V2024 SØK3514 V24 Assessment guidelines

This is guidelines for assessment. Thus, it is not a complete suggestion of solution. The presentation here is shorter than expected for a complete solution.

Question 1. (Weight 50%)

You want to uncover the causal effect of institutional quality on economic performance measured by GDP per capita using a regression model. You have available data on GDP per capita for a number of years for a sample of countries. You have available a continuous variable, R, that measures the institutional quality for the countries in addition to a number of control variables. Your worry is that institutional quality is an endogenous explanatory variable.

a) Give a short explanation of what is meant by institutional quality being an endogenous explanatory variable and discuss possible reasons for this and what consequences it has for your ability to identify the causal effect of institutional quality on GDP per capita.

b) Discuss how the problem of endogenous institutional quality can be dealt with using the instrumental variable approach and discuss the properties potential instruments must fulfill in order for the approach to give credible results and whether these properties can be investigated empirically.

c) Discuss to what extent panel data methods can be used to deal with the problem with endogenous institutional quality. Explain how the possibility to use panel data to identify the causal effect of institutional quality depends on the properties of the data available.

The article Acemoglu et al (2001) on the reading list investigates the relationship between GDP per capita in 1995 and institutional quality based data for GDP per capita in a sample of countries that were former European colonies. Their measure of institutional quality is an index of the *average protection against expropriation risk, 1985-1995* . Table 4 below taken from the article shows some of their estimation results.

d)Explain how you can use the estimated coefficient in column (1) in panel A in Table 4 to evaluate the contribution of the institutional quality differences to explain actual differences in GDP per capita. You are given the information that the actual *average protection against expropriation risk* is 7.8 and 5.6 in Chile and Nigeria, respectively, while the actual *log GDP per capita* is 9.3 and 6.8 in Chile and Nigeria, respectively.

e) Acemoglu et al. use the variable *log European settler mortality* at the time when these countries were colonized by Europeans in an instrumental variable approach . Explain how you can investigate the credibility of this approach in this setting. What does the regression results in Table 4 tell you about the credibility of the approach?

f) A student argues that the authors should exploit data for a number of years between, say 1965-1995, and panel data methods to get more credible results for the causal effect of institutional quality. Comment on this argument.

a)Institutional quality variable, R, is endogeneous in the regression model, where y denotes GDP per capita.

(1) $y = \beta_1 R + X'\gamma + u$ if $E(u|R) \neq 0$ which implies that u is correlated with R. If u and R is correlated, OLS estimator for the parameter of β_1 will be inconsistent, i.e the OLS estimate cannot be interpreted as the causal effect of institutional quality on GDP per capita.

Reasons:

- Omitted variables

R may be correlated with variables omitted from the regression equation. Since GDP per capita depends on a large number of economic, cultural, institutional and geographical factors in addition to institutional quality it is strong reason to believe that the R variable will be correlated with some of these other determinants, and thus will be correlated with the error term u in a cross section regression model even if we control for a large number of control variables X in the regression.

- Simultaneity/two-way causality

It is likely that there may be two-way causality in the sense that institutional quality may depend on GDP per capita, Rich countries may have better institutions because they are rich, i.e. there is a second regression equation

(2) $R = \alpha y + X'\delta + \theta Z + v$

This simultaneity implies that R is correlated with the error term in the structural equation (1) which is the equation we want to estimate, and thus OLS will be inconsistent.

In addition to omitted variables and simultaneity, measurement error in R leads to correlation between u and R (classical measurement error argument) which leads OLS estimator to be inconsistent and biased towards zero.

b)Suppose, there exist a variable Z or set of variables that affect R, but is excluded from the equation of interest (1) and is uncorrelated with u. Then the IV-2SLS method can be used to estimate the causal effect of R on y. Two important criteria must be met for Z to be a valid instrumental variable

(i)$E(u|\, Z) = 0$

(ii)$cov(R, Z) \neq 0$

(i) is the exclusion restriction, (ii) is the assumption that the instrument is relevant, i.e. it explains some of the variation in R. (i) is basically untestable, but we may test for overidentification restrictions if two or more instruments are available. The IV estimation strategy implemented by the 2SLS method, offers an opportunity to test whether the second assumption, (ii), is fulfilled. The first stage can be written as

(3)$R = X'\pi_1 + \pi_2 Z + e$ and since it only contains exogenous explanatory variables, its parameters can be consistently estimated by OLS

The second stage in the 2SLS/IV method consists of replacing the endogeneous R in (1), with its predicted value from (3), and estimating the equation by OLS.

