Quiz 5 – Chapter 5: instrumental variables

(Lucas Girard) – This version: 26 December 2022
(With a bit of delay, Merry Christmas! With some advance, Happy New Year!)

Solutions

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,

The question refers to **Corollary 1, slide 11**. Note that Corollary 1 considers the average causal effect on the treated δ^T while Theorem 1 (slide 9) is concerned with the average causal effect on the compliers δ^C (the so-called LATE).

Important point: under assumption (1) or (I) (independence) "$Z \perp\!\!\!\perp (D(0), D(1), Y(0), Y(1))$", we have, $\forall z \in \{0, 1\}$, (note: z is a free variable, "variable muette", not stochastic, any number 0 or 1)

$$\mathbb{E}[D \mid Z = z] = \mathbb{E}[D(Z) \mid Z = z] = \mathbb{E}[D(z) \mid Z = z] = \mathbb{E}[D(z)] = \mathbb{P}(D(z) = 1)\,,$$

where the first equality uses the definition of the *observed* treatment variable $D := D(Z)$, the second equality the conditioning event $\{Z = z\}$, the third equality uses assumption (I) to remove the conditioning, and the last equality holds because, in the setting of the first section of Chapter 5 (Randomized experiments with imperfect compliance) the *potential* treatment variable $D(z)$ is binary: $D(z) \in \{0, 1\}$, is a Bernoulli variable.

Therefore, under assumption (I), we have

$$\mathbb{E}[D \mid Z = 0] = 0 \iff \mathbb{E}[D(0)] = 0 \iff D(0) = 0 \text{ almost surely.}$$

The first equivalence uses the previous result directly. The second one holds because $D(0)$ is a non-negative variable (equal to 0 or 1): consequently, its expectation is null if and only if the variable is constant equal to 0 (almost surely).

If $\mathbb{E}[D \mid Z = 0] = 0$, then we have $D(0) = 0$ almost surely, or equivalently since $D(0)$ is binary, $\mathbb{P}(D(0) = 1) = 0$: there is no always taker, and also no defiers.[1]

Thus, the last condition $\mathbb{E}[D \mid Z = 0] = 0$ in the statement of Corollary 1 asserts there is no always taker nor defiers.

[1] This explains why we do not need assumption (2): $D(1) \geq D(0)$ almost surely (no defier) (slide 8) in the statement of Corollary 1 (and of this question 1): $D(0) = 0$ almost surely implies the monotonicity condition. See also slide 8, third bullet-point: "(2) holds in particular if $D(0) = 0$: no one, if allocated in the control group, manages to get the treatment." $D(0) = 0$ almost surely $\implies$ there is no defiers and no always taker (remember the 2×2 table depending on the value 0 or 1 of $D(0)$ and of $D(1)$ to define always taker, compliers, defiers, and never takers (see TD9, handwritten solutions, page 1, or below the solution for Question 2).

In the absence of always taker and defiers, the only possibility to be treated is to be a complier *and* to be allocated in the treatment initially $Z = 1$. In this case, the average causal effect of the treatment on the treated δ^T is equal to the average causal effect on the compliers who are treated, which is equal[2] to the average causal effect on the compliers δ^C.

This explains the result of Corollary 1 and the correct answer here: answer 3.
Indeed, saying "$D = 0$ when $Z = 0$" means $D(0) = 0$ or, equivalently, $\mathbb{E}[D \mid Z = 0] = 0$.

1. $\delta^T = \mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]$

   – **False**, the right-hand side would identify the new "treatment" = "was supposed to receive the treatment" (the effect of Z: allocation to the treatment, instead of D: effective treatment), but this effect of Z on Y is often without much interest (see bullet-points 4 and 5, slide 6).

2. $\delta^T = (\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0])/(\mathbb{E}[D \mid Z = 1] - \mathbb{E}[D \mid Z = 0])$

   – **False**, would be correct with δ^C instead of δ^T and if we also assume the monotonicity condition (2) and that the denominator is positive (see Theorem 1, slide 9).

3. $\delta^T = (\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0])/\,\mathbb{E}[D \mid Z = 1]$ if $D = 0$ when $Z = 0$

   – **True** (Corollary 1, slide 11).

4. none of the previous assertions

   – **False**.

Question 2 Definition of compliers

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

Important point: complier, always taker, never taker, and defier are defined with respect to the *potential* treatment variables $D(0)$ and $D(1)$. The actual realization of Z is irrelevant. These four groups are defined only by $D(0)$ and $D(1)$. As $D(0)$ and $D(1)$ are both binary, there are four possibilities:

- *always taker*: $D(0) = D(1) = 1$
- *never taker*: $D(0) = D(1) = 0$
- *complier*: $D(0) = 0$ and $D(1) = 1 \iff D(1) - D(0) = 1 \iff D(1) > D(0)$
  (because $D(1)$ and $D(0)$ are binary variables; note that the inequality must be strict)
- *defier*: $D(0) = 1$ and $D(1) = 0$

	D(1) = 0	D(1) = 1
D(0) = 0	never taker (NT)	complier (C)
D(0) = 1	defier (D)	always taker (AT)

For $z \in \{0, 1\}$, $D(z)$ is the potential treatment variable corresponding to $Z = z$. Similarly as in Chapter 5 with the potential *outcome* variables $Y(d)$, we only observe for each individual[3]

$$D := D(Z) = ZD(1) + (1 - Z)D(0) = \begin{cases} D(0) & \text{if } Z = 0, \\ D(1) & \text{if } Z = 1 \end{cases},$$

[2] Because Z is drawn randomly and independent of any other random variables, in particular, of $D(0)$ and $D(1)$: compliers with $Z = 1$ are the same on average as compliers with $Z = 0$ and the same on average as any compliers; formally, we can remove the conditioning $Z = 1$ in the conditional expectation (see page 7 of the handwritten solutions of TD9 for further details).

[3] Remark: as always we assume i.i.d. data $(Y_i, D_i, Z_i)_{i=1,\ldots,n} \sim (Y, D, Z)$. Therefore, we omit the index i, but all those variables are defined at the individual level. For each individual i, we have the variables $D(0)_i, D(1)_i$, etc.

D is the *observed* treatment variable while $D(0)$ and $D(1)$ are the *potential* treatment variables.

