Applied Econometrics Part 5

Endogeneity

Jose Angel Garcia Sanchez

Université Paris 1 Panthéon-Sorbonne
jagarsanc@gmail.com

November 2025

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Outline

1. Framework
2. Non-exogeneity
3. The IV estimator
4. IV in practice

Considering random explanatory variables

- In the basic framework, explanatory variables X are usually considered not random
- This amounts to reasoning conditionally on X, i.e. as if the X were fixed
- An important hypothesis was that X and u were independent, which is always the case if X is fixed
- But the X's can be considered as random variables
- All the previous results hold, because they were found conditionally on X

Assumptions for estimation

1. $E(u) = 0$
2. $\forall t, t'$, X_t is random and uncorrelated to $u_{t'}$
3. $Rk(X) = k$
4. $E(uu') = \sigma^2 I_N$
5. Error terms are iid $(0, \sigma^2)$
6. $plim[(X'X)]/N = V_X$ which is a positive definite matrix (when N goes to infinity, the X variables always keep some variance)

Properties of OLS: unbiasedness

- $\hat{b}_{ols}$ is still unbiased
- We know that $\hat{b}_{ols} = (X'X)^{-1}X'y = b + (X'X)^{-1}X'u$
- So $E(\hat{b}_{ols}) = b + E[(X'X)^{-1}X'u]$
- We get: $E[(X'X)^{-1}X'u] = E_X[E_u[(X'X)^{-1}X'u \,/\, X]]$
  $= E_X[(X'X)^{-1}X'E_u[u \,/\, X]]$ and $E_u[u \,/\, X] = 0$
- In the same way, we get an expression for the variance of $\hat{b}_{ols}$:
- $V(\hat{b}_{ols}) = \sigma^2 E[(X'X)^{-1}]$

Properties of OLS: consistency

$\hat{b}_{ols}$ is still consistent

$$plim(\hat{b}_{ols}) = b + [plim(\frac{X'X}{N})]^{-1} plim(\frac{X'u}{N}) = b + V_X^{-1}.0 = b$$

We will see in the next section that the asymptotic distribution of $\hat{b}$ is a normal, as expected.

Asymptotics: OLS estimators

We wish now to enlarge our framework to the case where u's are only iid (and not necessarily normal) and where X's can be random (and not necessarily fixed). In what follows, we will thus use the following asymptotic results, obtained under the usual model hypotheses:

$$plim(\hat{b}_{ols}) = b + [plim(\frac{X'X}{N})]^{-1} plim(\frac{X'u}{N}) = b + V_X^{-1}.0 = b$$

$$plim(\hat{\sigma}^2_{ols}) = \sigma^2$$

$$\sqrt{N}(\hat{b}_{ols} - b) \rightarrow \mathcal{N}(0, \sigma^2 V_X^{-1})$$

Asymptotic tests

Tests are run using the following formula:

$$f = \frac{(C\hat{b}_{ols} - Cb)'(C(X'X)^{-1}C')^{-1}(C\hat{b}_{ols} - Cb)}{\sigma^2_{ols}} \to \chi^2_r$$

Where, as before in the finite sample case, σ^2_{ols} is unknown. But since by assumption we work with a very large sample, $\hat{\sigma}^2_{ols}$ is a consistent estimate of σ^2_{ols}. We can thus replace σ^2_{ols} by its consistent estimate, without having to replace the χ^2 with a Fisher.

General properties

If we assume that the X_i are iid and the u's are normal, then:

- OLS estimators are normal and consistent
- We can thus run the usual T- and F-tests
- If the u are not normal but only iid, we need to make sure the sample is large, and use the asymptotic version of the usual T-test and F-tests
- In that case, Normal replaces Student and χ^2 replaces Fisher

Non-exogenous explanatory variables

- Let's use the following model: $y_t = X_t' b + u_t$
- Assume that $E(X_t' u_t) \neq 0$: some explanatory variables are correlated to the current error term
- In this case, $plim(X'u/N) \neq 0$
- OLS is biased and inconsistent

$$plim(\hat{b}_{ols}) = b + [plim(\frac{X'X}{N})]^{-1} plim(\frac{X'u}{N}) \neq b$$

Even if we increase the size of the sample, we won't get the right value for b

Exogeneity: definitions

- **Case 1**: Strong exogeneity: $\forall t, t'$, $E(X_t' u_{t'}) = 0$
- X are u are never correlated
- **Case 2**: Exogeneity: $\forall t$, $E(X_t' u_t) = 0$
- X and u are uncorrelated for any given time period
- **Case 3**: X predetermined: $\forall t' \geq t$, $E(X_t' u_{t'}) = 0$
- X uncorrelated to current and future u

This class focuses on cross-sections so on the second case.

