ECONOMETRICS FINAL EXAM

UNIVERSIDAD CARLOS III DE MADRID

JUNE 3, 2022

NAME:
NIA:
GROUP:

INSTRUCTIONS:

1. Write your name and group clearly in all sheets.
2. Leave an ID card with your picture on the desk.
3. Each of the four questions will be answered on (both sides of) the sheet where it is written. You cannot use the space on other sheets, or additional sheets.
4. You can use both sides of this sheet only for calculations which will not be evaluated.
5. All parts in each question have the same value.
6. The exam lasts 120 minutes:
   (a) Questions 1, 2 and 3 will be answered in the first 80 minutes using only a pen or pencil.
   (b) Then you can take your personal computer, where the Wooldridge database will have been downloaded in advance. 5 minutes will be given to boot the computer. Question 4 will be answered in 35 minutes using GRETL. Only GRETL can be visible on the computer screen, no other programs can be running. A personal calculator can be used if GRETL's one is not working. Critical values and p-values can be obtained in GRETL.

NAME: _____ GROUP: _____

QUESTION 1 (20%): Consider a linear model with an explanatory variable that may be endogenous. We have a random sample to estimate the causal relationship between the explained and explanatory variable using an instrument that satisfies the conditions of exogeneity and relevance.

a. With the information available, derive an expression for the slope parameter of the model in terms of the covariances of the instrumental variable with the explained and explanatory variable (70%). From this expression, provide a consistent estimator of the slope parameter of the model (30%).

b. Explain how the two-stage least squares estimator of the slope is obtained (50%), and show that the estimator obtained is algebraically identical to the estimator obtained in a. (50%).

c. Express the parameters of the structural form in terms of the parameters of the reduced forms of the explained and explanatory variable (50%). From the relation obtained, obtain a consistent estimator of the slope parameter of the model (10%), and show that it is identical to the estimator in b. (40%).

ANSWER:

a. In the model

$$Y = \beta_0 + \beta_1 X + u$$

we know that there exists and instrument Z so that $Cov(Z,u) = 0$ and $Cov(X,Z) \neq 0$. Then,

$$0 = Cov(Z,u) = Cov(Z, Y - \beta_0 - \beta_1 X) = Cov(Z,Y) - \beta_1 Cov(Z,X),$$

and using that $Cov(X,Z) \neq 0$,

$$\beta_1 = \frac{Cov(Z,Y)}{Cov(Z,X)}. \quad (70\%)$$

This suggests the IV estimate,

$$\hat{\beta}_1^{IV} = \frac{\widehat{Cov}(Z,Y)}{\widehat{Cov}(Z,X)}. \quad (30\%)$$

b. in the first stage, we estimate the reduced form of X,

$$X = \pi_0 + \pi_1 Z + v$$

by OLS and get the predicted values

$$\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i, \ i = 1, ..., n.$$

In the second stage, we substitute X_i by these predicted values in the structural form,

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + error.$$

The OLS estimate of this model,

$$\hat{\beta}_1^{TSLS} = \frac{\widehat{Cov}\left(\hat{X}, Y\right)}{\widehat{Var}\left(\hat{X}\right)},$$

is the Two Stage Least Squares estimate of β_1. (50%)

We show that this identical to the estimate in a.,

$$\begin{aligned}\hat{\beta}_1^{MC2} &= \frac{\widehat{Cov}\left(\hat{X},Y\right)}{\widehat{Var}\left(\hat{X}\right)} = \frac{\widehat{Cov}\left(\hat{\pi}_0+\hat{\pi}_1 Z,Y\right)}{\widehat{Var}\left(\hat{\pi}_0+\hat{\pi}_1 Z\right)} = \frac{\hat{\pi}_1\widehat{Cov}\left(Z,Y\right)}{\hat{\pi}_1^2\widehat{Var}\left(Z\right)} \\ &= \frac{\widehat{Cov}\left(Z,Y\right)}{\hat{\pi}_1\widehat{Var}\left(Z\right)} = \frac{\widehat{Cov}\left(Z,Y\right)}{\frac{\widehat{Cov}(Z,X)}{\widehat{Var}(Z)}\widehat{Var}\left(Z\right)} = \frac{\widehat{Cov}\left(Z,Y\right)}{\widehat{Cov}\left(Z,X\right)} = \hat{\beta}_1^{IV}. \quad (50\%)\end{aligned}$$

