5304 Econometrics Autumn 2023[1]

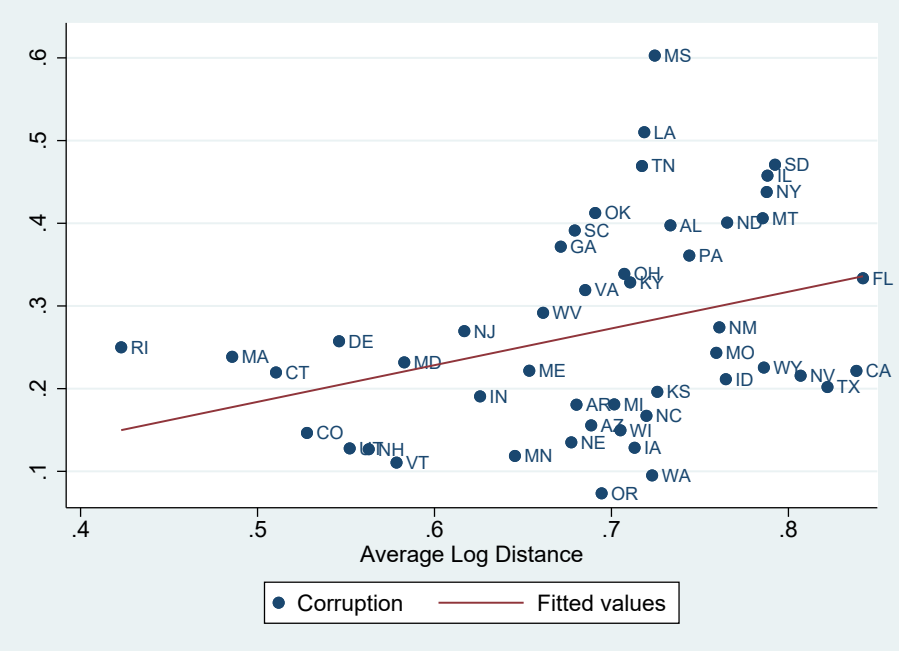
Problem set 3: Group 15

Part 1: Instrumental variables in practice (Campante and Do (2014))

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 and 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 "naive" specification?

Figure 1: State capital distance and corruption relationship



Slope of the line is 0.44392, the naive specification would therefore imply a positive relationship, one unit increase in the average log distance of the population from the state capital city being associated with a 0.44 unit increase of the average corruption rate.

However, the location of the capital city is an institutional decision and it affects the spatial distribution of population. Both of these could be correlated with omitted variables that are also associated with corruption. For instance, corruption and the location of the capital city could be jointly determined, with relatively corrupt states choosing to isolate their capital cities.

[1]42624@student.hhs.se, 42613@student.hhs.se, 42632@student.hhs.se, 25164@student.hhs.se

Alternatively, it could be the case that corruption affects the population flows that determine how isolated the capital city will ultimately be, by pushing economic activity and population away from the capital. For instance, the upsurge in street crime caused by corruption weakening the effectiveness of law enforcement within the capital may encourage families to settle outside of the capital in satellite towns.

Furthermore, a possible source of OVB a capital city's location and the spatial distribution (i.e. the institutional process of urban planning) may be influenced by presence of natural resources (e.g. oil fields / establishment of mining towns), with the natural resources and wealth it promises creating incentives for rent-seeking behaviour. One can argue that in mining towns / oil fields, there is an incentive for bureaucrats and the wealthy to establish the capital city further away to avoid the industrial pollution. Although historically, it depends on whether the technology / means existed to exploit these resources in the first place for us to avoid making any anachronistic assumptions.

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 of the state. Why?

Campante and Do (2014) argue that the centroid is an essentially arbitrary location and should not affect any relevant outcomes in and of itself once the territorial limits of each state are set. If their justification of the exclusion restriction holds, then the centroid should only affect corruption levels indirectly through determining the capital city location.

Yet, for the validity of the IV approach (exogenous condition specifically) to hold, we further require the instrument to be as good as randomly assigned (to mimic RCT conditions). Hence, the relationship between capital city centre and centroid should ideally be constant for all state types. However, this may not be true for states of different sizes and shapes and hence these controls should be included as well. The researchers rightly then controlled, in all of their specifications, for the geographical size of the state, to guard against the possibility that a correlation between omitted variables and the expansion or rearrangement of state borders, which might skew the results.

4. Report the following estimations in a table: 1) the naive OLS regression of corruption on $ALDmean1970$, 2) the first stage and 2SLS estimate of corruption on $ALDmean1970$, using $centr_A LDmean1970$ 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).