Omitted variables

Let's assume the true model is the following:

$$y_t = X_t'b + w_t d + v_t$$

If we omit w_t and estimate the following model instead:

$$y_t = X_t'b + u_t$$

- w_t is included in u_t
- If w_t is correlated to X, then $E(X_t'w_t) \neq 0$ and thus $E(X_t'u_t) \neq 0$: OLS is biased
- Remark: in the fixed X case, we would get the same result using the Frish-Waugh theorem
- When a variable that is correlated to the X's is omitted, estimators suffer from an *omitted variable bias*
- Estimators are inconsistent

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Variables measured with error

Let's use a very simple one-variable theoretical model (no constant if variables are centered):

$$y_t^* = x_t^*.b + v_t$$

Assume that x is measured with error. To run the estimation, we have access to values y_t and x_t: $y_t = y_t^*$ and $x_t = x_t^* + e_t$, e_t being a white noise.
The estimated model is thus:

$$y_t = x_t.b + u_t$$

With $u_t = v_t - be_t$. We thus have the following:

$$E(x_t u_t) = E[(x_t^* + e_t)(v_t - be_t)] = -b\sigma_e^2 \neq 0$$

OLS is thus inconsistent, and it can be shown that it is biased towards 0: the influence of x on y is underestimated.

Simultaneous equations

Say we have the following system of equations:

- $Y_t = a + bX_t + u_t$ (1)
- $X_t = Y_t + Z_t$ (2)
- Because of equation (1), Y_t is endogenous and because of equation (2), X_t is endogenous too
- At the system level, only Z_t is exogenous
- To get back to what we are used to, we should rewrite endogenous variables as functions of exogenous variables only

The system can be rewritten: $Y_t = \frac{a}{1-b} + \frac{b}{1-b}Z_t + \frac{1}{1-b}u_t$ and $X_t = \frac{a}{1-b} + \frac{1}{1-b}Z_t + \frac{1}{1-b}u_t$. Notice that X_t appears to be a function of u_t: in equation (1), it is thus correlated to the error term. A basic hypothesis of OLS is violated and if we estimate (1) without taking into account the information provided by (2), estimates will be non consistent.

Autocorrelation with lagged dependent variable

Assume we use the following model:

$$y_t = b_0 + b_1 x_t + b_2 y_{t-1} + u_t$$

With: $u_t = \rho u_{t-1} + \varepsilon_t$
We can then write:

- $y_t = b_0 + b_1 x_t + b_2 y_{t-1} + \rho u_{t-1} + \varepsilon_t$ (1)
- And at the same time: $y_{t-1} = b_0 + b_1 x_{t-1} + b_2 y_{t-2} + u_{t-1}$ (2)
- Equation (2) shows that y_{t-1} depends on u_{t-1}, so y_{t-1} is correlated to the error term in equation (2)
- The lagged outcome variable is thus not exogenous in the estimated model
- It can be shown that it is exogenous if errors terms are not autocorrelated
- How to test for this: Durbin's h statistic

General issue in policy evaluation

- Say we want to evaluate the impact of a policy on people's wages (ex: a training program)
- A model describing the wage outcome is $Y_i = a + bX_i + cP_i + u_i$, where X comprises individual characteristics and P is a dummy variable indicating whether the individual was assigned to the program
- If the policy is not *randomized*, i.e. if the fact of being assigned to the program depends on unobserved individual characteristics, then P is correlated to u and evaluation of the impact of the policy cannot be estimated consistently
- Intuitive reason: those who where assigned to the program are not comparable to the ones who were not, so the latter cannot be considered as a valid control group for the former
- This is called a *selection effect*

Consequences

In the case of non-exogeneity (also called endogeneity) of some explanatory variables:

- OLS estimators are biased: on average, we do not get the true value of parameters
- OLS are non consistent: even if we increase the size of the sample, the bias does not go to zero and we will never get the true value of parameters

We thus have to use another technique of estimation, using auxiliary variables: the instrumental variables technique.