A concrete example. Let us imagine we want to study the causal effect of preparing the TD on the final grade Y in Econometrics 1.[4] Imagine that Z is equal to 1 if you are explicitly asked to prepare the next TD (for instance, through an email with references to the corresponding Stata output and hints), and 0 otherwise; the treatment variable D is equal to 1 if you *effectively* prepare the TD, 0 otherwise:

- always takers: students who always prepare the TD, whatever they are asked to do
- never takers: students who never prepare the TD, whatever they are asked to do
- complier: students who prepare the TD if they are asked to do so, and who do not prepare the TD if they are not asked to do so
- defier: students who prepare the TD if they are *not* asked to do so, and who do *not* prepare the TD if they are asked to do so. Defiers are often a bit strange, and this is why assumption (2) or (M) (monotonicity) that rules out defiers is often rather credible, or at least a reasonable simplifying assumption.

1. Z – **False.**
2. D and Z – **False.**
3. $D(0)$ and $D(1)$ – **True.**
4. $D(0), D(1)$, and Z – **False.**

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

The question is related to TD9, question 2 (see also the solutions). There are two bricks to understand the answer.

First, under assumption (independence) ((1) or (I)), we have

$$\forall z \in \{0,1\}, \mathbb{E}[D \mid Z = z] = \mathbb{E}[D(z)] = \mathbb{P}(D(z) = 1)\,,$$

and, with i.i.d. data we can estimate consistently $\mathbb{E}[D \mid Z = z]$ by its empirical counterpart: the sample mean $n_z^{-1} \sum_{i:Z_i=z} D_i$, with $n_z := \sum_{i=1}^n \mathbb{1}\{Z_i = z\}$ (see also slide 14, Chapter 5).

Second, under monotonicity ((2) or (M)), we have

- $\{AT\} = \{D(0) = 1\}$ (equality of the two events);
- $\{NT\} = \{D(1) = 0\}$ (equality of the two events);
- $\mathbb{P}(C) + \mathbb{P}(AT) + \mathbb{P}(NT) = 1$, in other words, the three events $\{AT\}$, $\{AT\}$, and $\{C\}$ form a partition of the probability space.

[4] You can continue this example to understand better Chapter 5 (and the course in general). What assumption would you need on Z? In practice, how would you implement that? What average causal effect would you be able to identify? Would it be of interest to the head of studies of ENSAE? Is there an omitted variable bias in a simple naive regression of Y and D? Give the likely sign of the selection bias B (see Chapter 5). Etc.

Combining the two bricks, we obtain the following consistent estimators for the proportion of always takers (AT), never takers (NT), and compliers (C):

$$\frac{1}{n_0}\sum_{i:Z_i=0} D_i \xrightarrow[n\to+\infty]{P} \mathbb{E}[D\,|\,Z=0] = \mathbb{P}(D(0)=1) = \mathbb{P}(AT)\,,$$

$$1-\frac{1}{n_1}\sum_{i:Z_i=1} D_i \xrightarrow[n\to+\infty]{P} 1-\mathbb{E}[D\,|\,Z=1] = 1-\mathbb{P}(D(1)=1) = \mathbb{P}(D(1)=0) = \mathbb{P}(NT)\,,$$

$$\frac{1}{n_1}\sum_{i:Z_i=1} D_i - \frac{1}{n_0}\sum_{i:Z_i=0} D_i \xrightarrow[n\to+\infty]{P} \mathbb{E}[D\,|\,Z=1]-\mathbb{E}[D\,|\,Z=0] = 1-\mathbb{P}(AT)-\mathbb{P}(NT) = \mathbb{P}(C)\,.$$

Therefore, it is possible to estimate consistently under assumption (1) (independence) and (2) (monotonicity) the proportions of compliers, of always takers, and of never takers in the population. Hence, the correct answer 2.

Additional remark about answers 3 and 4.

Z and D are binary variables: there are four possibilities for the values of the couple of *observed* variables (Z, D). *Under assumption (2) (monotonicity)*, we have $\{AT\} = \{D(0) = 1\}$ and $\{NT\} = \{D(1) = 0\}$. This enables us to know that *some* individuals are NT (respectively AT). Indeed,

- if i is such that $D_i = 0$ and $Z_i = 1$, then $D_i = D(1)_i = 0$: i is a never taker;
- if i is such that $D_i = 1$ and $Z_i = 0$, then $D_i = D(0)_i = 1$: i is an always taker;
- if i is such that $D_i = 0$ and $Z_i = 0$, then $D_i = D(0)_i = 0$: i can be a never taker or a complier, but we cannot know which one between the two possibilities;
- if i is such that $D_i = 1$ and $Z_i = 1$, then $D_i = D(1)_i = 1$: i can be an always taker or a complier, but we cannot know which one between the two possibilities.

Hence, for *arbitrary* individuals (in the sense of without knowing ex-ante their values of D and Z), we cannot tell for sure their type in terms of C, AT, or NT. In particular, we can never say for sure whether a given individual is or not a complier. The fundamental reason behind is that a complier is defined by $D(0) = 0$ and $D(1) = 1$, but we only observe one of them $D = D(Z)$, the other potential treatment variable is counterfactual. What enables us to go beyond that for the special case of $D_i = 0$ and $Z_i = 1$ (respectively $D_i = 1$ and $Z_i = 0$) is the monotonicity assumption that rules out defiers.

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

Question 4 A sufficient condition for monotonicity

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

In the course, Chapter 5, you have the result that $D(0) = 0$ almost surely implies the monotonicity condition (2) (see bullet-point 3 of slide 8): “(2) holds in particular if $D(0) = 0$ almost surely: no one, if allocated in the control group, manages to get the treatment.” Indeed, it implies that there are no defiers nor always takers.

Here, it is a symmetric sufficient condition to obtain monotonicity: if $D(1) = 1$ almost surely, then assumption (2) of monotonicity holds because $D(0) \in \{0, 1\}$, hence $D(0) \leq D(1)$ almost surely.

Finally, to get the correct answer 3, it is the same important reasoning as in Question 1: $\mathbb{E}[D \mid Z = 1] = 1$ if and only if $\mathbb{P}(D = 1 \mid Z = 1) = 1$ because D is binary (the expectation is equal to the maximal value 1 if and only if the variable is constant equal to 1 conditional on $Z = 1$), that is, if and only if $D(1) = 1$ almost surely.

1. $\mathbb{E}[D \mid Z = 1] > \mathbb{E}[D \mid Z = 0]$ – **False,** this just says that Z and D are positively correlated.
2. $\mathbb{E}[D \mid Z = 0] = 1$ – **False**, on the contrary, this would say that $D(0) = 1$ almost surely, and the monotonicity condition $D(1) \geq D(0)$ would then require $D(1) = 1$ almost surely too, that is everyone is treated: it would be a weird case.
3. $\mathbb{E}[D \mid Z = 1] = 1$ – **True.**
4. $D = Y$ – **False**, nonsense: Y has nothing to do with the monotonicity condition $D(1) \geq D(0)$.

