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 pseudo causal effect that is known to equal zero. If based on a statistical 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, and Addressing Lack of, Overlap in Covariate Distributions

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 Assessing Overlap in Covariate Score Distributions

The first method to assess overlap is to report some summary statistics for all covariates. Specifically, it is useful to report the normalized difference in covariate means by treatment status:

$$\text{nor}-\text{dif} = \frac{\overline{X}_1 - \overline{X}_0}{S^2_{X,0} + S^2_{X,1}},$$

where

$$\overline{X}_w = \frac{1}{N_w} \sum_{i:W_i=w} X_i \qquad \text{and} \quad S^2_{X,w} = \frac{1}{N_w - 1} \sum_{i:W_i=w} \left(X_i - \overline{X}_w\right)^2.$$

Note that we do not report the t-statistic for the difference,

$$t = \frac{\overline{X}_1 - \overline{X}_0}{S^2_{X,0}/N_0 + S^2_{X,1}/N_1}.$$

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.

More generally one can 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 Through Matching

Once one determines that there is a lack of overlap one can attempt to construct a sample with more overlap. Here we discuss two methods for doing so. The first is particularly appropriate when the focus is on the average effect for the treated, and there is a relatively large number of controls.

First, the treated observations are ordered, typically by decreasing values of the estimated propensity score. The reason for this is that among units with high values of the propensity score there are relatively more treated than control units, and therefore treated observations with high values of the propensity score are relatively more difficult to match.

Then the first treated unit (e.g., the one with the highest value for the estimated propensity score) is matched to the nearest control unit. Next, the second treated unit is matched to the nearest control unit, excluding the control unit that was used as a match for the first treated unit. Matching without replacement all treated units in this manner leads to a sample of $2 \cdot N_1$ units, (where N_1 is the size of the original treated subsample), half of them treated and half of them control units. Note that the matching is not necessarily used here as the final analysis. We do not propose to estimate the average treatment effect for the treated by averaging the differences within the pairs. Instead, this is intended as a preliminary analysis, with the goal being the construction of a sample with more overlap. Given a more balanced sample, one can use methods discussed in these notes for estimating the average effect of the treatment, including regression, propensity score methods, or matching. Using those methods on the balanced sample is likely to reduce bias relative to using the simple difference in averages by treatment status.

3.3 Selecting a Sample with Overlap Through Trimming

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

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

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. Algorithm for Estimating the Propensity Score and Stratification

Many of the estimators discussed in this and the previous lecture are rely on estimators for the propensity score. Here I briefly describe one way of selecting a specification for the

propensity score. This particular procedure is a step-wise method, where increasingly flexible specifications are selected until the specification is deemed adequate. This is not the only way of doing, this, and in fact there are many such methods out there, some of which are undoubtedly more effective. The main point is that the common practice of including the full vector of covariates linearly, and not include any second order terms is not likely to be effective.

The algorithm starts with a K-dimensional vector of covariates X_i (these may already contain functions of the original covariates). The algorithm will selection a subset of the covariates to be included linearly, and based on that subset also select a number of second order terms (both quadratic terms and interactions).

The algorithm starts with a logistic model with no covariates. Next, logistic regression models are estimated with each of the covariates included separately. The covariate that improves the log likelihood function the most is included, as long as the increase in the log likelihood function is above some threshold $t_{\rm lin}$. Next, we select among the $K-1$ remaining covariates the one that improves the logistic model with the single covariate the most, again based on the increase in the log likelihood function. We repeat this till no additional covariate improves the log likelihood function by at least $t_{\rm lin}$. Suppose this leads to selecting $0 \leq K_{\rm lin} \leq K$ covariates out of the original set of K covariates.

In the second part we select among the $K_{\rm lin} \times (K_{\rm lin}+1)/2$ second order terms based on these $K_{\rm lin}$ covariates. Similar to the way we selected the linear terms, we keep adding second order terms, till no additional second order term improves the log likelihood by more than $t_{\rm qua}$. The tuning constants used below are $t_{\rm lin} = 0.5$ and $t_{\rm qua} = 1.35$, based on cutoffs for likelihood ratio test statistics (equal to twice the increaese in the log likelihood function) of 1 and 2.71, the latter corresponding to a 10% level test.

One may modify this algorithm by selecting a subset of the covariates to be included irrespective of the correlations with the treatment. In the analyses below, we selected the last pre-program earnings and the indicator for those earnings being positive to be included

in this way, prior to selecting further covariates.

