Quiz 5 – Chapter 5: instrumental variables

(L.G.) – This version: 7 December 2022

Questions

The quizzes are provided as training to help you check your knowledge and understanding of the course; the course and the TD remain the only reference. The quizzes are not necessary, all the less so sufficient, to study Econometrics 1 but might be helpful in your reviews.

Notation Throughout this document, the notations follow those of Chapter 5: absent contrary indications, the variables Z, $D(z)$, D, $Y(d)$, and Y have the same meaning as in the slides of the course.

Question 1 Identification of the average treatment effect on the treated

We consider a randomized experiment with a binary treatment and *imperfect compliance*, namely the dummy variable D, which is equal to 1 when the treatment is effectively received and 0 otherwise, is *not equal* to the binary variable Z of the initial random allocation to the treatment.

We assume $Z \perp\!\!\!\perp (D(0), D(1), Y(0), Y(1))$ and we recall that $\delta^T := \mathbb{E}[Y(1) - Y(0) \mid D = 1]$. Then,

1. $\delta^T = \mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]$
2. $\delta^T = \dfrac{\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]}{\mathbb{E}[D \mid Z = 1] - \mathbb{E}[D \mid Z = 0]}$
3. $\delta^T = \dfrac{\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]}{\mathbb{E}[D \mid Z = 1]}$ if $D = 0$ when $Z = 0$
4. none of the previous assertions

Question 2 Definition of compliers

A "complier" is defined by the following random variable(s):

1. Z
2. D and Z
3. $D(0)$ and $D(1)$
4. $D(0), D(1)$, and Z

Question 3 What can be learned about always takers (AT), compliers (C), and never takers (NT)?

Under the assumptions "$Z \perp\!\!\!\perp (D(0), D(1), Y(0), Y(1))$" (independence) and "$D(1) \geq D(0)$ almost surely" (monotonicity), with i.i.d. data, it is possible

1. to consistently estimate the proportion of compliers only
2. to consistently estimate the proportions of compliers, of always takers, and of never takers
3. to determine whether an arbitrary individual in the data is a complier or not
4. to determine the "type" (complier, always taker, or never taker) of an arbitrary individual in the data
5. none of the previous assertions

Question 4 A sufficient condition for monotonicity

The monotonicity condition "$D(1) \geq D(0)$ almost surely" is necessarily satisfied when

1. $\mathbb{E}[D \mid Z=1] > \mathbb{E}[D \mid Z=0]$
2. $\mathbb{E}[D \mid Z=0] = 1$
3. $\mathbb{E}[D \mid Z=1] = 1$
4. $D = Y$

Question 5 A practical example about monotonicity

We want to investigate the effect of fertility (the number of children) on women's labor participation. We restrict to women with at least one child, and we define D as the indicator of having two children or more, and Z as the indicator equal to 1 if a woman had twins ("jumeaux") in her first pregnancy ("grossesse"), and $Z = 0$ otherwise.

In this setting, the monotonicity condition (2) "$D(1) \geq D(0)$ almost surely"

1. is necessarily satisfied because $D(0) = 0$
2. is necessarily satisfied because $D(1) = 1$
3. is not satisfied because having twins may induce parents not to have other children
4. we cannot conclude here with certainty since we may have $D(0) = 1$ and $D(1) = 0$

Question 6 Two-Stage Least Squares (2SLS) estimator

In this question, D and Z are real random variables, not necessarily binary, and there are no control variables G. In this setting, the Two-Stage Least Squares (2SLS) estimator is obtained by doing the linear regression of (As usual, all regressions include a constant.)

1. Y on D and Z
2. Y on D, Z, and $D \times Z$
3. Y on $\widehat{D}$ where $\widehat{D}$ is the predicted value of D obtained in the regression of D on Z
4. Z on $\widehat{D}$ where $\widehat{D}$ is the predicted value of D obtained in the regression of D on Z
5. Y on $\widehat{Z}$ where $\widehat{Z}$ is the predicted value of Z obtained in the regression of Z on D

Question 7 Measurement errors

We assume model (4) of the course (see slide 29) but without control variables G and with the treatment D (not necessarily binary) assumed to be *exogenous*, that is: for any d in the support of D, we posit

$$Y(d) = \zeta_0 + \delta_0 d + \eta,$$

with $\mathbb{E}[\eta] = \mathbb{E}[D\eta] = 0$.