c. The two reduced forms are

$$X = \pi_0 + \pi_1 Z + v \tag{1}$$
$$Y = \gamma_0 + \gamma_1 Z + w. \tag{2}$$

We solve for Z in (1) and obtain

$$Z = -\frac{\pi_0}{\pi_1} + \frac{1}{\pi_1}X - \frac{1}{\pi_1}v,$$

which is substituted in (2),

$$\begin{aligned}Y &= \gamma_0 + \gamma_1\left(-\frac{\pi_0}{\pi_1} + \frac{1}{\pi_1}X - \frac{1}{\pi_1}v\right) + w \\ &= \underbrace{\left(\gamma_0 - \frac{\gamma_1\pi_0}{\pi_1}\right)}_{=\beta_0} + \underbrace{\frac{\gamma_1}{\pi_1}}_{=\beta_1} X + \underbrace{\left(w - \frac{\gamma_1}{\pi_1}v\right)}_{=u}. \quad (50\%)\end{aligned}$$

Therefore, as $\beta_1 = \gamma_1/\pi_1$, the estimate (known as indirect least squares) is:

$$\hat{\beta}_1^{ILS} = \frac{\hat{\gamma}_1}{\hat{\pi}_1}, \quad (10\%)$$

where $\hat{\gamma}_1$ and $\hat{\pi}_1$ are the OLS estimates of γ_1 and π_1 in (1) and (2), respectively. Therefore,

$$\hat{\beta}_1^{ILS} = \frac{\hat{\gamma}_1}{\hat{\pi}_1} = \frac{\frac{\widehat{Cov}(Z,Y)}{\widehat{Var}(Z)}}{\frac{\widehat{Cov}(Z,X)}{\widehat{Var}(Z)}} = \frac{\widehat{Cov}\left(Z,Y\right)}{\widehat{Cov}\left(Z,X\right)} = \hat{\beta}_1^{IV}. \quad (40\%)$$

NAME: _____ GROUP: _____

QUESTION 2 (20%): Consider a linear regression model with a dependent variable, Y, and two explanatory variables, X_1 and X_2.

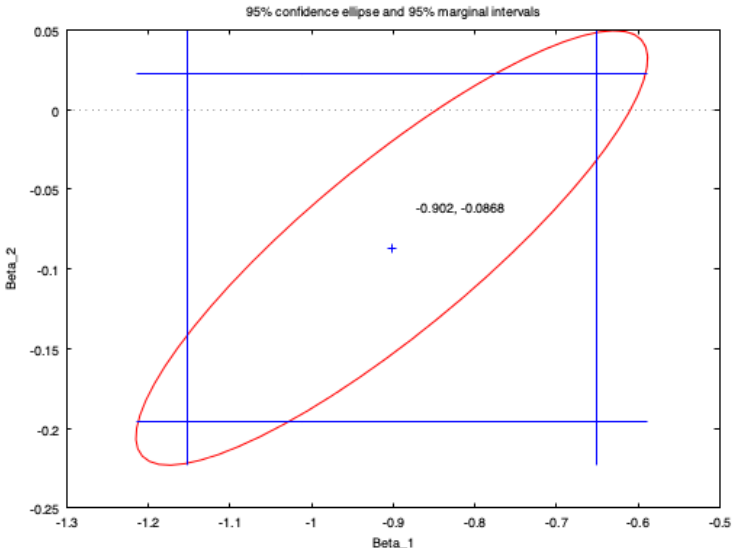
a. Suppose that you test for the global significance of the model at a $100\alpha\%$ significance level by means of a sequential test: we reject the null hypothesis of global significance when the individual significance hypothesis for some of the slopes is rejected at the $100\alpha\%$ significance level. Use a confidence ellipse at the $100(1-\alpha)\%$ and the corresponding individual confidence intervals for the two coefficients to show that, for a given significance level, the conclusion of the sequential test might be opposite to that of the test based on the F statistic.

b. Suppose that the conditional variance of the model errors is not constant and that we test the joint significance of the model coefficients using an F statistic that imposes the homoskedasticity assumption. What consequences would it have on the decisions resulting from this test?

c. Suppose we want to test whether the partial/marginal effect between Y and X_1 depends on the value that the variable X_2 takes. Explain how you would perform this test at a $100\alpha\%$ significance level.