Table 1: Corruption and Isolation of the Capital City

	(1) OLS	(2) 1st stage	(3) 2SLS	(4) 1st stage	(5) 2SLS
Average Log Distance of State Capital	0.444**		0.187		1.169**
	(0.140)		(0.132)		(0.417)
Average Log Distance of State Centroid		0.948***		1.185***	
		(0.0572)		(0.322)	
Controls	No	No	No	Yes	Yes
Observations	48	48	48	48	48
R-squared	0.114	0.710	0.0756	0.715	0.254
F-statistic	-	274.5	-	83.37	-

Robust standard errors in parentheses. Control variables: Average size and shape of the state.
Dependent variable: Federal convictions for corruption-related crime relative to population.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The specifications of the estimations are:

(1) OLS, no controls: $\widehat{corruptrate}_{avg,i} = -0.038 + 0.444 ALDmean_{1970,i}$

(2) 2SLS, 1st stage, controls:

$$\widehat{ALDmean}_{1970,i} = -0.003 + 1.185 centr\_ALDmean_{1970,i} - 0.0137 logarea_i - 0.0104 logMaxDistSt_i$$

(3) 2SLS, 2nd stage, controls:

$$\widehat{corruptrate}_{avg,i} = 0.270 + 1.169 \widehat{ALDmean}_{1970,i} - 0.0429 logarea_i - 0.0605 logMaxDistSt_i$$

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?

The coefficient is 0.948 without controls and 1.185 with controls, i.e. a one unit increase in average log distance of the population from the state centroid is estimated to cause a 0.948 unit increase in the average log distance of the population from the state capital when not controlling for state size and shape, and 1.185 unit when controlling. Intuitively, this means that states where the population lives further from the state centroid also tend to have it's population living far from the state capital. We can infer that state capital cities tend to be located near the centroid, possibly due to the incentive to centralize for ease of state governance (given historical means of transport / communications).
In order for the instrument $centr\_ALDmean1970$ to be relevant, $Cov(ALDmean1970, centr\_ALDmean1970) \neq 0$ must hold, which is the case if the coefficients are nonzero. This seems to be the case since both estimated coefficients have a P-value less than 0.001, so the probability that we falsely reject that $Cov(ALDmean1970, centr\_ALDmean1970) \neq 0$ from the estimates is less than 0.1%.

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?

The 2SLS estimates both have positive coefficients on Average Log Distance of State Capital and thus both estimate that a larger average distance increase corruption. The 2SLS estimate with control is statistically significant with t-stat = 2.80 > 1.96 for 5 % significance, but without controls is insignificant at 1.42.
The values are:

- OLS without controls $0.444^{**}(0.140) > 0.187(0.132)$ 2SLS without controls
- OLS without controls $0.444^{**}(0.140) < 1.169^{**}(0.417)$ 2SLS with controls

2SLS estimates (with controls) and the OLS estimate differ because:

$$\beta_{OLS} = \frac{Cov(y,x)}{Var(x)} = \frac{Cov(\alpha + \beta x + \epsilon, x)}{Var(x)} = \beta + \frac{Cov(\epsilon, x)}{Var(x)}$$

OLS estimate of 0.444 is less than 1.169 as estimated by 2SLS with controls. Since variance is always positive, we have reason to believe that $Cov(\epsilon, x) < 0$. Going back to our stories of sources of OVB in the "naive specification" (Question 2), Crudely illustrating, as we make the average log population distance the dependent variable of the omitted variable bias and true noise:

$$PopDistance = \alpha + \lambda_{OVB} NaturalResources + \epsilon_{noise}$$

λ_{OVB} argued to be negative since the presence of natural resources results in the creation of mining towns / clusters that decrease the average population distance, compared to states with fewer natural resources that are relatively spread out.
We also note the possibility of reverse causality at work (i.e. population outflow caused by corruption).
Moving on, why do these controls matter for IV give a causal estimate? Use the first stage of the IV (with controls and then without controls).
Specification A (2SLS without controls):

$$\widehat{ALDmean}_{1970,i} = \phi_0 + \phi_1 centr\_ALDmean_{1970,i} + \omega_i$$

Specification B (2SLS with controls):

$$\widehat{ALDmean}_{1970,i} = \beta_0 + \beta_1 centr\_ALDmean_{1970,i} + \beta_2 logarea_i + \beta 3 logMaxDistSt_i + \epsilon_i$$

We have found that $\beta_2 = -0.0137$ and $\beta_3 = -0.0104$ are negative (from Question 4). Intuitively, using specification B, given the same distance from the centroid, a larger state size and shape results in the average log distance of the population from the capital city decreasing. A plausible explanation is that the state population in larger states is less dispersed since the proportion of livable land relative to the total land is lower, due to the deserts in the West. In a cursory examination looking at the sizes of US states, the Eastern US states are considerably smaller than the Western US states.

