ARE 213 **Applied Econometrics**

UC Berkeley Department of Agricultural and Resource Economics

Selection on Observables Designs:

Part III, Matching, Dimensionality Reduction, the Propensity Score, and a Reality Check

This set of lecture notes discusses matching under the assumption of unconfoundedness (i.e., selection on observables). The idea behind matching is very simple. If $Y_i(0), Y_i(1) \perp D_i|X_i$, then we can estimate $\tau(x) = E[Y_i(1) - Y_i(0)|X_i = x]$ because the treatment is effectively randomly assigned after conditioning on X_i. In the last lecture, we learned a variety of techniques for nonparametrically estimating conditional expectations. These techniques — in particular, kernel regression — have a close linkage to a nonparametric technique known as "matching." The idea behind matching is to compare treated units ($D_i = 1$) to control units ($D_i = 0$) that have similar values of X_i. This guarantees that every treatment-control comparison is performed on units with identical (or close to identical) values of X_i, so we are literally conditioning on $X_i = x$. Given the selection on observables assumption, we know that D_i is as good as randomly assigned after conditioning on X_i, so we should get causal estimates.

1 Matching

See end of 02B notes

For every treated unit, i.e. every unit with $D_i = 1$, the goal of the matching estimator is to find a comparison unit among the controls that has similar values of observable characteristics X_i. It is important to note, however, that this "comparison unit" need not be a single unit — rather, it can be a composite (i.e., a weighted average) of several different control units that have similar values of X_i.[1] Assume that there are N_T treated units and N_C control units. Define N_T sets of weights, with N_C weights in each set: $w_i(j)$ $(i = 1, ..., N_T,\ j = 1, ..., N_C)$.

[1] There are obvious efficiency gains to doing this, particularly if there are more control units than treated units.

For each set of weights, let $\sum_j w_i(j) = 1$. Then the generic matching estimator is:

$$\hat{\tau}_M = \frac{1}{N_T} \sum_{i \in \{D=1\}} [y_i - \sum_{j \in \{D=0\}} w_i(j) y_j]$$

In other words, we are simply computing the average difference between the treated units and the composite comparison units. The key to this estimator is how you calculate the weights used to construct the composite comparison units, $w_i(j)$. For example, you could set $w_i(j) = \frac{1}{N_C}$. In that case, $\sum_j w_i(j) y_i$ simply equals the control group mean, $\bar{y}_C$, for all i, and $\hat{\tau}_M$ is just the difference in means between the treated and control groups. Obviously that is not very exciting estimator and it does not solve the selection on observables problems.

In general, we want to choose $w_i(j)$ so that it measures the "nearness" of X_j to X_i — $w_i(j)$ is what I will call the distance measure. If X is discrete, then in principle you could choose $w_i(j)$ such that it equals one if $X_i = X_j$ and zero otherwise (you would of course have to rescale $w_i(j)$ by $\sum_j w_i(j)$ so that it summed to one for each i). If X is continuous, then that particular measure won't work, but there are several common choices of distance measures. The most popular is probably "nearest-neighbor" matching. With nearest-neighbor matching, $w_i(j)$ is a function of the Euclidean distance between X_i and X_j. Specifically, $w_i(j)$ equals one for the control unit with the closest X_j to X_i — where closeness is measured by Euclidean distance ($(X_i - X_j)'(X_i - X_j)$) — and zero otherwise. Thus, $w_i(j)$ selects the "nearest (control) neighbor" j to treated unit i, and $\hat{\tau}_M$ computes the mean difference between each treated unit and its nearest control neighbor. This procedure should produce valid causal estimates under the selection on observables assumption, *assuming that there is sufficient overlap between the treated and control groups* (we will return to this issue shortly).

Of course, the choice of units for each component of X_i is arbitrary, so it may not make sense to weight each component equally when computing the distance between two points, as the Euclidean distance metric does. A popular alternative to the Euclidean metric is thus the Mahalanobis distance metric, $(X_i - X_j)'\Sigma_x^{-1}(X_i - X_j)$, where Σ_x is the covariance

matrix of X — note the parallel to GLS. Effectively, you are normalizing the components of $(X_i - X_j)$ by the root of the inverse covariance matrix.[2]

If there are substantially more control than treated units, then it seems somewhat inefficient to choose only N_T nearest neighbors, when $N_T < N_C$.[3] Why not use more of the information in the control group? This is the idea behind "kernel matching." With kernel matching, we set $w_i(j) = \frac{K(X_i - X_j)}{\sum_j K(X_i - X_j)}$, where $K(.)$ is one of the kernel functions that we discussed in the previous lecture, such as the triangle kernel or the Epanechnikov kernel. We could also include a bandwidth term, h, to control the kernel's "reach." The advantage of kernel matching over nearest-neighbor matching is that, if there are several control units with X_j's in the neighborhood of X_i, then it makes sense to take the average outcome for all of those control units rather than just to pick a single control unit — the former should be more efficient than the latter. Kernel matching allows us to do this kind of averaging and to create composite comparison control units for each match. The kernel weighting function puts the most weight on the closest control unit and less weight on further control units.

The link to kernel regression should now be clear: effectively, we are creating each comparison unit by estimating a kernel regression at X_i using the sample of controls. We can increase the number of controls used to create the comparison unit by increasing the reach of the kernel (i.e., raising h), and we know that doing so reduces the variance but increases the bias (you're extrapolating from units with X_j far from X_i).[4] However, the topic of kernel regression also brings up memories of our old nemesis, the Curse of Dimensionality.

The short story is, the Curse of Dimensionality strikes back with a vengeance in the case of matching. As with kernel regression, increasing the dimension of X increases the sparsity of the data. The more variables you have in X, the less likely you are to find a comparison

[2]The odd thing about Mahalanobis distance is that, depending on the covariance structure, you can end up in situations in which (10, 10) is closer to (0, 0) than (8, 2). I believe this is because the weight in the inverted covariance matrix can become negative.

[3]In fact, we will almost always choose less than N_T neighbors, because some comparison units will be picked more than once.

[4]While the variance reduction holds at any given point X_i, it's not as clear for the mean of all of the composite comparison units, because there's a lot of duplication in observations between composite comparison units as you increase the bandwidth.

control unit lying close to any given treatment unit — there are simply too many dimensions to match along. What can be done? Enter propensity score matching.

2 Methods of the Propensity Score

Assume that we have unconfoundedness: $(Y_i(0), Y_i(1)) \perp D_i|X_i$. Also assume that the overlap assumption holds: $0 < P(D_i = 1|X_i) < 1$. Combining these two assumptions, we say that the treatment assignment is "strongly ignorable." We know that if we condition on X_i, then we can get a consistent estimate of ATE by simply comparing the difference in means between treated and control units. In practice, however, it is hard to condition on X_i if X_i is high dimensional. Note that this is effectively because the overlap assumption fails in finite samples — for most observations, it is impossible to find a comparison unit with the opposite treatment assignment and the same value of X.

