Econometrics 1
Chapter 5: instrumental variables

Xavier D'Haultfœuille and Elia Lapenta

- A key condition for the OLS estimator to estimate consistently a causal effect is the absence of selection.

- Without control variables, this corresponds to $\text{Cov}(D, Y(d)) = 0$ for all d.

- With control variables, this corresponds to $\text{Cov}(D, Y(d)|G) = 0$ for all d.

- In many cases, not very credible.

- For instance, hard to be sure that we observe all the factors that explain both education and wages.

- Another approach here for causal inference : use special variables ("instruments") that affect D but not directly potential outcomes.

- In some cases (but not always !), we have natural candidates for such variables.

- We extend the OLS method to use such variables in a meaningful way.

- In medical randomized experiments :
    - all patients in the treatment group receive the treatment
    - while none in the control group get it.

- Simple situation b/c patients (and doctors) usually ignore who is treated and who is control.

- Not the case in social sciences.

- Even if they initially volunteered, participants may lose interest for the « treatment » and opt out.

- Sometimes, they do not even volunteer (e.g., when treatment is drawn at a larger scale, e.g. classrooms or unemployment centers).

    $\implies$ some people in the treatment group are eventually not treated.

- Sometimes also, individuals in the control group still manage to receive the treatment.

- ▶ Then, one should disentangle the initial allocation $Z \in \{0, 1\}$ from the treatment effectively received $D \in \{0, 1\}$ (in case of a binary treatment).

- ▶ We still let $Y(0)$ (resp. $Y(1)$) denote the potential outcome absent the treatment (resp. with the treatment).

- ▶ By construction, Z is "random", namely $Z \perp\!\!\!\perp (Y(0), Y(1))$.

- ▶ But in general, we do not have $D \perp\!\!\!\perp (Y(0), Y(1))$:
  - ▶ in the treatment group, those forgoing the treatment ($D = 0, Z = 1$) are potentially different from the others ($D = 1, Z = 1$) ;
  - ▶ in the control group, those managing to be treated ($D = 1, Z = 0$) are potentially different from the others ($D = 0, Z = 0$).

- Then, in general

$$E(Y|D = 1) - E(Y|D = 0) \neq \delta^T = E(Y(1) - Y(0)|D = 1).$$

- Therefore, the theoretical regression of Y on D does not identify the causal effect of D.

- Idea : since $Z \perp\!\!\!\perp (Y(0), Y(1))$, use Z rather than D.

- If we redefine the treatment as « was supposed to receive the treatment » , the theoretical regression of Y on Z identifies this new « treatment » .

- But often not a very interesting treatment !

- Nonetheless, we can identify an average causal effect of D, under additional conditions.

- Let $D(z)$ be the potential treatment corresponding to $Z = z$.

- For instance, $D(1) = 0$ if the individual forgo the treatment though she was affected to the treatment group.

- Let us suppose that
$$Z \perp\!\!\!\perp (Y(0), Y(1), D(0), D(1)) \qquad \text{(Indep.)}$$

- (Indep.) is very credible since allocation to groups is done randomly.

- Then :
$$E(Y|Z = z) = E[D(z)Y(1) + (1 - D(z))Y(0)|Z = z]$$
$$= E[D(z)(Y(1) - Y(0)) + Y(0)].$$

- Thus $E[Y|Z = 1] - E[Y|Z = 0] = E\left[(D(1) - D(0))(Y(1) - Y(0))\right]$.

▶ The second condition is monotonicity :

$$D(1) \geq D(0) \text{ almost surely.} \qquad \text{(Monoton.)}$$

▶ It holds if there are no « defiers » (« rebels ») : individuals who would get the treatment if in the control group, but would not get it if in the treatment group.

▶ (Monoton.) holds in particular if $D(0) = 0$: no one, if allocated in the control group, manages to get the treatment.

▶ Let us define

$$\delta^C = E[Y(1) - Y(0)|D(1) - D(0) = 1].$$

▶ The exponent C refers to « compliers » , as $D(1) - D(0) = 1$ is equivalent to complying to Z ($D(z) = z$ for $z \in \{0, 1\}$).

▶ We will show that we can identify δ^C (rather than δ^T) from (Indep.)-(Monoton.) and the distribution of (Y, D, Z).

