Vanderbilt University
Leadership, Policy and Organizations
Class Number 9553
Spring 2017

Methods for Matching

Matching is in many ways a non-parametric form of selection on observables. We're going to use all of the observed information we have and make some assumptions about the distribution of unobserved covariates.

For this discussion, we're going to work with the idea of the community college diversion effect. This is the impact of attending community college on the probability of receiving a bachelor's degree, among those students who said their goal is a bachelor's degree

Suppose that we divide groups into two different outcome groups.

Group y_1 denotes those students who were "treated"– those who attended a community college

Group y_0 denotes those students who were not treated– those who attended a four year college.

z is a binary variable, where $z = 1$ if a student is treated. $z = 0$ otherwise.

y_i for any individual is:

$$y_i = \begin{cases} y_{1i} \text{ if } z_i = 1 \\ y_{0i} \text{ if } z_i = 0 \end{cases} \quad = y_{0i} + (y_{1i} - y_{0i})z_i$$

The impact of participation is:

$$\begin{aligned} E[y_i|z_i = 1] - E[Y_i|z_i = 0] = E[y_{1i}|z_i = 1] - E[y_{0i}|z_i = 1] \\ + E[y_{0i}|z_i = 1] - E[y_{0i}|z_i = 0] \end{aligned}$$

Under random assignment, selection bias goes away. This happens because z_i is independent of potential outcomes:

$$\begin{aligned} E[y_i|z_i = 1] - E[Y_i|z_i = 0] &= E[y_{1i}|z_i = 1] - E[y_{0i}|z_i = 0] \\ &= E[y_{1I}|z_i = 1] - E[y_{0i}|z_i = 1] \end{aligned}$$

Question: why can we say that $E(y_{0i}|zi = 0) = E(y_{oi}|z_i = 1)$ in experiments?

In experiments, the second part simplifies to:

$$E[y_{0i}|z_i = 1] - E[y_{0i}|z_i = 1] = E[y_{1i} - y_{0i}|z_i = 1]$$
$$= E[y_{1i} - y_{0i}]$$

In regression and matching, $\boldsymbol{x}$ is a series (or vector) of characteristics of students.

We want to estimated treatment on the treated.

For each person, only Y_1 or Y_0 is observed, so we have a missing data problem.

$$\delta_{TOT} = E[y_{1i} - y_{oi}|z_i = 1]$$
$$= EE[y_{1i} - y_{0i}|\boldsymbol{x}_i, z_i = 1]|z_i = 1 \quad = EE[y_{1i}|\boldsymbol{x}_i, z_i = 1] - E[y_{0i}|\boldsymbol{x}_i, z_i = 1]|z_i = 1$$

We don't observe $E(y_{0i}|\boldsymbol{x}, z = 1)$—this is the counterfactual outcome. In an experimental study, we know that $E(y_0|\boldsymbol{x}, z = 1) = E(y_0|\boldsymbol{x}, z = 0)$ because of randomization. In observational studies, we need to make an assumption about the information contained in $\boldsymbol{x}$ and its relationship with y_0.

Our assumption is that the information in $\boldsymbol{x}$ must be sufficient to meet the conditional independence assumption:

$$(y_0, y_1 \perp z)|\boldsymbol{x}$$

This is an *extremely* strong assumption. What we're saying here is that once we've got two students with the same distribution of covariates, assignment to treatment or control is essentially random. You have to think that there's no additional information that's being supplied by whether the person gets the treatment. We can always demonstrate that the distribution of covariates between the treatment and control group looks similar after matching. This is a necessary but NOT SUFFFICIENT condition to meet the conditional independence assumption.

We can then use the information in x to create a propensity score p, which denotes the probability of selection into the treatment or control group. $p = pr(z = 1|\boldsymbol{x})$. This is estimated as the predicted value from a model for binary data, typically a probit model.

