"A Course in Applied Econometrics"

Lecture 2

Estimation of Average Treatment Effects

Under Unconfoundedness, Part II

Guido Imbens

IRP Lectures, UW Madison, August 2008

Outline

1. Assessing Unconfoundedness (not testable)
2. Overlap
3. Illustration based on Lalonde Data

5.I Assessing Unconfoundedness: Multiple Control Groups

Suppose we have 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

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

Suppose we extend the unconfoundedness assumption to independence of the potential outcomes and the three-valued group indicator given covariates,

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

Now 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.

5.II Assessing Unconfoundedness: Estimate Effects on Pseudo Outcomes

Partition the covariate vector into $X_i = (X_i^p, X_i^r)$, X_i^p scalar.

Unconfoundedness assumes

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \mid (X_i^p, X_i^r)$$

Suppose we are willing to assume X_i^r is sufficient:

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \mid X_i^r$$

and suppose X_i^p is a good proxy for $Y_i(0)$, then we can test

$$X_i^p \perp\!\!\!\perp W_i \mid X_i^r$$

Most useful implementations with X_i^p a lagged outcome.

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 $X_i = (X_i^p, X_i^r)$, with $X_i^p = Y_{i,-1}$ and $X_i^r = (Y_{i,-2}, \ldots, Y_{i,-T}, Z_i)$. Outcome is $Y_i = Y_{i,0}$.

Now consider the following two assumptions. The first is unconfoundedness given only $T-1$ lags of the outcome:

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

Then, under stationarity it seems reasonable to expect Then it follows that

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

which is testable.

6.I Assessing Overlap

The first method to detect lack of overlap is to look at summary statistics for the covariates by treatment group.

Most important here is the normalized difference in covariates:

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

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

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

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

The t-statistic partly reflects the sample size. Given the normalized difference, 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.

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 $\mu_0(x) = \mathbb{E}[Y_i|W_i = 0, X_i = x]$ and using this estimated regression function to predict the (missing) control outcomes for the treated units.

With a large difference between the two groups, linear regression is going to rely heavily on extrapolation, and thus will be sensitive to the exact functional form.

Assessing Overlap by Inspecting the Propensity Score Distribution

The second method for assessing overlap is more directly focused on the overlap assumption.

It involves inspecting the marginal distribution of the propensity score in both treatment groups.

Any difference in covariate distribution shows up in differences in the average propensity score between the two groups.

Moreoever, any area of non-overlap shows up in zero or one values for the propensity score.

6.II Selecting a Subsample with Overlap: Matching

Appropriate when the focus is on the average effect for treated, $\mathbb{E}[Y_i(1)-Y_i(0)|W_i=1]$, and when there is a relatively large pool of potential controls.

Order treated units by estimated propensity score, highest first.

Match highest propensity score treated unit to closest control on estimated propensity score, without replacement.

Only to create balanced sample, not as final analysis.

6.III Selecting a Subsample with Overlap: Trimming

Define average effects for subsamples $\mathbb{A}$:

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

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 .

The optimal set 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

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

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.

7. Applic. to Lalonde Data (Dehejia-Wahba Sample)

Data on job training program, first used by Lalonde (1986), See also Heckman and Hotz (1989), Dehejia and Wahba (1999).

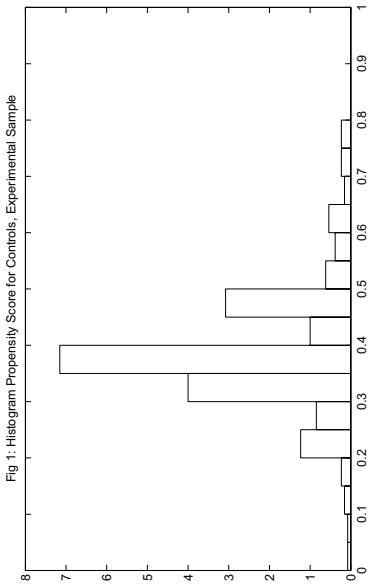
Small experimental evaluation, 185 trainees, 260 controls, group of very disadvantaged in labor market.

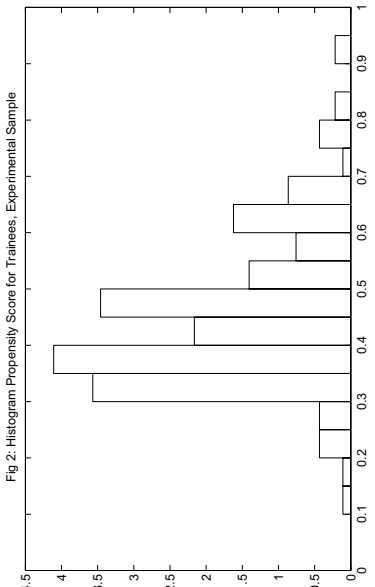
Large, non-experimental comparison group from CPS (15,992 observations). Very different in distribution of covariates.

