PSC 504 | Prof. Matthew Blackwell | Harkness 307 | m.blackwell@rochester.edu

# Homework 7[*]

Due Thursday April 18th, 2013 by email or hard copy to Jeffrey Marshall

## 1 Instrumental Variables

This question will have you revisit Acemoglu, Johnson, and Robinson (2001) the empirical analysis of country wealth as a function of political and economic institutions. The full citation is:

> Daron Acemoglu, Simon Johnson, and James A Robinson (Dec. 2001). "The Colonial Origins of Comparative Development: An Empirical Investigation". In: *The American Economic Review* 91.5, pp. 1369–1401

The goal of this paper is to find the effect of political and economic institutions on economic development. The authors argue that these institutions are not randomly assigned and so they take an instrumental variables approach to estimating this relationship, with the instrument being the mortality rates among early European settlers to the country. The story goes like this: in places where settler mortality was high, colonizers set up extractive institutions that sought to exploit the local populations to the benefit of the colonizer. In places with low settler mortality, colonizing powers were more likely to set up growth-promoting institutions like property right to protect the settler's investments. These institutions, the authors argue, tended to stick so that even after decolonization, settler mortality affects these institutions.

You will reproduce the main results of the paper. Here are the variables included in the dataset on the course website:

- `shortnam`: 3 letter country name

[*] Certain problem modified from Jens Hainmueller's Causal Inference course.

- africa: dummy=1 for Africa
- lat_bst: Absolute value of the latitude of capital divided by 90
- rich4: dummy=1 for neo-europes
- avexpr average protection against expropriation risk
- logpgp95: log PPP GDP pc in 1995, World Bank
- logem4 log settler mortality
- asia =1 for all of asia
- loghjypl log GDP per work, Hall&Jones
- baseco base sample Colonial Origins paper

a) In words, what are the key assumptions that must hold so that an IV approach is valid in this case? Do you find these assumptions plausible? Why or why not?

b) Replicate column 2 of Table 4 in Acemoglu, Johnson, and Robinson (2001), including the first stage, 2SLS, and OLS estimates. You may use the tsls function from the sem package to run 2SLS.

c) What additional assumptions do you have to make in order to interpret these results as an ATE?

d) Choose cutoff values for the treatment (avexpr) and the instrument (logem4) and dichotomize these variables. Then calculate the Wald estimate associated with this data. Substantively interpret this coefficient. Comment on the external validity of this estimate.

## 2 Regression Discontinuity Design

This question will have you investigate the research design and results of Lee's paper on the incumbency advantage:

> David S Lee (Feb. 2008). "Randomized experiments from non-random selection in U.S. House elections". In: *Journal of Econometrics* 142.2, pp. 675–697

The goal of this study was to investigate the effect of a party's incumbency status on the party's ability to win a seat in the U.S. House of Representatives. The dataset on the course website contains the following variables:

- `state`: state code
- `distnum`: congressional district number for each state
- `distid`: congressional district id (nationwide)
- `party`: party code (100 - Democrats and 200 - Republican)
- `partname`: party name
- `yearel`: year of election
- `origvote`: votes each candidate received
- `totvote`: total votes cast in each district in a given year
- `highestvote`: votes for candidate who received the largest votes in a district in a given year
- `sechighestvote`:votes for candidate who received second largest votes
- `officeexp`: terms a house representative have served.

a) Explain, both mathematically and substantively, what the key assumptions needed to make causal inferences in this setting. Is this a sharp RD design or a fuzzy RD design?

b) Generate the forcing variable and the treatment variables from the given data. Note that in its current form has rows for each candidate in a race. Plot the treatment as a function of the forcing variable to check the RD design.

c) For the Democrats, run a model of the Democratic vote share in the next election on the forcing variable (Democratic margin of victory), allowing for different slopes for Democratic incumbents and nonincumbents. Report the coefficients and standard errors from this analysis and give a substantive interpretation of the coefficient on the treatment variable.

d) Plot the fitted values from this analysis over a scatterplot of the relationship between the forcing and outcome variables.

e) Perform a placebo test of the effect by replacing the next election's Democratic vote share with the previous election's Democratic vote share. Do you see any potential violations of the identifying assumptions?

f) Perform a placebo test of the effect by replacing the next election's Democratic vote share with the number of previous victories the candidate has won (`officeexp`). Do you see any potential violations of the identifying assumptions?

g) Conduct an alternative placebo test with a fake threshold. Subset the data to only losers (with a Democratic margin of victory less than 0) and create a fake margin of victory threshold at -0.1. Run the RD model from above to see if there is an effect. Do the same thing with winners and a fake threshold of 0.1. What does this type of analysis hope to accomplish?