PS C236A / Stat C239A
Problem Set 5
Due: Nov. 12, 2012

Instructions

This assignment is due **4 pm Monday, Nov. 12.** You may submit your analytical work either electronically or in paper form. Electronic versions must be sent as a .pdf to <jahenderson[at]berkeley.edu>. Paper copies should be placed in my mailbox in 210 Barrows. For the computing portion of the assignment, you must submit a fully executable version of all .R code, along with any data used in the code (excepting that provided through the course webpage) to the email above. All files for each assignment sent electronically should be included in one omnibus email, with the subject line containing the course and homework number, and your last name (e.g., PS239A/STAT236A: HW5 - Pelosi).

You are encouraged to work together in groups to complete the assignments. However, you must hand in your own individual answers. Photocopies and other reproductions of someone else's answers are not acceptable. Please also list the names of everyone with whom you have collaborated on this assignment.

Problem 1 An alternative distance metric to the propensity score is Mahalanobis distance. This metric reduces the multidimensional problem of multivariate matching to a unidimensional problem. Although Mahalanobis distance was originally developed for use with multivariate Normal data, we often encounter covariates that are not normally distributed. This problem will explore the implications of these non-normal variables on this distance metric.

a. When including a binary variable in a Mahalanobis distance metric, will a binary variable with $p = 1/2$ or a binary variable with p near zero be given greater weight by this distance metric? Prove why this is true mathematically.

b. How will this distance metric treat covariates with outliers? How about covariates that have long-tailed distributions?

c. Should we or shouldn't we be concerned by the behavior of the Mahalanobis distance metric for the covariate distributions described in parts (a) and (b)? Why?

Problem 2 For this problem, you will critique "The Fox News Effect: Media Bias and Voting", by Stefano Della Vigna and Ethan Kaplan. The paper can be found here: http://sekhon.berkeley.edu/causalinf/papers/DellaVignaFoxNews.pdf. Please write a page or two addressing the following questions:

a. Describe and discuss the identification strategy of the paper. What are the weaknesses? What parts do you find convincing?

b. Explain the importance of section III.A in the article? Would you do it any differently?

c. Perform the following thought experiment: hold the estimation procedure in section III.B constant, and assuming that you have access to all existing data in the US, what data would you include to the improve the validity of the estimates? Now do the reverse. Holding the data constant, discuss what parts you would change and add to the estimation procedures to increase confidence in the validity of the results.

d. Overall, are you convinced that their conclusions are correct?

Problem 3 This question will analyze the following dataset: `http://sekhon.berkeley.edu/causalinf/data/cross_section_wfl.csv`. The variables are described in the following file: `http://sekhon.berkeley.edu/causalinf/data/codebook.water.txt`

The dataset has 435 observations and was used in the article "Water for Life: The Impact of the Privatization of Water Services on Child Mortality", by S. Galiani, P. Gertler, and E. Schargrodsky (2005, Journal of Political Economy, volume 113). The paper is here: `http://sekhon.berkeley.edu/causalinf/papers/GalianiWater.pdf`.

The units of observation are municipalities in Argentina, and the treatment under study is the privatization of municipal water services. All 435 municipalities in this sample had public water services in the year 1990, but by the year 1999, 123 municipalities had privatized their water services. Of the 123 municipalities which privatized between 1990 and 1999, 83 municipalities did so between 1998 and 1999. The original panel structure of the dataset has been simplified to a cross-section: for each municipality, the dataset you will be working with has the covariates for each year between 1990 and 1999.

The treatment indicator has been defined as equal to one if the municipality privatized its water services sometime between 1991 and 1999, and equal to zero if a municipality whose water services were public in 1990 never privatized between 1991 and 1999. The outcomes of interest are total child mortality and child mortality from infectious parasitic diseases, i.e., water-borne diseases. Perinatal mortality is also of interest for the reasons discussed in Galiani, Gertler, and Schargrodsky (2005).

For parts (b) - (e) below, be sure to explicitly set seeds to ensure that GenMatch recovers reproducible results, i.e. `set.seed` in general, and in GenMatch `unif.seed`, `int.seed`.

a. Select a set of covariates to condition on. Be sure to include higher order terms and interactions you think are appropriate. Using these variables, perform Mahalanobis distance matching, with the average treatment effect for the treated (ATT) as your estimand. Report your balance statistics before and after matching using `MatchBalance`.

b. With the same set of covariates, use GenMatch to generate weights that optimize balance. Use the default setting for the loss function, but feel free to adjust other parameters of the function. (Make sure you **do not** drop any treated units in matching here however.) Present balance before and after matching using `MatchBalance`. Produce two QQ-Plots illustrating improvement in balance on one important continuous covariate before and after matching.

c. Estimate the ATT of privatizing water services between 1991 and 1999 on child mortality in 1999, using your matched data from part (b). Is the ATT estimate statistically different from zero? Plot the density of the unit effects of treatment on the treated municipalities. Does your interpretation of the effect of privatization change when examining average versus unit effects?

d. Using the same covariates from (a) and (b), rerun GenMatch, this time dropping at most 10% of the treated units. Does your balance improve with respect to the balance you found in part (b)? Now what is the ATT of privatization in this matched data? Is the ATT significantly different from zero?

e. Now find the best balance with GenMatch using your own loss function. In doing so, retain every other specification you used in GenMatch in part (d). Explain the logic behind your choice of loss function. (An example loss function would maximize the median p-value from a vector of `t.test` and `ks.test` results). In your loss function, you may want to prioritize important selection variables, for instance pre-treatment mortality rates. Present balance statistics after matching using `MatchBalance`.

f. Overall, how do your results differ from those in Galiani, Gertler, and Schargrodsky (2005)? In particular, are your results in part (b) and (c), and their published findings comparable?