Instrumental variables

Let's use the following model, where (to make things simple) *all* X's are endogenous, except the constant (can't be endogenous by definition):

$$y = X.b + u$$

Let's consider a set of variables $Z_{(N,p)}$ different from the X's but including the constant, with the following properties:

- $E(Z_t' u_t) = 0$: variables Z are exogenous
- $Rk(Z) = p$
- $plim(Z'X/N) = V_{ZX}$ with V_{ZX} a non null matrix of dimension (p, k) and rank k
- $plim(Z'Z/N) = V_Z$ with V_Z a finite positive definite matrix of dimension (p, p)

Means: variables Z need to be both exogenous and correlated to X, and the number of IV, p, is greater than the number of explanatory variables X ($p \geq k$).

The estimator

The original model is:

$$y = X.b + u$$

- Assume we regress X on variables Z
- We compute their predictions $\tilde{X}$
- We use these new variables in the model instead:

$$y = \tilde{X}.b + u$$

- It should work because $\tilde{X}$ is only made up of Z that are exogenous: $\tilde{X}$ is thus exogenous too

Intuition (1)

- OLS on the original model $y = X.b + u$ are inconsistent because variables X are correlated to u
- We can't replace the X's by other variables because the model should remain unchanged
- We would like to get rid of this correlation, while keeping the information provided by X
- Imagine a regression where each variable in X would be explained by the set of variables Z, say: $\forall j,\ x_j = Zd + e$
- A prediction for x_j would be: $\hat{x}_j = Z\hat{d}$
- If this regression is good enough, $\hat{x}_j$ is close to the true x_j
- And $\hat{x}_j$ is a linear combination of variables Z that are exogenous: it is thus exogenous as well
- We can thus replace x_j by $\hat{x}_j$ in the original model

Transforming the model

The original model is:

$$y = X.b + u$$

The new model is:

$$y = \tilde{X}.b + u$$

- Problem: the new model is not directly equivalent to the original one: only X was modified
- How can we make the new model equivalent to the old one?
- Notice that $\tilde{X}$ is the prediction of X using the Z
- So that $\tilde{X} = P_Z X$
- We need to multiply every term of the original model by P_Z

The IV estimator

- Let's call P_Z the orthogonal projection matrix, that projects on $L(Z)$: $P_Z = Z(Z'Z)^{-1}Z'$
- P_Z is such that it is symmetric ($P'_Z = P_Z$) and idempotent ($P_Z.P_Z = P_Z$)
- It can be shown that $\tilde{y} = P_Z y$, $\tilde{u} = P_Z u$ and $\tilde{X} = P_Z X$
- It's as if we had premultiplied the original model by P_Z
- The IV estimator is: $\hat{b}_{iv} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} = (X'P_Z X)^{-1}X'P_Z y$

Remark

- To get to the estimated model, we multiply everything on the left y and X by P_Z
- What would happen if we multiplied only X and not y?
- $\hat{b}_{vi} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'y = (X'P_ZX)^{-1}X'P_Zy$
- It does not change anything as regards $\hat{b}_{vi}$: in practice, we transform only X and leave y unchanged

Intuition (2)

We can also explain this algebraically:

- To remove the correlation between X and u, we project the model onto the vector space $L(Z)$, spanned by the columns of Z, that are at the same time exogenous and correlated to X
- This amounts to keeping from X only the exogenous information, uncorrelated to the error terms
- The more the Z are correlated to the X (and the more numerous the Z's are), the more precise the estimator is because the information loss is minimal
- $P_Z y$ is the prediction from the regression of y on variables Z, same for $P_Z X$

Remarks

- If $p = k$ (same number of instruments Z and of explanatory variables X), we get $\hat{b}_{iv} = (Z'X)^{-1}Z'y$
- Proof: rewriting $\hat{b}_{iv}$, given that in this case, matrix $Z'X$ is square and invertible
- Even if this expression is simple, it is not recommended to choose this minimal number of intruments

Generalization (1)