Theorem 1

If D and Z are binary, (Indep.)-(Monoton.) *hold and* $E(D|Z = 1) > E(D|Z = 0)$,

$$\delta^C = \frac{E(Y|Z = 1) - E(Y|Z = 0)}{E(D|Z = 1) - E(D|Z = 0)}.$$

Proof : we already showed, under (Indep.), that

$$E[Y|Z = 1] - E[Y|Z = 0] = E\left[(D(1) - D(0))(Y(1) - Y(0))\right].$$

If (Monoton.) also holds,

$$E\left[(D(1) - D(0))(Y(1) - Y(0))\right]$$
$$= P(D(1) - D(0) = 1)E\left[Y(1) - Y(0)|D(1) - D(0) = 1\right].$$

Moreover, under (Indep.), $E(D|Z = z) = E(D(z))$. Then, using again (Monoton.),

$$E(D|Z = 1) - E(D|Z = 0) = E(D(1) - D(0)) = P(D(1) - D(0) = 1).$$

The result follows $\square$

- Intuition for the result :
  - $E(Y|Z=1) - E(Y|Z=0)$ is too small in absolute value b/c of untreated people in the treatment group, and treated people in the control group.
  - We then divide it by the difference of treatment rate b/w the two groups, $E(D|Z=1) - E(D|Z=0)$.
- We only measure the average treatment effect among those reacting to Z (the « compliers »).
- For the others, such that $D(0) = D(1)$, the information is not available in the data.
- We call those s.t. $D(0) = D(1) = 1$ the « always takers » and those s.t. $D(0) = D(1) = 0$ the « never takers » .

- δ^C is often called the « local average treatment effect » (or LATE). Local since it focuses on a subpopulation, the « compliers » .

- Link b/w δ^C and δ^T : we have

$$\delta^T = \lambda\delta^C + (1-\lambda)E[Y(1) - Y(0)|D(0) = D(1) = 1]$$

with $\lambda = \dfrac{P(D(1) > D(0), Z = 1)}{P(D(1) > D(0), Z = 1) + P(D(1) = D(0) = 1)}.$

- Then we have :

Corollary 1

If D and Z are binary, (Indep.) *holds and $E(D|Z = 1) > E(D|Z = 0) = 0$,*

$$\delta^T = \delta^C = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)}.$$

- If $D = Z$ (viz., perfect compliance), we recover Prop. 1 of Chapter 4

▶ Let D^* be the theoretical prediction of D by Z : $D^* = \gamma_0 + \gamma_1 Z$, with

$$(\gamma_0, \gamma_1) = \arg\min_{c_0, c_1} E[(D - c_0 - c_1 Z)^2].$$

Theorem 2

If D and Z are binary, (Indep.)-(Monoton.) *hold, and $V(D^*) > 0$, then*

$$\delta^C = \frac{Cov(D^*, Y)}{V(D^*)}.$$

▶ Thus, δ^C can be obtained by two (theoretical) regressions :
  ▶ We regress D on Z to obtain D^* ;
  ▶ We regress Y on $D^* \Rightarrow$ the coefficient of D^* is equal to δ^C.

Proof of Theorem 2 : since $D^* = \gamma_0 + \gamma_1 Z$, we have

$$\text{Cov}(D^*, Y) = \gamma_1 \text{Cov}(Z, Y),$$
$$V(D^*) = \gamma_1^2 V(Z).$$

Hence :

$$\frac{\text{Cov}(D^*, Y)}{V(D^*)} = \frac{\text{Cov}(Z, Y)}{\gamma_1 V(Z)}.$$

Moreover, by e.g. Proposition 5 in Chap. 1, $\gamma_1 = \text{Cov}(Z, D)/V(Z)$. Thus,

$$\frac{\text{Cov}(D^*, Y)}{V(D^*)} = \frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, D)}.$$

Since Z is binary, $\text{Cov}(A, Z)/V(Z) = E(A|Z = 1) - E(A|Z = 0)$ for all random var. A. The result follows by Theorem 1 $\square$

- If we have a sample $(D_i, Y_i, Z_i)_{i=1,\ldots,n}$, we can estimate δ^C by

$$\widehat{\delta}^C = \frac{1/n_1 \sum_{i:Z_i=1} Y_i - 1/n_0 \sum_{i:Z_i=0} Y_i}{1/n_1 \sum_{i:Z_i=1} D_i - 1/n_0 \sum_{i:Z_i=0} D_i},$$

  with $n_z = \sum_{i=1}^n \mathbb{1}\{Z_i = z\}$ for $z \in \{0, 1\}$.

- Equivalently, we can compute $\widehat{\delta}^C$ by :

  1. Regressing D on Z, so as to obtain the predictions $\widehat{D}_i$ for all i ;

  2. Regress Y on $\widehat{D}$. $\widehat{\delta}^C$ is the slope coeff. of this reg.

- If $E(|Y|) < \infty$, by the law of large numbers and the continuous mapping theorem,

$$\widehat{\delta}^C \xrightarrow{P} \delta^C.$$

Example : evaluation of a job training program

- We seek to evaluate the U.S. labor market program « Job Training Partnership Act » (JTPA) on future earnings.

- This program included a randomized experiment : treatment was *proposed* to participants drawn randomly.

- Only $\simeq 62\%$ of those drawn randomly indeed participated to the program.

- $\simeq 2\%$ of those not drawn still participate to it $\Rightarrow \delta^C \simeq \delta^T$.

- Conclusions from Table 1 ?