Some of the estimators discussed below also require an algorithm for choosing the number and boundaries for the blocks. Here is the algorithm used below. We start with a single stratum. The option is to split the stratum in two equal parts, with the new boundary point the median of the values of the estimated propensity score in the old stratum. The old stratum will be split if three conditions are satisfied. First, the t-statistic for testing equality of the average estimated propensity score among treated and controls is at least 1.96, the number of treated and control observations in both new strata is at least 3, and the number of obsersvations in each block is at least 3 plus the dimension of the covariate vector X_i. We then keep splitting the strata in the middle, until none of the strata satisfies the criteria for further division.

5. An Illustration Based on 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). These data are available on Dehejia's website (reference). 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 illustration we use three subsamples, the (experimental) trainees, the experimental controls, and a CPS comparison group. In both cases we focus on estimating the average effect of the treatment for the treated.

In the next three subsections we do the design part of the analysis. Without using the outcome data we first assess the overlap in covariate distributions, then assess whether strong ignorability has some credibility and finally create a matched sample and assess these issues there.

5.1 Summary Statistics

First we give some summary statistics

Table 1: Summary Statistics for Experimental Sample

	Trainees (N=260)		Controls (N=185)			CPS (N=15,992)		
	mean	(s.d.)	mean	(s.d.)	nor-dif	mean	(s.d.)	nor-dif
	260.00	0.00	185.00	0.00	0.00	0.00	0.00	0.00
Black	0.84	0.36	0.83	0.38	0.03	0.07	0.26	1.72
Hispanic	0.06	0.24	0.11	0.31	0.12	0.07	0.26	0.04
Age	25.82	7.16	25.05	7.06	0.08	33.23	11.05	0.56
Married	0.19	0.39	0.15	0.36	0.07	0.71	0.45	0.87
No Degree	0.71	0.46	0.83	0.37	0.21	0.30	0.46	0.64
Education	10.35	2.01	10.09	1.61	0.10	12.03	2.87	0.48
Earnings '74	2.10	4.89	2.11	5.69	0.00	14.02	9.57	1.11
Unempl '74	0.71	0.46	0.75	0.43	0.07	0.12	0.32	1.05
Earnings '75	1.53	3.22	1.27	3.10	0.06	13.65	9.27	1.23
Unempl. '75	0.60	0.49	0.68	0.47	0.13	0.11	0.31	0.84

In this table we report averages and standard deviations for the three subsamples. In addition we report for both the trainee/experimental-control and for the trainee/CPS-comparison-group pairs the normalized difference in average covariate values by treatment status, normalized by the standard deviation of these covariates:

$$\frac{\overline{X}_1 - \overline{X}_0}{\sqrt{S^2_{X,0} + S^2_{X,1}}}.$$

Again, it is not the statistical significance of this difference we are interested in, as much as the degree of difficulty of the statistical problem of adjusting for these differences. In Table 1 we see that in the experimental data set the difference in average age between treated and controls is 0.08 standard deviations. In the nonexperimental comparison the difference in age is 0.56 standard deviations.

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 0.21 standard deviations. In contrast, with the CPS comparison group the differences between the averages are up

to 1.23 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.

The results so far already strongly indicate that simple analyses such as least squares regression are unlikely to lead to credible estimates of the average causal effects of interest.

5.2 Assessing Unconfoundedness

First we use the experimental data. We analyze the data as if earnings in 1975 (Earn '75) is the pseudo outcome. This is in fact a covariate, and so it cannot be affected by the treatment, and we are looking for estimates that are substantially close to zero, and statistically indistinguishable from zero. Table 2 reports the results for nine estimators.

1. The first is the simple difference in average outcomes:

$$\hat{\tau} = \overline{Y}_1 - \overline{Y}_0.$$

2. The second estimator is based on least squares regression using all ten covariates:

$$Y_i = \alpha + \tau \cdot W_i + \beta' X_i + \varepsilon_i.$$

3. The third estimator is based on least squares regression using all ten covariates and their interaction with the treatment indicator:

$$Y_i = \alpha + \tau \cdot W_i + \beta' X_i + \gamma'(X_i - \overline{X}_1) \cdot W_i + \varepsilon_i.$$

The interaction is based on deviations from the average covariate values for the treated in order for the least squares estimator for τ to estimate the average effect on the treated.

