Midterm Exam

Tao Wu

October 29, 2023
Professor Suleyman Taspinar
ECON 387: Advanced Econometrics

1 Question 1

(20 pts) Islam [1995] studies cross-country growth using a panel of 96 countries from 1960 to 1985. Read the highlighted sections in the paper.

(a) Using a panel data model as opposed to using a cross-sectional model as suggested by the author has decisive reasons. Before delving into the reasons, let's have some context of the problem facing the author. He wants to study the convergence of countries' economies that requires an assumption that the countries included in the sample are in their steady states. A steady state is hard to define unless he uses a workaround way i.e. the correlation between initial levels of income and subsequent growth rates. The idea is that if the relationship between these two variables is negative, poorer countries with an initially lower level of income would have a higher growth rate and thus catch up to richer countries with an initially higher level of income and lower growth rate. Okay, now on the question of why he prefers a panel data model, for starters, panel data allows us to estimate more efficiently by controlling for hidden factors that cause omitted variable biases. In looking for convergence, it is **necessary** to control for the differences in steady states of various countries by the inclusion of differences in preference and technology across countries that have dimensions that are not readily measurable or observable according to the author who states emphatically, "In the framework of cross-section regression, it is **not possible** to take account of such unobservable or unmeasurable factors. Only a panel data approach can overcome this problem."

(b) Why the fixed effects estimation results are more plausible as claimed by the author is because we can see changes directly in Table 4 in comparison to the results in Table 1. *lamda* that represents the convergence rate increases nearly 10fold and α, namely values of the elasticity of output with respect to capital, drops down to a more

realistic level that conforms to growth theory, and thus more plausible. Both of these are the desirable results of the panel data approach which helps us eliminate omitted variable bias by controlling for time-invariant and unobserved factors.

2 Question 2

(20 pts) Acemoglu et al. [2001] study whether or not differences in institutions can help to explain observed economic outcomes. Read the highlighted sections in the paper.

(a) Even though their results in Table 2 show a strong positive relationship between institutional quality and economic performance, they should not be taken as causal. First, the authors exemplify this point by comparing the estimates between Nigeria and Chile on a weighted scale. In reality, the GDP between these two countries is much greater than what the estimates have shown, thus diminishing its accuracy and predictive power. Also, it's a reverse casual problem. Correlation is not causation. We don't actually know if it's the richer countries that prefer better institutions. Secondly, our regressor institutions simply correlated with so many determinants of income that are not accounted for and unobserved, which creates a problem for the authors to speak about causality confidently. Moreover, there are errors in the measurement of the institution variable. Positive bias is introduced in the OLS estimates after samples are collected. All in all, without an instrumental variable to "purge" the endogeneity of the institution variable, causality cannot be established.

(b) The mortality rates expected by the first European settlers in the colonies more than 100 years ago are proposed for their use as an instrument variable to fix all of the aforementioned problems. The authors think this IV is well-chosen because relevance and exogeneity conditions are both met. Let's examine why. The authors effectively explain why the mortality rate 100 years ago has no effect on GDP per capita today, other than their effect through institutional development. A vast majority of European deaths back in colonial times were diseases that are almost eradicated nowadays and are not fatal at all in most poor Asian and African countries, which limits its effect **only** through institutional development. And why does such an IV have a strong correlation with institutions? The theory goes: mortality rates determine if there would be settlements; settlements were a major determinant of early institutions; and there is a strong correlation between early institutions and institutions today. Plus, when the authors run a regression on the IV and institutions, they show mortality rates explain over 25% of the variation in current institutions. Lastly, the mortality rate (IV) provides a solution to the

violation of the zero mean condition assumption and gives us better estimates. By taking a quick look at Table 4, TSLS estimates based on this IV exhibit better estimates that are more in line with data about GDP in practice and conform to the economic theories that authors hold.