Table 1 – Results from the JTPA experiment

	$E(Y\|D=1)$ $-E(Y\|D=0)$	$E(Y\|Z=1)$ $-E(Y\|Z=0)$	δ^C
Men	3 970 (555)	1 117 (569)	1 825 (928)
Women	2 133 (345)	1 243 (359)	1 942 (560)

Notes : results from Abadie et al. (2002), « Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings ». Std. err. under parentheses.

▶ The previous analysis extends outside randomized experiments, through « natural » experiments.

▶ We seek to identify the causal effect of D on Y, but D is not drawn randomly $\Rightarrow D$ may not be independent from $Y(0)$.

▶ Idea : find a variable Z affecting D but independent of $(Y(0), Y(1), D(0), D(1))$.

▶ Example : effect of fertility on female participation to the labor market $(\Rightarrow Y \in \{0, 1\}$ here).

▶ Idea of Angrist and Evans (1998, « Children and Their Parents' Labor Supply : Evidence from Exogenous Variation in Family Size ») : use :

   (i) the information on the sex of the first 2 children ;

   (ii) the fact that parents may value "diversity", viz. prefer to have both boys and girls.

- Let us restrict ourselves to parents with at least two 2 children. Let $D = 1$ if parents have a 3rd child, $D = 0$ otherwise.

- Let $Z = 1$ if the first 2 children are of same sex, 0 otherwise.

- $D(1) =$ have a 3rd child if the first two are of same sex.

  $D(0) =$ have a 3rd child if the first two are of different sex.

- Z is independent of $(Y(0), Y(1), D(0), D(1))$

- The monotonicity condition holds if parents :
    - either are not affected by the sex composition of their first 2 children in their decision to have a third child $(D(0) = D(1))$;
    - or want children from both sex, so that we may have $D(1) > D(0)$.

- Monotonicity is violated if, e.g., some parents want two boys.

- Angrist and Evans use the US census data from 1980 and 1990.

- In 1980, $E(D|Z=1) - E(D|Z=0) \simeq 0,06$, $P(D=1|Z=0) \simeq 0,372$ and $P(D=0|Z=1) \simeq 0,568$.

- Thus, 6% of compliers, 37,2% of always takers and 56,8% of never takers.

Table 2 – Effects of fertility on female labor market participation

	$E(Y|D=1)$ $-E(Y|D=0)$	$E(Y|Z=1)$ $-E(Y|Z=0)$	δ^C
1980	-0.121 (0.0016)	-0.0079 (0.0016)	-0.132 (0.026)
1990	-0.128 (0.0016)	-0.0055 (0.0015)	-0.087 (0.024)

Notes : own computations from Angrist and Evans data. Standard errors under parentheses.

- Is the difference b/w $E(Y|D=1) - E(Y|D=0)$ and δ^C equal to the sole selection bias ?

- ▶ We now only assume that $D \in \mathbb{R}$, $Z \in \mathbb{R}$, so they may not be binary.

- ▶ For simplicity, we assume that the treatment effect is linear and homogeneous :

$$Y(d) - Y(d_0) = (d - d_0)\delta_0. \qquad \text{(Lin. model 1)}$$

Equivalently, $Y(d) = \gamma_0 + d\delta_0 + \eta$ with $E(\eta) = 0$.

- ▶ Issue : D may be correlated to $Y(d_0)$ (or, equivalently, to η). Then :

$$\begin{aligned}
\frac{\text{Cov}(D, Y)}{V(D)} &= \frac{\text{Cov}(D, Y(d_0) + \delta_0(D - d_0))}{V(D)} \\
&= \frac{\text{Cov}(D, Y(d_0))}{V(D)} + \delta_0 \\
&\neq \delta_0 \quad \text{in greneral.}
\end{aligned}$$

- ▶ Thus, the regression of Y on D will not identify δ_0 in general.

The failure of usual regressions

- Two main reasons why D is not independent of $Y(d_0)$:
    1. Omitted variables ;
    2. Simultaneity.

- Example for 1) : let D be class size and $Y(d)$ a student's potential achievement at a test if she is in classroom of size d.

- $Y(d)$ depends not only on d, but also on the initial ability A of the student, and other factors summarized in η' :
$$Y(d) = \gamma_0 + \delta_0 d + \underbrace{\beta_0 A + \eta'}_{\eta}.$$

  Moreover, $\mathrm{Cov}(D, A) \neq 0$. Thus, in general $\mathrm{Cov}(D, Y(d_0)) \neq 0$.

- If we observe A, we will identify δ_0 with a regression of Y on A and D. Otherwise, we will get a bias.

- We seek to identify the demand function $p \mapsto Y_d(p)$ (where p is price) for a given good.

- We assume $Y_d(p) = \gamma_0 + \delta_0 p + \eta_d$.

- We also consider the supply function $Y_o(p) = \alpha_0 + \beta_0 p + \eta_o$.