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"

The monotonicity condition (2) "$D(1) \geq D(0)$ almost surely" (Chapter 5, slide 8) is a crucial assumption to estimate the average causal effect on the compliers δ^C. Indeed, by ruling out the existence of defiers, this condition guarantees that the only individuals whose behavior regarding the treatment status (namely $D = 0$ or $D = 1$) is affected by the instrument Z (the initial allocation) are the compliers. In other words, the only way the instrument Z can change an individual's behavior is by inducing him or her to follow the treatment (it cannot be the contrary effect). Intuitively, this is why (coupled with the independence assumption (1) and $\mathbb{E}[D \mid Z = 1] > \mathbb{E}[D \mid Z = 0]$) we can recover δ^C (Chapter 5, Theorem 1).

Slide 8 of Chapter 5 (third bullet-point) provides a sufficient condition for the monotonicity assumption to hold: $D(0) = 0$ almost surely. Concretely, this condition means the experiment's organization is such that people allocated to the control group ($Z = 0$) cannot follow the treatment effectively. It is an interesting condition because it can be a reasonable assumption in some settings. $D(0) = 0$ almost surely rules out the existence of both defiers and always takers. This is why, in this case, we have $\delta^T = \delta^C$ (see Corollary 1, slide 11). It is an interesting condition but not the unique sufficient condition to guarantee (2).

Indeed, the condition $D(1) = 1$ almost surely also implies $D(1) \geq D(0)$ almost surely since $D(0) \in \{0, 1\}$ (see Question 4 of this quiz).

$D(1) = 1$ almost surely may happen automatically depending on the definition of D and Z. In this example[5], it is the case: $Z = 1$ means the woman had twins in her first pregnancy, therefore has at least two children, which means $D = 1$.

1. is necessarily satisfied because $D(0) = 0$ almost surely – **False**, there is no reason $D(0) = 0$ almost surely: a woman can have a single child in her first pregnancy ($Z = 0$) and other children in later pregnancies afterward.
2. is necessarily satisfied because $D(1) = 1$ – **True**.

[5] See the article "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size", by Joshua D. Angrist and William N. Evans, published in *The American Economic Review* in 1998 – `https://www.jstor.org/stable/116844?seq=1`.

3. is not satisfied because having twins may induce parents not to have other children – **False**, irrelevant here because $D = 1$ if there are two children or more, hence automatically satisfied if $Z = 1$.

4. we cannot conclude here with certainty since we may have $D(0) = 1$ and $D(1) = 0$ – **False**, precisely, here we have $D(1) = 1$ almost surely, therefore we cannot have $D(1) = 0$.

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.)

It is directly the explanations of slide 24 following Theorem 3:

1st LS Regression of D on Z. We then obtain the predicted value $\widehat{D}$;

2nd LS Regression of Y on $\widehat{D}$.

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

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

$$Y(d) = \zeta_0 + \delta_0 d + \eta, \tag{1}$$

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

This is precisely the setting of the two slides 27 and 28 of Chapter 5: "A particular case: measurement error on D" (see also TD8, Question 5 for an extension).

Remark that this type of computations and proofs are good preparation for a theoretical exercise at the exam. It is important that you understand and can do the following solution by yourself. You can also review the proofs of the main results, Theorems, and Propositions of the course.

We know (Chapter 1, Proposition 5) that, given i.i.d. sampling and the standard moment conditions (implicitly assumed absent contrary indications), the limit in probability of the OLS estimator of the slope in the simple linear regression of Y on $\widetilde{D}$ is equal to

$$\widetilde{\delta}_0 = \frac{\mathbb{C}\text{ov}(\widetilde{D}, Y)}{\mathbb{V}[\widetilde{D}]}.$$

From there, the idea is to replace $\widetilde{D}$ and Y by their respective expressions and compute $\widetilde{\delta}_0$ thanks to the bilinearity of the covariance and the assumptions.

Equation (1) in the statement of the question is implicitly written for any d in the support of D. It also holds for $Y = Y(D)$ the observed outcome, so that

$$Y \overset{\text{definition of } Y}{:=} Y(D) \overset{\text{Model (1)}}{=} \zeta_0 + \delta_0 D + \eta, \quad \mathbb{E}[\eta] = \mathbb{E}[D\eta] = 0.$$

The exogeneity assumptions about η and D implies that $\mathbb{C}\mathrm{ov}(D, \eta) = \mathbb{E}[D\eta] - \mathbb{E}[D]\,\mathbb{E}[\eta] = 0$.

We also have, by assumption about the measurement error ν, $\mathbb{C}\mathrm{ov}(\nu, D) = \mathbb{C}\mathrm{ov}(\nu, \eta) = 0$. Therefore replacing $\widetilde{D}$ and Y by their expression, we obtain

$$\widetilde{\delta}_0 = \frac{\mathbb{C}\mathrm{ov}(\overbrace{D+\nu}^{=\widetilde{D}}, \overbrace{\zeta_0 + \delta_0 D + \eta}^{=Y})}{\mathbb{V}[\widetilde{D}]} = \delta_0 \frac{\mathbb{V}[D]}{\mathbb{V}[\widetilde{D}]}.$$

Furthermore, since D and ν are assumed to be uncorrelated, the variance of their sum is equal to the sum of their variances:

$$\mathbb{V}[\widetilde{D}] = \mathbb{V}[D + \nu] = \mathbb{V}[D] + \mathbb{V}[\nu].$$

Hence,

$$\widetilde{\delta}_0 = \delta_0 \times \frac{\mathbb{V}[D]}{\mathbb{V}[D] + \mathbb{V}[\nu]} = \delta_0 \times \frac{1}{1 + \mathbb{V}[\nu]/\mathbb{V}[D]}.$$

Therefore $|\widetilde{\delta}_0| \leq |\delta_0|$ with strict inequality whenever $\mathbb{V}[\nu] > 0$, that is, when ν is not a constant. In general, ν is not a constant (otherwise, this is not a measurement error but rather a change of units in the measurement); hence correct answer 3.

1. $\widetilde{\delta}_0 = \delta_0$ – **False**.
2. $\widetilde{\delta}_0 \neq \delta_0$ in general, and $|\widetilde{\delta}_0| > |\delta_0|$ when they differ – **False**.
3. $\widetilde{\delta}_0 \neq \delta_0$ in general, and $|\widetilde{\delta}_0| < |\delta_0|$ when they differ – **True**, this is why we often talk about "attenuation bias" in case of measurement errors.
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 – **False**.