ANSWER:

a. This figure represents a confidence ellipse for both coefficients (in red) and the corresponding confidence intervals (in blue).We can observe that the 5% significance individual tests of the hypotheses



$$H_0 : \beta_1 = -1 \text{ vs } H_1 : \beta_1 \neq -1,$$

and

$$H_0 : \beta_2 = 0 \text{ vs } H_1 : \beta_2 \neq 0,$$

cannot be rejected, because these values are inside the blue intervals. However, the hypothesis

$$H_0 : \beta_1 = -1 \text{ and } \beta_2 = 0 \text{ vs } H_1 : \beta_1 \neq -1 \text{ or } \beta_2 \neq 0$$

is rejected, because the point $(-1, 0)$ is outside the red ellipse. This illustrates that a sequential test on individual parameters can lead to opposite conclusions that those based on a joint test.

b. The approximation of the distribution of the F statistic with a $\chi^2_2/2$ variable would be incorrect and the resulting test would be invalid because the type I error is out of control.

c. We would introduce an interaction variable in the model. That is, the model to be considered would be

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \cdot X_2 + u.$$

Then we would test

$$H_0 : \beta_3 = 0 \text{ vs } H_1 : \beta_3 \neq 0$$

by means of a t statistic.

QUESTION 3 (20%): Consider a linear model with one endogenous explanatory variable and r exogenous explanatory variables.

a. Suppose that we have m possible instrumental variables. Explain which conditions have to satisfy to be valid instruments.

b. Explain which it means that the m instruments are weak (25%). Which are the consequences of using weak instruments on the inferences performed? (25%) How would you test that these instruments are weak? (50%)

c. How many potential instruments do you need to be able to test that they are exogenous? (25%) Explain how would you execute the test. (75%)

ANSWER:

a. Suppose that we have m possible instrumental variables. Explain which conditions have to satisfy to be valid instruments.

.

The possible instruments have to satisfy the exogeneity and relevance conditions:

Exogeneity: all instrumental variables Z_j, $j = 1, \ldots, m$ have to be uncorrelated with the error term U of the structural model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 W_1 + \cdots + \beta_{1+r} W_r + U,$$

i.e.

$$Cov\,(Z_j, U) = 0, \quad \text{for all } j = 1, \ldots, m.$$

Relevance: in the reduced form of the endogenous explanatory variable X_1 (the population model of the first stage regression) at least one of the coefficients corresponding to the instrumental variables Z_j, $j = 1, \ldots, m$, have to be different from zero, i.e. in

$$X_1 = \pi_0 + \pi_1 Z_1 + \cdots + \pi_m Z_m + \pi_{m+1} W_1 + \cdots + \pi_{m+r} W_r + V,$$

with $Cov\,(Z_j, V) = 0$ and $Cov\,(W_j, V) = 0$ all j,

$$\text{at least one } \pi_j \neq 0, \quad j = 1, \ldots, m,$$

or the hypothesis

$$H_0 : \pi_1 = \pi_2 = \cdots = \pi_m = 0$$

must be false.

.

b. Explain which it means that the m instruments are weak (25%). Which are the consequences of using weak instruments on the inferences performed? (25%) How would you test that these instruments are weak? (50%)

.

The m instruments are weak when the coefficients $\pi_1, \pi_2, \ldots, \pi_m$ of the instrumental variables $Z_1, Z_2, \ldots, Z_m$ in the reduced form of X_1 are all zero or very small, so they explain very little of the variation of X_1 (in addition to $W_1, \ldots, W_r$). (25%)

In this case the sampling distribution of the TSLS estimates and the corresponding t statistics is far from normal and usual inference rules do not control the size of the tests. (25%)

To check for weak instruments we calculate the (robust) F statistics of test stage regression

$$X_1 = \pi_0 + \pi_1 Z_1 + \cdots + \pi_m Z_m + \pi_{m+1} W_1 + \cdots + \pi_{m+r} W_r + V$$

for joint significance of the instruments $Z_1, Z_2, \ldots, Z_m$ testing the null hypothesis

$$H_0 : \pi_1 = \pi_2 = \cdots = \pi_m = 0$$

and use the rule-of-thumb that says that instruments are weak if $F < 10$. (50%)

.

c. How many potential instruments do you need to be able to test that they are exogenous? (25%) Explain how would you execute the test. (75%)

.

To be possible to test for exogeneity of the instruments $Z_1, Z_2, \ldots, Z_m$ we need overidentification, which is this is the condition $m > 1$ as in this case as we have a single endogenous regressor ($k = 1$). (25%)

Then, if we have more than one instrument, the procedure for the instruments exogeneity test is as follows:

1. Estimate the model by TSLS with the m instruments and compute the predictions

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 W_1 + \cdots + \hat{\beta}_{1+r} W_r$$

   and the residuals $\hat{U}_i = Y_i - \hat{Y}_i$, $i = 1, \ldots, n$.

