Political Science 207
Problem Set 3

Professor: Daniel Masterson

Due Thursday, April 29th, at 5:00 p.m.

Submit your completed assignment as a single PDF or HTML file (including your write-up and all R code) on GauchoSpace under the assignment for Week 3. I strongly recommend that you write your problem sets in R Markdown. (I recommend this tutorial.)

Make sure you follow good coding style (see tutorial in link), and show your work for problems that require calculations.

You are encouraged to work in groups, but you should write up the problem set alone, and you should note at the top of your problem set who you worked with.

For this assignment, we will analyze data from the National Supported Work Demonstration, a subsidized work program implemented in the mid-70's. You can read more about it here.

The data set contains an experimental sample from a randomized evaluation of the NSW program (`nsw_exper.dta`) and a non-experimental sample from the Population Survey of Income Dynamics (`nsw_psid_withtreated.dta`). In both datasets, the variable `nsw` is the treatment, the variables `re78` and `u78` are outcomes, and the rest are covariates.

Variable Definitions:	
`nsw`	=1 for NSW participants, =0 otherwise
`age`	age in years
`educ`	years of education
`black`	=1 if African American, =0 otherwise
`hispanic`	=1 if Hispanic, =0 otherwise
`married`	=1 if married, =0 otherwise
`re74`	real (inflation adjusted) earnings for 1974
`re75`	real (inflation adjusted) earnings for 1975
`re78`	real (inflation adjusted) earnings for 1978
`u74`	=1 if unemployed in 1974, =0 otherwise
`u75`	=1 if unemployed in 1975, =0 otherwise
`u78`	=1 if unemployed in 1978, =0 otherwise

1) **Experimental Baseline**

   a) Using the **experimental** data (`nsw_exper.dta`), obtain an unbiased estimate of the ATE of NSW on 1978 earnings, its SE and a 95% confidence interval. Make sure to use heteroskedasticity-consistent standard errors.

   b) Using the experimental data, use OLS to estimate the ATE controlling for age, education, race/ethnicity, marital status, and employment and earnings in 1974 and 1975. Report the estimate and its SE and compare it to the one obtained in 1a), and explain how they compare with each other and why.

2) **Observational Data: Regression**

   a) File `nsw_psid_withtreated.dta` contains treated units taken from the experiment, but with control units replaced by the non-experimental sample from the PSID. We will refer to this file as the **non-experimental** dataset. Check and report the covariate balance in the non-experimental dataset using appropriate statistics. How do the treatment and control group differ? Among the observed covariates, what seem to be the most important factors that determine selection into the program (the "treatment")?

   b) Using the non-experimental data, compute the (naive) estimate of the ATE of the program on 1978 earnings *without* adjusting for any of the covariates. Report the estimate and a standard error.

   c) Still using the non-experimental data, use OLS to estimate the ATE controlling for age, education, race/ethnicity, marital status, and employment and earnings in 1974 and 1975. Report the estimate and its SE and compare these results to those in 2b). Did your point estimate change? Why?

   d) When would it be the case that the ATT is not identified (and hence cannot be computed)? (hint: think about what variation on the treatment indicator exists within each strata/subgroup)

3) **Observational Data: Matching**

   a) Install and load the `Matching` package. With the non-experimental data (`nsw_psid_withtreated.dta`), use the following covariates to match on: "age", "educ", "black", "hisp", "married", "re74", "re75", "u74", and "u75". Use the matching approach, with one match per treated unit and using normalized Euclidean distance (default in the `Match()` function) as your distance metric, to estimate the ATT–the average effect of the program (`nsw`) on earnings (`re78`) for those who are treated.

   The syntax for the `Match` function is as follows: `Match(Y, Tr, X)`, where `Y` is the outcome variable, `Tr` is the treatment variable, and `X` is the set of covariates to match on. Make sure to include bias adjustment (`BiasAdjust = T`).

   b) Using the matched data, report the balance statistics for the covariates that you matched on. How does the balance look in the matched data compared to the balance in the full non-experimental data set?

c) Re-estimate the ATT using the same matching approach as in 3a), except do *not* include bias adjustment (`BiasAdjust = F`), and compare the results. What is the importance of the bias adjustment? When is it most important to use the bias adjustment?

d) Compare the number of treated and untreated units in the non-experimental data. Comment on whether it is good or bad in terms of estimating the ATT.

e) Now, match again on the same covariates, but this time use genetic matching. Make sure to install and load the `rgenoud` package. (You may want to read up on genetic matching at: http://sekhon.berkeley.edu/papers/GenMatch.pdf. Use `GenMatch` to obtain the weight matrix, which you then pass to `Match` as the "`Weight.matrix`" parameter.) Report your estimate of the ATT and its standard error.

4) **Matching with Propensity Scores**

a) Estimate propensity scores using a logistic regression with the non-experimental data (the full set, not the matched set). Plot the distributions of propensity scores for treated and control groups and comment on the overlap.

b) Now, use the experimental data and estimate propensity scores using the same model. Again, plot and compare the distributions of the propensity scores for treated and control groups here. What do you observe?

c) Back to the non-experimental dataset: remove observations that have propensity scores lower than 0.1. Then report balance statistics for the trimmed data. How does the balance look in this trimmed dataset compared to the balance in the full dataset? Do results differ? Why or why not?

d) Now, match on the estimated propensity scores from 4c) to estimate the ATT. Report your point estimate and standard error.

e) How closely were you able to replicate the experimental results using matching estimators with the non-experimental data? Which version of your matching estimator seemed to perform the best?