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

This is the case in point of the failure of usual regressions due to **simultaneity**: demand and supply curves. Chapter 5 deals with this example in slides 22 and 26 (see also the second exercise of TD10 and its solutions for further details).

Here, we are interested in *the demand function*: the quantity asked as a function of the price, $p \mapsto Y_d(p)$ with the notations of Chapter 5. That is why, here, the quantity $Q = Y$ is the outcome or explained variable, while the price P is the explanatory variable or regressor.

Following Chapter 5, we assume a linear demand curve $Y_d(p) = \gamma_0 + \delta_0 p + \eta_d$ and we want to estimate the slope δ_0.[6]

As explained in slide 22, due to simultaneity issues, there is no reason that the simple linear regression of the quantity on the price identifies δ_0: we need an instrumental variable Z.

To satisfy the exogeneity condition, Z cannot have a direct effect on the demand: *we seek an instrument that has an effect on the supply, hence on the price that equalizes supply and demand, but not directly on demand.*

[6] Remark: if we use a log-log model instead of this level-level model, the coefficient δ_0 would be the elasticity of the demand with respect to price, assumed to be constant with a linear log-log model.

Thinking about the standard curves of supply and demand, we are looking for Z that shifts the supply curve only, not the demand curve, so that we will be able to obtain the slope of the demand curve. This explains why the correct answer is answer 3.

1. we do not need any instrument variable: a regression of the quantity Q on the price P works – **False**.

2. we do not need any instrumental variable: a regression of the price P on the quantity Q works – **False**.

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 – **True**.

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 – **False**.

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, univariate) 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

In Chapter 5, there are several expressions of the relevance condition that correspond to particular or more general settings depending on

- the dimension of endogenous regressor(s) D,
- the dimension of instrument(s) Z,
- and the presence or absence of exogenous controls G.

You should know them well, for sure the first two with a univariate D (classical questions in exams); the third case with multivariate D is more advanced.

– Univariate D, univariate Z, without controls G.
Without controls G and with univariate endogenous regressor D and univariate instrumental variable Z, the relevance condition is simply $\mathbb{C}\text{ov}(D, Z) \neq 0$ (see slide 23). Remember that the slope coefficient in the theoretical simple linear regression (SLR) of D on Z is

$$\lambda := \frac{\mathbb{C}\text{ov}(D, Z)}{\mathbb{V}[Z]} \overset{\text{if } Z \text{ binary}}{=} \mathbb{E}[D \,|\, Z = 1] - \mathbb{E}[D \,|\, Z = 0].$$

Hence, in this setting, the relevance condition $\mathbb{C}\text{ov}(D, Z) \neq 0$ is equivalent to $\lambda \neq 0$ with λ the theoretical slope coefficient in the SLR of D on Z, which is the first-stage regression in this setting, without control variables.

The relevance condition can (and should) be tested with a standard simple bilateral t-test (see Chapter 2) $H_0 : \lambda = 0$ against $H_1 : \lambda \neq 0$. A rejection of the null hypothesis H_0 supports the relevance condition.

When Z is binary, the relevance condition is equivalent to $\mathbb{E}[D \,|\, Z = 1] \neq \mathbb{E}[D \,|\, Z = 0]$. This explains why the condition $\mathbb{E}[D \,|\, Z = 1] > \mathbb{E}[D \,|\, Z = 0]$ in Theorem 1 (slide 9) is the expression of the relevance condition in the particular case of a binary instrument Z.[7]

[7] With this ranking/sense of the inequality because under the monotonicity assumption (2) (no defiers): the only way to have $\mathbb{E}[D \,|\, Z = 1] \neq \mathbb{E}[D \,|\, Z = 0]$ is to have $\mathbb{E}[D \,|\, Z = 1] > \mathbb{E}[D \,|\, Z = 0]$. Besides, as regards Theorem 1, the independence condition (1) implies the exogeneity condition, which is the second condition, along with relevance, for the instrument Z to be valid.

– Univariate D, multivariate (or univariate) Z, with controls G.

The idea behind the relevance of Z remains the same: Z needs to have an effect on D, but the idea is that Z needs to affect D *net of the influence of G on D; in other words, conditional on G*. In this more general setting (starting in slide 29 of Chapter 5) with exogenous controls G and with $Z \in \mathbb{R}^q$ that can be a multivariate random variable (same result if Z is a scalar, $q = 1$), the relevance condition writes: $\mathbb{E}[X^* X^{*\prime}]$ is invertible, with $X^* := (1, G', D^*)'$ where D^* is the theoretical linear prediction of D from G and Z (see the notation of slide 30).

Why is this expression related to the intuitive idea of Z affecting D "net of G" or, in other words, conditionally on G? Imagine Z does not affect D when we control by G, namely α_2 is equal to 0, where α_2 is the coefficients associated with Z in the first-stage regression of the endogenous regressor D on the controls G and the instruments Z (see the notation of slide 30, equation (5)). Note that, in general, α_2 *is a vector* of dimension q, the dimension of Z.

If $\alpha_2 = 0$, the predicted value D^* is simply a linear combination of G since Z does not intervene because $\alpha_2 = 0$. Therefore, there is a linear relationship in $X^* := (1, G', D^*)'$ and $\mathbb{E}[X^* X^{*\prime}]$ is not invertible.

This explains why, if we assume that 1, G and Z are not linearly dependent (no perfect colinearity between controls G and instruments Z), the condition $\mathbb{E}[X^* X^{*\prime}]$ invertible is equivalent to $\alpha_2 \neq 0$ (this is the result of Proposition 1, slide 31).

In practice, the expression $\alpha_2 \neq 0$ is the one used to test the relevance condition in the general framework of multivariate instrument Z and with controls G. The relevance condition can (and should) be tested thanks to the multiple F-test (see Chapter 2) $H_0 : \alpha_2 = 0 \ (= 0_{\mathbb{R}^q})$ against $H_1 : \alpha_2 \neq 0$. A rejection of the null hypothesis H_0 supports the relevance condition.

Note that this test is not the same as the test of joint nullity of *all* the coefficients, including also those of G whose null hypothesis is $H_0 : \alpha_1 = \alpha_2 = 0$ (with the notation of slide 30). To test the relevance condition, we need to focus on the coefficients associated with the instruments Z only, although we need to include the controls G in the first-stage regression.

The previous results remain unchanged in the particular case of a *univariate* instrument Z with controls G. Formally, $q = 1$ and α_2 is just a real number instead of a q-dimensional vector. This explains the correct answer 5. Regarding the test of the relevance condition, with a univariate Z, α_2 is a real number, and the previous F-test is the same as doing the simple bilateral t-test $H_0 : \alpha_2 = 0$, $H_1 : \alpha_2 \neq 0$.