We can test the hypothesis that $\pi_2 = 0$, in (3) and if the F-statistic for this hypothesis rejects by a sufficient clear margin (rule of thumb: F>10) it indicates that the relevance criterion (ii) is fulfilled.

c)In principle, using panel data in combination with the inclusion of fixed country effects to estimate (1) offers a way to deal with the omitted variable problem. However, it requires that there is sufficient variation within countries over time in institutional quality, R. Institutions often evolve slowly over time. Thus, the within country variation exploited when using fixed country effects is likely to be insufficient to generate precise and credible results using FE-approach. In addition, the FE approach does not deal with the simultaneity problem. This suggests that a IV/2SLS method may be necessary to give credible estimates of the causal effect of R.

d) Compare Nigeria (R=5.6) and Chile (R=7.8), Column(1) result in panel A implies that the predicted difference in logGDP per capita between Chile and Nigeria is

$0.94 \cdot (7.8 - 5.6) = 2.06$. It implies a predicted relative difference between GDP per capita in the two countries at

$e^{2.06} - 1 = 6.8$

Actual difference in GDP between Chile and Nigeria is

$e^{9.3-6.8} - 1 = 11.2$

Thus, Institutional quality, R, explain a substantial share of the actual difference in GDP per capita in the two countries.

e)As discussed in b), the credibility of the IV/2SLS approach in terms of the relevance criterium for the instrument can be judged by testing the null hypothesis that the coefficient in front of the instrument in the first stage equation is zero. The t-statistic for this hypotheses is -0.61/0.13=-4.7 which implies an F-value at 22, clearly above the rule of thumb, 10.

f)Good candidates should see the relevance of the arguments presented in c) where the limitations of the FE, panel approach was discussed and refer to them.

Table 4—IV Regressions of Log GDP per Capita

	Base sample	Base sample	Base sample without Neo-Europes	Base sample without Neo-Europes	Base sample without Africa	Base sample without Africa	Base sample with continent dummies	Base sample with continent dummies	Base sample, dependent variable is log output per worker
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
					Panel A: Two-Stage Least Squares				
Average protection against expropriation risk 1985–1995	0.94	1.00	1.28	1.21	0.58	0.58	0.98	1.10	0.98
	(0.16)	(0.22)	(0.36)	(0.35)	(0.10)	(0.12)	(0.30)	(0.46)	(0.17)
Latitude		−0.65		0.94		0.04		−1.20	
		(1.34)		(1.46)		(0.84)		(1.8)	
Asia dummy							−0.92	−1.10	
							(0.40)	(0.52)	
Africa dummy							−0.46	−0.44	
							(0.36)	(0.42)	
"Other" continent dummy							−0.94	−0.99	
							(0.85)	(1.0)	
					Panel B: First Stage for Average Protection Against Expropriation Risk in 1985–1995				
Log European settler mortality	−0.61	−0.51	−0.39	−0.39	−1.20	−1.10	−0.43	−0.34	−0.63
	(0.13)	(0.14)	(0.13)	(0.14)	(0.22)	(0.24)	(0.17)	(0.18)	(0.13)
Latitude		2.00		−0.11		0.99		2.00	
		(1.34)		(1.50)		(1.43)		(1.40)	
Asia dummy							0.33	0.47	
							(0.49)	(0.50)	
Africa dummy							−0.27	−0.26	
							(0.41)	(0.41)	
"Other" continent dummy							1.24	1.1	
							(0.84)	(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	0.47	0.49	0.47	0.48	0.47	0.42	0.40	0.46
	(0.06)	(0.06)	(0.08)	(0.07)	(0.07)	(0.07)	(0.06)	(0.06)	(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 a higher score means more protection against risk of expropriation of investment by the government, from Political Risk Services. Panel A reports the two-stage least-squares estimates, instrumenting for protection against expropriation risk using log settler mortality; Panel B reports the corresponding first stage. Panel C reports the coefficient from an OLS regression of the dependent variable against average protection against expropriation risk. Standard errors are in parentheses. In regressions with continent dummies, the dummy for America is omitted. See Appendix Table A1 for more detailed variable descriptions and sources.