However, instead of observing D, we only observe $\widetilde{D} = D + \nu$, with $\mathbb{C}\text{ov}(\nu, D) = \mathbb{C}\text{ov}(\nu, \eta) = 0$.

Then, assuming the standard moment conditions (see Proposition 5, Chapter 1), the OLS estimator of the slope in the linear regression of Y on $\widetilde{D}$ converges in probability to the non-stochastic scalar quantity $\widetilde{\delta}_0$ with

1. $\widetilde{\delta}_0 = \delta_0$
2. $\widetilde{\delta}_0 \neq \delta_0$ in general, and $|\widetilde{\delta}_0| > |\delta_0|$ when they differ
3. $\widetilde{\delta}_0 \neq \delta_0$ in general, and $|\widetilde{\delta}_0| < |\delta_0|$ when they differ
4. $\widetilde{\delta}_0 \neq \delta_0$ in general, and we may have $|\widetilde{\delta}_0| > |\delta_0|$ or the reverse ranking $|\widetilde{\delta}_0| < |\delta_0|$ depending on the application

Question 8 Simultaneity

We seek to identify the demand function from the observation of several markets where prices equalize demand and supply.

To do so

1. we do not need any instrumental variable: a regression of the quantity Q on the price P works
2. we do not need any instrumental variable: a regression of the price P on the quantity Q works
3. we can use as an instrumental variable for the price P a variable Z that has an effect on the supply but not directly on the demand
4. we can use as an instrumental variable for the price P a variable Z that has an effect on the demand but not directly on the supply

Question 9 Relevance condition

We consider model (4) of the course (see slide 29) with exogenous control variables G and a scalar treatment D (that is, D is a real random variable) possibly endogenous, namely

$$Y(d) = \zeta_0 + G'\gamma_0 + \delta_0 d + \eta, \quad \mathbb{E}[\eta] = \mathbb{E}[G\eta] = 0$$

but we do *not* assume $\mathbb{E}[D\eta] = 0$. We consider Z, a real random variable, as an instrument for D.

(a) Then, the relevance condition is satisfied if

1. $\mathbb{E}[Z\eta] = 0$
2. $\mathbb{C}\text{ov}(D, Z) \neq 0$
3. $\mathbb{C}\text{ov}(G, Z) \neq 0$
4. The theoretical coefficient of Z in the simple linear regression of D on Z is not null
5. The theoretical coefficient of Z in the multiple linear regression of D on G and Z is not null

(b) Can that condition be tested in the data? If so, explain how to perform the test.

Question 10 Inference for 2SLS estimators

We consider model (4) of the course (see slide 29) without control variables G and a scalar treatment D possibly endogenous, namely

$$Y(d) = \zeta_0 + \delta_0 d + \eta, \text{ with } \mathbb{E}[\eta] = 0,$$

but we do *not* assume $\mathbb{E}[D\eta] = 0$.

On the other hand, we assume to have an instrument Z for D that satisfies the exogeneity and relevance conditions. We also assume that $\mathbb{V}[\eta \mid Z] = \sigma^2$, a constant that does not depend on Z.

Then, all other things being equal, the asymptotic variance of the 2SLS estimator decreases when

1. the prediction of D by Z worsens (that is, Z is less useful to predict D)
2. $\mathbb{E}[Z]$ increases
3. the sample size n increases
4. σ^2 decreases

Question 11 Valid instrument and 2SLS estimators

We consider model (4) of the course (see slide 29) without control variables G, where D is a real random variable ($p = \dim(D) = 1$) possibly *endogenous* (thus, it also corresponds to model (3) on slide 20), namely

$$Y(d) = \zeta_0 + d\delta_0 + \eta, \quad \mathbb{E}[\eta] = 0,$$

(*remark: implicitly, this type of equation is always written for any d in the support of D*) and we do not assume $\mathbb{E}[D\eta] = 0$.