– Multivariate D, multivariate Z, with controls G.

This is the most general case with possibly multivariate endogenous regressors D, multivariate instruments Z, and with controls G. The relevance condition still writes $\mathbb{E}[X^* X^{*\prime}]$ is invertible, with $X^* := (1, G', D^{*\prime})'$ where $D^* = (D_1^*, \dots, D_p^*)'$, and D_j^* is the theoretical linear prediction of D_j from G and Z (see the notation of slide 33).

Proposition 2 (slide 34) provides an equivalent formulation if we assume that 1, G, and Z are not linearly dependent: the matrix A_2 of the coefficients of Z in the separate first-stage regressions (one for each of the components of $D = (D_1, \dots, D_p)'$) must be of rank p.

Remark that this setting encompasses the simpler case with a univariate endogenous regressor D since, in this case, $A_2 = \alpha_2$ and the rank of this vector, seen as a (column) matrix, is not null if and only if the vector is not the null vector, that is, $\alpha_2 \neq 0$.

1. $\mathbb{E}[Z\eta] = 0$ – **False**, on the contrary, this is the exogeneity condition for the instrument Z.

2. $\mathbb{C}\text{ov}(D, Z) \neq 0$ – **False**, 2. and 4. are equivalent, but it is the relevance condition *without* controls, for univariate D and Z.

3. $\mathbb{C}\text{ov}(G, Z) \neq 0$ – **False**, it has nothing to do here. Nonetheless, remark that to have 1, G and Z not linearly dependent, the correlation between G and Z cannot be exactly one or minus one.

4. The theoretical coefficient of Z in the simple linear regression of D on Z is not null – **False**.

5. The theoretical coefficient of Z in the multiple linear regression of D on G and Z is not null – **True**.

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

Yes, it is possible to test the relevance condition. It is a major difference compared with the other condition required for Z to be a valid instrument: exogeneity.

To do so, in the first-stage regression of D on G and Z, we can perform a F-test of the joint nullity of the coefficients α_2 associated with the instruments Z: $H_0 : \alpha_2 = 0$ against $H_1 : \alpha_2 \neq 0$ (see above). See slide 39 of Chapter 5 about testing the relevance condition.

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 $\mathbb{V}[\eta \,|\, Z] = \sigma^2$, a constant that does not depend on Z.

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

Remark: this question is a bit more advanced; you can focus on the ideas only; the result is written in the course slides, though.

If we assume[8] $\mathbb{E}[\eta^2 \,|\, Z] = \sigma^2$, it is possible to show in this setting (no control variables, univariate endogenous regressor D and instrument Z) that the asymptotic variance of the 2SLS estimator of the slope, that is, of δ_0 here, is equal to

$$V_a(\widehat{\beta}_{\text{2SLS, slope}}) = \frac{\sigma^2}{\mathbb{V}[D^*]},$$

with D^* the theoretical prediction of D by Z.

This result is written in slide 44 of Chapter 5. To prove it, it is the same computation as in Chapter 2, equation (3), slide 10, using the asymptotic normality of the 2SLS estimator (Theorem 6, Chapter 5, slide 37), and the homoscedasticity-type condition $\mathbb{E}[\eta^2 \,|\, Z] = \sigma^2$ for the instrument Z. You can check that, as well as the result of equation (3) of Chapter 2; it is good training in the spirit of the theoretical exercises of the exams.

Answer 1 is false because, on the contrary, if Z is less useful to predict D, the variance of the prediction D^* is lower. To see that, consider the limit case where Z is useless to predict D; in this case, D^* is simply the expectation of D, a constant, and $\mathbb{V}[D^*]$ is null.

1. the prediction of D by Z worsens (that is, Z is less useful to predict D) – **False**.
2. $\mathbb{E}[Z]$ increases – **False**, nothing to do. We could translate Z or change the units of Z, and it would not modify the reasoning of using Z as an instrument of D and should not impact the inference.
3. The sample size n increases – **False**, note that, by definition, the *asymptotic* variance cannot depend on the sample size n.
4. σ^2 decreases – **True**.

[8] Rather than $\mathbb{V}[\eta \,|\, Z] = \sigma^2$ in fact. If we assume a stronger type of exogeneity condition: $\mathbb{E}[\eta \,|\, Z] = 0$ (which implies $\mathbb{E}[Z\eta] = 0$), then $\mathbb{V}[\eta \,|\, Z] = \mathbb{E}[\eta^2 \,|\, Z]$ and the two assumptions are equivalent.

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

$$Y(d) = \zeta_0 + \delta_0 d + \eta, \quad \mathbb{E}[\eta] = 0, \tag{2}$$

(*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 are in the setting of a univariate endogenous regressor D, a univariate instrument Z,[9] and without exogenous controls G. In this setting (see Question 9 of this quiz for details) **the relevance condition** simply writes

- $\boxed{\mathbb{C}\text{ov}(Z, D) \neq 0}$: Z “has an effect” on D, meaning is correlated with D; Z is said to be *relevant* (“instrument pertinent” in French).

The exogeneity condition (i) rules out an effect of Z on Y other than through the regressor D, (ii) asserts that Z is uncorrelated with the potential outcome variables. It writes

- $\boxed{\mathbb{C}\text{ov}(Z, Y(d_0)) = 0}$ where $Y(d_0)$ is the potential outcome variable evaluated at d_0 (for any d_0 – free variable, “variable muette” – in the support of D; Z is said to be *exogenous with reference to or in Equation* (2) (“instrument exogène” in French).

Implicitly, the condition needs to hold for any d_0 in the support of D (because we can choose without loss of generality any d_0 to write the linearity of the effect – see slide 20, Chapter 5).

Another way to write the exogeneity condition is to remark that, given the assumed model (2) on the potential outcomes, we have

$$Y(d_0) = \zeta_0 + \delta_0\, d_0 + \eta, \tag{3}$$

and, *in this equation, the only stochastic term is η (and $Y(d_0)$ therefore, of course)*.

The random variable η represents unobserved factors that affect the potential outcomes and are specific to each individual (**“unobserved individual heterogeneity”**). Remember that, as always in the course, we assume i.i.d. sampling, and therefore, to lighten notations, the individual indices i are omitted. If we use them, for any individual i, Equation (3) rewrites

$$Y(d_0)_i = \zeta_0 + \delta_0\, d_0 + \eta_i. \tag{4}$$

As η is the only stochastic term in the right-hand side expression of (3), we have

$$\mathbb{C}\text{ov}(Z, Y(d_0)) = \mathbb{C}\text{ov}(Z, \eta).$$

Hence, the exogeneity condition can be written equivalently as $\boxed{\mathbb{C}\text{ov}(Z, \eta) = 0}$.