An important result is that, under strongly ignorable treatment assignment, it is sufficient to condition simply on $p(X_i) = E[D_i|X_i]$, also known as the *propensity score*. Formally, if we assume $(Y_i(0), Y_i(1)) \perp D_i|X_i$, then

$(Y_i(0), Y_i(1)) \perp D_i \mid p(X_i)$

Proof

We will show that $P(D_i = 1 \mid Y_i(0), Y_i(1), p(X_i)) = P(D_i = 1 \mid p(X_i)) = p(X_i)$. This implies independence of D_i and $(Y_i(0), Y_i(1))$ after conditioning on $p(X_i)$.

$$P(D_i = 1 \mid Y_i(0), Y_i(1), p(X_i)) = E[D_i \mid Y_i(0), Y_i(1), p(X_i)]$$

$$= E[E[D_i \mid Y_i(0), Y_i(1), p(X_i), X_i] \mid Y_i(0), Y_i(1), p(X_i)]$$

$$= E[E[D_i \mid Y_i(0), Y_i(1), X_i] \mid Y_i(0), Y_i(1), p(X_i)]$$

cond. indep on X

$$= E[E[D_i \mid X_i] \mid Y_i(0), Y_i(1), p(X_i)]$$

$$= E[p(X_i) \mid Y_i(0), Y_i(1), p(X_i)]$$

$$= p(X_i)$$

For completeness, note that:

$$P(D_i = 1 \mid p(X_i)) = E[D_i \mid p(X_i)]$$

$$= E[E[D_i \mid p(X_i), X_i] \mid p(X_i)]$$

$$= E[E[D_i \mid X_i] \mid p(X_i)]$$

$$= E[p(X_i) \mid p(X_i)]$$

$$= p(X_i)$$

So $P(D_i = 1 \mid Y_i(0), Y_i(1), p(X_i)) = p(X_i) = P(D_i = 1 \mid p(X_i))$. Since D_i is binary, this implies independence of D_i and $(Y_i(0), Y_i(1))$ after conditioning on $p(X_i)$. In other words, it is sufficient to merely condition on $p(X_i)$ — we don't have to condition on X_i.

Why is it sufficient to condition on the propensity score? Our concern is that units selecting into treatment differ in some meaningful way from units that do not select into treatment, and that this difference is consistently related to the probability of entering treatment. If, however, we only compare units with the exact same probability of treatment, then it is impossible for the differences to be consistently related to the probability of treatment.[5] After conditioning on the propensity score, the units are "as good as randomly assigned."

[5] If they were, then we would be using them to estimate the propensity score, or so our unconfoundedness assumption claims.

2.1 Estimating the Propensity Score

Before you can condition on the propensity score, $p(X_i) = E[D_i|X_i]$, you have to estimate it. There are several ways to do this — it's not clear that one method is uniformly superior, so your choice may be context dependent. The easiest way — suggested by Rubin and Rosenbaum (1983) — is to use a flexible logit specification (flexible in the sense that there are interactions between the various components of X_i). Alternatively, one could use the kernel regression methods that we discussed in the previous lecture. Finally, Hirano, Imbens, and Ridder (2003) suggest a variant on a series estimator — the series logit estimator. This is similar to Rubin and Rosenbaum, but it also contains higher order terms — how high the order becomes is a function of the sample size and the dimension of X.[6]

2.2 Regression Adjusting on the Propensity Score

Once you've estimated the propensity score, the next question is what to do with it, i.e. how to condition on it. One obvious candidate is to simply include it as a regressor, i.e. run the regression:

$$Y_i = \alpha + \delta D_i + \beta p(X_i) + u_i$$

As we saw in an earlier lecture, controlling for the conditional expectation of D — which is equivalent to controlling for the propensity score when D is binary — is sufficient to generate consistent estimates of δ under unconfoundedness and the assumption of a *constant* (i.e., homogeneous) treatment effect. More generally, we may want to interact the propensity score with the treatment indicator if we believe that the treatment effects may be heterogenous (and that the heterogeneity may vary with X in some consistent manner). To see how this type of heterogeneity might affect our estimator, consider the following model:

$$Y_i(0) = \alpha + \beta X_i + u_i$$

[6]The rule for choosing the order is rather arcane — it should be less than $N^{\frac{1}{9}}$ and more than $N^{\frac{1}{24}}$. You are probably best off simply choosing something reasonable.

$$Y_i(1) = Y_i(0) + \delta_1 + \delta_2 X_i$$

Then

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0) = D_i(Y_i(0) + \delta_1 + \delta_2 X_i) + (1 - D_i) Y_i(0)$$

$$= \alpha + \delta_1 D_i + \delta_2 X_i D_i + \beta X_i + u_i$$

The interaction term between D and X suggests that we should interact the propensity score with treatment status in order to capture the treatment effect heterogeneity. It is thus preferable to run the following specification:

$$Y_i = \alpha + \delta_1 D_i + \delta_2 D_i p(X_i) + \beta p(X_i) + u_i$$

Note that if you run this regression, the estimated treatment effect for any given X_i is $\delta_1 + \delta_2 p(X_i)$, and the average treatment effect is $\delta_1 + \delta_2 \overline{p}(X_i)$. However, there is little to be gained from using propensity score methods in this manner. There is no guarantee that the even the interacted propensity score regression produces consistent estimates of meaningful average treatment effects when there is treatment effect heterogeneity.[7] And it may not produce estimates that are much different from a normal linear regression with X_i as controls. In fact, if we estimated the propensity score using a linear probability model, then we would get *the exact same estimate* of δ from the first regression (i.e., non-interacted) that we would get from running Y_i on D_i and all of the terms of X_i that go into the linear probability model (and at least in the latter case we would get the standard errors right). The advantage in using the propensity score is really that it enables us to apply less parametric estimators even if X_i is high-dimensional. Of course, we still have to estimate $p(X_i)$, and our ability to do this nonparametrically is limited if X_i is high-dimensional. So, in practice,

[7] I believe the treatment effect heterogeneity has to be a function of $p(X_i)$, rather than of X_i, to be guaranteed consistent estimates with the interacted propensity score regression model.

conditioning on the propensity score, even with superior methods like blocking and weighting (see below), is often not as advantageous as you might initially believe.[8]

2.3 Blocking on the Propensity Score

Another candidate for using the propensity score is to "block," or stratify, on the propensity score. That is to say, divide the range of the propensity score into K blocks (Dehejia and Wahba use 20 blocks of width 0.05) and place observations in each block according to their estimated propensity scores, $\hat{p}(X_i)$. Within each block k, compute, $\hat{\tau}_k$, the difference in means between treated and untreated observations. Finally, combine all K treatment effect estimates as follows:

$$\hat{\tau} = \sum_{k=1}^{K} \hat{\tau}_k \cdot \frac{N_{1k} + N_{0k}}{N}$$

In other words, the average treatment effect is a weighted sum of the block-level treatment effects, with the each block's weight equal to the number of observations contained in that block. Choosing the number of blocks is at the researcher's discretion. One popular algorithm is to start with a given number of blocks (e.g., 10), and check whether the covariates are balanced within each block. If they are not, then split the blocks and check again. Continue until the covariates are balanced.[9] If the covariates remain unbalanced within blocks even when the propensity score is balanced, then you may need to estimate the propensity score more flexibly.