But we assume Z, a real random variable ($q = \dim(Z) = 1$), is a valid instrument for D.

(a) State the two conditions satisfied by Z to be a valid instrument of D and give their names.

-
-

We denote by D^* the theoretical prediction obtained from a linear regression of D on Z.

(b) Then, the limit in probability of $\widehat{\delta}_{2SLS}$, the 2SLS estimator of δ_0 obtained by instrumenting D by Z in the regression of Y on D, is equal to

1. $\dfrac{\mathbb{C}\text{ov}(D^*, Y)}{\mathbb{V}[D]}$
2. $\dfrac{\mathbb{C}\text{ov}(Z, Y)}{\mathbb{V}[Z]}$
3. $\dfrac{\mathbb{C}\text{ov}(Z, Y)}{\mathbb{C}\text{ov}(Z, D)}$
4. None of the previous expressions; if so, indicate a correct one below

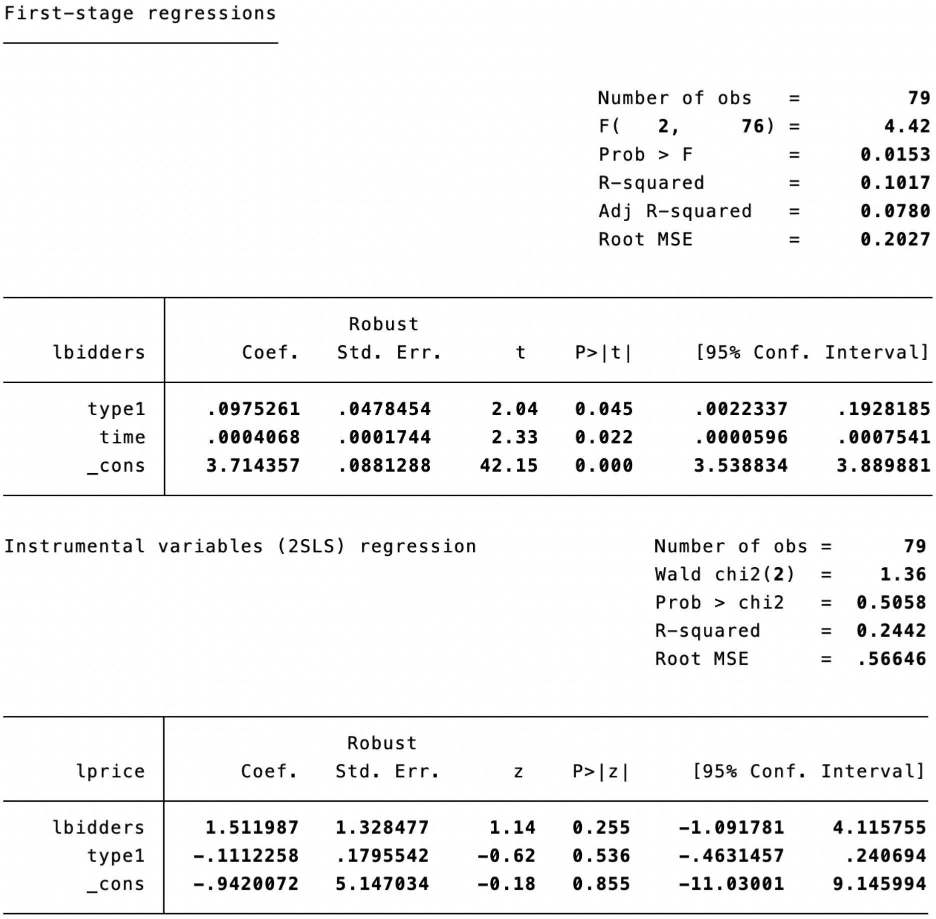
(c) Based on your previous answer, compute the limit in probability of $\widehat{\delta}_{2\text{SLS}}$ as a function of δ_0, $\mathbb{C}\text{ov}(\eta, Z)$, and $\mathbb{C}\text{ov}(Z, D)$.

