Problem Set 5

Professor: Christopher Lucas

Due Monday, February 24

You are required to use either `knitr` or `Markdown`, these are fully compatible with `R` and LaTeX.

Please (1) bring a printed version of your answers to class and (2) upload an electronic version on Canvas. Hard copies of your solutions should be handed in at the start of class and electronic copies are due before class begins at 4:00 pm. Late submissions will not be accepted. Please, comment our code and show your work in questions that require calculations.

Please set your seed using set.seed(02139) when the question requires simulations.

1 Curse of Dimensionality

The *curse of dimensionality* makes it difficult to work in a situation where there are many pre-treatment covariates to condition on. Suppose we have covariates X_k for $k = 1, \ldots, P$, where P is the number of pre-treatment covariates (i.e., the dimensionality of the covariate space). Then x_i is a vector of covariate values for observation i, $x_i = [x_{i1}, \ldots, x_{iP}]^T$.

(a) Write an expression that gives the Euclidean distance between observations i and j in terms of their covariates x_i and x_j, respectively.

(b) Create a dataset X with 500 observations and 20 covariates, where each covariate should be independently drawn from $\mathcal{N}(0, 1)$. Now, consider a new observation with covariates $x^\star$ all equal to zero. Find the new observation's nearest neighbor in the dataset you created, using your expression for Euclidean distance and only the first covariate. That is, find the point i in the dataset whose first covariate x_{1i} is closest to $x_1^\star$. Record the Euclidean distance between x_{1i} and $x_1^\star$. Repeat this process, each time adding an additional covariate: next using two covariates, then three, and so on through all 20. Use your results to plot the Euclidean distance to the new observation's nearest neighbor as a function of dimensionality.

(c) What do your results demonstrate about matching?

2 Matching

We will now 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.dta`). In both datasets, the variable `nsw` is the treatment, the variables `re78` and `u78` are outcomes,

and the rest are covariates:[1]

Variable Definitions:

`nsw`	=1 for NSW participants, =0 otherwise
`age`	age in years
`educ`	years of education
`black`	=1 if African American, =0 otherwise
`hisp`	=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, obtain an unbiased estimate of the effect of NSW on 1978 earnings and its standard error. Then estimate this effect again using a linear regression that controls for age, education, race, ethnicity, marital status, employment in 1974 and earnings in 1974. Compare these two estimates and comment.

b) Now let's use the non experimental data. File `nsw_psid_withtreated.dta` contains treated units taken from the experiment, but control units are replaced by the non-experimental sample from the PSID. Calculate the (naive) ATE of employment program on trainee's by the same two methods you used in 1(a) (controlling for the same covariates). Briefly but concretely describe what are you estimating? Do these methods recover the experimental results? Why or why not?

c) Using the non experimental dataset, check covariate balance in the unmatched dataset using the `Matching` package's `MatchBalance()` function, for all covariates. Your output should be in the form of a balance table. Make sure to present statistical tests of the similarity of means and similarity of distributions. Based on your table, which of the observed covariates seem to be the most important factors in selection into the program?

d) Estimate propensity scores using logistic regression for both the experimental and non-experimental data. Report the distributions of propensity scores for treated and control groups. Comment on the overlap for both data sets. How do they differ and why?

[1]You may be interested in reading up on this very famous experiment in `http://isites.harvard.edu/fs/docs/icb.topic1141086.files/Lalonde\%201986\%20-\%20Evaluating\%20the\%20Economietric\%20Evaluations\%20of\%20Training\%20Programs\%20with\%20Experimental\%20Data.pdf`LaLonde (1986), or in `http://isites.harvard.edu/fs/docs/icb.topic1311568.files/dehejia_wahba\%201999.pdf`Dehejia and Wahba (1999), or more specifically on matching in `http://www.sciencedirect.com/science/article/pii/S030440760400082X`Smith and Todd (2003)

e) Choose some covariates on which to match, and then do so using the `Match()` function. Briefly justify your choice of covariates. Be sure to carefully check the options available to you in this function. For now, find only one match for each treated unit, use the Mahalanobis distance metric to determine weights, and do not (yet) use exact matching. Apply the matching estimator to estimate the average effect of the employment program on trainee earnings i.e., the ATT. Report your estimate and standard error, as well as balance statistics for the matched data.

f) Re-estimate the ATT using exact matching on education, race, ethnicity and married. Report your estimate, its standard error, and produce a balance table as before. In general, do your results differ from previous results? Why or why not?

g) Use `Match()` to estimate the average effect of the treatment on the treated for the NSW program using bias-corrected matching estimators and $M = 1, 4$, and 10 matches (and Mahalanobis distance). Repeat the analysis using matching estimators that are not bias-corrected. In this part you should get 6 estimates with corresponding standard errors. What kind of differences do you observe across different M sizes and across the bias-corrected and non-bias corrected estimates you produced? What accounts for them? Which results would you trust more and why?

h) Now let's use the propensity scores we calculated before to match on the estimated propensity scores (from part 3(d)) using Match() to obtain an estimator of the average treatment effect on the treated for the NSW program.

i) Use the weighting on the propensity score to estimate the average effect of the treatment on the treated for the NSW program. Do a bootstrap to get a standard error of the estimate (be sure to bootstrap the entire process, including the estimation of the propensity scores).

j) Now match again, but this time use genetic matching. 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`. Comment on your results. Which type of matching do you prefer?

k) Consider the work you've done on DAGs - Part 2 from last week's problem set. Under what assumptions is the ATT you estimated identified? Does matching make any identification assumption more plausible? Discuss.