The overlap assumption becomes prominent when blocking on the score. When a block

[8] Imbens, for example, writes: "Although [propensity-score methods] avoid the high-dimensional nonparametric estimation of the two conditional expectations, they require instead the equally high-dimensional nonparametric estimation of the propensity score. In practice the relative merits of [propensity score vs. doing nonparametric regression] will depend on whether the propensity score is more or less smooth than the regression functions, or whether additional information is available about either the propensity score or the regression functions."

[9] Note that if you have many covariates and many blocks, you should not expect 100% of the covariates to have no significant relationship to the treatment status in every block - some coefficients should be significant simply by chance. A more realistic target would be, for example, to find that only 10% of the covariates are significantly related to treatment status at the 10% level.

contains either zero treated units or zero control units, no estimate of the treatment effect exists for that block, and it must be discarded. Furthermore, because the logit specification forces $0 < \hat{p}(X_i) < 1$, it may appear that the overlap assumption is satisfied for all units when in fact it is not. To be safe, one should discard all control units with $\hat{p}(X_i)$ less than the minimum $\hat{p}(X_i)$ in the treated group and all treated units with a $\hat{p}(X_i)$ greater than the maximum $\hat{p}(X_i)$ in the control group.[10]

Note that blocking on the score is analogous to matching on the score in that you are only comparing observations with propensity scores that are close to one another. One could formally implement a matching estimator, however, using one of the methods discussed in Section 1. Dehejia and Wahba, for example, use nearest-neighbor matching as an alternative estimator to blocking on the propensity score (both estimators give similar results in most, but not all, cases).

2.4 Weighting with the Propensity Score

Hirano, Imbens, and Ridder (2003) advocate weighting by the (inverse of) the propensity score as a method to adjust for differences between treated and control units. To understand this estimation procedure, first consider a simple estimator that takes the difference in means between treated and control units (no conditioning on the propensity score or any regressors):

$$\hat{\tau}_{naive} = \overline{y}_T - \overline{y}_C = \frac{\sum D_i Y_i}{\sum D_i} - \frac{\sum (1 - D_i) Y_i}{\sum (1 - D_i)}$$

This method is biased because $E[Y_i(0)|D_i = 1] \neq E[Y_i(0)]$ — the units that select into treatment have different (unobserved) control outcomes than the entire population of units (which includes units that do not select into treatment). Hence using the observed control units to estimate the unobserved control outcomes of the treated units gives is an invalid

[10] I am assuming that the minimum $\hat{p}(X_i)$ occurs in the control group and the maximum $\hat{p}(X_i)$ occurs in the treated group. If not, perform the trimming so that the minimum $\hat{p}(X_i)$ is (virtually) the same for both groups and the maximum $\hat{p}(X_i)$ is (virtually) the same for both groups.

strategy. Suppose, however, that we knew the propensity score, $p(X_i)$. If we weighted each treated observation by the inverse of $p(X_i)$, we would find:

$$E[\frac{D_i Y_i}{p(X_i)}] \;=\; E[\frac{D_i(D_i Y_i(1) + (1-D_i)Y_i(0))}{p(X_i)}] \;=\; E[\frac{D_i Y_i(1)}{p(X_i)}]$$

$$= E[E[\frac{D_i Y_i(1)}{p(X_i)}|X_i]]$$

$$= E[\frac{E[D_i|X_i]E[Y_i(1)|X_i]}{p(X_i)}]]$$

$$= E[\frac{p(X_i)E[Y_i(1)|X_i]}{p(X_i)}]]$$

$$= E[E[Y_i(1)|X_i]]$$

$$= E[Y_i(1)]$$

CIA

given $D_i | X_i \perp\!\!\!\perp Y_i(k)$

Likewise, weighting each control observation by the inverse of $1 - p(X_i)$ gives us:

$$E[\frac{(1-D_i)Y_i}{1-p(X_i)}] \;=\; E[Y_i(0)]$$

We can implement the weighting scheme with the following estimator:

$$\hat{\tau}_{p(X)} = \frac{1}{N}\sum_{i=1}^{N}(\frac{D_i Y_i}{p(X_i)} - \frac{(1-D_i)Y_i}{1-p(X_i)})$$

What's the intuition behind this reweighting scheme (which, to me, is the least transparent of the methods)? Recall the problem: the propensity score is not balanced across treated and control groups. Treated observations are, on average, those with X_i's that make them more likely to be treated (which is, from our perspective, bad - covariates are not balanced across treated and controls), as well as those that randomly got treated (which is, from our perspective, good).

Consider any given observation i. Suppose that $p(X_i) = 0.80$. That means that there is an 80% chance that this observation would end up in the treated group and a 20% chance that it would end up in the control group — i.e., it is four times more likely to be in the treated group relative to the control group. Therefore, on average, there are four of these observations in the treated group for every one that is in the control group. Our weighting scheme fixes this imbalance. When this observation with $p(X_i) = 0.80$ is in the treated group, we weight it by $\frac{1}{0.8}$, effectively upweighting it by a factor of 1.25. When it is in the control group, however, we weight it by $\frac{1}{0.2}$, effectively upweighting it by a factor of 5. The control group weight is thus 4 times (5/1.25) the treated group weight, and the observation's frequency in the control group is increased by a factor of four relative to its frequency in the treated group. Our weighting scheme thereby ensures that this observation is equally represented (in expectation) in the treated and control groups. The same analysis holds true for any $p(X_i)$ such that $0 < p(X_i) < 1$, so the weighting scheme balances the propensity score across treated and control groups.[11]

The problem with the estimator given above is that there is no guarantee that the weights will sum to one ($\frac{1}{N}\sum \frac{D_i}{p(X_i)}$ equals one in expectation, but it need not equal one in any given sample). We can instead normalize each weighted sum by the actual sum of the weights. Doing this, and plugging in the estimated score in place of the true score, gives us our weighting estimator:

$$\hat{\tau} = (\sum_{i=1}^{N} \frac{D_i Y_i}{\hat{p}(X_i)} / \sum_{i=1}^{N} \frac{D_i}{\hat{p}(X_i)}) - (\sum_{i=1}^{N} \frac{(1-D_i)Y_i}{1-\hat{p}(X_i)} / \sum_{i=1}^{N} \frac{1-D_i}{1-\hat{p}(X_i)})$$

The nice thing about this weighting estimator is that it is, according to Hirano, Imbens, and Ridder, efficient.

[11] Note again the importance of the overlap assumption.

2.5 Dual Methods: Two Are Better Than One

Doubly Robust

One can enhance most of the methods above by combining them with regression. The advantage of this strategy is that incorporating regression can reduce any remaining bias and potentially enhance the precision of the estimator. Furthermore, it produces an estimator that is "doubly robust" — that is to say, if either the propensity score *or* the regression function is correctly specified, then the estimator will be consistent. In Guido Imbens' words, "you have two chances to get lucky."