- Suppose that the observed price P equalizes supply and demand. Then :

$$\gamma_0 + \delta_0 P + \eta_d = \alpha_0 + \beta_0 P + \eta_o.$$

- As a result,

$$P = \frac{\alpha_0 - \gamma_0 + \eta_o - \eta_d}{\delta_0 - \beta_0}.$$

- Therefore

$$\mathsf{Cov}(Y_d(p_0), P) = \frac{\mathsf{Cov}(\eta_d, \eta_o - \eta_d)}{\delta_0 - \beta_0}$$

- Hence, $\mathsf{Cov}(Y_d(p_0), P) \neq 0$ in general.

▶ We now suppose to have an "instrumental" variable (or simply an "instrument"), namely a variable Z satisfying

    1. $Cov(Z, D) \neq 0$ (relevance);

    2. $Cov(Z, Y(d_0)) = 0$ (exogeneity).

▶ Then we have the following result, with D^* defined as in Slide 12 :

Theorem 3

Suppose (Lin. model 1) *and that Points 1 and 2 above hold. Then :*

$$\delta_0 = \frac{Cov(D^*, Y)}{V(D^*)}.$$

Proof : we already know from the proof of Theorem 2 that

$$\frac{Cov(D^*, Y)}{V(D^*)} = \frac{Cov(Z, Y)}{Cov(Z, D)}.$$

Moreover, by (Lin. model 1) and exogeneity of Z,

$$\frac{Cov(Z, Y)}{Cov(Z, D)} = \frac{Cov(Z, Y(d_0) + \delta_0(D - d_0))}{Cov(Z, D)} = \delta_0 \; \square$$

▶ We thus recover the causal effect of D on Y through the same two regressions as before :

    1. Regression of D on Z. We obtain the predicted value D^* ;

    2. Regression of Y on D^*. We obtain δ_0.

▶ Intuition : we create a "proxy" of D (i.e., D^*) that varies in an exogenous way. The regression of Y on D^* is then causal.

▶ As with a binary D, we can obtain a similar result even if $[Y(d) - Y(d_0)]/[d - d_0]$ is random, under a monotonicity assumption on $z \mapsto D(z)$.

▶ Condition 1 (i.e. the relevance assumption) is testable in the data : we can test with a regression $H_0 : \text{Cov}(Z, D) = 0$. Condition 2 (i.e. the exogeneity assumption) is not testable.

Example of instruments when facing omitted variables

- Effect of class size on students' achievement.

- Idea : use local demographic shocks and rule on class sizes (Angrist and Lavy, 1999, Piketty, 2004).

- Example : suppose that classrooms should be ≤ 30 and there are N students entering in 1st grade in a given school. Then :
    - If $N = 60$, there could be only 2 classes, each of size 30 ;
    - If $N = 61$, there will be at least 3 classes, of average size 20.3.

- We can then consider $Z = N/\lceil N/30 \rceil$, with $\lceil x \rceil = \min\{n \in \mathbb{N} | n \geq x\}$. Namely, $Z =$average class size if schools have the lowest possible number of classes but still apply the ceiling rule.

- We will have $\text{Cov}(Z, D) \neq 0$ if the ceiling rule is at least partially applied.

- We will also have $\text{Cov}(Z, Y(d_0)) = 0$ if local demographic shocks (N) are not correlated with students' achievement.

▶ Suppose that supply $Y_o(p)$ depends on p but also on cost production factors Z :
$$Y_o(p) = \alpha_0 + \beta_0 p + \lambda_0 Z + \eta_o \,, \text{Cov}(Z, \eta_d) = \text{Cov}(Z, \eta_o) = 0.$$

▶ Then :
$$P = \frac{\alpha_0 + \lambda_0 Z - \gamma_0 + \eta_o - \eta_d}{\delta_0 - \beta_0}.$$

▶ Thus, $\text{Cov}(Z, P) \neq 0$ as soon as $\lambda_0 \neq 0$.

▶ The condition $\text{Cov}(Z, \eta_d) = 0$ holds if Z is not correlated with the demand level.

▶ Example for Z : price of raw material.

- Problem : suppose $Y(d) = \gamma_0 + d\delta_0 + \eta$ with $\text{Cov}(D, \eta) = 0$.

- If we observed D, a regression of Y on D would identify δ_0.

- But what if D is measured with error ? I.e., if we only observed $\widetilde{D} = D + \nu$ instead of D ?

- Example : effect of parental income on students' achievement, with parental income reported with error by the students.

- Then, if we assume $\text{Cov}(\nu, D) = \text{Cov}(\nu, \eta) = 0$, we get :

$$\frac{\text{Cov}(\widetilde{D}, Y)}{V(\widetilde{D})} = \frac{\text{Cov}(\widetilde{D}, D\delta_0)}{V(\widetilde{D})} = \delta_0 \frac{V(D)}{V(\widetilde{D})}.$$

- Thus, $\text{Cov}(\widetilde{D}, Y)/V(\widetilde{D}) \neq \delta_0$ except if $V(\nu) = 0$ or $\delta_0 = 0$.