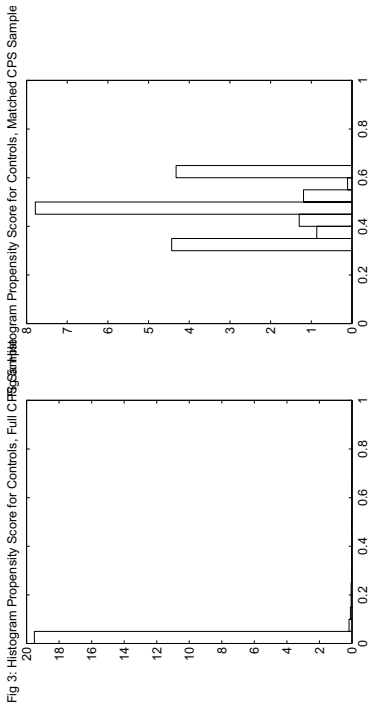
How well do the non-experimental results replicate the experimental ones? Is non-experimental analysis credible? Would we have known whether it was credible without experiments results?

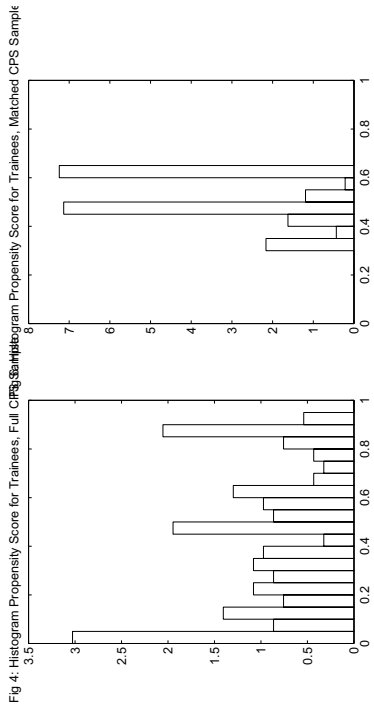
Table 1: Summary Statistics for Lalonde Data

	Trainees (N=260)		Controls (N=185)			CPS (N=15,992)		
	mean	(s.d.)	mean	(s.d.)	n-dif	mean	(s.d.)	n-dif
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	25.8	7.2	25.1	7.1	0.08	33.2	11.1	0.56
Married	0.19	0.39	0.15	0.36	0.07	0.71	0.45	0.87
No Deg	0.71	0.46	0.83	0.37	0.21	0.30	0.46	0.64
Educ	10.4	2.0	10.1	1.6	0.10	12.0	2.9	0.48
Earn '74	2.10	4.89	2.11	5.69	0.00	14.02	9.57	1.11
U '74	0.71	0.46	0.75	0.43	0.07	0.12	0.32	1.05
Earn '75	1.53	3.22	1.27	3.10	0.06	13.65	9.27	1.23
U '75	0.60	0.49	0.68	0.47	0.13	0.11	0.31	0.84









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.

Next, let us assess unconfoundedness in this sample using earnings in 1975 as the pseudo outcome.

We report results for 9 different estimators, including the simple difference, parallel and separate least squares regressions, weighting and blocking on the propensity score, and matching, with the last three also combined with regression.

Both for experimental control group and for cps comparison group.

Specification for propensity score, and block choice are based on algorithm, see notes for details.

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
Weight and Regr	0.21	0.22	0.99	-1.58	0.23	-6.83
Block and Regr	0.12	0.21	0.59	-1.13	0.21	-5.42
Match and Regr	-0.01	0.25	-0.02	-1.34	0.34	-3.96

With the cps comparison group, results are discouraging. Consistently find big "effects" on earnings in 1975, with point estimates varying widely.

The sensitivity is not surprising given substantial differences in covariate distributions.

Next, create a matched sample to improve balance.

Order treated observations on estimated propensity score.

Starting with the highest propensity score, match each treated observation to the closest control, without replacement. Match on the propensity score.

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

In the matched sample the normalized differences are comparable to those in the experimental sample.

Now we revisit the analysis on earnings in 1975, and also carry out analysis on earnings in 1978 (the actual outcome).

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
Weight and Regr	0.02	0.25	0.07	1.32	0.78	1.69
Block and Regr	0.00	0.25	0.01	1.77	0.76	2.33
Match and Regr	-0.22	0.37	-0.60	1.41	0.95	1.49

Now results are consistently small and statistically insignificant for earnigns in 1975, so unconfoundedness seems reasonable, and analyses potentially credible.

Estimates for earnings in 1978 are robust accross all nine estimators, with the exception of the simple difference in average outcomes by treatment status.

Estimates are consistent with experimental estimates (1.77).

Conclusion

Important to assess and address lack of overlap.

In reasonably balanced samples choice of estimator is less important.

Combining regression and matching or propensity score blocking is preferred method for robustness properties.