We then can estimate treatment on the treated by substitution:

$$E(y_{0i}|\boldsymbol{x}_i, z_i = 0) = E(y_{oi}|\boldsymbol{x}_i, z_i = 1)$$

We can use this insight to get to the matching estimand,

$$E[y_{1i} - y_{0i}|z_i = 1] = \sum_x \delta_x P(Xi = x|z_i = 1)$$

P is the probability mass of $\boldsymbol{x}_i$ given $z_i = 1$.

What does this mean in practice? We need a GREAT model for p, the probability of selection. In fact, most of the original work on this topics assumed p was known, rather than estimated. Once we have an estimate of p, we need to think hard about how to match units based on this propensity score.

There are multiple ways to create a matched control group to a treatment group:

Exact Matching Exact matching involves finding a unit in the control group that has *exactly* the same values of the covariates as the treated unit. This works best in very large datasets with only a few covariates. Exact matching is typically the most rigorous approach, but can suffer from a lack of efficiency. As more variables are introduced, this approach also suffers from what's called the "curse of dimensionality".

Nearest Neighbor Matching In nearest neighbor matching, the analyst finds an untreated unit that is most similar to the treated unit based on a measurement. In many cases, this measure is the propensity score, which is the predicted value of treatment for both the unmatched and matched unit, based on a regression for binary outcome with selection as the dependent variable. Nearest neighbor matching can use multiple controls per treatment. It can also use the same control twice for different treatment cases.

Regression conditional on the propensity score Another technique is to estimate the propensity score, then include it as a control in the regression. Technically, this should get rid of selection bias in the parameter estimates.

Propensity Score Stratification When stratifying based on the propensity score, the analyst estimates the propensity score as described above, then splits up the data based on levels of the propensity score. The analyst then runs the regression/does the comparison within each stratum, then averages the results, with weights (sometimes) based on the number of units in each stratum.

We'll go over each of these methods and their advantages and disadvantages. What you'll learn very quickly is that estimates vary depending on the choice of estimator. The best practice for now is to use a couple of different estimators as robustness checks.

Required Checks

For the selection model, you need to demonstrate good model fit, usually through the ROC curve stat, c.

Any time you undertake matching, you must demonstrate that you have achieved balance between the treatment and the control group. The whole point of matching is that the treatment and control group are very similar in terms of the *distribution* of all observed covariates. This is an empirical question that should be checked.

The standard comparison is to look at the mean and the variance of covariates in the treatment and control group, both before and after matching. Remember that the technical question you're asking is whether the two groups are similar in terms of distributions. Sometimes t-tests on the covariates between the treatment and control group are used. These should be used quite carefully, as a lack of significance in a t-test can come about because of small sample sizes. One of the standard things that matching algorithms do is to reduce sample size, so it's pretty easy to get rid of statistical significance that way. Instead, use the ratio of the variances, or the standardized difference. Whatever you use, think about the problem of establishing similar distributions– for continuous variables, kernel density plots can really help here.

Sensitivity Analysis

Propensity score matching has become widely used in education, and most analysts are using it incorrectly. It helps to reduce bias in many cases, but it rarely meets the conditions that would be required for it to be used for the purposes of causal inference. Once very useful check on your results is to create bounds based on the introduction of a theoretical, unobserved covariate with certain characteristics.

The idea here is to introduce a new variable, call it c. This variable represents an unobserved covariate that can impact both selection into treatment $pr(z|x,c)$ and the outcome $(y|x,c)$. By creating a variable with known impacts on both selection and outcome, we can come up with a sense of how "powerful" a missing covariate would have to be in terms of both selecting into treatment and the outcome. If it would require a very "powerful" covariate to change the estimate, that's useful information. On the flip side, if a single covariate with modest impacts on selection into treatment and the outcome can cause results to vary, that should make us doubt the results.

Final Thoughts

The promise of matching is that it can re-create experimental estimates by creating "as good as randomly assigned" treatment and control groups. The reality falls far short of this promise. It's much better to think of matching as a means by which we can reduce bias in our estimates due to a lack of balance and/or overlap in the sample. In the end, there's never any such thing as a free lunch.