Problem Set 8

1. Suppose we're interested in the effects of the generosity of public assistance on labor supply. Historically, U.S. states have offered widely-varying welfare payments to poor unmarried mothers. Labor economists have long been interested in the effects of such income maintenance policies - how much of an increase in living standards they facilitate, and whether they make work less attractive. Could there be any potential problems using a DiD approach to look at this question? (Hint: people can move...) How about the example we covered in class looking at the employment of teenagers?

2. Use the data in 'eitc.dta' for this question (you'll have to use the function read.dta(), in the 'foreign' library, to get it into R). The Earned Income Tax Credit (EITC) is a refundable tax credit for low income working individuals, particularly those with children. A major change to the EITC occurred in 1993. In his first State of the Union Address, President Clinton said, "The new direction I propose will make this solemn, simple commitment: By expanding the refundable earned income tax credit, we will make history; we will reward the work of millions of working poor Americans by realizing the principle that if you work forty hours a week and you've got a child in the house, you will no longer be in poverty." The main result of this change was that single, working, women with children get more money from the government. The dataset contains information on single US women around the time of the expansion of EITC.

   (i) Create a dummy variable to indicate post-treatment (assume that 1993 counts as pre-treatment). Also create a dummy variable to indicate whether a woman was 'treated'.

   (ii) Compute the four averages that you need in order to calculate the DiD estimate of whether the expansion of EITC has lead single mothers to work more. What is the estimate, how do you interpret it, and is this what you expected? Can you say anything about its statistical significance at this point?

   (iii) Answer the same question as part (ii) but this time, use a regression approach. Has this given you any more information?

(iv) Include the following controls into your regression from part (iii): *nonwhite*, *age*, *ed* (where *ed* denotes education). What do you notice? What do you conclude about the causal interpretation from part (iii)?

(v) Finally, to the regression in part (iv), add unearned income to your regression, *unearn* (this is income that you get from anything other than working). Now what happens? Why?

3. The National Supported Work Demonstration (NSW) was a temporary employment program in the USA designed to help disadvantaged workers lacking basic job skills move into the labor market by giving them work experience and counseling in a sheltered environment. Unlike other federally sponsored employment and training programs, the NSW program assigned qualified applicants to training positions randomly. Those assigned to the treatment group received all the benefits of the NSW program, while those assigned to the control group were left to fend for themselves.
There is a long standing debate about whether social programs or other interventions can be evaluated without use of data from a randomized control experiment. Lalonde (1986) used the NSW data to investigate the differences between experimental and observational data. The experimental dataset is given by those who were treated as the treatment group and those who were eligible but not treated as the control group. He created an observational dataset by taking those who were treated as the treatment group, and then taking a random sample of the normal population (i.e. not those who were eligible) as the control group. So there is now definitely a selection process to be treated. But this is usually how policies work - they are designed to be given to a certain subset of the population based on some 'qualifications'.
The experiment happened between 1975 and 1977. So, data from 1974 and 1975 are pre-treatment, and data from 1978 are pos-treatment.

(i) Using the data in 'Lalonde_experimental', test if the means of each explanatory variable (i.e. excluding the outcome variable $earn78$, and the treatment itself) differ between treatment and control group. (Hint: you can do this by running a simple regression...) Why are these tests helpful in establishing the credibility of the experiment?

(ii) Estimate the treatment effect from the experiment using the outcome $earn78$ (earnings in 1978). Do this with the extra regressors in the model and with them out of the model. Should the treatment effects be significantly different between these two specifications? What happens to the R-squared in the second regression? Why does this matter?