To combine the weighting estimator with regression adjustment, simply run a weighted least squares regression of:

$$Y_i = \alpha + X_i\beta + \tau D_i + u_i$$

$\beta = (X'WX)^{-1}X'Wy$

Literature on estimated regressors

→ how will not including est. of uncertainty of $\hat{p}$ in the regression affect SEβ or bias(β)?

where the regression weights are defined as:

$$w_i = \sqrt{\frac{D_i}{\hat{p}(X_i)} + \frac{1 - D_i}{1 - \hat{p}(X_i)}}$$

→ bootstrap SE... valid?

This estimator has the "double robustness" property. Alternatively, we could combine the blocking procedure with regression by running the following regression within each of the K blocks:

$$Y_i = \alpha_k + X_i\beta_k + \tau_k D_i + u_i$$

Then combine all of the $\hat{\tau}_k$ estimates together as:

$$\hat{\tau} = \sum_{k=1}^{K} \hat{\tau}_k \cdot \frac{N_{1k} + N_{0k}}{N}$$

2.6 Regression Revisited

Recall that in a previous lecture we demonstrated that if you assume unconfoundedness and a constant treatment effect, then you can consistently estimate τ with the following regression:

$$Y_i = \alpha + \tau D_i + \delta E[D_i|X_i] + u_i$$

$E[D_i|X_i]$ is, of course, simply the propensity score. If $E[D_i|X_i]$ is linear, then we can estimate it by simply including X_i as regressors.[12] If $E[D_i|X_i]$ is of unknown functional form, then we can in principle estimate it nonparametrically using kernel or series regression, although in practice we may face the Curse of Dimensionality. What if we drop the assumption of constant treatment effects, however? In that case, we need to account for the possibility that the treatment effect heterogeneity may be related to the covariates in some manner (e.g., perhaps some program has differential effects on high school dropouts and college graduates). Consider the simplest case, in which the treatment effect heterogeneity is only a function of X_i:

$$Y_i(0) = \overline{\alpha} + g_0(X_i) + \alpha_i$$

$$Y_i(1) = Y_i(0) + \overline{\beta} + g_1(X_i)$$

Then we get the following regression model:

$$Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$$

$$= (Y_i(0) + \overline{\beta} + g_1(X_i))D_i + Y_i(0)(1 - D_i)$$

$$= (\overline{\beta} + g_1(X_i))D_i + Y_i(0)$$

[12] Note that, strictly speaking, $E[D_i|X_i]$ is only likely to be linear when you have a saturated model.

$$= \overline{\alpha} + \overline{\beta} D_i + g_0(X_i) + g_1(X_i) D_i + \alpha_i$$

So with treatment effect heterogeneity, you have to estimate separate series or kernel estimators for treated observations and nontreated observations, even if you assume unconfoundedness. Conceptually, it may be easier to split the sample into treated and untreated observations and estimate the functions separately on each sample. Let $f_0(X_i) = \overline{\alpha} + g_0(X_i)$ and $f_1(X_i) = \overline{\alpha} + g_0(X_i) + \overline{\beta} + g_1(X_i)$. Then

$$E[Y_i(0)|D_i = 0, X_i] = f_0(X_i)$$

$$E[Y_i(1)|D_i = 1, X_i] = f_1(X_i)$$

$$E[Y_i(1) - Y_i(0)|X_i] = f_1(X_i) - f_0(X_i)$$

In practice, we would estimate this quantity using:

$$\frac{1}{N} \sum \hat{f}_1(X_i) - \hat{f}_0(X_i)$$

$\hat{f}_1(X_i) = \begin{cases} Y_i & D_i = 1 \\ \hat{Y}_1(X_i) = \text{kernel reg est. at } X_i \text{ of } Y \text{ on } X \text{ of treated obs} & D_i = 0 \end{cases}$

$\hat{f}_0(X_i) = \begin{cases} \hat{Y}_0(X_i) = \text{kernel reg est. at } X_i \text{ of } Y \text{ on } X \text{ of control obs} & D_i = 1 \\ Y_i & D_i = 0 \end{cases}$

Since we observe $Y_i(1)$ when $D_i = 1$ and $Y_i(0)$ when $D_i = 0$, we set $\hat{f}_1(X_i) = Y_i(1)$ when $D_i = 1$ and $\hat{f}_0(X_i) = Y_i(0)$ when $D_i = 0$. When $D_i = 0$, our estimate of $f_1(X_i)$ is generated by plugging X_i into the regression on the treated subsample, i.e. $\hat{f}_1(X_i)$. When $D_i = 1$, our estimate of $f_0(X_i)$ is generated by plugging X_i into the regression on the control subsample, i.e. $\hat{f}_0(X_i)$. Thus our regression estimator is:

$$\hat{\tau} = \frac{1}{N} \sum (D_i Y_i + (1 - D_i)\hat{f}_1(X_i)) - ((1 - D_i)Y_i + D_i \hat{f}_0(X_i))$$

Though it may not be readily apparent, the estimator above highlights the importance of overlap in the distributions of the X_i for treated and control groups. This is probably easiest to see if you consider estimating f_0 and f_1 using kernel regressions. Consider $\hat{f}_1(X_i)$ for some control observation i. $\hat{f}_1(X_i)$ is estimated by running a kernel regression at X_i (which,

2 big concerns: 1) does CIA hold?
2) Is there covariate overlap

remember, comes from a control observation) using the treated data. But if the treated and control distributions of X do not have overlap, then there may be no data in the treated group near X_i, and the kernel regression will need to extrapolate from data points that are far from X_i (i.e., it will need to have a large bandwidth).

If you recall from the initial lecture on regression adjustment, we noted that overlap is important in determining whether the precise specification of a regression will impact our estimated treatment effects. Here, we show that overlap is also important if you are using nonparametric estimators. We have also seen that overlap is important for matching (i.e., if you don't have overlap of the covariates, then you can't find matches) and for the propensity score (if $p(X_i)$ gets close to zero or one, then the weights get enormous if you're doing the weighting procedure, or it becomes hard to find matches if you are blocking or matching). The dominant theme is thus that the estimation technique itself is probably not as important as:

(1) Whether the unconfoundedness assumption holds.

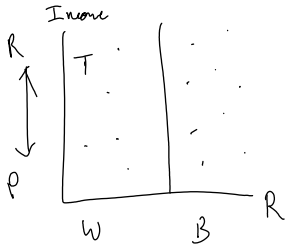
(2) Whether there is overlap in the treatment and control distributions of the covariates.

2.7 Assessing Overlap: LaLonde Revisited

We have seen that all the estimation procedures are sensitive to whether there is overlap in the distributions of the X's for the treated and control groups. In cases in which we have only one or two variables in X, it is fairly easy to assess the overlap of the covariates — simply plot the distribution of X for the treated group and the control group, and see whether they have similar support. This could be done with a histogram (just don't make the bins too wide), or you could use a kernel density estimator.

