Homework 3 Due Wednesday April 25th

Problem 1

In this problem and the next, we will analyze data from the National Supported Work Demonstration, a subsidized work program implemented in the mid-70's. The data set contains an experimental sample from a randomized evaluation of the NSW program (`nsw_exper.dta`), and also a non-experimental sample from the Population Survey of Income Dynamics (PSID) (`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

(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.

(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 (a), and explain how they compare with each other and why.

(c) File `nsw_psid_withtreated.dta` contains treated units taken from the experiment, but control units replaced by the non-experimental sample from the PSID. We will refer to this file as the non-experimental dataset. Check the covariate balance in the non-experimental dataset. Decide on a few sensible balance statistics and report them in a table. 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")?

(d) Using the non-experimental data, estimate the (naive) ATE of the program on 1978 earnings without

adjusting for any of the covariates. Report the estimate and a standard error.

(e) 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 part (d). Did your point estimate change? Why?

(f) Using the non-experimental data, condition (only) on the marital status of individuals, and manually compute the subclassification estimator of the ATT.

(g) Now write your own function in R that takes a dataset, a dependent variable, a treatment, and a vector of discrete covariates, and then computes and returns the subclassification point estimate of the ATT. What is the estimated ATT when conditioning on marriage status, race, unemployment status in 1975, and reported earnings in 1975? For this question, you should make earnings discrete by binning it by tercile (i.e. make it a discrete variable with 3 levels, which indicate the tercile).

(h) When would it be the case that the ATT is not identified (and hence cannot be computed)? If you have not already done so, build into the function you just wrote a safeguard whereby the function will break and return a warning message in such a case. Test this by conditioning on education (including all levels of education in the data) in addition to the variables you conditioned on in part (g).

Problem 2

In this problem, we will continue to use the non-experimental data (`nsw_psid_withtreated.dta`) and implement matching estimators.

(a) 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. Include bias adjustment.

(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 part (a) except do not include bias adjustment, 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 to untreated units. Comment on the result, and whether it is good

or bad in your view.

(e) Now, match again on the same covariates, but this time use genetic matching. Report your estimate of the ATT and its standard error. 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`.

(f) 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.

(g) 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? Compare your results with part (f).

(h) Back to the non-experimental dataset: trim off 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?

(i) Now, match on the estimated propensity scores from part (f) to estimate the ATT. Report your point estimate and standard error.

(j) 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?

Problem 3

The curse of dimensionality makes it difficult to work in high-dimensional covariate spaces. Here we will see for ourselves what happens as the number of covariates grows large. Suppose you have covariates $X_1, X_2, \ldots, X_P$ where P is the number of dimensions of your data (i.e. the number of covariates). Each observation's x_i will be a (column) vector describing the covariate values for observation i. That is, $x_i = [x_{i,1}, x_{i,2}, \ldots x_{i,P}]^T$.

a) Write a mathematical expression that gives the Euclidean distance between observations i and j, with covariates x_i and x_j respectively.

b) Suppose you have many covariates, each one uniformly distributed on the interval $[0, 1]$. You want to draw a square (or cube, or hypercube) in the covariate space that in expectation would capture 5% of the observations. If you have a 2-dimensional covariate space, how large would the sides of the square have to be to include 5% of the observations? Now consider a 3-dimensional covariate space. Again, if you want to capture 5% of the observations, how long would the side of that cube

have to be?

c) Derive a general expression that gives the side-length, r, that you would need for a (hyper)cube to capture a percentage of the data, v, when the number of dimensions is p. Create a graph that shows the r needed to capture $v = 0.05$ of the total observations as the number of dimensions goes from 1 to 20. What do you notice? What does this tell you about matching in high-dimensional spaces?

d) Create a simulated dataset, X. It should have 100 observations, and 20 dimensions. Each dimension of X should be drawn from $unif(-1, 1)$. We will use this as the base dataset, but will not always use all of the dimensions. Consider a single point, $x^\star$, whose covariates are all equal to zero. You want to find this point's closest neighbor in the dataset.

   (i) Start with a one dimensional version of X, using just the first dimension of the simulated dataset that you made. Find the point in the dataset i whose covariate value x_i is closest to $x^\star$ in a Euclidean sense. Record the (Euclidean) distance from this point x_i to $x^\star$,

   (ii) Repeat the step above, but using 2 dimensions, then 3, and so on up to all 20 dimensions. In each case, get the Euclidean distance between the x_i that was the nearest neighbor and $x^\star$.

   (iii) Plot the Euclidean distance as a function of the number of dimensions.

   (iv) What do you conclude from this? What does this teach us about matching?