the averages are up to 1.77 standard deviations from zero, suggesting there will be serious issues in obtaining credible estimates of the average effect of the treatment.

In Figures 1 and 2 we present histogram estimates of the distribution of the propensity score for the treatment and control group in the experimental Lalonde data. These distributions again suggest that there is considerable overlap in the covariate distributions. In Figures 3 and 4 we present the histogram estimates for the propensity score distributions for the CPS comparison group. Now there is a clear lack of overlap. For the CPS comparison group almost all mass of the propensity score distribution is concentrated in a small interval to the right of zero, and the distribution for the treatment group is much more spread out.

4.2 Assessing Unconfoundedness

First we use the experimental data. We analyze the data as if earnings in 1975 (Earn '75) is the outcome. This is in fact a covariate, and so it cannot be affected by the treatment. Table 2 reports the results for eleven estimators .

Figure 1: histogram propensity score for controls, exper full sample

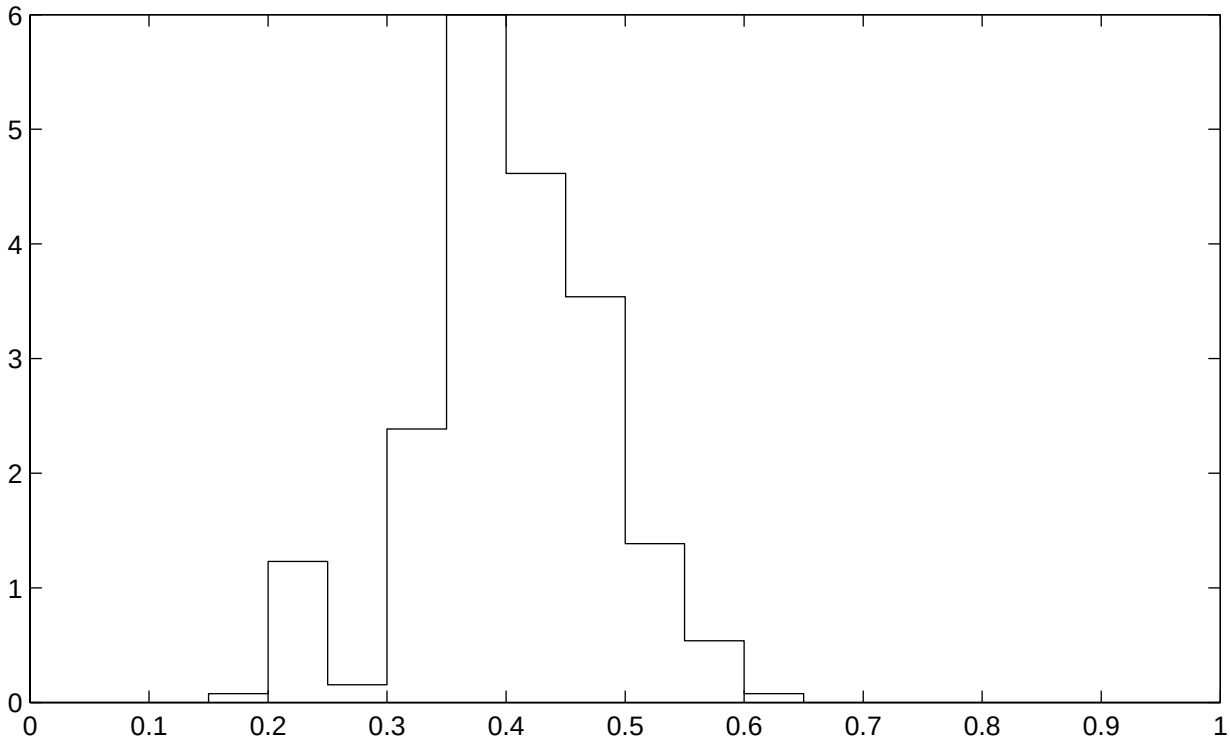

Figure 2: histogram propensity score for treated, exper full sample

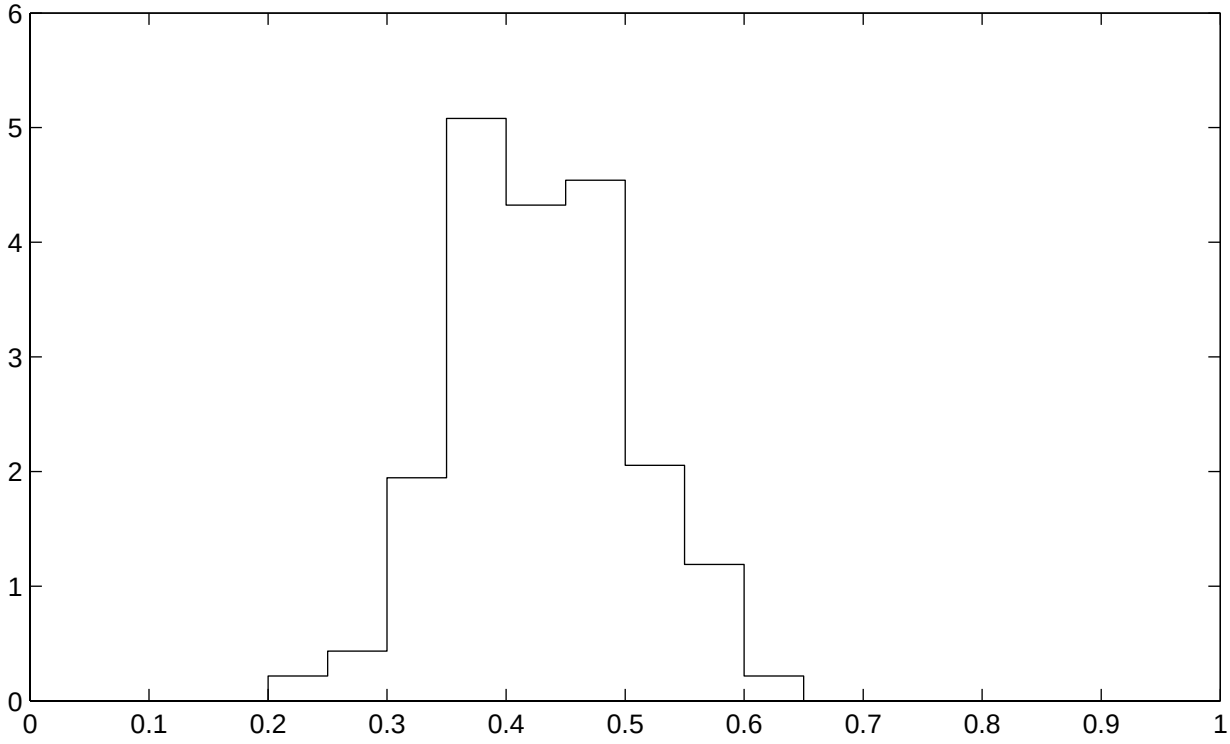

Figure 3: hist p−score for controls, cps full sample

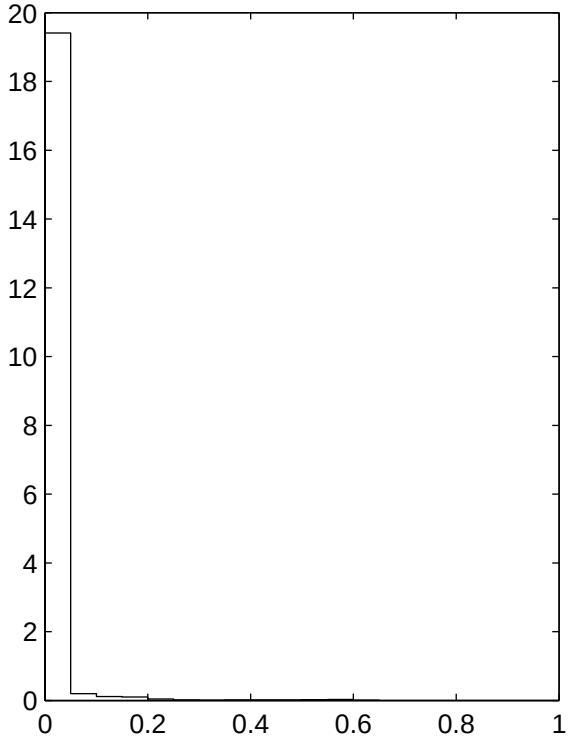

Figure 4: hist p−score for treated, cps full sample

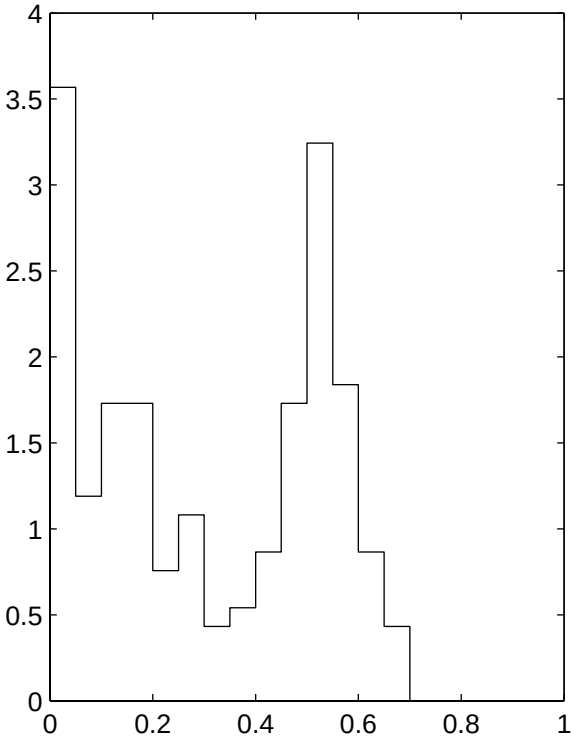

Figure 5: hist p−score for controls, cps selected sample

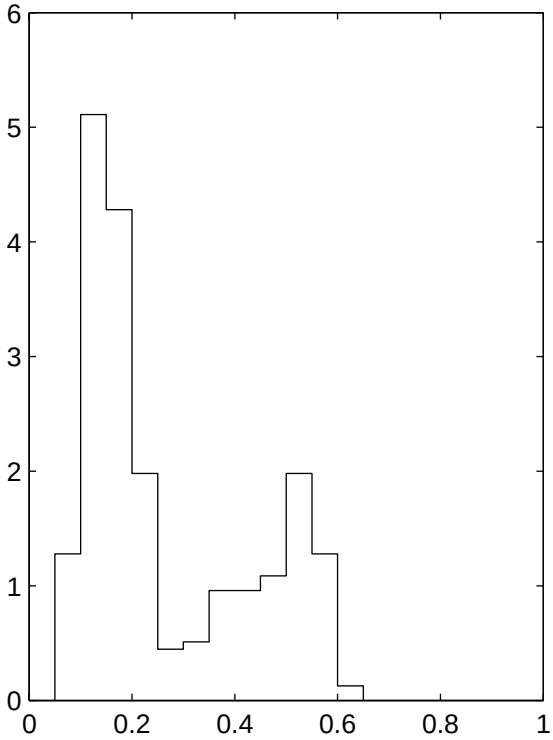

Figure 6: hist p−score for treated, cps selected sample

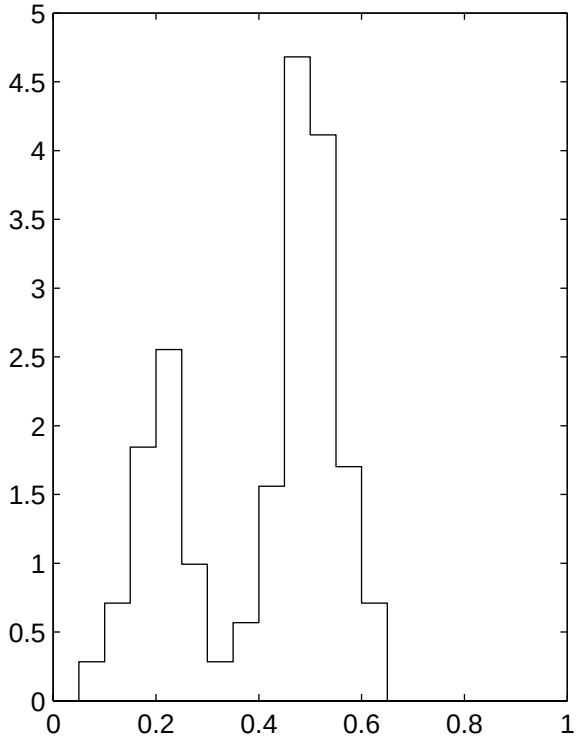

Table 2: Estimates for Lalonde Data with Earnings '75 as Outcome

	Experimental Controls			CPS Comparison Group		
	mean	(s.e.)	t-stat	mean	(s.e.)	t-stat
Simple Dif	0.27	0.30	0.9	-12.12	0.68	-17.8
OLS (parallel)	0.15	0.22	0.7	-1.15	0.36	-3.2
OLS (separate)	0.12	0.22	0.6	-1.11	0.36	-3.1
Propensity Score Weighting	0.15	0.30	0.5	-1.17	0.26	-4.5
Propensity Score Blocking	0.10	0.17	0.6	-2.80	0.56	-5.0
Propensity Score Regression	0.16	0.30	0.5	-1.68	0.79	-2.1
Propensity Score Matching	0.23	0.37	0.6	-1.31	0.46	-2.9
Matching	0.14	0.28	0.5	-1.33	0.41	-3.2