TABLE 4—IV REGRESSIONS OF LOG GDP PER CAPITA

	Base sample (1)	Base sample (2)	Base sample without Neo-Europes (3)	Base sample without Neo-Europes (4)	Base sample without Africa (5)	Base sample without Africa (6)	Base sample with continent dummies (7)	Base sample with continent dummies (8)	Base sample, dependent variable is log output per worker (9)
Panel A: Two-Stage Least Squares									
Average protection against expropriation risk 1985–1995	0.94 (0.16)	1.00 (0.22)	1.28 (0.36)	1.21 (0.35)	0.58 (0.10)	0.58 (0.12)	0.98 (0.30)	1.10 (0.46)	0.98 (0.17)
Latitude		−0.65 (1.34)		0.94 (1.46)		0.04 (0.84)		−1.20 (1.8)	
Asia dummy							−0.92 (0.40)	−1.10 (0.52)	
Africa dummy							−0.46 (0.36)	−0.44 (0.42)	
"Other" continent dummy							−0.94 (0.85)	−0.99 (1.0)	
Panel B: First Stage for Average Protection Against Expropriation Risk in 1985–1995									
Log European settler mortality	−0.61 (0.13)	−0.51 (0.14)	−0.39 (0.13)	−0.39 (0.14)	−1.20 (0.22)	−1.10 (0.24)	−0.43 (0.17)	−0.34 (0.18)	−0.63 (0.13)
Latitude		2.00 (1.34)		−0.11 (1.50)		0.99 (1.43)		2.00 (1.40)	
Asia dummy							0.33 (0.49)	0.47 (0.50)	
Africa dummy							−0.27 (0.41)	−0.26 (0.41)	
"Other" continent dummy							1.24 (0.84)	1.1 (0.84)	
R^2	0.27	0.30	0.13	0.13	0.47	0.47	0.30	0.33	0.28
Panel C: Ordinary Least Squares									
Average protection against expropriation risk 1985–1995	0.52 (0.06)	0.47 (0.06)	0.49 (0.08)	0.47 (0.07)	0.48 (0.07)	0.47 (0.07)	0.42 (0.06)	0.40 (0.06)	0.46 (0.06)
Number of observations	64	64	60	60	37	37	64	64	61

Notes: The dependent variable in columns (1)–(8) is log GDP per capita in 1995, PPP basis. The dependent variable in column (9) is log output per worker, from Hall and Jones (1999). "Average protection against expropriation risk 1985–1995" is measured on a scale from 0 to 10, where

3 Question 3

(20 pts) Card and Krueger [2000] study the effect of a rise in the minimum wage on employment. To this end, they consider the rise in minimum wage from $4.25 to $5.05 in New Jersey in 1992. Read the highlighted sections in the paper.

(a) These authors were not content with the simple analysis that just calculates the change in employment in New Jersey using employment numbers before and after the rise in minimum change because a simple differences estimator won't give them an accurate estimate and here is why. First, they consider the assignment of participants. The setting takes place in the real world rather than in a lab where every factor can be controlled for causal effect estimation. Without perfect randomization, some differences might remain between the treatment and control groups even after having control variables in the model. This is not an ideal situation to just take the differences. Furthermore, the authors think that a group of fast-food stores in eastern Pennsylvania is naturally suited for being a control group for comparison with the experiences of restaurants in New Jersey and

that seasonal patterns of employment are similar in New Jersey and eastern Pennsylvania, as well as across high- and low-wage stores within New Jersey, which provides ground for the common trend assumption in their DiD approach.

(b) To estimate the employment effects of the minimum wage increase, they opt for finding the Difference-in-differences estimator in this quasi-experimental study. They find the relative change, namely the difference-in-differences, between New Jersey and Pennsylvania stores, is virtually identical, which indicates zero effect that the minimum wage rise brings on sample restaurant employment after accounting for any pre-existing differences between the treatment group (NJ) and control group (PA). This is new evidence that contradicts conventional economic theory stating that a minimum wage increase reduces employment, which is most certainly surprising to the authors. Later on, the authors go a step further and conclude that the rise in New Jersey's minimum wage actually increases employment at fast-food restaurants in the state.

4 Question 4

(20 pts) Angrist and Lavy [1999] study the effect of class size on academic achievement. Read the highlighted sections in the paper.