In higher dimensional cases, inspecting marginal distributions of single covariates is not as informative. It could easily be the case that the marginal distributions overlap for each covariate, and yet the joint distributions do not overlap. For example, suppose that X contains two covariates: income and race. Further suppose that both the treated and control

• Want to est $\hat{p}$ w/ less smoothing, i.e. use higher order terms of Xs so when you chop off low and high p-score obs, you'll catch important variation near the edges



groups contain a mix of blacks and whites and a mix of rich and poor. When inspecting the marginal distributions of income and race, we appear to have overlap. Suppose, however, that all rich whites appear in the treated group, and that all poor blacks appear in the control group. Obviously, we do not overlap in important parts of the joint distribution. However, because both the treated and control groups contain rich blacks and poor whites, we appear to have overlap when examining each of the marginal distributions.

One nice thing about the propensity score methods is that they allow you to assess overlap using the propensity score itself — recall that the overlap assumption for the propensity score is $0 < p(X_i) < 1$. Simply plot the distribution of the propensity scores for the treated and control groups, and check to see whether there is sufficient overlap in those distributions (insufficient overlap would be neighborhoods of $p(X)$ that contain many estimated scores from one group and few of the other — often these areas will be near the extreme values of 0 and 1). Of course, the accuracy of your assessment will depend on the accuracy of your specification of the score. In this case, it is better to use procedures that do less smoothing (e.g., lower bandwidth with kernel estimators, more higher order terms, interactions, and knots with series estimators) — otherwise you may artificially produce overlap in the propensity score distributions by extrapolating from areas in which the propensity score has positive support to areas in which it does not.

If there are some points in the propensity score distributions of the two samples that do not overlap, then you should trim one or both of the samples to address the problem. In practice, this entails discarding some observations with propensity scores above or below a certain level. Imbens performs a reanalysis of the LaLonde (1986) data to demonstrate how this is done. He begins with a table that summarizes the values of key covariates for the treated group, the control group, and the CPS sample from which LaLonde draws his simulated control groups. Table 1 summarizes the values of several of the covariates.

Note that, as expected (given the random assignment), the differences between the control group and the treated group are small — less than 0.2 standard deviations in all cases.[13]

[13]To compute the standard deviation for each covariate x, Imbens recommends using the formula

Table 1: Summary Statistics

	Controls		Treated			CPS		
	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	13.65	9.27	-1.30
Unempl '74	0.75	0.43	0.71	0.46	-0.09	0.12	0.32	1.77
Unempl '75	0.68	0.47	0.60	0.49	-0.18	0.11	0.31	1.54

Source: Imbens (2007).

would expect failure of overlap assumption

The differences between the CPS sample and the treated group, however, are substantial. All but one of them are greater than 0.5 standard deviations, and in one case the difference reaches 2.8 standard deviations. Roughly speaking, any difference in means in excess of 0.25 standard deviations is considered large (i.e., imbalanced) — estimation methods relying on linear regression will require substantial extrapolation. Note that, unlike the t-statistic, this metric is not a function of the sample size. Using a t-statistic as a measure of balance can be misleading since, as the sample grows, any nontrivial difference in means will necessarily generate a large t-statistic. This property of the t-stat is somewhat deceptive for our purposes because a larger sample is actually more desirable (in that it allows us more freedom to discard observations for which there is no overlap).

Table 2 presents results for a wide variety of estimators using the treated data and the full CPS comparison data set, with 1975 earnings as the dependent variable. Since the program did not begin until after 1975, we can be sure that any significant result in this table is simply due to selection bias. Though all the estimators — OLS, propensity score methods, matching, and dual methods — do much better than a simple difference in means between the treated group and the full CPS group (the estimated treatment effect decreases by 77 to 91 percent, depending on the method), they all incorrectly reject the null hypothesis of no

$\sqrt{(s_{xc}^2 + s_{xt}^2)/2}$, where s_{xt}^2 is the sample variance of x in the treated group and s_{xc}^2 is the sample variance of x in the control group.

treatment effect at a high level of statistical significance. Ironically, the two OLS procedures actually perform better than the more sophisticated procedures in this case. Nevertheless, the take-away message is that **none** of the procedures performs well if there is not overlap in the covariate distributions between the treated and control samples. In this case, all methods are performing substantial extrapolation in some form or another.

Table 2: Estimates with Earnings '75 As Outcome

	Effect	S.E.	t-stat
Simple Diff	-12.12	0.68	-17.8
OLS (parallel)	-1.15	0.36	-3.2
OLS (separate)	-1.11	0.36	-3.1
Propensity Score Weighting	-1.17	0.26	-4.5
Propensity Score Blocking	-2.80	0.56	-5.0
Propensity Score Regression	-1.68	0.79	-2.1
Propensity Score Matching	-1.31	0.46	-2.9
Matching	-1.33	0.41	-3.2
Weighting and Regression	-1.23	0.24	-5.2
Blocking and Regression	-1.30	0.50	-2.6
Matching and Regression	-1.34	0.42	-3.2

Source: Imbens (2007).

Figures 1 through 6, taken from Imbens (2007), plot the distributions of the (estimated) propensity scores for each of the samples. Figures 1 and 2 plot histograms of the propensity scores for the control sample and the treated sample respectively.[14] As expected, given the random assignment, there is a high degree of overlap between these two propensity score distributions.

Figures 3 and 4 plot histograms of the propensity scores for the CPS comparison sample and the treated sample respectively. In this case, the propensity score is generated by running a logit regression in which the dependent variable is an indicator that equals unity

[14] In Dehejia and Wahba (1999) the "propensity score" is generated by running a logit regression in which the dependent variable is an indicator that equals unity if a unit is in the treated group and zero if a unit is in the CPS sample, and the covariates are the ones listed in the table plus other unlisted covariates, with higher order and interaction terms. In Figures 1 and 2, I believe that Imbens may actually estimate the propensity score using only the experimental treated and experimental control groups, which seems strange because the treatment status was randomly assigned. However, one could think of it as being similar to controlling for covariates even when you have random assignment. Regardless, the random assignment explains why virtually all of the p-scores fall between 0.2 and 0.6.

if a unit is in the treated group and zero if a unit is in the CPS sample. Note the marked lack of overlap between the two distributions — very few of the CPS comparison units have propensity scores exceeding 0.05, while the treated units have propensity scores ranging from 0 to 0.70.

To address this clear lack of overlap, Imbens implements a simple "0.1 rule" in which he drops all observations in either group with propensity scores less than 0.1 or greater than 0.9. This rule reduces the treated population by 24%, from 185 to 141. It reduces the simulated control population by 98%, from 15,992 to 313. Figures 5 and 6 plot the propensity score distributions from the trimmed CPS sample and the trimmed treatment group. Note that the overlap of the two distributions is substantially improved.[15]

Table 3 presents summary statistics for the treated group and the trimmed CPS sample. Note that there is now much better balance of the covariates between the two groups. All but two of the covariates now have differences in means that are less than 0.25 standard deviations.