4. The fourth estimator uses the estimated propensity score to weight the observations:

$$\hat{\tau} = \frac{1}{N_1} \cdot \sum_{i:W_i=1} Y_i - \sum_{i:W_i=0} Y_i \cdot \frac{\hat{e}(X_i)}{1-\hat{e}(X_i)} \Big/ \sum_{j:W_j=0} \frac{\hat{e}(X_j)}{1-\hat{e}(X_j)}.$$

The weights here are modified from those discussed previously to take account of the focus on the average effect for the treated.

5. Here the propensity score is used to create strata. Within the J strata the average effect is estimated as the difference in average outcomes between treated and controls, and the within-stratum estimates are averaged, weighted by the number of treated units in each strata. The number of strata is choosen in a data-dependent way, as described in Section 4.

6. Here all the treated observations are matched to the closets control, with replacement. The matching is on all covariates, weighted by the diagonal matrix with the inverse of the variances on the diagonal.

7. The seventh estimator is based on weighted least squares regression of the regression function

$$Y_i = \alpha + \tau \cdot W_i + \beta' X_i + \varepsilon_i,$$

with weights

$$\lambda_i = W_i + (1-W_i) \cdot \frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}.$$

(The fourth estimator is a special case of this where β is set equal to zero.)

8. The eight estimator is based on the same blocks as the fifth estimator, but now within blocks linear regression is used to estimate the average effect.

TABLE 2: ESTIMATES FOR LALONDE DATA WITH EARNINGS '75 AS OUTCOME

	Experimental Controls			CPS Comparison Group		
	est	(s.e.)	t-stat	est	(s.e.)	t-stat
Simple Dif	0.27	0.31	0.87	-12.12	0.25	-48.91
OLS (parallel)	0.22	0.22	1.02	-1.13	0.36	-3.17
OLS (separate)	0.17	0.22	0.74	-1.10	0.36	-3.07
Weighting	0.29	0.30	0.96	-1.56	0.26	-5.99
Blocking	0.26	0.32	0.83	-12.12	0.25	-48.91
Matching	0.11	0.25	0.44	-1.32	0.34	-3.87
Weighting and Regression	0.21	0.22	0.99	-1.58	0.23	-6.83
Blocking and Regression	0.12	0.21	0.59	-1.13	0.21	-5.42
Matching and Regression	-0.01	0.25	-0.02	-1.34	0.34	-3.96

9. The ninth estimator uses the same matching as the sixth estimator. Then linear regression is used on the 185 matches to estimate

$$Y_i = \alpha + \beta' X_i + \varepsilon_i,$$

and the estimated regression coefficients $\hat{\beta}$ are used to adjust the matched outcomes based on the Abadie-Imbens estimator.

For all nine estimators the estimated effect is close to zero and statistically insignificant at conventional levels. The results suggest that unconfoundedness is plausible for the experimental data set. This is not surprising, as the randomization implies unconfoundedness.

With the CPS comparison group the results are very different. All nine 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 particular estimators, is not very credible for this sample. This is not surprising, because the treated and control samples

are so far apart, as measured by the normalized differences, that the estimates were very unlikely to be robust.

5.3 Creating a Matched Sample

Now let us consider the matched CPS sample. Matching is done on here the estimated propensity score, without replacement, for all the treated observations, starting with the treated unit with the highest value for the estimated propensity score. This leads to a matched sample with 185 treated (as before), and 185 controls. First we assess the balance by looking at the summary statistics.

Table 4: Summary Statistics for Matched CPS Sample

	Trainees (N=185)		Controls (N=185)		
	mean	(s.d.)	mean	(s.d.)	nor-dif
Black	0.84	0.36	0.85	0.35	-0.02
Hispanic	0.06	0.24	0.06	0.25	-0.02
Age	25.82	7.16	25.88	7.65	-0.01
Married	0.19	0.39	0.25	0.43	-0.10
No Degree	0.71	0.46	0.57	0.50	0.20
Education	10.35	2.01	10.91	2.93	-0.16
Earnings '74	2.10	4.89	2.81	5.61	-0.10
Unempl '74	0.71	0.46	0.66	0.47	0.07
Earnings '75	1.53	3.22	1.82	3.79	-0.06
Unempl. '75	0.60	0.49	0.50	0.50	0.14

These suggest that the balance is much improved, with the largest differences now on the order or 0.20 of a standard deviation, where before they difference was as high as 1.12. Now the normalized differences are comparable to those in the experimental sample.

