Gov 2001: Problem Set 7

Due Wednesday, April 13th at 6pm

Instructions

You should submit your answers and R code to the problems below using the Quizzes section on Canvas.

Problem 1 - Estimating an ATE

In this Problem Set, we'll look at comparing matching methods for estimating a causal effect to an experimental benchmark. The classic comparison of the two is the Lalonde (1986) evaluation of the National Supported Work (NSW) job training program, which examines the effect of a job training program on individual's wages in two settings: an experiment where individuals were randomly assigned to the job program, and in a hypothetical observational setting, where we observed individuals assigned to the program and a series of potential controls drawn from the overall US population. While Lalonde shows that nonexperimental approaches fail to meet the experimental benchmark or are highly sensitive to modeling choices, subsequent work by Dehejia and Wahba (1999) shows that matching estimators can help over-come this model sensitivity.

We're going to replicate some of these analyses here.

The datasets should be pulled from the `causalsens` R package (we're not going to use any functions in this package, only the datasets). Install `causalsens` from CRAN and load it into your workspace. Load the **experimental** benchmark data using the command `data(lalonde.exp)`. This will load the `lalonde.exp` dataset into your workspace. Do the same for the `lalonde.psid` dataset (the **observational** dataset) using the command `data(lalonde.psid)`. Each of the dataset has the same set of variables:

- `age` - age in years.
- `education` - number of years of schooling.
- `black` - 1 if black, 0 otherwise.
- `hispanic` - 1 if Hispanic, 0 otherwise.
- `married` - 1 if married, 0 otherwise.
- `nodegree` - 1 if no high school degree, 0 otherwise.
- `re74` - earnings in 1974.
- `re75` - earnings in 1975.
- `re78` - earnings in 1978.

- `u74` - if unemployed in 1974, 0 otherwise.
- `u75` - 1 if unemployed in 1975, 0 otherwise.
- `treat` - 1 if treated, 0 otherwise.

1A

The causal effect of interest is the the effect of assignment to the job training program `treat` on earnings in 1978 `re78`. The variable `treat` takes on a value of 1 if the individual was assigned to the job program and 0 if they were not.

Use a simple difference-in-means estimator to estimate the average treatment effect in the **experimental** dataset. Form a 95% confidence interval for the average treatment effect.

1B

Now apply the same estimator to the **observational** dataset, `lalonde.psid`. Compute the point estimate and large-sample 95% confidence interval

1C

How does your estimated treatment effect differ between 1A and 1B and why do you think this is the case?

1D

In order for the difference-in-means estimator $E[Y_i|T_i = 1] - E[Y_i|T_i = 0]$ to be unbiased for the true causal effect, what is the key assumption that must be made about the relationship between treatment T_i and the potential outcomes $Y_i(1)$ and $Y_i(0)$? For which of the two datasets is this assumption more likely to hold and why?

1E

A common way of adjusting for observed confounders when estimating a causal effect in observational data is to fit a regression model for outcome with treatment and your confounding confounding covariates as regressors.

Run a linear regression of earnings in 1978 (`re78`) on treatment and your 10 covariates using the **observational** dataset. Assume that there are no interactions or polynomial terms in the model and that all of the covariates enter into the model additively. What is your estimate of the average treatment effect using this regression model (the coefficient on treatment)?[1]

[1]Note that the coefficient is not *quite* an estimate of the ATE. See Aronow and Samii (2015) "Does Regression Produce Representative Estimates of Causal Effects?" for a discussion of this regression weighting problem. Assuming constant effects, however, these two quantities are equal.

How does this estimated treatment effect compare to your experimental benchmark in 1A?

Problem 2 - Assessing Balance

When making causal inferences from observational data, we typically want to approximate a hypothetical randomized experiment. The goal of randomization is to generate balance between treated and control groups on potentially confounding covariates. Typically, when first approaching an observational dataset to estimate a causal effect, we will investigate the degree of balance or imbalance on relevant covariates.

2A

For each of the ten covariates, calculate the absolute standardized difference in means (that is, the difference divided by the sample standard deviation) between treated and control observations using the **observational** dataset.

What is the value of the largest absolute standardized difference? For which variable is it the largest?

2B

Now repeat 2A, but instead, use the **experimental** dataset.

What is the value of the largest absolute standardized difference? For which variable is it the largest?

2C

Now, make a plot comparing the level of imbalance in the observational dataset to the imbalance in the experimental one. Your y-axis should be the absolute standardized difference in means. For each of your ten covariates, plot the imbalance measures in the experimental data and the observational data side-by-side as two vertical columns of points. Add a line for each variable that connects the two points for each variable (to visualize the differences in imbalance between observational and experimental data).

Note: This is meant to be challenging and to get you to think about useful visual methods of conveying balance. If you are stuck on how to do this, note that your plot should look similar in style to Figure 2 from Stuart (2010) "Matching Methods for Causal Inference: A Review and a Look Forward" which you can download by going here Hint: use `type = "b"` to get R's plot command to plot both lines and points.

Problem 3 - Propensity score matching

Now we'll use matching to try to get better balance on our covariates in the observational data and try to approximate balance levels in the experimental data. First we'll use nearest neighbor propensity score matching to create a dataset matched on all ten covariates. The propensity score is defined as $\pi_i = Pr(T_i = 1|X_i)$.