(iii) Now, instead of using the experimental data we will use the observational data, it is in 'Lalonde_observational'. To be able to use the regression approach to get our DiD estimate we need to manipulate our

data... create two datasets that are just a copy of the observational data; one will be the before treatment data, the other will be the after treatment data (label the two identical datasets appropriately). In each dataset create a new variable called 'outcome', for the before treatment dataset, this will be $earn75$, for the after treatment dataset, this will be $earn78$. Also construct a dummy variable for each dataset which indicates if the data is the before treatment data or the after treatment (call this new variable post - it'll take 1 for the after data and 0 for the before data). Finally, combine these two datasets using the rbind() function.
You are now in a position to run your DiD regression. Ensure the result you get is the same as simply calculating the mean outcome in each case and taking the difference of the differences. How do these results compare to the experimental results from part (ii)? (Do not worry about trying to explain what is going on; that's for the next couple of questions!)

(iv) To construct this observational sample, Lalonde tried to pick a comparable group of individuals as the control group. What would he do to test that? Carry out this test yourself, what do you find? Is this a problem for our analysis. (Hint: Think carefully about what assumptions we need for our DiD analysis to have a causal interpretation.)

(v) Compare the difference in the pre-training incomes by constructing a difference between $earn74$ and $earn75$. Do the same comparison of means for this variable (like you did in part (i)). What do you find? And why are we doing this?

4. Find an academic article that uses a difference-in-differences approach. The easiest way to do this is through google scholar. Search for 'difference in difference' (but make sure you put it in quote marks - this tells google to look for that exact phrase), and use the advanced search option to look in specific economics journals. I recommend the following journals: American Economic Review, Review of Economic Studies, Economic Journal, American Economic Journal: Applied Economics, Review of Economics and Statistics, Journal of Applied Econometrics, Journal of the European Economic Association.) Discuss the question they want to answer, what they find, and any criticisms you have.

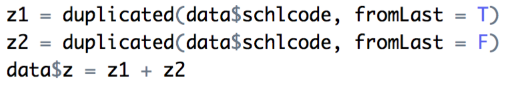
5. 'Class_size' contains data on class size, and average math and verbal test scores for 2 019 5th grade classes in 1,002 public schools in Israel, as well as enrollment data for these schools and percentage disadvantaged pupils. In Israel, schools face a rule which states that classes cannot be larger than 40 pupils. When enrollment is 41, schools are supposed to open a second classroom, and then open a third classroom at 81 pupils, etc. This causes discontinuous drops in class size at multiples of 40.

   (i) Estimate the effect of class size on math scores using OLS without any controls. Now add the percentage of disadvantaged students in the class and enrollment as controls (enrollment is the total size of students in that year in that school). Are these results what you expected, both in the naive OLS and with the controls?

(ii) Limit the sample to schools with enrollment between 20 and 60 students and generate a ”large cohort” dummy based on the first discontinuity at 40 students, i.e. create a dummy which equals 1 if enrollment is above 40. Estimate a sharp RDD for the effect of being in a larger class on math scores. (Hint: think carefully about what your running variable is and where your cutoff point is along it). What do you conclude? Now also control for the percentage of disadvantaged students in the class (to do this, you need to include the option covar=disadv in the rdd_data() function, and also include the option covariates="disadv" in the rdd_reg_lm() function). What do you conclude from a comparison of the two results? (Hint: to answer this fully, I recommend calculating the standard deviation of the math score).

(iii) Repeat the first RDD estimation from part (ii), i.e. without the control, but this time, use a second order polynomial to pick up any nonlinearities. Now try with a third order and a fourth order polynomial. Now try a nonparametric approach. Which model do you have most faith in?

(iv) It turns out that not every school who crosses the 40 student threshold splits their class, and it also happens that some schools who don't cross the 40 student threshold do split their class. With this in mind, the following code creates a dummy variable, z, for whether a particular school appears twice in our sample.



```
z1 = duplicated(data$schlcode, fromLast = T)
z2 = duplicated(data$schlcode, fromLast = F)
data$z = z1 + z2
```

Why do we need this variable, and what are we going to do with it? To perform the analysis that I'm cryptically describing here, you will need to include the option z=data$z into the rdd_data() function. Describe the result that you get and compare it to your previous estimates.