Table 3: Summary Statistics after trimming using 10-90 p-score

	CPS Controls (313)		Treated (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

much more balanced than table 1

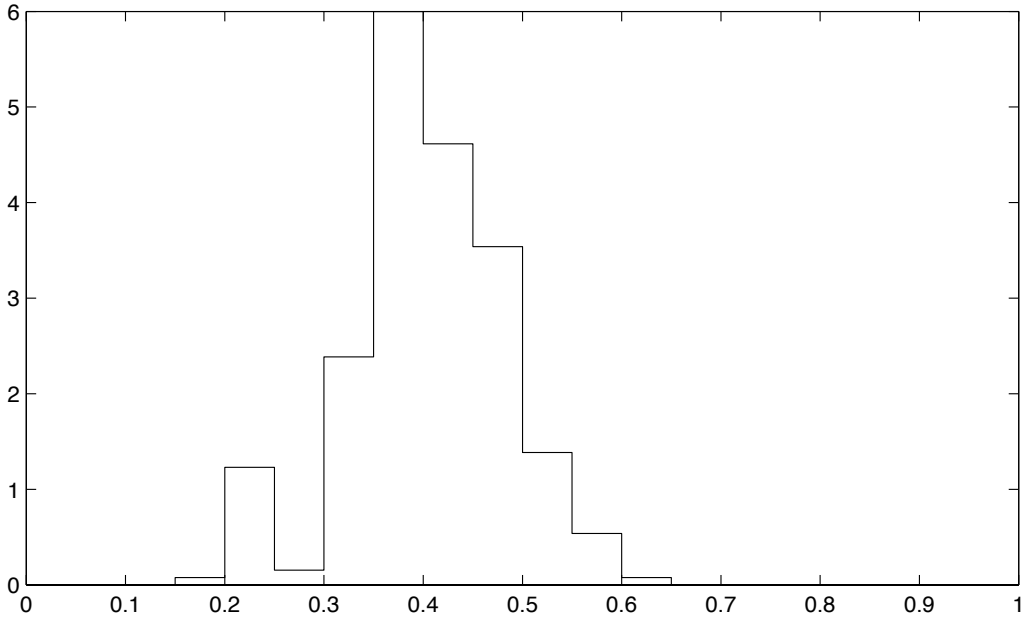
Source: Imbens (2007).

Table 5 presents results for a wide variety of estimators using the trimmed data in which 24% of treated observations and 98% of CPS observations have been discarded. Two models

[15]Both groups appear to have some propensity scores that fall below 0.1, which one would think would be impossible given the trimming rule. It's likely that Imbens reestimated the propensity scores after he did the trimming — then it would be possible to get estimated propensity scores below 0.1.

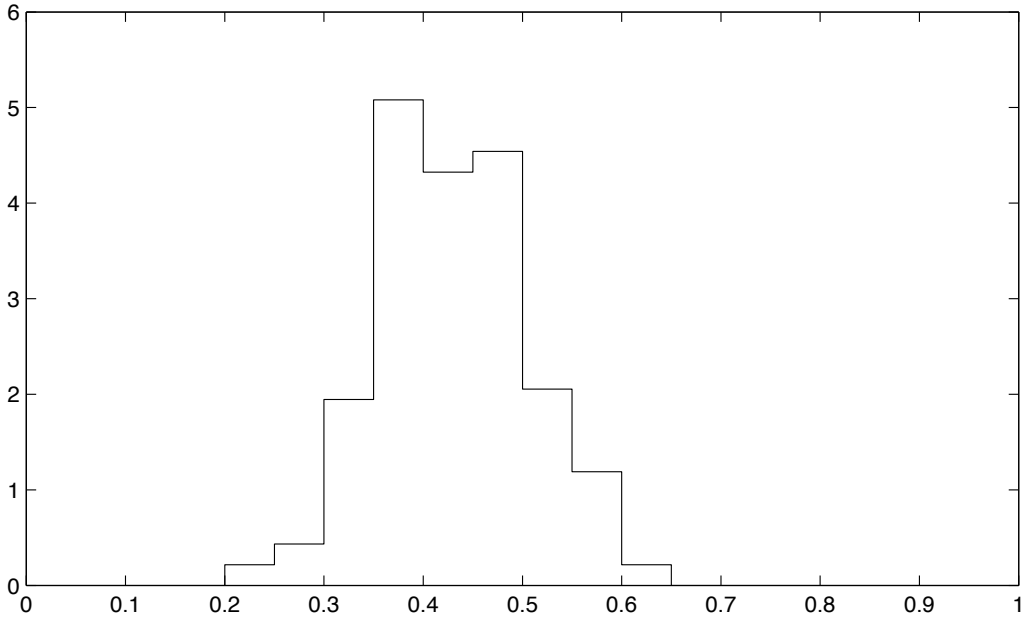
actual controls in experiment

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



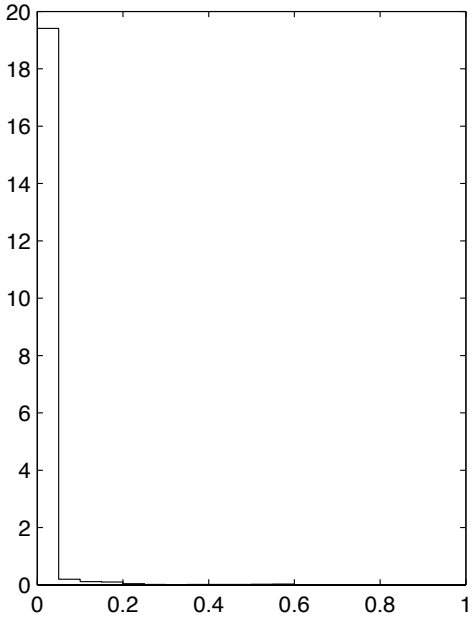
all treated are in all samples

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



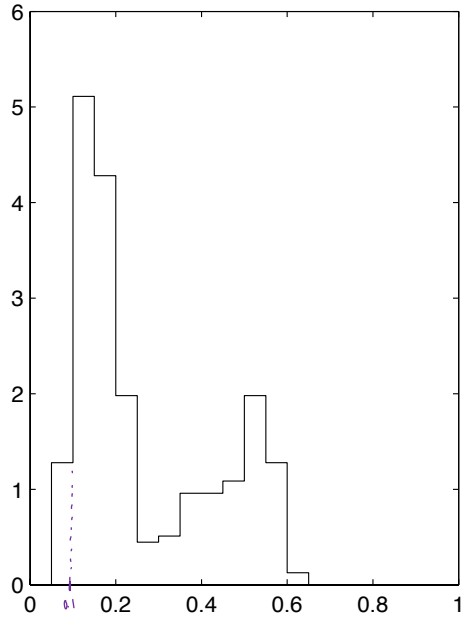
$N_C = 15,992$

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



$N_C = 313$

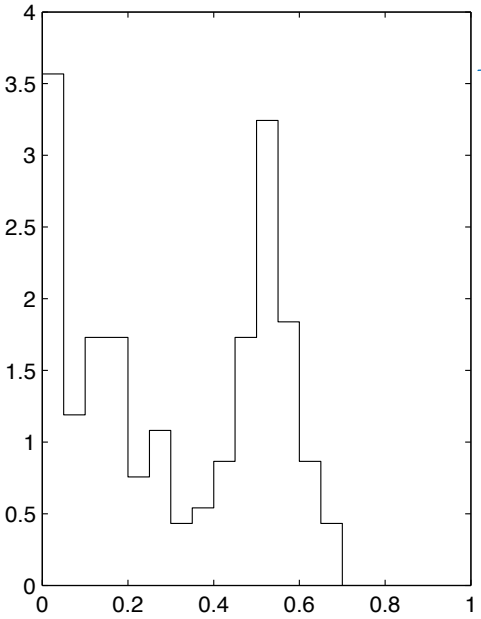
Figure 5: hist p–score for controls, cps selected sample