Finally, when η is centered ($\mathbb{E}[\eta] = 0$), which is without loss of generality since model (2) includes a constant ζ_0, the exogeneity condition is also equivalent to $\boxed{\mathbb{E}[Z\eta] = 0}$.

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

[9] But note that Z and D are not necessarily binary.

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

Remark: this question is important and good training for theoretical exercises.

As explained in the course at several places in Chapter 5 (see notably slides 12, 24, and 36), in this setting without control variables with a scalar treatment D, the 2SLS estimator is obtained by doing:

1. first, the regression of D on Z and get the predicted value $\widehat{D}$ (first-stage),
2. second, the regression of Y on $\widehat{D}$ (second-stage).

To determine the limit in probability of $\widehat{\delta}_{2\text{SLS}}$ under usual i.i.d. sampling and moment conditions, it is directly the proof of Theorem 2 (slides 12 and 13).

By definition of the slope coefficient in a simple linear regression (Proposition 5, Chapter 1),

$$\widehat{\delta}_{2\text{SLS}} \xrightarrow[n\to+\infty]{P} \frac{\mathbb{C}\text{ov}(D^*, Y)}{\mathbb{V}[D^*]}.$$

where D^* is the limit in probability of $\widehat{D}$; it is the theoretical error term and is equal to $D^* = \gamma_0 + \gamma_1 Z$, where (again from Proposition 5, Chapter 1 – one of the most important propositions of the course!) $\gamma_1 = \mathbb{C}\text{ov}(Z, D) \,/\, \mathbb{V}[Z]$.

Therefore, using the bilinearity of the covariance and the properties of the variance, we obtain

$$\begin{aligned}
&\mathbb{C}\text{ov}(D^*, Y) = \mathbb{C}\text{ov}(\gamma_0 + \gamma_1 Z, Y) = \gamma_1 \mathbb{C}\text{ov}(Z, Y)\\
&\mathbb{V}[D^*] = \mathbb{V}[\gamma_0 + \gamma_1 Z] = \gamma_1^2 \mathbb{V}[Z].
\end{aligned}$$

Using the previous computations and the expression of γ_1, we get

$$\frac{\mathbb{C}\text{ov}(D^*, Y)}{\mathbb{V}[D^*]} = \frac{\gamma_1 \mathbb{C}\text{ov}(Z, Y)}{\gamma_1^2 \mathbb{V}[Z]} = \frac{\mathbb{C}\text{ov}(Z, Y)}{\gamma_1 \mathbb{V}[Z]} = \frac{\mathbb{C}\text{ov}(Z, Y)}{\mathbb{C}\text{ov}(Z, D)}.$$

Hence, the correct answer 3.

1. $\dfrac{\mathbb{C}\text{ov}(D^*, Y)}{\mathbb{V}[D]}$ – **False**, it would be correct with $\mathbb{V}[D^*]$ instead in the denominator (Theorem 3).
2. $\dfrac{\mathbb{C}\text{ov}(Z, Y)}{\mathbb{V}[Z]}$ – **False**, it would correspond to the limit in probability of the OLS estimator of the slope in the simple linear regression of Y on Z, but it is not what we want here.
3. $\dfrac{\mathbb{C}\text{ov}(Z, Y)}{\mathbb{C}\text{ov}(Z, D)}$ – **True**.
4. None of the previous expressions; if so, indicate a correct one below – **False**.

(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 $\mathbb{C}\text{ov}(Z, D)$.

The causal model (2) is valid for any d (d is a free variable, "variable muette" in French) in the support of D (D is a real random variable, the treatment variable). Therefore, "applied" in the random variable D, Equation (2) yields

$$Y \overset{\text{definition of } Y}{:=} Y(D) \overset{\text{from model (2)}}{=} \zeta_0 + \delta_0 D + \eta.$$

We replace the *observed outcome* Y by the previous expression in the answer to question (b). Using the bilinearity of covariance, we obtain

$$\begin{aligned}\widehat{\delta}_{2\text{SLS}} \xrightarrow[n\to+\infty]{P} \frac{\mathbb{C}\text{ov}(D^*,Y)}{\mathbb{V}[D^*]} &= \frac{\mathbb{C}\text{ov}(Z,Y)}{\mathbb{C}\text{ov}(Z,D)}\\ &= \frac{\mathbb{C}\text{ov}(Z,\zeta_0+\delta_0 D+\eta)}{\mathbb{C}\text{ov}(Z,D)}\\ &= \frac{\delta_0\mathbb{C}\text{ov}(Z,D)+\mathbb{C}\text{ov}(Z,\eta)}{\mathbb{C}\text{ov}(Z,D)}\\ &= \delta_0 + \frac{\mathbb{C}\text{ov}(Z,\eta)}{\mathbb{C}\text{ov}(Z,D)}.\end{aligned}$$

This is the same result as the one of slide 41 of Chapter 5 (except for the change of notations, D instead of X).

As a comparison, the OLS slope estimator in the simple linear regression of Y on D converges in probability[10] to

$$\begin{aligned}\frac{\mathbb{C}\text{ov}(D,Y)}{\mathbb{V}[D]} &= \frac{\mathbb{C}\text{ov}(D,\zeta_0+D\delta_0+\eta)}{\mathbb{V}[D]}\\ &= \frac{\delta_0\mathbb{C}\text{ov}(D,D)+\mathbb{C}\text{ov}(D,\eta)}{\mathbb{V}[D]}\\ &= \delta_0 + \frac{\mathbb{C}\text{ov}(D,\eta)}{\mathbb{V}[D]}.\end{aligned}$$

(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.

The previous computation on the probability limit of $\widehat{\delta}_{2\text{SLS}}$ requires the relevance condition, which writes $\mathbb{C}\text{ov}(Z,D)\neq 0$ in this setting. Otherwise, D^* is a constant variable, and $\mathbb{V}[D^*]$ is null: the probability limit is not correctly defined as the denominator is null.

The bias term (in the econometrics sense:[11] the difference of the limit in probability of the estimator minus the target parameter we want to estimate) for $\widehat{\delta}_{2\text{SLS}}$ is equal to

$$\underbrace{\frac{\mathbb{C}\text{ov}(D^*,Y)}{\mathbb{V}[D^*]}}_{\text{plim } \widehat{\delta}_{2\text{SLS}}} - \underbrace{\delta_0}_{\text{targeted parameter}} = \frac{\mathbb{C}\text{ov}(Z,\eta)}{\mathbb{C}\text{ov}(Z,D)}.$$

If the exogeneity condition holds, $\mathbb{C}\text{ov}(Z,\eta)=0$, this bias term is null: we get back to the result of Theorem 3 (slide 23), namely the 2SLS estimator identifies the causal effect δ_0.