3A

Estimate the propensity score for each observation (π_i) in the **observational** dataset using a logit regression model on all ten covariates (with no interactions or polynomial terms – just an additive function of the ten covariates).[2] Your outcome variable in the propensity score model should be whether or not the observation is treated. Use the model to predict $\hat{\pi}_i$ for each observation in your dataset.

Plot two histograms of the estimated propensity scores, one for the treated units and one for the control units. Make sure the x-axes of the two histograms are the same so you can compare the distributions of the propensity scores.

3B

Compare the two plots from 3A. How well do the propensity scores seem to predict whether the unit is in the treatment or control group? What does that tell you about the data?

3C

Now we're going to use matching to try to improve balance on the covariates. Start by pruning the data so as to drop all control units with propensity scores that lie outside of the support of the propensity scores for treated units. How many units did you drop?

3D

Now we're going to try to estimate the average treatment effect on treated units (ATT) by searching for appropriate counterfactuals to use for each treated observation. We'll start by doing one-to-one matching *with* replacement. Again, we are using the **observational** dataset.

Start by subsetting out all of the treated observations. Then, for each treated observation, search through the set of all controls (that you didn't prune) to find the control observation with a propensity score closest to that of the treated unit (in terms of absolute distance). Store that control observation. Combine the treated and matched control observations to make your "matched" dataset.

How many individual rows (observations) are in your matched dataset?

[2]Note, you may get a warning from R regarding possible separation in the model - for now, just ignore the warning for the purposes of this exercise. However, do take note of possible issues that can arise when fitting logit models with near-perfect separation.

3E

Create the same histograms as you did in 2A, but now using the matched dataset. Compare the distribution of propensity scores. What did matching do to the distribution?

3F

Using the same standardized difference-in-means that you used in Problem 2, calculate the imbalance in the matched dataset for each of your ten covariates. What is the largest difference-in-means that you calculated? For which variable is it the largest?

3G

Calculate a difference-in-means estimate of the treatment effect using your matched dataset.[3]

Compare the point estimate to your original estimates on the observational and experimental data from part 1. What has matching done to your estimate? Did matching alleviate the problem of unobserved confounding?

Problem 4 - Mahalanobis matching

Now we'll compare the balance changes using an alternative distance metric for matching – Mahalanobis matching.

The Mahalanobis distance between two covariate column vectors X_i and X_j is defined as

$$M(X_i, X_j) = [(X_i - X_j)^T \mathbf{S}^{-1} (X_i - X_j)]^{1/2}$$

where $\mathbf{S}$ is the sample variance-covariance matrix of X (which you can compute using `cov()` on a the matrix of covariates).

4A

Write a function that takes two vectors and the S variance-covariance matrix and returns the Mahalanobis distance between those two vectors.[4] Use it to calculate the Mahalanobis distance between observation 1 and observation 350 in the raw obserational data on the 10 covariates (that is, variables that are not treatment or outcome).

4B

We're going to again do one-to-one matching *with* replacement using the observational data, but instead of matching each using the propensity score, we're going to match using

[3]Note that standard errors for these types of matching estimators are a bit more complicated to work out so we're not having you calculate them them here. Intuitively, since we have repeated observations that are being matched, we would need to take into account some non-independence across units.

[4]Hint, you may find it useful to convert the vectors and S to matrix form in order to get R to correctly do matrix multiplication. Pay attention to what is a column and what is a row vector.

the Mahalanobis distance between the covariates of treated and control units.[5]

Start by subsetting out all of the treated observations. Then, for each treated observation, calculate the Mahalanobis distance between that unit's covariates and the the covariates of all of the control observations. Pick the control observation with the smallest Mahalanobis distance. Store that control observation. Combine the treated and matched control observations to make your "matched" dataset.

Using your matched dataset, calculate the difference-in-means estimate of the average treatment effect on the treated.

4C

Finally, we're going to compare the balance on covariates between the propensity score matched dataset from Problem 3 and the Mahalanobis matched dataset. Calculate the average of the absolute standardized differences-in-means between treated/control for each of the covariates in your propensity score matched dataset. Do the same thing for the matched Mahalanobis dataset. Based on your results, which matching metric produced better covariate balance? From what we learned in class, why do you think it worked better?

[5]Note that we're not doing any pruning before matching here (though we could via common support approaches like the convex hull).

Problem 5 - Coarsened Exact Matching

In this part you will use Coarsened Exact Matching to estimate the causal effect. You'll do this by using the `cem()` function in the `cem` package.

For now, use the default settings for generating the coarsenings. Also, we're going to allow CEM to prune both treated and control observations.
Consult the documentation for more details about how to use the `cem()` function.

5A

Use CEM to prune the **observational** dataset using the default coarsening settings.

How many treated and control observations. observations does your matched dataset contain after you prune with CEM?

5B

Estimate the treatment effect by using a difference-in-means estimate from your CEM-matched data using the `att()` function. Report the point estimate of the treatment effect.

5C

Let's assess the improvement in balance from the CEM match in 5A. Take your matched dataset and calculate the weighted absolute standardized difference in means across all of your ten covariates (using the observation weights provided by `cem()`).**Hint:** Use the `weighted.mean()` function to calculate a weighted average.

What is the average weighted absolute standardized difference in means? How does this compare to the propensity score and Mahalanobis distance matches?

R Code

Please submit all your code for this assignment as a .R file. Your code should be clean, commented, and executable without error.