- The regression does not identify δ_0, even if $\text{Cov}(\widetilde{D}, Y(d_0)) = 0$.

- Suppose we observe Z such that $\mathrm{Cov}(Z, D) \neq 0$ and $\mathrm{Cov}(Z, \eta) = \mathrm{Cov}(Z, \nu) = 0$.

- Then :

$$\frac{\mathrm{Cov}(Z, Y)}{\mathrm{Cov}(Z, \widetilde{D})} = \frac{\mathrm{Cov}(Z, \gamma_0 + D\delta_0 + \eta)}{\mathrm{Cov}(Z, D + \nu)}$$

$$= \delta_0.$$

- Previous example : $Z =$ average income in the occupation of the parents.

- Valid if the parental occupation has no effect on the students' achievement, once their income is accounted for ($\mathrm{Cov}(Z, \eta) = 0$).

Adding controls : motivation

- We now consider a model including controls G, as in Chap. 3 :
  $$Y(d) = \zeta_0 + G'\gamma_0 + d\delta_0 + \eta, \; E(\eta) = E(G\eta) = 0, \qquad \text{(Lin. model 2)}$$
  but we do not assume here $E(D\eta) = 0$.

- On the other hand, we suppose $E(Z\eta) = \text{Cov}(Z, \eta) = 0$, with $Z \in \mathbb{R}^q$.

- The inclusion of G is useful if $\text{Cov}(Z, Y(d)) \neq 0$.

- Example : quarter of birth and returns to schooling.

- Idea of Angrist and Krueger (1991) : quarter of birth is exogenous but it affects education.

- Those born at the beginning of the year start school more lately. If they stop at the age of 16 (the minimal age to quit school), they will have stayed less than those born at the end of the year...

- ... But they are also older on average, and thus have worked more. This increases their wage.

$\Rightarrow$ we add age in the regression.

Adding controls : main result

- Let D^* be the theoretical prediction of D from (G, Z) : $D^* = \alpha_0 + G'\alpha_1 + Z'\alpha_2$, with

$$(\alpha_0, \alpha_1', \alpha_2') = \arg \min_{(a_0, a_1', a_2')} E[(D - a_0 - G'a_1 - Z'a_2)^2]. \qquad (1)$$

- We also let $X^* = (1, G', D^*)$. The following result generalizes Theorem 3 to models with controls :

Theorem 4

Assume (Lin. model 2), $E(Z\eta) = 0$ *and* $E(X^*X^{*\prime})$ *invertible. Then* $\beta_0 = (\zeta_0, \gamma_0', \delta_0)'$ *can be obtained by the theoretical regression of* Y *on* X^*.

Proof : we have $D = D^* + \nu$ with $E(D^*\nu) = E(G\nu) = 0$. Thus, (Lin. model 2) implies :

$$Y = \zeta_0 + G'\gamma_0 + D^*\delta_0 + \varepsilon,$$

with $\varepsilon = \nu\delta_0 + \eta$ and $E(G\varepsilon) = E(D^*\varepsilon) = 0$. Because $E(X^*X^{*\prime})$ is invertible, the theoretical regression of Y on X^* recovers β_0 $\square$

- Same principle as before :
    1. we regress D on G and Z. We deduce D^* ;
    2. we regress Y on G and D^*.

- The result relies, as in Theorem 3, on the exogeneity of Z ($E(Z\eta) = 0$), but also on its relevance, through the invertibility condition on $E(X^*X^{*\prime})$:

Proposition 1

Assume that 1, G and Z are not linearly dependent. Then $E(X^*X^{*\prime})$ is invertible iff $\alpha_2 \neq 0$, with α_2 defined in (1).

Proof of the "if" part : $E(X^*X^{*\prime})$ is invertible iff for all $\lambda = (\lambda_0, \lambda_1', \lambda_2)$, $X^{*\prime}\lambda = 0$ implies $\lambda = 0$. Now, by definition of D^*,

$$X^{*\prime}\lambda = 0 \Rightarrow \lambda_0 + G'\lambda_1 + (\alpha_0 + G'\alpha_1 + Z'\alpha_2)\lambda_2 = 0$$
$$\Rightarrow \lambda_0 + \alpha_0\lambda_2 + G'(\lambda_1 + \alpha_1\lambda_2) + Z'\alpha_2\lambda_2 = 0$$
$$\Rightarrow \lambda_0 + \alpha_0\lambda_2 = \lambda_1 + \alpha_1\lambda_2 = \alpha_2\lambda_2 = 0$$
$$\Rightarrow \lambda_0 = \lambda_1 = \lambda_2 = 0 \text{ because } \alpha_2 \neq 0 \; \Box$$

▶ Finally, we consider the same set-up as above, but with $D = (D_1, ..., D_p)' \in \mathbb{R}^p$, $p > 1$:

$$Y(d) = \zeta_0 + G'\gamma_0 + d'\delta_0 + \eta, \ E(\eta) = E(G\eta) = 0, \qquad \text{(Lin. model 3)}$$

▶ Examples :
   1. Experiment with imperfect compliance and multiple treatments (as in the STAR project) ;
   2. $Y = \log(\text{wage})$, $D = (\text{education}, \text{education}^2)'$. If education is endogenous, so will education^2.
   3. $Y = \log(\text{wage})$, $D = (\text{education}, \text{experience})'$ (Mincer model). Not only education, but also experience, is potentially endogenous.