Figures 5 and 6 present histograms of the propensity score for this matched sample. Note that we re-estimate the propensity score for this sample. If the matching had been perfect, the estimated propensity score would be equal to 0.5 for all units. It is not, and there is still considerable variation in the propensity score, but not to the extent that simple analyses

could not adjust for the covariate differences between the treatment and control samples.

These normalized differences suggest that given unconfoundedness, the matched sample is well balanced, and likely to lead to robust estimates. They do not directly reflect, however, on the question whether unconfoundedness itself is plausible. In order to address that, we return to the analysis with earnings in 1975 as the pseudo outcome. Again we report estimates for nine estimators.

Table 5: Estimates on Selected CPS Lalonde Data

	Earn '75 Outcome			Earn '78 Outcome		
	est	(s.e.)	t-stat	est	(s.e.)	t-stat
Simple Dif	-0.29	0.37	-0.79	0.87	0.80	1.08
OLS (parallel)	0.01	0.26	0.02	1.40	0.77	1.81
OLS (separate)	0.05	0.26	0.20	1.26	0.77	1.64
Weighting	-0.01	0.37	-0.02	1.20	0.80	1.49
Blocking	-0.04	0.37	-0.10	1.16	0.82	1.41
Matching	-0.10	0.37	-0.28	1.53	0.95	1.61
Weighting and Regression	0.02	0.25	0.07	1.32	0.78	1.69
Blocking and Regression	0.00	0.25	0.01	1.77	0.76	2.33
Matching and Regression	-0.22	0.37	-0.60	1.41	0.95	1.49

Here we find that all nine estimators find only small (relative to the experimental benchmark of about 1.7, in thousands of dollars), and statistically 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 actually use the outcome data. Note that with the exclusion of the simple difference $\overline{Y}_1 - \overline{Y}_0$, the estimates are all between 1.16 and 1,77, and thus relatively insensitive to the choice of estimator. The benchmark estimate from the experimental sample is $\overline{Y}_1 - \overline{Y}_0 = 1.79$, very similar to these non-experimental estimates.

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., (2008), "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.