But, as long as $\mathbb{C}\text{ov}(Z,\eta)\neq 0$, the closer $\mathbb{C}\text{ov}(Z,D)$ to 0, the larger the bias term in absolute value (indeed, we have a non-null numerator and a denominator close to 0). Hence the potential issues of so-called “weak instruments”: instruments Z with $\mathbb{C}\text{ov}(Z,D)\neq 0$ but close to 0. In other words, those weak instruments do have an effect on D but only a small one. For such instruments, it is enough that the exogeneity condition is not *exactly* satisfied to have a large bias. This discussion refers to the last point of slide 41 of Chapter 5.

Besides, as regards inference (as opposed to estimation only) even if the exogeneity condition is satisfied so that there is no bias (estimation is fine), the weaker the instrument, the lower the accuracy of the estimation of δ_0 basically (see Question 10).

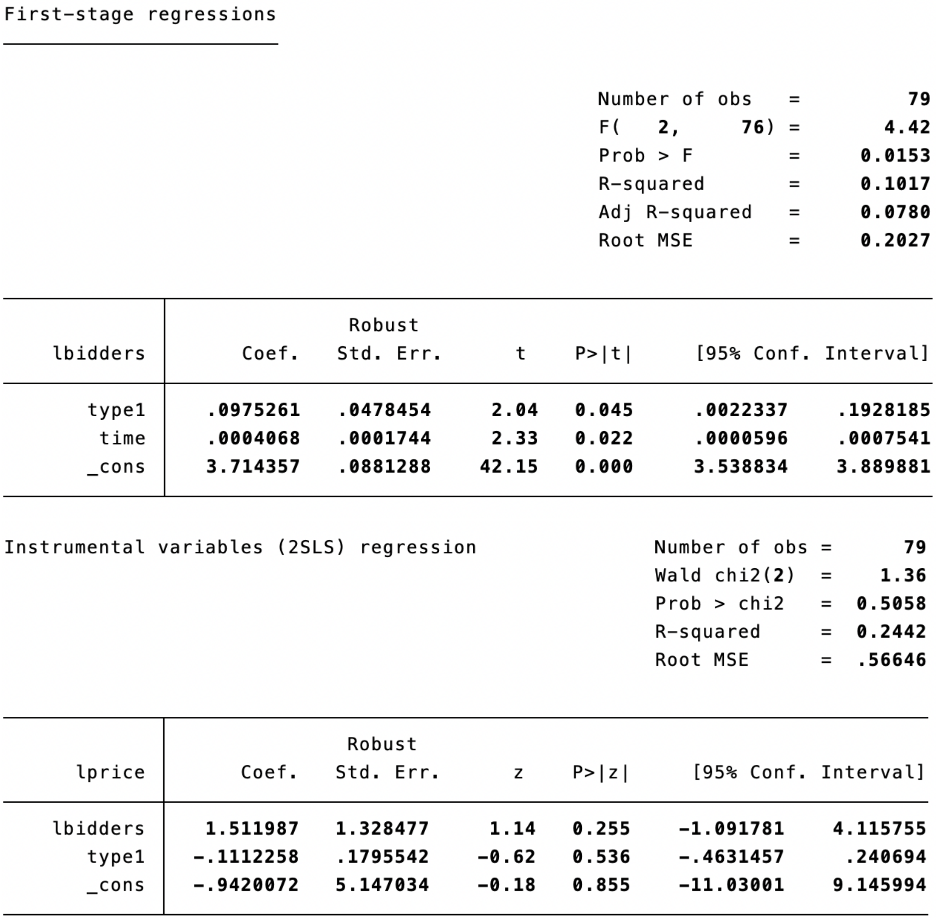
[10] Assuming i.i.d. sampling and the classical moment conditions: $\mathbb{E}[|Y|^2]<+\infty$, $\mathbb{E}[|D|^2]<+\infty$, $\mathbb{V}[D]>0$.

[11] Unlike the Statistics 1 sense where being biased or unbiased is a *non-asymptotic* = *finite-sample* = *valid for any sample size* n of an estimator. The finite-sample bias of an estimator $\widehat{\theta}_n$, based on a n-sample, of a parameter θ is defined as $\mathbb{E}_\theta[\widehat{\theta}_n]-\theta$. In Econometrics 1, the approach is asymptotic.

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.

The Stata command that generates this output is:

```
ivregress 2sls lprice (lbidders = time) type1, robust first.
```

The command `ivregress 2sls` performs 2SLS estimation.

The first variable written next is the outcome variable Y.

Then the syntax expects between parenthesis the instrumented endogenous variable(s) D, an equal sign "=", followed by the instrument(s) Z. Hence, here, $D =$ `lbidders` is the endogenous regressor, and $Z =$ `time` is the instrument. Then, afterward, outside the parenthesis, the syntax requires the exogenous control variable(s) G in the notation of Chapter 5. Here, there is a single control $G =$ `type1`.

Finally, we use the command with two options, written after the comma ",".

The option `first` tells Stata to display in the console the first-stage regression: the upper table ("`First-stage regressions`")[12] By default, without option `first`, Stata only reports the second-stage regression: the lower table, "`Instrumental variables (2SLS) regression`"

The option `robust` computes the standard errors without assuming homoscedasticity, robust to heteroscedasticity. With the syntax used, as revealed by the "Robust Std. Err." in the second column of the two tables, the robust standard errors are computed, both in first-stage and second-stage regressions. The option `robust` is used to stick to the standard set-up seen in your course (see Theorem 6 for 2SLS and, more generally, Chapter 2 about inference).

[12] Stata writes First-stage regressions with an "s" as there could be several endogenous regressors possibly; if so, Stata reports each separate first-stage regression. Here, there is only one for the only endogenous variable `lbidders`.

(b) Explain what that command did by specifying

1. the outcome variable Y: `lprice` (logarithm of the unit price paid in the auction).

2. the possibly endogenous regressor(s) D: `lbidders` (logarithm of the number of bidders).

3. the exogenous control(s) G: `type1` (an indicator/dummy variable that indicates whether the begonia belongs to a particular type – like a certain specific color or another quality).

4. the instrument(s) Z: `time` (the time at which the auction took place, measured in seconds elapsed from half-past six).