▶ Intuition : in such cases, we need at least as many instruments as the number of endogenous variables ($q \geq p$) to have enough exogenous variations.

- Let D_j^* be the theoretical prediction of D_j from (G, Z) :
  $D_j^* = \alpha_{0j} + G'\alpha_{1j} + Z'\alpha_{2j}$, with

  $$(\alpha_{0j}, \alpha_{1j}', \alpha_{2j}') = \arg \min_{(a_{0j}, a_{1j}', a_{2j}')} E[(D_j - a_{0j} - G'a_{1j} - Z'a_{2j})^2]. \qquad (2)$$

- Let $D^* = (D_1^*, ..., D_p^*)'$ and $X^* = (1, G', D^{*'})'$. The following result generalizes Theorem 4 to the multivariate case :

Theorem 5

Assume (Lin. model 3), $E(Z\eta) = 0$ *and* $E(X^*X^{*'})$ *invertible. Then* $\beta_0 = (\zeta_0, \gamma_0', \delta_0')'$ *can be obtained by the theoretical regression of* Y *on* X^*.

- The proof is exactly the same as that of Theorem 4.

- Same logic again, but in the 1st step, we now need to run p regressions instead of just one.

▶ The relevance condition still involves the 1st step regressions :

Proposition 2

Let $A_2 = (\alpha_{21}, ..., \alpha_{2p})$ with α_{2j} defined by (2). Assume that $1, G$ and Z are not linearly dependent. Then $E(X^* X^{*\prime})$ is invertible iff rank$(A_2) = p$.

Proof (of the "if" part) : Let $\lambda = (\lambda_0, \lambda_1', \lambda_2')$ with $\lambda_2 = (\lambda_{21}, ..., \lambda_{2p})'$. Then :

$$X^{*\prime}\lambda = 0 \Rightarrow \lambda_0 + G'\lambda_1 + \sum_{j=1}^{p}(\alpha_{0j} + G'\alpha_{1j} + Z'\alpha_{2j})\lambda_{2j} = 0$$

$$\Rightarrow \lambda_0 + \sum_{j=1}^{p}\alpha_{0j}\lambda_{2j} = \lambda_1 + \sum_{j=1}^{p}\alpha_{1j}\lambda_{2j} = A_2\lambda_2 = 0$$

$$\Rightarrow \lambda_0 = \lambda_1 = \lambda_2 = 0 \text{ because rank}(A_2) = p \;\square$$

_____

▶ To apply Thm 5, we must have indeed $q \geq p$: otherwise rank$(A_2) < p$.

▶ Also : if $q = p$, each Z_i must affect at least one of the D_j.

Definition

- Assume to have an i.i.d. sample $(Y_i, G_i', D_i', Z_i')_{i=1\ldots n}$, with the same distribution as (Y, G', D', Z').

- Let $X = (1, G', D')'$, $X^* = (1, G', D^{*'})'$, $k = \dim(X) = \dim(X^*)$.

- If we assume (Lin. model 3) and note $\beta_0 = (\zeta_0, \gamma_0', \delta_0')'$, we have :

$$Y = X'\beta_0 + \eta.$$

- We then consider the empirical counterpart of the previous regressions :

  1. We regress each component D_j of D on G and Z. We then obtain

  $$\widehat{D}_j = \widehat{\alpha}_{0j} + G'\widehat{\alpha}_{1j} + Z'\widehat{\alpha}_{2j}$$

  2. We regress Y on $\widehat{X} = (1, G', \widehat{D}_1, ..., \widehat{D}_p)'$. We then obtain an estimator $\widehat{\beta}_{2SLS}$ of $\beta_0 = (\zeta_0, \gamma_0', \delta_0')'$.

$\Rightarrow$ Two-stage least squares (2SLS) estimator.

