5304 Econometrics Autumn 2023[1]

Problem set 3
OLS & IV

Part 1: Instrumental variables in practice

The dataset from this assignment comes from Campante and Do (2014), which investigates whether corruption increases if the capital city of a U.S. state is "isolated", in the sense of being located far from the most populous regions of the state.[2] The idea here, is that state capitals that are geographically far from the populous regions are subject to less scrutiny and accountability, which increases the likelihood of corruption. In this assignment, we are going to go through their identification strategy step by step. A trimmed version of the data set used is found in *PS3_ Campante2014.dta.*

1. The data is a panel starting in 1970. In this analysis, we are going to use geographical information about city location, which has not changed since the start of the panel. Hence, we restrict our current analysis only to the year of 1970 – drop all other years.

2. Plot the relationship between corruption and *ALDmean1970*, which is the average log distance of the population from the state capital city, in a scatter plot with a fitted line. What is the slope of the line? What does this imply about the relationship between state capital isolation and corruption, and what might be possible sources of omitted variable bias in this "naïve" specification?

In order to address endogeneity issues, the author proposes the following IV strategy: while the exact location of state capital cities are likely endogenous, the norm was to place them close to the center of the state for political reasons unrelated to other factors. If the location of the center of a state (called the state centroid) is exogenous, there is a plausible instrument: the average log distance of the population from the state centroid, *centr_ ALDmean1970.*

3. In the paper, the author makes the point that it is crucial for the validity of the instrument to control for the average size and shape[3] of the state. Why?

4. Report the following estimations in a table: 1) the naïve OLS regression of corruption on *ALDmean1970*, 2) the first stage and 2SLS estimate of corruption on *ALDmean1970*, using *centr_ ALDmean1970* as an instrument, with and without controlling for state size (*logarea*) and shape (*logMaxDistSt*). Write out the specifications of the OLS (without controls), as well as the first and second stage equations of the 2SLS (with controls).

5. Interpret the coefficient on isolation of the state centroid in the first stage (what does it mean intuitively?). Does the instrument seem to be relevant?

6. Interpret the 2SLS estimates (focus on the signs). How do they differ from the OLS estimate? Does the controls seem to matter in this case?

[1] TA: Petter Berg, petter.berg@phdstudent.hhs.se

[2] Think for example about New York or California, where Albany and Sacramento are state capitals and not the much larger NYC, Los Angeles or San Francisco!

[3] Here, shape is captured by *logMaxDistst*, the maximum distance from the state capital to any of the counties in the state. For example, if a state is quite small but very slim (think Tennessee), the distance from the state centroid could still be large for certain cities!

7. A friend points out that an IV analysis with only 48 observations might not be very credible – in particular, the result could be very sensitive to outliers. Conduct a leave-one-out analysis: make a loop of 48 iterations that, for each iteration, excludes one of the states and runs the full IV specification (with controls) and stores the point estimate. What is the range of point estimates that you obtain? Does the result seem to be sensitive to single outliers?

Part 2: Interpreting published results

In the QJE paper *Do Political Protests Matter? Evidence from the Tea Party Movement* (Madestam et al. 2013), the authors estimate the effect of protest turnout (i.e. how many people show up) for Tea Party rallies[4] on subsequent outcomes such as policy action and vote shares for Republicans in elections. To overcome endogeneity issues, the authors use the random incidence of rainfall at the day of the protest as an instrument for turnout, since less people show up when weather is bad. In summary, the paper finds that the Tea Party protests were quite effective in spurring political change.

1. Why would it probably not be a good idea to simply regress e.g. the vote share of Republicans in the next local election on the number of people that showed up to the Tea Party protests in a given county?

2. Data on outcomes, protest turnout and rainfall is available at the county level. A stylized version of the first stage is given by:

$$Protesters_c = \alpha + \beta RainyRally_c + \gamma ProbabilityOfRain_c + \epsilon_c \tag{1}$$

   where $RainyRally_c$ is a dummy for whether it rains at the day of the protest, and $ProbabilityOfRain_c$ is the forecasted probability of rain for the same day. Can you explain why it is important to include the forecasted probability of rain? *Hint: is rainfall equally likely across counties?*

3. Various first stage estimations are shown in Table 3 from Madestam et al. (2013), reproduced below. Let's focus on column (1). Interpret the estimated coefficient to answer the following question: if a county had $1,000$ protesters in absence of rain, how many would have turned out in case of rain (according to the model's prediction)?

[4] On April 15, 2009, the Tea Party movement held coordinated "Tax Day Rallies" all around the U.S.