(a) This paper is heavy to read on my own without following the colored text. Three pages into the reading, I quickly realized it was a mistake and became an obedient student who looks to capture the essence in highlight. I also want to say these heavy-duty econometric research papers have a lot of theories and some elementary math, both of which give them a nice flavor blend of concreteness and story-telling elements. Overall, the paper is dry, but nonetheless fun when you hit the analysis section.

Speaking of analysis, the authors think that simply regressing test scores on a measure of class size controlling for some school characteristics may not be sufficient to identify the effect of class size which has proved very difficult to measure. Let's look at the model here.

$$Y_{ise} = X_s'\beta + n_{sc}\alpha + \mu_c + n_s + \epsilon_{isc}$$

α is the effect estimator of interest. Since the size of class n_{sc} determined by an alternative assignment - as opposed to the class-size function derived from Maimonides' rule - is not truly randomly assigned, α would not be the weighted average response and is likely to be correlated with potential outcomes and the error term.

(b) As Campbell[1969] noted and I quote, "...to identify the causal effect of a treatment that is assigned as a deterministic function of an observed covariate that is also related to the outcomes of interest." This sentence sums up nicely about the proposed methodology by the authors of this paper. They find an instrumental variable that is uncorrelated with the error term and highly correlated with the regressor, namely class size, along with the use of fuzzy regression discontinuity. Okay, let me try my best to explain the methodology.

$$f_{sc} = e_s/[int((e_s - 1)/40) + 1]$$

This is called the class-size function derived from Maimonides' rule. It takes students' enrollment as input and spits out the average class size as its output. Due to the discontinuities or nonlinearities nature of this assignment rule (CS function), it is suited for becoming an instrumental variable for causal effect estimation while the correlation between enrollment and the outcome variable correlations are **adequately** controlled for by other "smooth" functions. In other words, f_{sc} is identified to construct instrumental variable estimates of class size. Additional assumptions are required for the model to work including enrollment which has zero effect on the student's academic performance except through f_{sc} function determining class size and non-meddling parents who we saw in the previous problem set like to play politics and get their children into smaller size class just because they know the rule behind and have the means.

Lastly, a few things to point out about why it is a good idea to have the assignment rule (I call it), f_{sc} function, or class size/enrollment size relationship induced by Maimonides' rule as an instrumental variable. The fact that the alternating pattern that f_{sc} exhibits matches a similar pattern in the outcome variable (test scores) increases the credence of the IV's exogeneity. After all, the likelihood of a factor that can also generate such a pattern as a smooth function is very small. Moreover, when reviewing Figures la and Ib on page 541, despite that the relationship between class size and enrollment size involves many factors, the degree of correlation is visually undeniable and thus demonstrates a high degree of the IV's relevance to the regressor.

(c) OLS estimates of the effect of class size on students' performance are definitely inaccurate. When it comes to the casual effect in complex settings, it's impossible to get any good results without controlling for confounding variables that can influence the outcome variable and other variables on which the regressor might depend. When comparing the results from Table 2 and Table 4, the finding is clear: instrumental variables estimates of the effect of class size on the reading and math scores of elementary school children improve the result accuracy and have shown a negative correlation between class size

and student academic performance.

5 Question 5

(20 pts) In this question, you will replicate some results from Acemoglu et al. [2001]. To this end, we will use two data sets provided by the authors: maketable1.dta and maketable4.dta. Note that .dta extension indicates a Stata formatted data file. To import Stata datasets, you can use read_dta() function from haven package.

(a) The first graph is Figure 3 from the article Acemoglu et al. [2001] and is used to justify the relevance of the instrument the authors are using. The second graph is a scatter plot I generated in R to replicate the results showing the relevance of the IV to the regressor of average expropriation risk, a proxy variable for institutions.