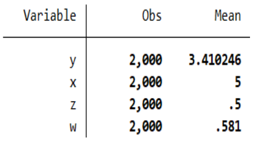
Question 2. (Weight 50%)

a)Explain what is meant by the Regression Discontinuity (RD) design. Explain the difference between Sharp and Fuzzy RD Design.

b) You have a dataset available with a continuous outcome variable y and a treatment variable w that takes the value 1 for treatment and 0 for non-treatment. You want to estimate the causal effect of w on the outcome variable y. In addition to w, the outcome is also affected by the continuous variable x. You want to exploit the fact that the variable x makes a jump for $x \geq 5$.

The variable z is a dummy variable taking the value 1 when $x \geq 5$. Table 2 contains descriptives statistics for the data set.

Table 2. Descriptive statistics , number of observations and mean values.



```
    Variable |        Obs        Mean
-------------+-----------------------
           y |      2,000    3.410246
           x |      2,000           5
           z |      2,000          .5
           w |      2,000        .581
```

c) What is the share of observations with $x \geq 5$? What is the share of observations in the treatment group?

d) Formulate a simple econometric model inspired by the RD approach for the estimation of the causal effect of treatment in this case. Explain how you would estimate the causal effect of treatment in this case. Also explain how you would check the credibility of the approach.

a) RDD.

Suppose the variable of interest is a dummy variable for treatment,

w=1 if treated, 0 if not treated

Treatment is determined by whether an observed “assignment” variable, x, exceeds a known cutoff point, c. The outcome variable of interest, y , might also depend on x , but we assume that the relationship between y and x in the absence of the treatment would be continuous.

We can use a regression to estimate the effect of treatment on the outcome variable and if there is a jump in x at x=c, we can attribute the jump to the treatment.

Sharp RDD design:

The treatment, w is fully determined by x.

That is $w = 1\ if\ x \geq c, w = 0\ if x < c$

The corresponding regression , allowing for nonlinear effects of x on the outcome variable is

$$y = \alpha + \tau w + f(x) + u$$

Where τ is the treatment effect, and $f(x)$ is the potential nonlinear effect of x.

This regression can be estimated by OLS.

Fuzzy RDD design:

The probability of treatment makes a jump for x = c but does not go from 0 to 1. More formally

$$P(w = 1|x) = \gamma + \delta T + g(x)$$

where $T = 1\ if\ x \geq c, T = 0\ if\ x < c$

We come back to the estimation procedure for this case under d)

b) -c) The share of observations in the treatment group, w=1 is 0.58, the share of observations with $x \geq c$ is the share with z=1, which is 0.5. Thus, according to the information given we should use a fuzzy RD design because treatment is not fully determined by the jump in x as explained in a).

d) a two-equation system should be formulated

Structural equation

(1)$y = \alpha + \tau w + f(x) + u$

First stage equation

(2) $w{=}\gamma + \delta T + g(x) + e$

Here, we use a IV/2SLS method. Estimate the first stage equation by OLS and find the predicted probability of treatment, $\hat{w}$ as a function of the instrument, T, and the exogeneous variables.

In the first stage equation (2), we can test the relevance of the instrumental variable by testing the hypothesis that $\delta = 0$. Should reject the hypothesis with a clear margin if treatment really jumps at x=c. Can extend the model in different ways to check the validity of the design. Check for nonlinearities in the continuous effects of x on the outcome variable as discussed lectures and in Lee and Lemieux on the reading list.