5. and the 2SLS estimator we obtained here: with the formalization of the course from model (4) (slide 29), we have

$$Y(d) = \zeta_0 + \gamma_0 G + \delta_0 d + \eta, \ \mathbb{E}[\eta] = \mathbb{E}[G\eta] = 0, \tag{5}$$

where $Y(d)$ is the potential logarithm of the price paid if the logarithm of the number of bidders who attend the auction were equal to d; G is an exogenous univariate control, the variable `type1`. The logarithm of the number of bidders D is possibly endogenous: we do *not* assume $\mathbb{E}[D\eta] = 0$. But we assume the univariate instrument $Z =$ `time` is exogenous: $\mathbb{E}[Z\eta] = 0$. We perform the 2SLS estimation of the vector of parameters $(\zeta_0, \gamma_0, \delta_0)'$ with Z instrumenting D. We obtain for the estimator $\widehat{\beta}_{2SLS} = (\widehat{\zeta}, \widehat{\gamma}, \widehat{\delta})'$ with

- $\widehat{\zeta} = -0.942$
- $\widehat{\gamma} = -0.111$
- $\widehat{\delta} = 1.512$.

If we assume that the instrument Z is valid (*validity of the instrument*) and that the previous model (5) on $Y(d)$ holds (*linear homogeneous causal effects*), this figure is interpreted as a consistent estimator of the causal effect of the number of bidders on the price paid.

More specifically, it is a log-log model here, so that δ_0 is the elasticity of the price with respect to the number of bidders – the elasticity is assumed constant by the linearity of model (5). Hence, the interpretation would be: we estimate that an increase of 1% of the number of bidders causes an increase of the price paid by 1.512%. Thus, the estimate would confirm the microeconomic theory explained in the Problem Set.

Yet, it would be a too quick, if not wrong, answer since we need to take care of the statistical significance of δ_0. Indeed, notice the large standard error for $\widehat{\delta}$, the broad 95% confidence interval, and the large p-value of the statistical significance test with null hypothesis $H_0 : \delta_0 = 0$ against $H_1 : \delta_0 \neq 0$. The p-value of this test is available in the lower table "`Instrumental variables (2SLS) regression`". **N.B.: for 2SLS Stata output tables, it is the same interpretation and reading as in the standard OLS case of Chapter 5 with OLS `regress` Stata output tables.** Here, the p-value is $0.255 = 25.5\%$: we do not reject H_0 at any usual level (1%, 5%, nor 10%). Consequently, with this data, we cannot reject that the effect of the number of bidders on the price is not null, and we do not confirm the microeconomics theory: we should collect more data ($n = 79$ only here) or search for a stronger instrument.

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

To discuss and test the relevance condition, be careful of the setting: univariate or multivariate endogenous regressor D? Same question for the instrument Z? Are there exogenous controls G or not? (See Question 9 of this quiz for details.). Here, we have a univariate endogenous regressor D, a univariate instrument Z, and exogenous controls G.

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.

This is directly the first part of slide 39 about the relevance condition.

We consider the first-stage regression of D on Z and G whose theoretical coefficients are (see slide 30, equation (5) of Chapter 5)

$$(\alpha_0, \alpha_1, \alpha_2) = \arg \min_{(a_0,a_1,a_2)\in\mathbb{R}^3} \mathbb{E}[(D - a_0 - a_1 G - a_2 Z)^2],$$

and we test the null hypothesis of the nullity of the coefficients of the instruments Z: $H_0 : \alpha_2 = 0$ against $H_1 : \alpha_2 \neq 0$. In general (with possibly multivariate instruments), it will be done by an F-test. Here, it happens that Z is univariate, so that we can perform the test with a simple bilateral t-test of the statistical significance of $Z =$ `time` in the first-stage regression (upper table (“`First-stage regressions`”))

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

   The p-value of this test is directly available in the upper table (“`First-stage regressions`”) at the row corresponding to the variable `time` and is equal to $0.022 = 2.2\%$.

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%?

   $2.2\% < 5\%$ but $> 2\%$ and 1%. Hence, we reject the null hypothesis at 5% but not at 2% nor, a fortiori, at 1%. Therefore, we are relatively confident in the relevance condition (because we reject its opposite H_0 at 5%), yet, we cannot validate it at the 1% level.

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?

The regression of the endogenous regressor D on a constant As usual, by default, there is a constant in any regression we do. , the exogenous controls G and the instruments Z is called the *first-stage regression* (*“régression de première étape”* in French).

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

It is directly the result of slide 46. The augmented regression is the regression of Y on X and $\widehat{\nu}$, where $X := (1, G', D')'$. Therefore, the augmented regression is the regression of Y on a constant, the endogenous regressor D, the exogenous controls G, and $\widehat{\nu}$.

1. a constant, D, G, Z, and $\widehat{\nu}$ – **False**, Z is wrongly here.

2. a constant, D, G, and $\widehat{\nu}$ – **True**.

3. a constant, D, and $\widehat{\nu}$ – **False**, the exogenous controls G are missing.

4. None of the previous answers; if so, indicate the correct one below – **False**.

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

1. No – **False**.
2. Yes. In this case, explain how. – **True**, the OLS estimator of the augmented regression yields the 2SLS estimator.

It is directly the result of Lemma 1 (slide 45) and the following comments (slide 46).

It is possible to prove that the OLS estimator of the augmented regression, denoted $(\widehat{\beta}_{\text{aug}}, \widehat{\rho})$ ($\widehat{\rho}$ is the estimator of the coefficient associated with the residual of the first-stage regression) coincides with the 2SLS estimator obtained when instrumenting D by Z in the sense that, putting apart $\widehat{\rho}$ (which is not computed in the 2SLS estimation method), we have $\widehat{\beta}_{\text{aug}} = \widehat{\beta}_{\text{2SLS}}$.

This approach is an alternative to the standard 2SLS estimation seen in the course. The 2SLS is a "projection" approach. In contrast, the augmented regression strategy is often called "a control variable approach"[13]. The basic idea is to add ν ($\widehat{\nu}$ in practice, which is recovered from the first-stage regression, since ν is unknown) in the initial equation: ν is this so-called "control variable" in the sense that adding it enables to solve the issue of endogeneity of D. The intuition behind this is the following. Given the exogeneity of Z (and of G), $\nu = D - D^* = D - (\alpha_0 + \alpha_1 G + \alpha_2 Z)$ contains all the endogenous part of D. Therefore, once we control for ν in the regression, it is as if D was exogenous, and there is no more endogeneity issue. You can remember that idea as we will use it again in Econometrics 2 next semester.

[13] Remark: this name has nothing to do with the presence or absence of exogenous controls G. It would work the same without additional exogenous controls G.