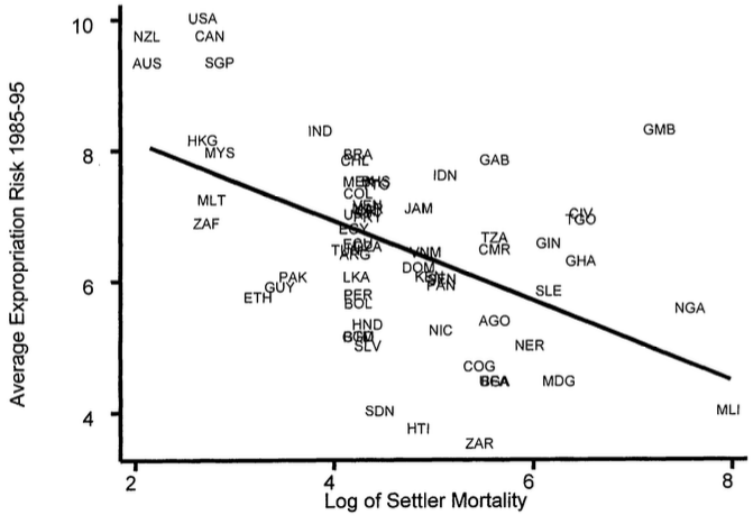
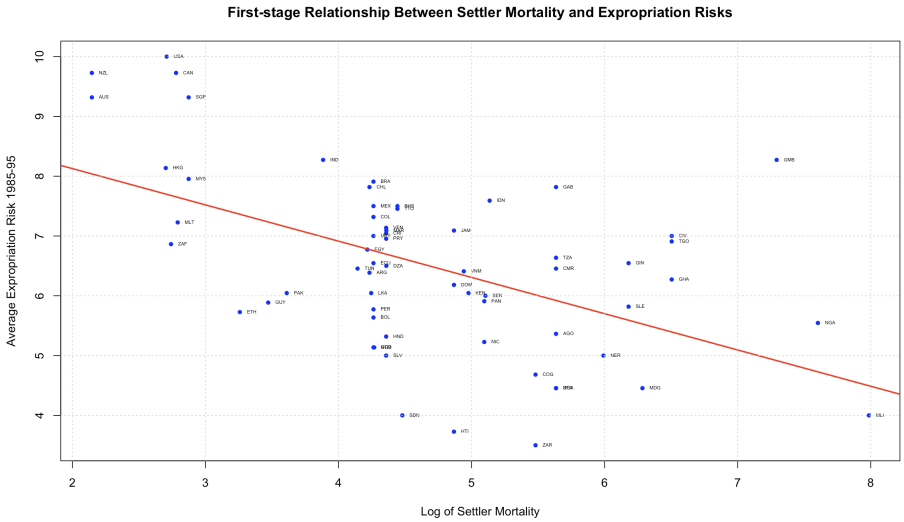