Weighting and Regression	0.15	0.21	0.7	-1.23	0.24	-5.2
Blocking and Regression	0.09	0.15	0.6	-1.30	0.50	-2.6
Matching and Regression	0.06	0.28	0.2	-1.34	0.42	-3.2

For all eleven estimators the estimated effect is close to zero and statistically insignificant at conventional levels. The results suggest that unconfoundedness is plausible. With the CPS comparison group the results are very different. All estimators suggest substantial and statistically significant differences in earnings in 1975 after adjusting for all other covariates, including earnings in 1974. This suggests that relying on the unconfoundedness assumption, in combination with these estimators, is not very credible for this sample.

4.3 Selecting A Subsample

Next we consider the effects of trimming the sample. We use the simple 0.1 rule where we drop observations with the propensity score outside of the interval $[0.1, 0.9]$. Table 3 we report the subsample sizes by treatment status and propensity score block.

Table 3: Sample Sizes for CPS Sample

	$\hat{e}(X_i) < 0.1$	$0.1 \leq \hat{e}(X_i) \leq 0.9$	$0.9 < \hat{e}(X_i)$	All
Controls	15679	313	0	15992
Trainees	44	141	0	185
All	15723	454	0	16177

Dropping observations with a propensity score less than 0.1 leads to discarding most of the controls, 15679 to be precise, leaving only 313 control observations. In addition 44 out of the 185 treated units are dropped. Nevertheless, the improved balance suggests that we obtain more precise estimates for the remaining sample.

Now let us consider the selected CPS sample. First we assess the balance by looking at the summary statistics.

Table 4: Summary Statistics for Selected CPS Sample

	Controls (N=313)		Trainees (N=141)		
	mean	(s.d.)	mean	(s.d.)	diff / sd
Age	26.60	10.97	25.69	7.29	-0.09
Black	0.94	0.23	0.99	0.12	0.21
Education	10.66	2.81	10.26	2.11	-0.15
Hispanic	0.06	0.23	0.01	0.12	-0.21
Married	0.22	0.42	0.13	0.33	-0.24
Earnings '74	1.96	4.08	1.34	3.72	-0.15
Earnings '75	0.57	0.50	0.80	0.40	0.49
Unempl '74	0.92	1.57	0.75	1.48	-0.11
Unempl. '75	0.55	0.50	0.69	0.46	0.28

These suggest that the balance is much improved, with the largest differences now on the order or 0.5 of a standard deviation, where before they difference was as high as 1.7.