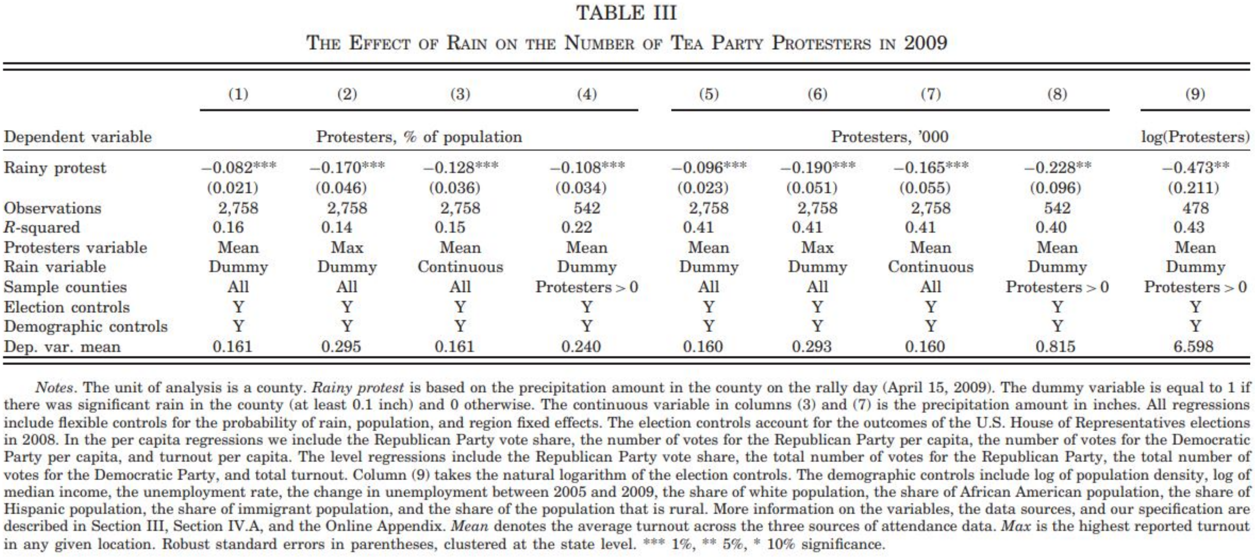

TABLE III

The Effect of Rain on the Number of Tea Party Protesters in 2009

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent variable	Protesters, % of population				Protesters, '000				log(Protesters)
Rainy protest	−0.082***	−0.170***	−0.128***	−0.108***	−0.096***	−0.190***	−0.165***	−0.228**	−0.473**
	(0.021)	(0.046)	(0.036)	(0.034)	(0.023)	(0.051)	(0.055)	(0.096)	(0.211)
Observations	2,758	2,758	2,758	542	2,758	2,758	2,758	542	478
R-squared	0.16	0.14	0.15	0.22	0.41	0.41	0.41	0.40	0.43
Protesters variable	Mean	Max	Mean	Mean	Mean	Max	Mean	Mean	Mean
Rain variable	Dummy	Dummy	Continuous	Dummy	Dummy	Dummy	Continuous	Dummy	Dummy
Sample counties	All	All	All	Protesters > 0	All	All	All	Protesters > 0	Protesters > 0
Election controls	Y	Y	Y	Y	Y	Y	Y	Y	Y
Demographic controls	Y	Y	Y	Y	Y	Y	Y	Y	Y
Dep. var. mean	0.161	0.295	0.161	0.240	0.160	0.293	0.160	0.815	6.598

Notes. The unit of analysis is a county. *Rainy protest* is based on the precipitation amount in the county on the rally day (April 15, 2009). The dummy variable is equal to 1 if there was significant rain in the county (at least 0.1 inch) and 0 otherwise. The continuous variable in columns (3) and (7) is the precipitation amount in inches. All regressions include flexible controls for the probability of rain, population, and region fixed effects. The election controls account for the outcomes of the U.S. House of Representatives elections in 2008. In the per capita regressions we include the Republican Party vote share, the number of votes for the Republican Party per capita, the number of votes for the Democratic Party per capita, and turnout per capita. The level regressions include the Republican Party vote share, the total number of votes for the Republican Party, the total number of votes for the Democratic Party, and total turnout. Column (9) takes the natural logarithm of the election controls. The demographic controls include log of population density, log of median income, the unemployment rate, the change in unemployment between 2005 and 2009, the share of white population, the share of African American population, the share of Hispanic population, the share of immigrant population, and the share of the population that is rural. More information on the variables, the data sources, and our specification are described in Section III, Section IV.A, and the Online Appendix. *Mean* denotes the average turnout across the three sources of attendance data. *Max* is the highest reported turnout in any given location. Robust standard errors in parentheses, clustered at the state level. *** 1%, ** 5%, * 10% significance.

References

Campante, F. R. and Q.-A. Do (2014): "Isolated Capital Cities, Accountability, and Corruption: Evidence from US States," *American Economic Review*, 104, 2456–81.

Madestam, A., D. Shoag, S. Veuger, and D. Yanagizawa-Drott (2013): "Do Political Protests Matter? Evidence from the Tea Party Movement," *The Quarterly Journal of Economics*, 128, 1633–1685.