Main result

Theorem 6

Assume $(Y_i, G_i, D_i, Z_i)_{i=1...n}$ are i.i.d., distributed as (Y, G, D, Z), (Lin. model 3) holds, $E(Z\eta) = 0$, $E(X^*X^{*\prime})$ invertible, $E(\|X^*\|^2) < \infty$ and $E(\eta^2\|X^*\|^2) < \infty$. Then :

$$\sqrt{n}\left(\widehat{\beta}_{2SLS} - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, E(X^*X^{*\prime})^{-1}E(\eta^2 X^* X^{*\prime})E(X^*X^{*\prime})^{-1}\right).$$

Proof (sketch) : we have

$$\widehat{\beta}_{2SLS} = \widehat{E}\left(\widehat{X}\widehat{X}'\right)^{-1}\widehat{E}\left(\widehat{X}Y\right) = \widehat{E}\left(\widehat{X}\widehat{X}'\right)^{-1}\widehat{E}\left(\widehat{X}(X'\beta_0 + \eta)\right).$$

Now, $\widehat{E}(\widehat{X}X') = \widehat{E}(\widehat{X}\widehat{X}')$. Thus, $\widehat{\beta}_{2SLS} = \beta_0 + \widehat{E}(\widehat{X}\widehat{X}')^{-1}\widehat{E}(\widehat{X}\eta)$. Hence,

$$\sqrt{n}\left(\widehat{\beta}_{2SLS} - \beta_0\right) = \left(\frac{1}{n}\sum_{i=1}^{n}\widehat{X}_i\widehat{X}_i'\right)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{X}_i\eta_i\right).$$

Since $\widehat{X}_i - X_i^* \xrightarrow{P} 0$, we can show that

$$\frac{1}{n} \sum_{i=1}^{n} \left(\widehat{X}_i \widehat{X}_i' - X_i^* X_i^{*\prime} \right) \xrightarrow{P} 0,$$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\widehat{X}_i - X_i^*) \eta_i \xrightarrow{P} 0.$$

Therefore,

$$\sqrt{n} \left(\widehat{\beta}_{2SLS} - \beta_0 \right) = \left(\frac{1}{n} \sum_{i=1}^{n} X_i^* X_i^{*\prime} \right) \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i^* \eta_i \right) + o_P(1).$$

We conclude the proof as with the OLS estimator $\square$

Relevance condition

▶ Suppose $p = \dim(D) = 1$ and 1, G and Z not linearly dependent.

▶ Reminder : α_2 is the vector of coeffs. of Z in the theoretical reg. of D on G and Z.

▶ Then (see Prop. 1) $E(X^* X^{*\prime})$ is invertible iff $\alpha_2 \neq 0$.

▶ We can easily test this conditon or, rather, its opposite, by a F-test of joint nullity of the coefficients α_2 of Z.

▶ Rejection of H_0 supports the hypothesis $\alpha_2 \neq 0$, and thus the invertibility of $E(X^* X^{*\prime})$.

▶ Note : we will nearly always be able to compute $\widehat{\beta}_{2SLS}$, even if $E(X^* X^{*\prime})$ is singular : in general $\widehat{\alpha}_2 \neq 0$, even if $\alpha_2 = 0$.

▶ But in such a case, $\widehat{\beta}_{2SLS}$ is not consistent.

⇒ Important to test $\alpha_2 = 0$ in the 1st stage regression.

- ▶ Effect of smoking for pregnant women on babys' weight.

- ▶ We consider the regression of `lweight`=log of weight on `packs`= nb of packs smoked per day.

- ▶ `packs` a priori correlated with other factors affecting the baby's weight.

- ▶ Instrument : average price of cigarettes (`cigprice`) in the state of residence.

- ▶ Using a sample from Wooldridge (with 1,388 obs.), we obtain in the 1st step :

$$\widehat{\texttt{packs}} = \underset{(0.103)}{0.067} + \underset{(0.0008)}{0.0003} \ \texttt{cigprice}$$

- ▶ Conclusion ? Since $\widehat{\alpha}_2 \neq 0$, we can compute $\widehat{\beta}_{2SLS}$:

$$\widehat{\log(\texttt{birthweight})} = \underset{(0.91)}{4.45} + \underset{(8.70)}{2.99} \ \widehat{\texttt{packs}}.$$

- ▶ What can we think of the coefficient ? Conclusion ?

- Theorem 6 also relies on the assumption $E(Z\eta) = 0$.

- This condition is not testable if $\dim(D) = p = q = \dim(Z)$.

- If $E(Z\eta) \neq 0$, $\widehat{\beta}_{2SLS}$ does not converge in general.

- Without controls G and when $p = q = 1$, we get :

$$\widehat{\beta}_{2SLS} \xrightarrow{P} \beta_0 + \frac{\mathrm{Cov}(Z, \eta)}{\mathrm{Cov}(Z, X)}.$$

- Besides, $\widehat{\beta}_{OLS} \xrightarrow{P} \beta_0 + \mathrm{Cov}(X, \eta)/V(X)$. Hence,

$$|\mathrm{plim}\widehat{\beta}_{2SLS} - \beta_0| < |\mathrm{plim}\widehat{\beta}_{OLS} - \beta_0| \text{ iff } |\mathrm{corr}(Z, \eta)| < |\mathrm{corr}(X, Z) \, \mathrm{corr}(X, \eta)|.$$

$\Rightarrow$ If n is large and Z is "weak" (viz., $\mathrm{corr}(X, Z) \simeq 0$), we need $|\mathrm{corr}(Z, \eta)|$ to be very small for $\widehat{\beta}_{2SLS}$ to be better than $\widehat{\beta}_{OLS}$.