FIGURE 3. FIRST-STAGE RELATIONSHIP BETWEEN SETTLER MORTALITY AND EXPROPRIATION RISK



(b) Present my estimation results in a table

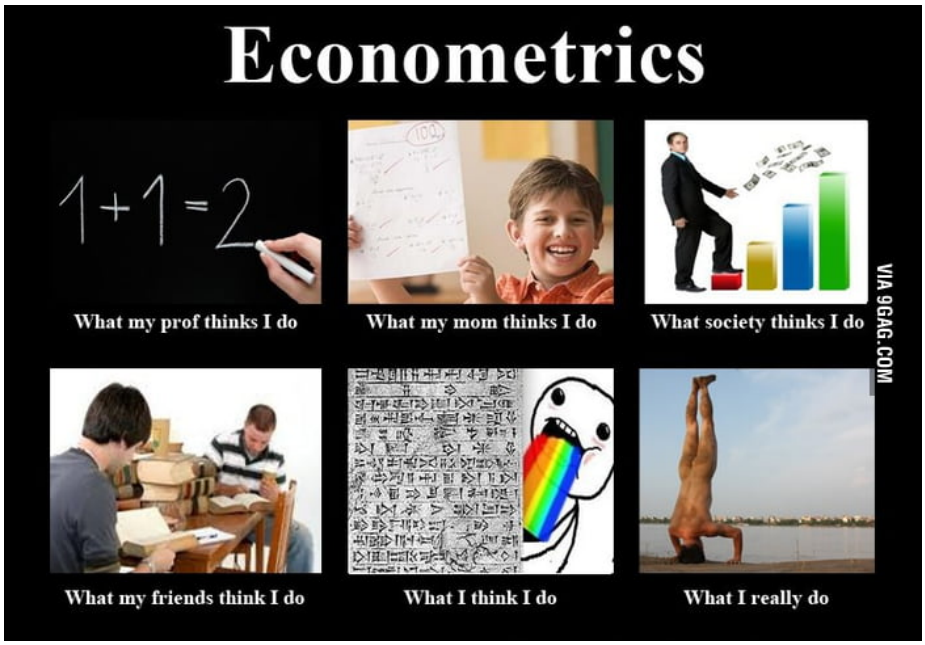
```
library(readr)
library(AER)
library(stargazer)
library(sandwich)
library(lmtest)
library(car)
library(haven)
df1 = read_dta("maketable1.dta")
df1 = df1[df1$baseco == 1, ]

# generate a scatter plot of the relationship between
# average expropriation risk and log of settler mortality
plot(avexpr ~ logem4, data = df1, col = "blue", pch = 20, xlab = "Log of Settler Mortality", ylab = "Average Expropriation Risk 1985-95")
text(df1$logem4, df1$avexpr, label = df1$shortnam)
abline(lm(avexpr~logem4, data=df1), col = "red", lwd = 2)
grid()

# Run an instrumental variable estimation of log GDP per capita on average expropriation risk
# using log of settler mortality as the instrument
df2 = read_dta("maketable4.dta")
df2.01 = df2[df2$baseco == 1, ]

# Model 01 OLS
# Presenting the first stage results of TSLS
m01=lm(avexpr ~ logem4 + africa + lat_abst + rich4 + asia + loghjypl, data=df2.01)
m01_vcov=vcovHC(m01, type = "HC1")
m01_se=sqrt(diag(m01_vcov))

# Model 02 TSLS
# Presenting the second stage results of TSLS using logem4 as the IV
m02=ivreg(logpgp95 ~ avexpr + africa + lat_abst + rich4 + asia + loghjypl | logem4 + africa + lat_abst + rich4 + asia + loghjypl, data=df2.01)
m02_vcov=vcovHC(m02, type = "HC1")
m02_se=sqrt(diag(m02_vcov))

stargazer(m01, m02, se=list(m01_se,m02_se),
          column.labels = c("simple lm", "1st Stage", "2nd Stage"),
          title= "Estimation Results", type = "text",
          keep.stat = c("n","rsq","adj.rsq", "ser", "f"),
          dep.var.labels.include = TRUE,
          model.names = TRUE)
```

codes in R used to generate an estimation table

Estimation Results

	Dependent variable:	
	avexpr	logpgp95
	OLS	instrumental variable
	1st Stage	2nd Stage
	(1)	(2)
logem4	0.261	
	(0.218)	
avexpr		-0.282
		(0.358)
africa	0.626*	0.267
	(0.337)	(0.315)
lat_abst	-0.896	-0.805
	(1.136)	(0.826)
rich4	2.312***	0.889
	(0.371)	(0.794)
asia	1.209**	0.220
	(0.472)	(0.438)
loghjypl	1.271***	1.289***
	(0.251)	(0.414)
Constant	7.335***	12.360***
	(0.928)	(2.961)
Observations	61	61
R2	0.631	0.725
Adjusted R2	0.590	0.694
Residual Std. Error (df = 54)	0.958	0.565
F Statistic	15.397*** (df = 6; 54)	
Note:		*p<0.1; **p<0.05; ***p<0.01

In the paper, the authors state, "mortality rates faced by the settlers more than 100 years ago explain over 25 percent of the variation in current institutions." Since protection against "risk of expropriation" index from Political Risk Services as a proxy for current institutions, the effect of logem4 on avexpr is 0.261 which aligns with what the authors describe.


Econometrics
1+1=2
What my prof thinks I do
What my mom thinks I do
What society thinks I do
What my friends think I do
What I think I do
What I really do
VIA 9GAG.COM