- What if some variables in X are endogenous, and others exogenous?
- Let's call X_1 the set of the exogenous ones and X_2 the set of the endogenous ones: the matrix notation would be $X = (X_1, X_2)$ (again, the constant is always exogenous)
- In that case, variables X_1 can be used as instruments in addition to the Z
- Let's call W this extended set of instruments: $W = (Z, X_1)$
- To compute the VI estimator, we premultiply X by P_W
- $P_W X = P_W(X_1, X_2) = (P_W X_1, P_W X_2) = (X_1, P_W X_2)$
- Indeed, since X_1 belongs to W, it's being projected onto itself and remains unchanged

Generalization (2)

- Conclusion: we will use X_1 as additional instruments because they will remain unchanged in the tranformed model, and will help to instrument the X_2
- We only need to make sure that there are enough "new" instruments Z as endogenous variables X_2 and don't rely only on exogenous X_1
- Z are sometimes called *excluded instruments*
- We realize here that the constant, exogenous by nature, belongs to the exogenous X_1 and that we had been right to use it as an instrument in the first place
- That's what we did when describing the first stage regression, that contain a constant by default

Important: final notation

- Since the exogenous variables from the set of variables X will always be used as instruments, we don't usually bother mentioning it
- **We usually call Z the global matrix of instruments that contains both the excluded instruments and the variables from X that are exogenous**
- So that the hypotheses and formulae are the same as the ones presented in the beginning
- We only need to make sure that there are at least as many excluded instruments as endogenous variables

Properties of the IV estimator (1)

- $\hat{b}_{iv}$ is biased in a small sample
- We thus cannot derive a general expression for its variance-covariance matrix
- But it is consistent (when sample size goes to infinity)
- It is asymptotically normal

$$\sqrt{N}(\hat{b}_{iv} - b) \rightarrow \mathcal{N}(0, \sigma^2(V'_{ZX} V_Z^{-1} V_{ZX})^{-1})$$

Properties of the IV estimator (2)

There is no general expression for its variance-covariance matrix, but we can derive its *asymptotic* variance-covariance matrix that is a consistent estimator of its "true" variance-covariance matrix:

$$\widehat{V}_{asympt}(\sqrt{N}(\hat{b}_{iv} - b)) = \hat{\sigma}^2_{iv}(\frac{X'P_Z X}{N})^{-1}$$

with $\hat{\sigma}^2_{iv} = \frac{SSR_{iv}}{N}$ and $\hat{u}_i = y - X\hat{b}_{iv}$

- When OLS is consistent, IV is less precise than OLS
- Precision increases with the number of instruments

Tests on regression parameters

- We do not have convenient properties of the estimator in the finite sample case
- We can only run asymptotic tests: Wald tests, similar to F-tests, but when sample size is large
- Example of a test for r linear constraints on parameters:

$$f = \frac{(C\hat{b}_{iv} - Cb)'[C(X'P_Z X)^{-1}C']^{-1}(C\hat{b}_{iv} - Cb)}{\hat{\sigma}^2_{iv}} \to \chi^2_r$$

With r the number of linear restrictions: C is of size (r, k).
Proof: taking the expression of the asymptotic normality of $\hat{b}_{iv}$, taking its quadratic form and "dividing" it by its variance will give a χ^2 distribution (the "true" variance is replaced by its consistent estimator).

How to run IV estimation

- Python: statsmodels
- This amounts to running two-stage least squares
- Intuition: first, regress the y and X's on variables Z, then use the predictions in the model instead of the original values

Two-stage least squares (1)

- Run $k+1$ regression, to get $P_Z y$ and $P_Z X$
- Estimate OLS on the transformed model $P_Z y = P_Z y b + u$
- We thus get $\hat{b}_{iv} = (X' P_Z X)^{-1} X' P_Z y$
- The first $k+1$ regressions can be used to assess the conveniency of instruments (they have to be correlated enough to the X's)
- Remark: this provides the same values if we do not replace y by $P_Z y$

Two-stage least squares (2)

Warning: if we do this procedure "by hand", running 2 OLS regressions, instead of running the convenient procedure with the software, the s.e.'s of the second regression cannot be used for tests on the coefficients

Reason: in the second stage equation, residuals are computed as: $\hat{u} = P_Z y - P_Z X \hat{b}_{iv}$

Whereas they should be computed as $\hat{u} = y - X \hat{b}_{iv}$

Exogeneity test (1)