trimming low & high p-score. Then re-est $\hat{p}$

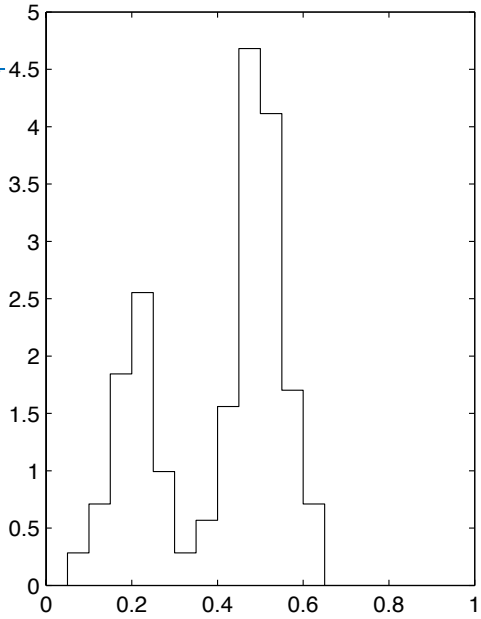
$N_T = 185$

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



$N_T = 141$

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



At least: • drop all C with $\hat{p}(x_i) < \min_{i \in T} \hat{p}(x_i)$

• drop all T with $\hat{p}(x_i) > \max_{i \in C} \hat{p}(x_i)$

are run, one using the 1975 earnings data, and one using the 1978 earnings data. For the former outcome, we expect no treatment effect (the program did not begin until after 1975). For the latter outcome, Imbens doesn't give the experimental benchmark, but it appears that he is using Dehejia and Wahba's "RE74 Earnings Sample," in which case the experimental benchmark is approximately $1,600 to $1,800.

With the trimmed data, virtually all of the estimators — simple differences, OLS, propensity score methods, matching, and dual methods — do very well. No estimator, including the simple difference in means between the two samples, finds a significant difference in the pre-treatment data, and the estimated treatment effects using the 1978 earnings data range from 1.73 to 2.23 (the experimental benchmark is approximately 1.7). The one exception is the p-score matching estimator, which estimates a treatment effect of only 0.65 using the 1978 earnings data. So, in a sample with a high degree of overlap in the two propensity score distributions, we see that almost all estimators perform well. The important step is thus trimming the samples to achieve overlap — the choice of estimators after that point is a second order consideration. Of course, this is all still conditional on the selection on observables assumption...

Table 4: Estimates Using Trimmed Data

	Earn '75 Outcome			Earn '78 Outcome		
	Mean	S.E.	t-stat	Mean	S.E.	t-stat
Simple Diff	-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

Source: Imbens (2007).

3 Additional Studies of Experimental vs. Observational Data

We have now reviewed a variety of estimation procedures that are valid under the selection on observables assumption: linear regression, nonparametric regression, matching, and propensity score methods. We have seen that the most important feature for these designs is that the treatment and control groups (either the original ones, or the ones you construct through trimming) have good overlap in terms of their distributions of covariates or propensity scores. So, if you have a rich set of covariates, and you have good overlap, then almost surely you will be able to reproduce the results of an experiment, right? *Hahahahahahaha*. No.

3.1 Arceneaux, Gerber, and Green: The NSW Euphoria Antidote

It was discovered — by political scientists, no less — that the National Supported Work Demonstration was in fact *not* the only randomized experiment ever run in the history of humankind (shocking, I know). Arceneaux, Gerber, and Green (2006) (henceforth AGG) perform an exercise similar to LaLonde's and Dehejia and Wahba's exercises using data from a large-scale voter mobilization effort (this type of effort is often referred to as a "Get Out the Vote" campaign). In this effort, households are randomly called and encouraged to vote. Although the calling assignment is random, whether a household is actually contacted is non-random — people often do not answer their phones. Regressing a household's voting behavior on whether or not that household was contacted can thus give biased estimates of the causal effect of encouragement on voting. One way to correct for this bias is to use the original random calling assignment as an instrument for actual contact/encouragement. This estimator will consistently estimate the causal effect of encouragement on voting for households for whom the original calling assignment changed whether or not they were contacted (i.e., households that actually got contacted). In this case, that means that the instrumental variables estimator will estimate TOT, the effect of the treatment (being contacted and

encouraged) on the treated (those who were contacted and encouraged). I refer to these estimates as the "experimental estimates."

Alternatively, however, we could try to use a matching estimator to condition on the observed covariates and, in that manner, estimate TOT. Specifically, we could find a match for every treated unit, and compare the difference in voter participation for the treated units and their matched pairs. We could then benchmark the results of this matching estimator, which should be valid under the selection on observables assumption, against the experimental estimates. This is exactly what AGG do.

The AGG data have at least one important advantage over the NSW data: the AGG sample size is massive. There are approximately 60,000 treated individuals and almost two million control individuals. All individuals (treated and control) were taken from voter registration lists, which contain detailed information on voting histories and demographic characteristics. Once included in the study, individuals were randomly assigned to treatment or control groups. Obviously, most people (97%) were assigned to the control group.

The first column of Table 5 reports experimental benchmark estimates (i.e., IV estimates). These estimates suggest that voter encouragement raises the probability of voting by approximately 0.3 to 0.5 percentage points. These estimates are precisely estimated ? and are not significantly different than zero — voter encouragement appears to have no appreciable effect on voting behavior, at least for this population.

Table 5: "Effect" of Voter Encouragement on Voting

	Experimental	OLS	Matching
Sample w/o Unlisted No.	0.5	2.7	2.8
	(0.4)	(0.3)	(0.3)
N	1,905,320	1,905,320	22,711
Sample w/ Unlisted No.	0.3	4.4	4.4
	(0.5)	(0.3)	(0.3)
N	2,474,927	2,474,927	23,467

Source: Arceneaux, Gerber, and Green (2006). Parentheses contain standard errors.

Exact

age
gender
county FE
2 years of previous voting

need to match these to controls

— matched exactly so covariate overlap probably not the issue

— focus on conditional independence

→ don't have enough Xs to eliminate bias?

