Econ 184b, Econometrics, Assignment 7

The raw .rmd file for this assignment are available at: https://raw.githubusercontent.com/f1kidd/Econ184/main/Assignment/Assignment_7.Rmd

Note: For the math problems, you can either (1) type your solution and compile it with RMarkdown; (2) type your solution use another word-processing software and convert it into a PDF file; (3) write your solution on paper, scan it and make it into a legible PDF file. TAs can decide, at their discretion, that the submitted file is illegible and thus give zero credit to a question or the entire problem set. If you type your solutions (options 1 and 2 above), you will get **a bonus equal to 5% of your performance** on the problem set. No credit will be given if you only report the final answers without showing intermediate steps whenever appropriate.

For the programming problems, you need to submit both the compiled pdf/html file and the RMarkdown (.rmd) file on LATTE. You will get **an additional bonus equal to 5% of your performance** if your RMarkdown file is completely reproducible with minimal altercation (install packages, etc.). That is, our team (or anyone else) should be able to recompile your Rmarkdown file and reach *the same* result. You need to explicitly write out your answers - just showing programming outputs will receive zero credit.

You will be receiving a total of 15% bonus if you submit one single pdf/html compiled directly from Rmarkdown, along with the .rmd file.

Question 1:

Suppose you want to test whether girls who attend a girls' high school do better in math than girls who attend coeducation schools. You have a random sample of senior high school girls from a state in the United States, and *score* is the score on a standardized math test. Let *girlhs* be a dummy variable indicating whether a student attends a girls' high school.

a. What other factors would you control for in the equation? (You should be able to reasonably collect data on these factors.)
b. Write an equation relating score to *girlhs* and the other factors you listed in part a.
c. Suppose that parental support and motivation are unmeasured factors in the error term in part b. Are these likely to be correlated with *girlhs*? Explain.
d. Discuss the assumptions needed for the number of girls' high schools within a 20-mile radius of a girl's home to be a valid IV for *girlhs*.
e. Suppose that, when you estimate the reduced form for *girlshs*, you find that the coefficient on *numghs* (the number of girls' high schools within a 20-mile radius) is negative and statistically significant. Would you feel comfortable proceeding with IV estimation where numghs is used as an IV for girlshs? Explain.

Question 2: Simultaneous equations models: using instrumental variables

This exercise replicates a well-known paper by Steven Levitt ("The Effect of Prison Population Size on Crime Rates: Evidence from Prison Overcrowding Litigation," Quarterly Journal of Economics, 1996) to illustrate the use and estimation of simultaneous equations models with panel data. A subset of the data set (prison.dta/prison.csv) is posted on Latte, along with the original article if you are interested in reading it.

The paper answers the following question: do increased incarceration rates reduce violent crimes? Levitt considers the following fixed effects model to answer this question:

$$ln(crime_{it}) = \beta_1 ln(prison_{it}) + \alpha_i + \lambda_t + \text{other controls} + u_i t \quad (1)$$

where:

- $crime_{it}$: the number of violent crimes per 100,000 population in state i in year t, which is the variable *gcriv* in the data set (in logs)
- $prison_{it}$: the prison population per 100,000 population (variable *gpris*, in logs)
- α_i: state fixed effect
- λ_t: year fixed effect

Other controls include the number of police per capita (*gpolpc*), the log of income per capita (*gincpc*), the state unemployment rate (*cunem*), the proportion living in metropolitan areas (*cmetro*), and the age distribution of the population (*cag0_14, cag15_17, cag18_24, cag25_34*).

First-differencing equation (1) gives the equation estimated by Levitt (this eliminates the state fixed effect α_i):

$$\Delta ln(crime_{it}) = \beta_1 \Delta ln(prison_{it}) + \lambda_t + \Delta \text{other controls} + \lambda_t + \Delta u_i t \quad (2)$$

a. The above equation will suffer from the simultaneity problem if estimated using OLS. Why?
b. Levitt proposes to use prison overcrowding litigation as an instrument to identify (i.e. estimate) equation (2). Why is prison overcrowding litigation a valid instrument? (Overcrowding litigation are cases brought by civil rights groups that argues for better prisoner conditions in state prisons. This limits the ability for the government to put convicted criminals into prisons.)
c. Estimate equation (2) using time fixed effect, **using data from 1981-1993** (year 1980 is the base year in the first-difference transformation).
d. Estimate equation (2) using 2SLS with *final1* as an instrument. This is a binary variable for whether a final decision was reached on overcrowding litigation in the current year. What happens to the coefficient on prison? What happens to the standard error? Is the coefficient on prison statistically significant?
e. Estimate equation (2) using 2SLS with *final1* and *final2* as instruments. These are binary variables for whether a final decision was reached on overcrowding litigation in the current year (*final1*) or in the previous two years (*final2*). What happens to (i) the coefficient on prison and (ii) the standard error for the coefficient on prison as compared to the same regression estimated using only *final1* as an instrument?
f. Are *final1* and *final2* strong instruments?
g. Are *final1* and *final2* exogenous?

Your code can start with this:

```
url = "https://raw.githubusercontent.com/f1kidd/Econ184/main/Data/prison.csv"
data = read_csv(url)
```

Question 3:

Does Islamic political control affect women's empowerment? Many countries have seen Islamic parties coming to power through democratic elections in recent years, and due to strong support among religious conservatives, constituencies with Islamic rule often tend to exhibit poor women's rights. Does this reflect a causal relationship? Erik Meyerson examines this question by using data on a set of Turkish municipalities from 1994 when an Islamic party won multiple municipal mayor seats across the country. We will use this data and implement a regression discontinuity design to compare municipalities where this Islamic party barely won or lost elections.

The islamic_women.csv dataset includes the following variables:

- margin - The margin of the Islamic party's win or loss in the 1994 election (numbers greater than zero imply that the Islamic party won, and numbers less than zero imply that the Islamic party lost. 0 is an exact tie.)
- school_men – the secondary school completion rate for men aged 15-20

- school_women – the secondary school completion rate for women aged 15-20
- log_pop – log of the municipality population in 1994
- sex_ratio – sex ratio of the municipality in 1994
- log_area – log of the municipality area in 1994

a. Calculate the mean secondary school completion for women in cities where Islamic parties won ($margin>0$) and where they lost ($margin<0$).
b. Do you think part (a) presents a credible estimate of the causal effect of Islamic party control? Why or why not?
c. Use a regression discontinuity model, estimate the causal effect of Islamic party control on secondary school completion for women. How large is this effect?
d. Estimate the causal effect of Islamic party control on secondary school completion for men. How large is this effect?

Your code can start with this:

```
url = "https://raw.githubusercontent.com/f1kidd/Econ184/main/Data/islamic_women.csv"
data = read_csv(url)
```