Problem Set 1
MGMT 737

Spring 2023 Due: Tuesday, January 1

1. Randomization. This analysis will use the Dehijia and Wahba sample from the Lalonde dataset of the NSW experiment. The dataset is `lalonde_nsw.csv`. The outcome variable is `re78` (real earnings in 1978). The treatment indicator is `treat`. The remaining variables are potential covariates. Assume for the purposes of this problem set that `treat` is completely randomly assigned.

   (a) Calculate the average treatment effect of the policy $E(\tau_i)$ using a simple difference in means.

   (b) Calculate the average treatment effect on the treated of the policy $E(\tau_i|\text{treat} = 1)$. How does it compare to part a?

   (c) Test the null of $\tau_i = 0$ for all i using a randomization test.
   N.B. Hold fixed the number of treated and control (e.g. assume the treatment count would be held fixed) and permute the labels randomly 1000 times – you do not need to fully do every permutation (there would be too many). Report the quantile that your estimate from the previous question falls.

   (d) Run a regression using robust standard errors (you may use canned software) of the outcome on the treatment dummy, and compare the p-values from this test to the previous answer.

2. Propensity Scores. This analysis will use the dataset from Problem 1 as well as the PSID dataset from Dehijia and Wahba, `lalonde_psid.csv`. These datasets have identical variables. The new dataset is a sample of observations from the Panel Survey of Income Dynamics that can be used as alternative control observations. Importantly, these observations were not in the initial randomization.

   (a) Using `age`, `education`, `hispanic`, `black`, `married`, `nodegree`, `RE74` and `RE75`, construct a propensity score using the *treated* group in `lalonde_nsw.csv` and the control sample of `lalonde_psid.csv`. Use a logit regression model to do so (you may use a canned routine to run the regression). Report the average p-score for the treated and control samples, and plot the propensity score densities for the treatment and control groups.

   (b) Using your p-score estimates, calculate the IPW and SIPW estimate for control and treated mean of the outcome, and the average treatment effect. Contrast these estimates to the control mean of the outcome from the NSW sample, and the treatment effect from last week's problem set.

   (c) Compare the ATE in the previous question to the treatment effect estimated using a linear regression using the PSID and NSW treatment sample, with `age`, `education`, `hispanic`, `black`, `married`, `nodegree`, `RE74` and `RE75` as controls.

   (d) Now revisit your estimates from part a and b, and following Crump et al. (2009), discard all units with estimated propensities outside the range of $[0.1, 0.9]$. Reestimate the IPW and SIPW estimator of the ATE from part b using this trimmed sample.

   (e) Finally, calculate the IPW and SIPW estimates for the ATE using this trimmed sample for Black and non-Black individuals. Compare this estimate to the ATE for Black and non-Black individuals using the full randomized sample.