2. Regress by OLS the residuals $\hat{U}_i$ on all exogenous variables,

$$\hat{U}_i = \gamma_0 + \gamma_1 Z_1 + \cdots + \gamma_m Z_m + \gamma_{m+1} W_1 + \cdots + \gamma_{m+r} W_r + error.$$

3. Compute the F statistic of global significance of $Z_1, Z_2, \ldots, Z_m$ in this regression testing the null hypothesis

$$H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_m = 0.$$

4. Compute the statistic $J = mF$.

5. Compare the J statistic to the critical value from a χ^2_{m-1} distribution, so that when J is larger than this critical value we reject the null hypothesis of exogeneity of all instruments at the given significance level. (75%)

.

QUESTION 4 WITH GRETL (40%): Use the data in **KIELMC** database of Wooldridge. Data are for houses that were sold during 1981 in North Andover, Massachusetts; 1981 was the year a garbage incinerator was built in that town. To study the effect of the garbage incinerator on house prices, consider a regression model which explains $\ln(price)$, where *price* is the price of the house in dollars, in terms of $\ln(dist)$, $\ln(intst)$, $\ln^2(intst)$, $\ln(area)$, $\ln(land)$, *rooms*, *baths*, and *age*, where *dist* is the distance to the incinerator in feet, *intst* is the distance to the interstate in feet, *area* is the area of the house in squared feet, *land* is the area of the lot in squared feet, *rooms* is the number of bedrooms, *baths* is the number of bathrooms and *age* is the age of the house in years.

a. Provide an estimate of the elasticity of the relation of house prices and distance to the interstate road for houses 1000 feet away of this road and a confidence interval for this elasticity, **using only data for 1981**.

b. Estimate the effect on the house price of transforming a bedroom into a bathroom (25%). Give the appropriate interpretation to your estimates (25%). Is this effect significative? (50%)

c. To evaluate whether the new incinerator had an effect on house prices it is decided to compare the estimates of the coefficient of $\ln(dist)$ obtained in two separated regressions, one with only 1978 data and another one with only 1981 data. Assuming that both samples are independent, perform an statistical test about whether the incinerator construction has an effect on the nearby houses.

d. Explain how to run the previous test using a single OLS regression with the binary variable $y81$.

ANSWER:

a. The elasticity is

$$\begin{aligned}\frac{\partial \mathbb{E}\left[price|\, intst, dist, \ldots\right]/price}{\partial intst/intst} &\approx \frac{\partial \mathbb{E}\left[\ln(price)|\, intst, dist, \ldots\right]}{\partial \ln(intst)} \\ &= \beta_{lintst} + 2\beta_{lintst^2}\ln(intst)\end{aligned}$$

and its estimate is

$$\begin{aligned}\hat{\beta}_{lintst} + 2\hat{\beta}_{lintst^2}\ln(1000) &= 2.07280 + 2\times(-0.119320)\times\ln(1000) \\ &= 0.424\,33\end{aligned}$$

which is interpreted as: when the distance to the interstate increases in 1%, prices increases approximately in 0.42% on average for this type of houses (with $intst = 1000$).

For the confidence interval we need the SE of the estimate,

$$\begin{aligned}& SE\left(\hat{\beta}_{lintst} + 2\hat{\beta}_{lintst^2}\ln(1000)\right) \\ = & \sqrt{\widehat{Var}\left(\hat{\beta}_{lintst} + 2\hat{\beta}_{lintst^2}\ln(1000)\right)} \\ = & \sqrt{\widehat{Var}\left(\hat{\beta}_{lintst}\right) + 4\ln^2(1000)\,\widehat{Var}\left(\hat{\beta}_{lintst^2}\right) + 4\ln(1000)\,\widehat{Cov}\left(\hat{\beta}_{lintst}, \hat{\beta}_{lintst^2}\right)} \\ = & \sqrt{SE\left(\hat{\beta}_{lintst}\right)^2 + 4\ln^2(1000)\,SE\left(\hat{\beta}_{lintst^2}\right)^2 + 4\ln(1000)\,\widehat{Cov}\left(\hat{\beta}_{lintst}, \hat{\beta}_{lintst^2}\right)} \\ = & \sqrt{0.510971^2 + 4\ln^2(1000)\times 0.0294736^2 + 4\ln(1000)\times(-0.0149547)} \\ = & \ 0.116\,97\end{aligned}$$

so the 95% confidence interval is

$$0.424\,33 \pm 1.96\times 0.116\,97 \rightarrow (0.195\,07, 0.653\,59)\,.$$

b. In this case the effect over ln(*price*) is estimated with 1981 data as $\hat{\beta}_{baths} - \hat{\beta}_{rooms} = 0.149550 - 0.0381091 = 0.111\,44$, (25%) i.e. the house price increases approximately in 11.14% (25%).