- We test H_0: $E(X'u) = 0$
- This is called the "Hausman test" or "Durbin-Wu-Hausman test"
- If H_0 is true, then both OLS and IV estimators are consistent
- If H_0 is false, only the IV estimator is consistent
- The test is based on the difference between $\hat{b}_{iv}$ and $\hat{b}_{ols}$
- They are asymptotically normal: if we compute the difference between the two, take its quadratic form and "divide" it by its variance matrix, we will get a χ^2 distribution

Exogeneity test (2)

One could think that the parameter of this χ^2 is the number of tested variables (those potentially endogenous), just like a Wald test. But looking close at the test statistic:

$$H = (\hat{b}_{vi} - \hat{b}_{ols})'(V(\hat{b}_{vi}) - V(\hat{b}_{ols}))^{-1}(\hat{b}_{vi} - \hat{b}_{ols})$$

This parameter is in fact equal to the rank of the following matrix: $(V(\hat{b}_{vi}) - V(\hat{b}_{ols}))$
So sometimes the software can't run the test ($H < 0$, small sample, etc)

The augmented regression

- Consider model $y = Xb + u$, where a subset of variables belonging to X might be endogenous: x
- Let's call Z the instruments, some belonging to X (in fact the X without the x) and some not
- Consider the augmented model: $y = Xb + M_Z xc + \varepsilon$
- $M_Z x$ are the residuals of the regression of x on Z
- The $\hat{b}$ of this "augmented" regression is equal to the IV estimator of the original model
- Testing $c = 0$ amounts to testing exogeneity of the x (it is equivalent to the Hausman test)

Proofs

- $\hat{b}_{aug} = \hat{b}_{iv}$ and equivalence of tests: using the Frish-Waugh theorem
- Remark: this augmented model has no theoretical meaning, it's only a tool

Selecting convenient instruments

Sargan test: H_0: $E(Z'u) = 0$
Also called: test of overidentifying restrictions
Under H_0:

$$\frac{\hat{u}' P_Z \hat{u}}{s^2} \to \chi^2_{p-k}$$

with $\hat{u} = y - X\hat{b}_{iv}$ and $s^2 = \frac{\hat{u}'\hat{u}}{N}$.
$\hat{u}' P_Z \hat{u}$ is the sum of the predicted value of the regression of $\hat{u}$ on Z, squared.
Remark: when $p = k$, the statistic is always zero because $\hat{b}_{iv} = (Z'X)^{-1}Z'y$ and $Z'\hat{u} = 0$. So we cannot run the test with the minimum number of instruments.

The problem with weak instruments

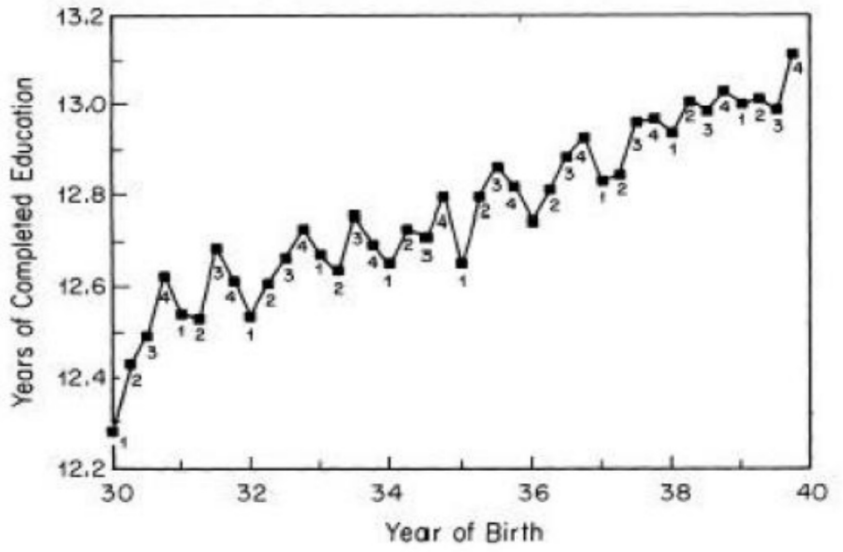
- If instruments are too weakly correlated to the X's, even if we increase the number of observations, there can be an important bias in estimations
- Plus, the estimator has a nonnormal sampling distribution which makes statistical inference meaningless
- The weak instrument problem is increased with many instruments, so drop the weakest ones and use the most relevant ones
- A way to measure how instruments are correlated to potentially endogenous variables is to run the regression explaining the former by the latter and check its goodness of fit
- A criterion can be the global F statistic: if $F<10$, then instruments are weak