Fig 1: Histogram Propensity Score for Controls, Experimental Sample

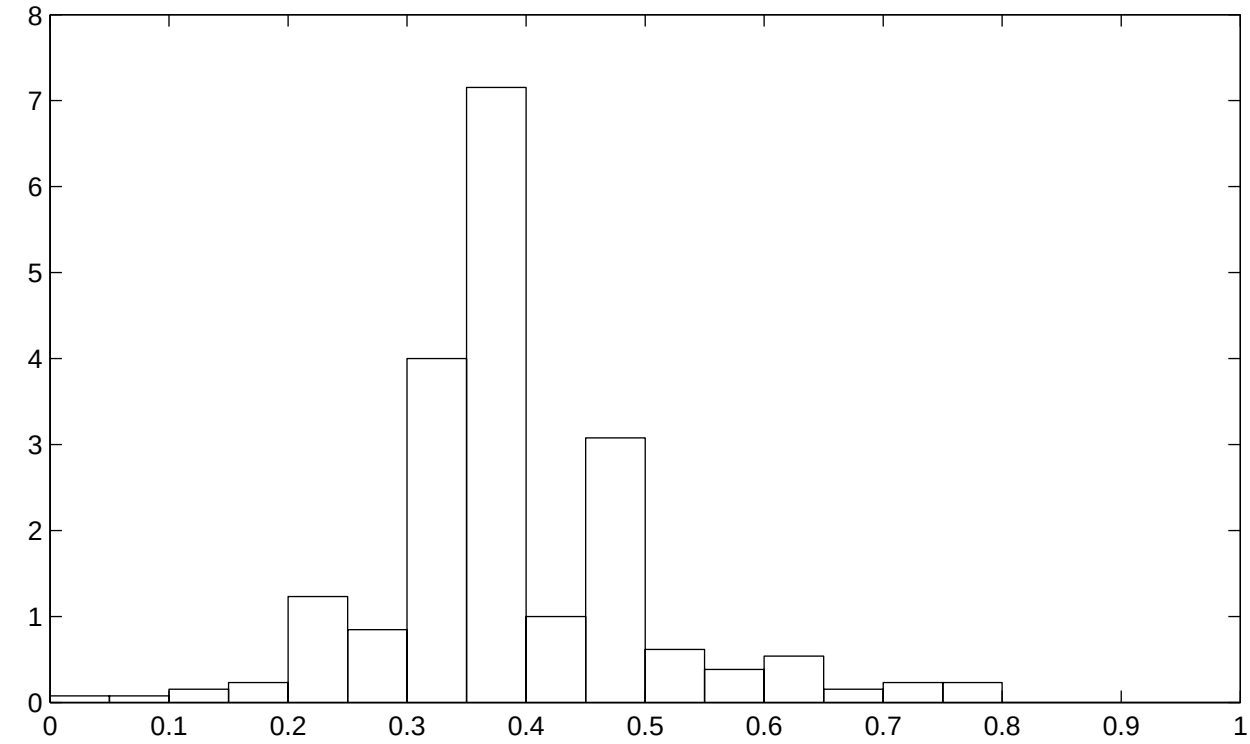

Fig 2: Histogram Propensity Score for Trainees, Experimental Sample

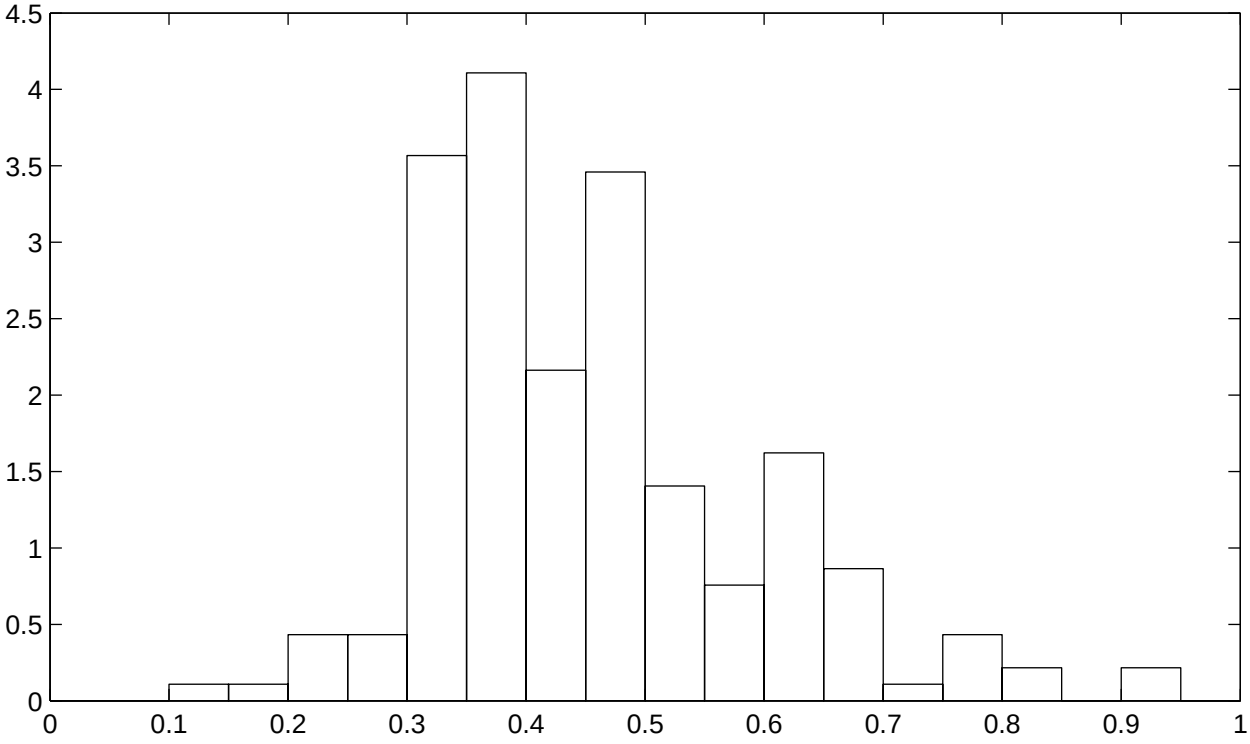

Fig 3: Histogram Propensity Score for Controls, Full CPS Sample