[1st option.] To check the significance of the effect we test

$$\begin{aligned} H_0 &: \beta_{baths} - \beta_{rooms} = 0 \\ H_1 &: \beta_{baths} - \beta_{rooms} \neq 0 \end{aligned}$$

so we need to calculate the standard error

$$\begin{aligned} SE\left(\hat{\beta}_{baths} - \hat{\beta}_{rooms}\right) &= \sqrt{\widehat{Var}\left(\hat{\beta}_{baths} - \hat{\beta}_{rooms}\right)} \\ &= \sqrt{\widehat{Var}\left(\hat{\beta}_{baths}\right) + \widehat{Var}\left(\hat{\beta}_{rooms}\right) - 2\widehat{Cov}\left(\hat{\beta}_{baths}, \hat{\beta}_{rooms}\right)} \\ &= \sqrt{SE\left(\hat{\beta}_{baths}\right)^2 + SE\left(\hat{\beta}_{rooms}\right)^2 - 2\widehat{Cov}\left(\hat{\beta}_{baths}, \hat{\beta}_{rooms}\right)} \\ &= \sqrt{.043^2 + 0.026^2 - 2 \times (-2.46 \times 10^{-4})} \;=\; 0.055 \end{aligned}$$

and

$$t = \frac{0.111\,44}{0.055} = 2.026$$

which is significative at 5% (50%).

[2nd option.] Alternatively we can do a direct test of H_0 : b[baths]-b[rooms]=0,

Test statistic: Robust F(1, 133) = 4.11301, with p-value = 0.0445517.

[3rd option.] Or reparametrize the model with $\theta := \beta_{baths} - \beta_{rooms}$ so that $\beta_{baths} = \theta + \beta_{rooms}$ and the model is specified as

$$\begin{aligned} \log(price) &= \beta_0 + \beta_{baths} baths + \beta_{rooms} rooms + \cdots + u \\ &= \beta_0 + (\theta + \beta_{rooms})\, baths + \beta_{rooms} rooms + \cdots + u \\ &= \beta_0 + \theta baths + \beta_{rooms}\,(rooms + baths) + \cdots + u \end{aligned}$$

where now $H_0 : \theta = 0$, where θ is the coefficient of *baths* : baths 0.111441 0.0549498 2.028 0.0446 **

c. We want to test

$$\begin{aligned} H_0 &: \beta_{ldist}^{1978} = \beta_{ldist}^{1981} \\ H_1 &: \beta_{ldist}^{1978} \neq \beta_{ldist}^{1981} \end{aligned}$$

for which we can use a t-test

$$t = \frac{\hat{\beta}_{ldist}^{1978} - \hat{\beta}_{ldist}^{1981}}{SE\left(\hat{\beta}_{ldist}^{1978} - \hat{\beta}_{ldist}^{1981}\right)} = \frac{0.0832611 - 0.185237}{0.123\,8} = -0.823\,71$$

because using independence of samples in the two years,

$$\begin{aligned} SE\left(\hat{\beta}_{ldist}^{1978} - \hat{\beta}_{ldist}^{1981}\right) &= \sqrt{\widehat{Var}\left(\hat{\beta}_{ldist}^{1978} - \hat{\beta}_{ldist}^{1981}\right)} \\ &= \sqrt{\widehat{Var}\left(\hat{\beta}_{ldist}^{1978}\right) + \widehat{Var}\left(\hat{\beta}_{ldist}^{1981}\right) - 2\widehat{Cov}\left(\hat{\beta}_{ldist}^{1978}, \hat{\beta}_{ldist}^{1981}\right)} \\ &= \sqrt{SE\left(\hat{\beta}_{ldist}^{1978}\right)^2 + SE\left(\hat{\beta}_{ldist}^{1981}\right)^2 - 2 \times 0} \\ &= \sqrt{0.0660953^2 + 0.104685^2} \;=\; 0.123\,8, \end{aligned}$$

which is not significative at usual significance levels.

d. For that we use the whole sample and add to the usual regressors, the dummy variable $y81$ and its interactions with all of them, so that $\beta_{ldist}^{1978} = \beta_{ldist}$ and $\beta_{ldist}^{1981} = \beta_{ldist} + \beta_{ldist \times y81}$, so now

$$\beta_{ldist}^{1978} - \beta_{ldist}^{1981} = \beta_{ldist} - \left(\beta_{ldist} + \beta_{ldist \times y81}\right) = -\beta_{ldist \times y81}$$

and the previous test is equivalent to testing the significance of the interaction of $\log(dist)$ and $y81$ in the enlarged model.