Next we estimate the pseudo treatment effect on earnings in 1975.

Table 5: Estimates on Selected CPS Lalonde Data

	Earn '75 Outcome			Earn '78 Outcome		
	mean	(s.e.)	t-stat	mean	(s.e.)	t-stat
Simple Dif	-0.17	0.16	-1.1	1.73	0.68	2.6
OLS (parallel)	-0.09	0.14	-0.7	2.10	0.71	3.0
OLS (separate)	-0.19	0.14	-1.4	2.18	0.72	3.0
Propensity Score Weighting	-0.16	0.15	-1.0	1.86	0.75	2.5
Propensity Score Blocking	-0.25	0.25	-1.0	1.73	1.23	1.4
Propensity Score Regression	-0.07	0.17	-0.4	2.09	0.73	2.9
Propensity Score Matching	-0.01	0.21	-0.1	0.65	1.19	0.5
Matching	-0.10	0.20	-0.5	2.10	1.16	1.8
Weighting and Regression	-0.14	0.14	-1.1	1.96	0.77	2.5
Blocking and Regression	-0.25	0.25	-1.0	1.73	1.22	1.4
Matching and Regression	-0.11	0.19	-0.6	2.23	1.16	1.9

Here we find that all estimators find only small and insignificant effects of the treatment on earnings in 1975. This suggests that for this sample unconfoundedness may well be a reasonable assumption, and that the estimators considered here can lead to credible estimates. Finally we report the estimates for earnings in 1978. Only now do we use the outcome data. Note that with the exclusion of the propensity score matching estimator the estimates are all between 1.73 and 2.23, and thus relatively insensitive to the choice of estimator.

References

Blundell, R. and M. Costa-Dias (2002), "Alternative Approaches to Evaluation in Empirical Microeconomics,"Institute for Fiscal Studies, Cemmap working paper cwp10/02.

Crump, R., V. J. Hotz, V. J., G. Imbens, and O.Mitnik, (2006), "Moving the Goalposts: Addressing Limited Overlap in Estimation of Average Treatment Effects by Changing the Estimand,"unpublished manuscript, Department of Economics, UC Berkeley.

Crump, R., V. J. Hotz, V. J., G. Imbens, and O.Mitnik, (2007), "Nonparametric Tests for Treatment Effect Heterogeneity,"forthcoming, *Review of Economics and Statistics*.

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.

Hirano, K., and J. Porter, (2005), "Asymptotics for Statistical Decision Rules,"Working Paper, Dept of Economics, University of Wisconsin.

Imbens, G. (2000), "The Role of the Propensity Score in Estimating Dose-Response Functions,"*Biometrika*, Vol. 87, No. 3, 706-710.

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 J. Wooldridge., (2007), "Recent Developments in the Econometrics of Program Evaluation,"unpublished manuscript, department of economics, Harvard University.

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

Lechner, M., (2001), "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption,"in Lechner and Pfeiffer (eds.), *Econometric Evaluations of Active Labor Market Policies in Europe*, Heidelberg, Physica.

Robins, J., and Y. Ritov, (1997), "Towards a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-parametric Models,"*Statistics in Medicine* 16, 285-319.

Rosenbaum, P., (1987), "The role of a second control group in an observational study", *Statistical Science*, (with discussion), Vol 2., No. 3, 292–316.

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", emphBiometrika, 70, 41-55.

Rosenbaum, P., and D. Rubin, (1983b), "Assessing the Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome," *Journal of the Royal Statistical Society*, Ser. B, 45, 212-218.

Rubin, D. B., (1978), "Bayesian inference for causal effects: The Role of Randomization", *A*nnals of Statistics, 6:34–58.

Rubin, D., (1990), "Formal Modes of Statistical Inference for Causal Effects", emphJournal of Statistical Planning and Inference, 25, 279-292.