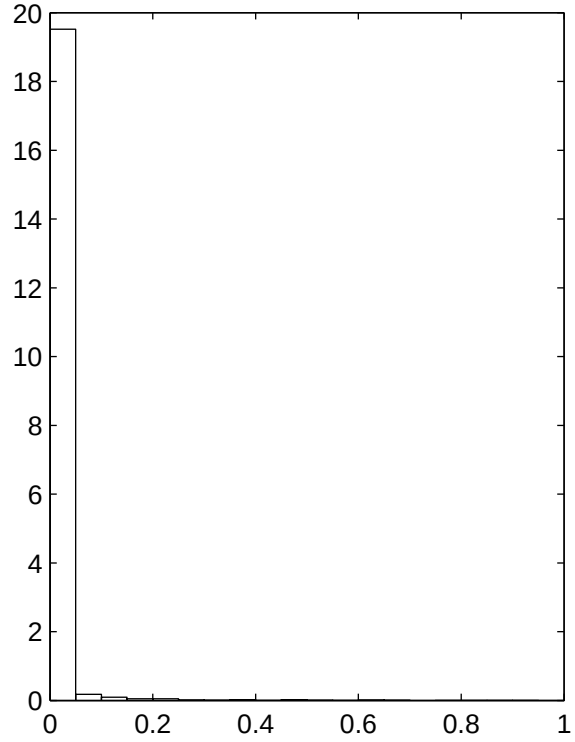

Fig 3: Histogram Propensity Score for Controls, Matched CPS Sample

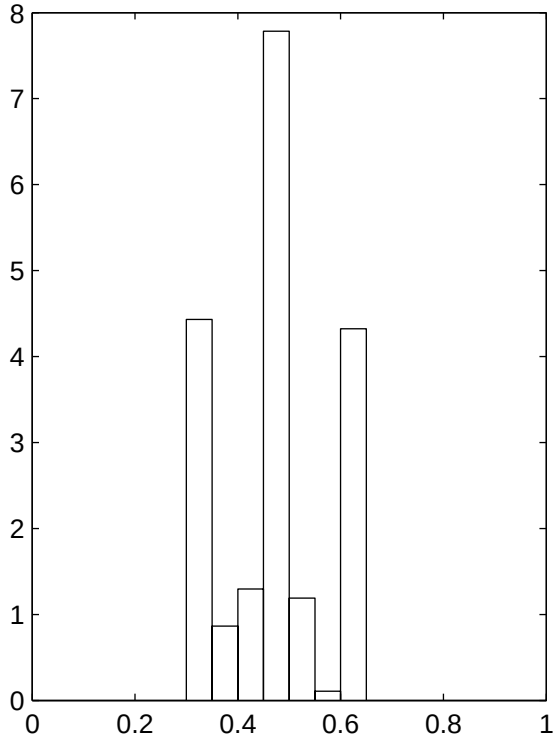

Fig 4: Histogram Propensity Score for Trainees, Full CPS Sample

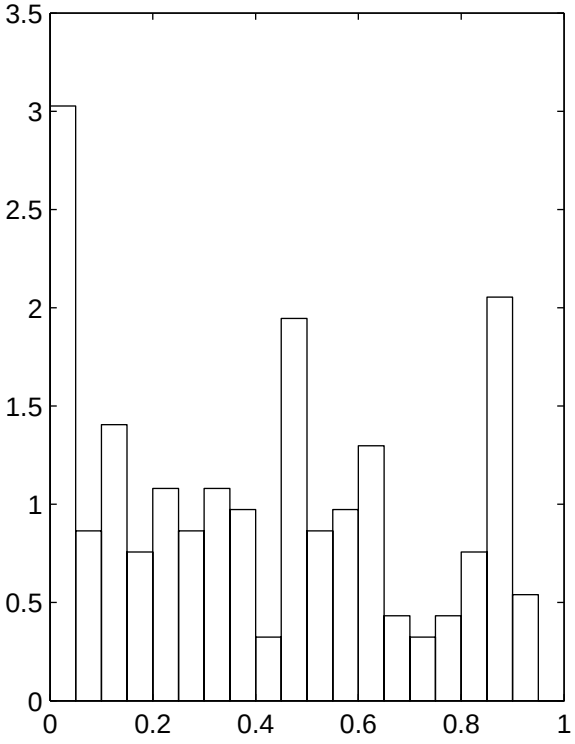

Fig 4: Histogram Propensity Score for Trainees, Matched CPS Sample