→ need enough pretreatment data to control for bias
① does X have explanatory power for y? for D?
②

in NSW data, pretreatment data → Y and D
employment, earnings → earnings, employment help

in this data, pretreatment outcome data → Y but not D
voting voting
called, and answered to encourage to vote

The second column of Table 5 reports OLS estimates that regress voting behavior on whether an individual was contacted, conditioning on a variety of covariates. These covariates include age, household size, gender, contest indicators, county indicators, and two years of previous voting behavior.[16] This is roughly comparable to LaLonde (1986) and Dehejia and Wahba (1999), who have age, education, marital status, gender, race, and two years of prior earnings. In particular, both studies include two years worth of pre-treatment outcomes, which DW stress as being very important. The OLS estimates range from 2.7 to 4.4 and are highly significant (t-statistics of 9 to 14). These estimates imply that voter encouragement raises the probability of voting by 2.7 to 4.4 percentage points. Clearly the OLS estimates are biased — does this bias occur because of a lack of overlap in the covariates between the treated and untreated groups?

The third column of Table 5 reports matching estimates that find an *exact* match in the control group for each treated unit (this is possible because all covariates are discrete and the control group is enormous). They are able to match about 91% of observations using exact matches and 99.9% of observations using slightly less exact matches (e.g., coding age in 3 year intervals and dropping some geographic indicators). The overlap assumption therefore appears to be satisfied, in the sense that close matches can be found for virtually all observations. Nevertheless, matching estimates range from 2.8 to 4.4 and are highly significant. Matching therefore does not appear to solve the selection bias problem, even with excellent overlap in the covariate distributions of the treated and control observations.

3.2 Shadish, Clark, and Steiner: Some Balance

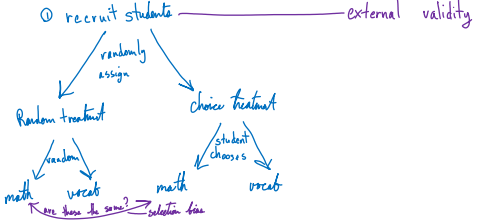
Shadish, Clark, Steiner (2008) raise a somewhat different, but important, issue with LaLonde's NSW paper (and others like it). Their complaint is that LaLonde's NSW exercise confounds the assignment mechanism (random assignment versus observational data) with other factors — for example, sites, times, variable measurements, missing outcome data, etc. The

[16] Though AGG do not show the covariate balance across treated versus untreated individuals, there must be substantial imbalance (i.e., covariates can predict whether an individual answers her phone) because controlling for covariates has a strong effect on the OLS estimates.

core of Shadish, Clark, and Steiner's (henceforth SCS) argument is that the randomly assigned "experimental" control units come from one data set — the NSW data — while the non-randomly assigned "observational" control units come from other data sets — CPS and PSID. Thus it is possible that factors specific to the different data sets may be contributing to the observed differences in the estimates generated by using the CPS/PSID controls instead of using the experimental NSW controls. Simply put, LaLonde's study of confounding in the context of observational data may itself by confounded by other factors that correlate with the assignment mechanism.

SCS's response is, naturally, to randomly assign the assignment mechanism. Their study proceeds in three basic steps. First, they recruit students as test subjects. Next, they randomly assign students to either have a treatment randomly assigned to them or to be given a choice about which treatment they would like. In the random assignment arm, students are randomly assigned to either math training or vocabulary training. In the choice arm, students either choose math training or vocabulary training. In all cases students are tested on math and vocabulary after completing the training. Finally, SCS estimate the "effects" of training in the choice arm using various selection on observable techniques (regression adjustment and several propensity score methods). These estimates adjust for a rich set of covariates that SCS collect from the students: sex, age, marital status, race, pretreatment vocabulary and math scores, number of math courses taken, stated preferences regarding math and literature, a personality measure, parental education, math intensity of major, ACT scores, and GPAs. The outcome of interest is the difference between a student's post-training math performance and her post-training vocabulary performance (or vice versa). SCS compare these estimates to the "true" effects of training that they estimate using the data from the random assignment arm. This is similar in spirit to LaLonde's exercise, but SCS can be confident that no other factors are correlated with whether a student ends up in the random assignment arm or the choice arm.

SCS present their results in Table 1 of their paper (p. 1338). I summarize the notable patterns from their results here. First, students randomly assigned to math training do bet-



ter at math relative to vocabulary, and students randomly assigned to vocabulary training do better at vocabulary relative to math. Second, selection bias turns out to be modest in this experiment. The unadjusted estimate in the non-randomly assigned data is only 25% higher than the true treatment effect (from the randomly assigned data) for math. It is only 9% higher than the true treatment effect for vocabulary. Third, of the various selection on observables designs that SCS implement, OLS (aka "ANCOVA") does as well or better than anything else. Regression adjustment (OLS) achieves an 84% bias reduction in the math estimate using non-randomly assigned data and a 94% bias reduction in the vocabulary estimate using non-randomly assigned data. Among the propensity score methods, SCS use blocking, including the propensity score as a regressor, and weighting. They also implement some "doubly robust" methods. Blocking works well for both math and vocabulary, but weighting only works well in the vocabulary case. Including the propensity score as a regressor results in less bias reduction than just controlling directly for the covariates that go into the score.

The most interesting part of SCS's study is that they also try propensity score blocking using only "predictors of convenience" — sex, age, marital status, and race. This purposely omits a lot of rich covariates that they have at their disposal, and it simulates a scenario in which a researcher has limited controls available in the data. Propensity score methods that only use predictors of convenience do much worse than propensity score methods (or simple linear regressions) that use the full set of predictors — bias reduction is in the range of only 0% to 40%. Thus SCS provide evidence that *if* you have a rich set of relevant covariates, you may get decent estimates using non-experimental data. However, in many cases it will be difficult to say when you have a sufficiently rich set of covariates!

3.3 Conclusions

The AGG experiment demonstrates that even in cases that seem well-suited to the selection on observables design, estimates can be biased pretty badly. The main problem that AGG face (or would face, if they didn't have the experimental estimates) is that it's hard to make

the case that they observe all of the important factors in determining whether a caller makes contact with an individual or whether an individual turns out to vote. Of course, *one could say the same thing about the NSW data*, so it's difficult for the real-world econometrician to determine when the selection on observables design does or does not hold.[17] The SCS conclusions are somewhat more optimistic than AGG, but again it's hard to know when you have observed "enough" selection factors. An alternative to the selection on observables design is, of course, the selection on unobservables design. The next section of the course focuses on this category of research designs.

4 Additional References

Hirano, Keisuke, Guido Imbens, and Geert Ridder. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica*, 2003, 71, 1161-1189.

Millimet, Daniel and Rusty Tchernis. "On the Specification of Propensity Scores, With Applications to the Analysis of Trade Policies." *Journal of Business and Economic Statistics*, 2009, 27, 397?415.

• Sequential treatment effects

[17] My personal belief is that it's most plausible when the selection was performed by an individual or body that has observes the same data that is available to the researcher — e.g., a college admissions officer who does not conduct interviews or read long personal essays. In cases in which units are self-selecting (which, to be honest, is most cases), it's far less plausible.