- By Theorem 6, $\sqrt{n}(\widehat{\beta}_{2SLS} - \beta_0) \xrightarrow{d} \mathcal{N}(0, V_a)$, with
$$V_a = E(X^* X^{*\prime})^{-1} E(\eta^2 X^* X^{*\prime}) E(X^* X^{*\prime})^{-1}$$

- We can estimate consistently V_a by
$$\widehat{V}_a = \left(\frac{1}{n} \sum_{i=1}^n \widehat{X}_i \widehat{X}_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \widehat{\eta_i^2} \widehat{X}_i \widehat{X}_i' \right) \left(\frac{1}{n} \sum_{i=1}^n \widehat{X}_i \widehat{X}_i' \right)^{-1},$$
with $\widehat{\eta_i} = Y_i - X_i' \widehat{\beta}_{2SLS}$ (careful, $\widehat{\eta_i} \neq Y_i - \widehat{X}_i' \widehat{\beta}_{2SLS}$)

- t- and F-tests and the corresponding confidence intervals/regions can then be built exactly as with the OLS estimator.

- Consider the simple case without G and $\dim(D) = 1$. Let us also assume that $E(D\eta) = 0$ and $V(\eta|D) = V(\eta|Z) = \sigma^2$.

- Then $\widehat{\beta}_{OLS} = (\widehat{\beta}_{1,OLS}, \widehat{\beta}_{2,OLS})'$ and $\widehat{\beta}_{2SLS} = (\widehat{\beta}_{1,2SLS}, \widehat{\beta}_{2,2SLS})'$ are both consistent, and

$$V_a(\widehat{\beta}_{2,OLS}) = \frac{\sigma^2}{V(D)}, \ V_a(\widehat{\beta}_{2,SLS}) = \frac{\sigma^2}{V(D^*)} \text{ with } V(D^*) \leq V(D).$$

$\Rightarrow$ The OLS estimator is more precise asymptotically.

- Hence, if $E(D\eta) = 0$, we should rather use the OLS estimator. This actually also holds with covariates and $\dim(D) > 1$.

- We thus consider the test of $E(D\eta) = 0$, under the maintained hyptothesis that $E(\eta) = E(G\eta) = E(Z\eta) = 0$.

▶ We start from

$$\begin{cases} Y & = & X'\beta_0 + \eta, \\ D & = & D^* + \nu. \end{cases}$$

▶ Let $\rho_0 = \text{Cov}(\eta, \nu)/V(\nu)$ and $\xi = \eta - \rho_0 \nu$. Then :

$$Y = X'\beta_0 + \rho_0 \nu + \xi. \tag{3}$$

▶ Moreover, $E(D\eta) = E(D^*\eta) + E(\nu\eta) = E(\nu\eta)$. Thus, $E(D\eta) = 0$ iff $\rho_0 = 0$.

▶ Idea, if we knew ν : estimate (β_0, ρ_0) by OLS of Y on (X, ν).

Lemma 1

The theoretical regression of Y on X and ν yields β_0 and ρ_0.

Proof : we just have to check that $E(X\xi) = E(\nu\xi) = 0$. We have
$E(G\eta) = E(G\nu) = 0$, thus $E(G\xi) = 0$. Similarly, $E(Z\xi) = 0$. Moreover,
$E(\nu\xi) = \text{Cov}(\nu, \eta) - \rho_0 V(\nu) = 0$. Hence,

$$E(D\xi) = E(D^*\xi) + E(\nu\xi) = 0 \; \square$$

- ν is unobserved but can be estimated by regressing D on G and $Z \Rightarrow \widehat{\nu}$.

- We then consider the *augmented regression* of Y on X and $\widehat{\nu}$.

- One can show that (i) $\widehat{\beta}_{\text{aug}} = \widehat{\beta}_{2SLS}$; (ii) under $H_0 : \rho_0 = 0$, the t-statistic t_{ρ_0} tends to a $\mathcal{N}(0, 1)$.

$\Rightarrow$ we can ignore measurement error on ν for the test of $\rho_0 = 0$.

- Example : experience JTPA :

Table 3 – Augmented regression in the JTPA experience

	Coeff of D	Coeff of ν
Men	1 825	3 259
	(943)	(1 162)
Women	1 942	302
	(569)	(715)

Notes : std. err. under parentheses.

- Conclusion on endogeneity ? Could we expect the coefficients on D ?

- Sources of endogeneity : omitted variable bias, simultaneity, measurement errors.

- Instrumental variables : exogeneity and relevance conditions.

- Principle of the two-stage least squares.

- Identification of LATE by 2SLS when D and Z are binary.

- Identification of constant/homogeneous causal effects in multiple regressions, possibly with several endogenous variables.

- Inference very close to that developed for OLS.

- Exogeneity test.