(d) Using the expression you found, explain the potential issues caused by a "weak" instrument, namely such that $\mathbb{C}\text{ov}(Z, D)$ is close to 0.

Question 12 Some practice with Stata

We consider the setting of Begonia auctions from one of the Problem Sets (you can prepare the exercise or review your notes to refresh your memory).

We run a Stata command, and we obtain the following output made of these two tables:



```
First-stage regressions
-----------------------

                                              Number of obs   =         79
                                              F(   2,     76) =       4.42
                                              Prob > F        =     0.0153
                                              R-squared       =     0.1017
                                              Adj R-squared   =     0.0780
                                              Root MSE        =     0.2027

------------------------------------------------------------------------------
             |               Robust
    lbidders |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       type1 |   .0975261   .0478454     2.04   0.045     .0022337    .1928185
        time |   .0004068   .0001744     2.33   0.022     .0000596    .0007541
       _cons |   3.714357   .0881288    42.15   0.000     3.538834    3.889881
------------------------------------------------------------------------------

Instrumental variables (2SLS) regression          Number of obs =         79
                                                  Wald chi2(2)  =       1.36
                                                  Prob > chi2   =     0.5058
                                                  R-squared     =     0.2442
                                                  Root MSE      =     .56646

------------------------------------------------------------------------------
             |               Robust
      lprice |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lbidders |   1.511987   1.328477     1.14   0.255    -1.091781    4.115755
       type1 |  -.1112258   .1795542    -0.62   0.536    -.4631457     .240694
       _cons |  -.9420072   5.147034    -0.18   0.855    -11.03001    9.145994
------------------------------------------------------------------------------
```

(a) Write the Stata command that generates that output.

(b) Explain what that command did by specifying

1. the outcome variable Y:
2. the possibly endogenous regressor(s) D:
3. the exogenous control variable(s) G:
4. the instrument(s) Z:
5. and the 2SLS estimator we obtained here:

(c) Discuss the relevance condition in this setting. To do so:

1. Write the statistical test whose null hypothesis H_0 is the opposite of the relevance condition (in other words, H_1 is true $\iff$ H_0 is false $\iff$ the relevance condition is satisfied). Of course, you will specify which regressions you are considering for this test.

2. In the previous tables, indicate the relevant p-value you need to look at to assess the result of that test.

3. In particular, here in this application, do you reject the null hypothesis of the negation of the relevance condition at 5%? at 2%? at 1%?

Question 13 Augmented regression

We consider model (4) of the course with exogenous controls G and a scalar treatment D possibly endogenous (see slide 29), where D is a real random variable ($p = \dim(D) = 1$), namely

$$Y(d) = \zeta_0 + G'\gamma_0 + \delta_0 d + \eta, \quad \mathbb{E}[\eta] = \mathbb{E}[G\eta] = 0,$$

and we do not assume $\mathbb{E}[D\eta] = 0$.

We assume that Z, a random vector ($q = \dim(Z) \geq 1$), is a valid instrument for D: it is relevant and $\mathbb{E}[Z\eta] = 0$.

As in the course, we denote by $\widehat{\nu}$ the estimated residual in the linear regression of D on G and Z (and a constant as usual).

(a) How is that regression of D on a constant, G, and Z, called?

(b) In such a setting, the so-called *augmented regression* is the regression of Y on

1. a constant, D, G, Z, and $\widehat{\nu}$
2. a constant, D, G, and $\widehat{\nu}$
3. a constant, D, and $\widehat{\nu}$
4. none of the previous answers; if so, indicate the correct one below

(c) Can we recover the 2SLS estimator from the augmented regression?

1. No
2. Yes. In this case, explain how.