PHP 2610 - HW 2
Due Friday December 21 at 5pm

Please work alone on this assignment. If you have questions, please email me or set an appointment to see me in my office. I will be around Friday December 14 and all of next week.

Objectives

- To use matching estimators to estimate causal treatment effect
- To use G estimation to estimate causal treatment effect

Exercises
Please use the dataset `lalonde`. It contains the following variables

```
# outcome variable:   r78 (real earnings in 1978)
# treatment varible:  treat (1=yes, 0=no; job training program)
# age:  age in years
# educ: education in years
# black:  1 if black, 0 if not
# hispan:   1 if latino/a, 0 if not
# married:  1 if yes, 0 if no
# nodegree:  1 if no college degree, 0 if not
# re74:  real earnings in 1974
# re75:  real earnings in 1975
```

This dataset is included in the `MatchIt` package. To load it, you just need to issue these commands:

```
library(MatchIt)
data(lalonde)
```

Once it's loaded you will have a dataset called `lalonde` containing the variables listed above.

These data come from a paper by Lalonde (1986), which sought to compare the results of an observational study to those from a randomized trial. The original randomized trial was a randomized study of worker training (treatment) on real earnings in 1978. The trial found that job training increased earnings by about $1800. The dataset above contains data from the *treatment group* of the original study, but has controls from two different surveys. Hence the dataset as it is constructed here can be viewed as an observational study where we have a sample of individuals who received the treatment of interest, and a pool of controls from which to match or otherwise use for comparison. Variables other than the treatment and outcome variable `re78` should be considered potential confounders. The idea here is to see whether causal inference methods can be used to reproduce the findings from the randomized study.

1. Let Y denote real income in 1978 and let T denote treatment group. Fit the model

$$E(Y \mid T) = \beta_0 + \beta_1 T$$

(a) Report the estimates of β_0 and β_1.

(b) What does the coefficient β_1 represent?

(c) Can it be interpreted as a causal effect? Why or why not?

2. Use propensity score matching to estimate the causal effect of job training. Select and justify the propensity score model and the method of matching that you ultimately decide to use. Please hand in the following

   (a) A description of the propensity score method, matching method, and analysis method that you use to estimate the causal effect. A few sentences is fine here.

   (b) The chunk of R code (or other code) that performs the matching and carries out the analysis. Not the output here, just the code.

   (c) A table that shows the numbers matched and not matched, and a summary of covariate distributions in the treated and control groups. For continuous variables, the summary could be (n, mean, standard deviation), or it could be (n, median, quantiles). For binary variables it should just be n and proportion. Please no graphs for this one.

   (d) Output from the regression model or analysis method that you use to estimate the causal effect.

   (e) Report of the estimated causal effect, its standard error, and an interpretation of what the causal effect represents.

3. Use G estimation to estimate the effect of job training among those who received job training (effect of treatment among the treated). For this, the idea is to calculate the mean of $Y_1 - Y_0$ over the distribution of confounders in the group where $T = 1$. In terms of a formula, this will require you to calculate

$$(1/n_1) \sum_{i:T_i=1} (\widehat{Y}_{1i} - \widehat{Y}_{0i}),$$

   where n_1 is the number of individuals in the treatment arm, $\widehat{Y}_{1i}$ is the predicted outcome of Y_1 for person i, and $\widehat{Y}_{0i}$ is the predicted outcome of Y_0 for person i.

   (a) Denote the potential confounders by V. Fit this model to the data

$$E(Y \mid V, T = 1) \quad = \quad \text{regression of } Y \text{ on } V \text{ among treated}$$

   Describe the method you use to select the regression model.

   What to hand in for this: For each model, the chunk of R code that you used to fit the model (no intermediate output please), a regression table (this can be copy/paste of the R output), and a residual-versus-fitted plot.

   (b) Use the model to generate individual-level predictions of Y_1 and Y_0 for individuals in the *treatment arm*.

   What to hand in for this: A line listing of covariates and predicted outcomes for 10 observations in the dataset.

(c) Calculate the causal effect.

What to hand in for this: The chunk of R code that you used to calculate the causal effect, and the estimated value of the causal effect.

(d) Bonus item #1 (not required): Use the bootstrap to calculate standard error.

(e) Bonus item #2 (not required): This question asks you to estimate the average treatment effect among the treated. Use the method of your choice to estimate the *average treatment effect*.