Returns on education

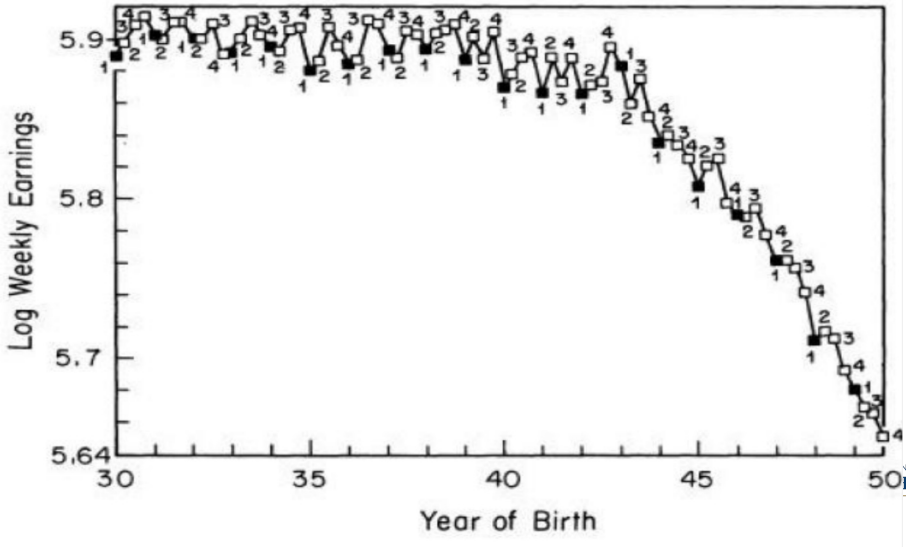
Angrist-Krueger (1991), "Does Compulsory School Attendance Affect Schooling and Earnings", Quarterly Journal of Economics

- Goal: find causal link between education and wages, and we know that such a wage equation is subject to endogeneity
- In the US, compulsory education starts the year pupils turn 6, and ends when pupil turns 16
- Pupils born in January: begin school at 6 years 8 months old
- Pupils born in December: begin school at 5 years 9 months old
- For pupils who stop school at 16, pupils born in December study almost one year more than those born in January
- Quarter of birth is thus correlated to schooling, but random so uncorrelated to individual ability: good instrument for schooling

Quarter of birth and education



Quarter of birth and earnings



Results

Independent variable	(1) OLS	(2) TSLS	(3) OLS	(4) TSLS
Years of education	0.0711	0.0891	0.0711	0.0760
	(0.0003)	(0.0161)	(0.0003)	(0.0290)
Race (1 = black)	—	—	—	—
SMSA (1 = center city)	—	—	—	—
Married (1 = married)	—	—	—	—
9 Year-of-birth dummies	Yes	Yes	Yes	Yes
8 Region-of-residence dummies	No	No	No	No
Age	—	—	−0.0772	−0.0801
			(0.0621)	(0.0645)
Age-squared	—	—	0.0008	0.0008
			(0.0007)	(0.0007)
χ^2 [dof]	—	25.4 [29]	—	23.1 [27]

Method

- Authors use as instruments first quarter, interacted with regional dummies
- Result: OLS underestimates the impact of education on earnings
- Even better method: use of randomized trials
- Ex: Moving to Opportunity, RAND Health Insurance Experiment
- Many other methods in policy evaluation: difference in difference, matching, etc

Are IV estimates reliable?

- The goal of IV estimation is to bring to light real causality and not mere correlation between outcome y and explanatory variables X
- In treatment evaluation, we want to assess the Average Treatment Effect
- So we instrument the dummy indicating whether the individual was assigned to the treatment group because we suspect that choosing a treatment is not exogenous: it is likely to be correlated to unobserved heterogeneity
- However, it can be shown that in doing so, we estimate a Local Average Treatment Effect, i.e. we estimate the impact of the treatment only on the individuals for which the instrument indeed has an impact (called *compliers*)