However, another historical explanation is that Eastern states were carved out much earlier by European settlers than the Western states. Given that the cost of resettling a state capital is not negligible, then the current state capital may be partly determined by these historical factors. Using this historical narrative, this matters because the historical context (failing to account for subgroups between the Eastern and the Western states) can jointly determine corruption levels and the assignment of the instrument.

Prima facie, specification A already violates the "as good as randomly assigned" condition (i.e. $Cov(z_A, \epsilon_{2ndstage}) \neq 0$), meaning our estimates are no longer causal. This affirms the researchers' narrative that average state size and state shape are good controls crucial for the validity of the instrument (as mentioned in Question 3). Crudely speaking:

$$\widehat{\beta_{IV}} = \frac{\widehat{Cov}(y,z)}{\widehat{Cov}(x,z)} = \frac{\widehat{Cov}(\alpha + \beta x + \epsilon, z)}{\widehat{Cov}(x,z)} = \frac{\beta_{true}\widehat{Cov}(x,z)}{\widehat{Cov}(x,z)} + \frac{\widehat{Cov}(\epsilon,z)}{\widehat{Cov}(x,z)} = \beta_{true} + \frac{\widehat{Cov}(\epsilon,z)}{\widehat{Cov}(x,z)}$$

From first stage, we know that $Cov(x,z) > 0$ and from violation of random assignment condition, $Cov(\epsilon, z) < 0$ such that 2SLS without controls would be negatively biased (i.e. $0.187 < 1.169$).

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?

The range of point estimates is $1.37 - 0.96 = 0.41$. The 95% CI of the mean estimate is $1.15 - 1.19$ which contains the estimate presented in column five of Table 1 and we can therefore conclude that the result is robust to single outliers.

Part 2: Interpreting published results (Madestam et al. (2013))

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.

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?

   Unobservable political preferences are likely to determine both the number of protesters and policy outcomes. A naive regression of policy on protest size is therefore unlikely to reflect a causal effect.

   Clearly, attendance at the Tea party rallies is not randomly assigned. Due to the existence of confirmation bias and the political echo chambers omnipresent in American politics, it is probable that those who attend Tea Party rallies are more conservative and come from a specific socio-economic and ethnic background. Coincidentally, these characteristics (omitted variables) make them more likely to vote Republican in the first place. Selection bias of Tea Party attendance in the naive specification then overestimates the true effect of attending Tea Party protests on Republican support (i.e., positive OVB bias likely).

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

$$Protesters_c = \alpha + \beta RainyRally_c + \gamma ProbabilityOfRain_c + \epsilon_c$$

   where $RainyRallyc$ is a dummy for whether it rains at the day of the protest, and $ProbabilityOfRainc$ 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?

For a rainy rally to be a valid instrumental variable, the "as good as randomly assigned" condition and the exclusion restriction both need to be satisfied. We believe the exclusion restriction argument is sound since rain on a protest day is very unlikely to affect Republican vote share apart from through affecting protest attendance, leaving us to satisfy the other validity requirement.

For the "as good as randomly assigned" condition, there needs to be zero correlation between the instrument and the error term in the second stage. What this intuitively means is that to achieve a golden standard of randomized controlled trials whereby treatment is randomly assigned, there needs to be no fundamental differences in raining probabilities across the different states assigned the instruments. However, given the sheer expanse of North America, it is likely that coastal states (e.g. Seattle with 153 average number of rainy days annually) have higher rainfall probabilities than states with deserts (e.g. Colorado with 93 average number of rainy days annually) with arid climates.

Moreover, since plans to attend a rally are made ex-ante to the realization of rain on rally day itself, projected probability of rain may affect actual rally attendance. For example, if it is projected to rain but it does not rain in actuality, rally attendance may be hypothetically lower than if it was not projected to rain.

Thus, without inclusion of projected probability of rain as a control, non-zero correlation between the instrument and unobserved variables (projected rain probability) in the second stage is unlikely. Probability of rain should therfore be included as a control in the first stage.

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)?

Using Column (1), we find that a rainy rally (dummy variable) results in a -0.082% decrease in protest attendance as a percentage of the population. We further know that the dependent variable mean for Column (1) is 0.161.

Given 1,000 protesters attended without rain, we simply find that protest attendance with rain is given by:

$$1000[\frac{0.161 - 0.082(1)}{0.161}] = 490.68 \approx 491$$

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.