Endogeneity. Instrumental variables estimation. Properties of instrumental variables.

Jakub Mućk
SGH Warsaw School of Economics

Least squares estimator

- **Least squares estimator** :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \varepsilon \tag{1}$$

where
  - y is the (outcome) dependent variable;
  - $x_1, x_2, \ldots, x_K$ is the set of independent variables;
  - ε is the error term.
- The dependent variable is explained with the components that vary with the **the dependent variable** and **the error term**.
- β_0 is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_K$ are the coefficients (slopes) on $x_1, x_2, \ldots, x_K$.

- **Least squares estimator** :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \varepsilon \tag{1}$$

where
  - y is the (outcome) dependent variable;
  - $x_1, x_2, \ldots, x_K$ is the set of independent variables;
  - ε is the error term.
- The dependent variable is explained with the components that vary with the **the dependent variable** and **the error term**.
- β_0 is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_K$ are the coefficients (slopes) on $x_1, x_2, \ldots, x_K$.

$\beta_1, \beta_2, \ldots, \beta_K$ measure the effect of change in $x_1, x_2, \ldots, x_K$ upon the expected value of y (*ceteris paribus*).

- **Assumption #1**: true DGP (data generating process):

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon. \tag{2}$$

- **Assumption #2**: the expected value of the error term is zero:

$$\mathbb{E}(\varepsilon) = 0, \tag{3}$$

and this implies that $\mathbb{E}(y) = \mathbf{X}\beta$.

- **Assumption #3**: Spherical variance-covariance error matrix.

$$var(\varepsilon) = \mathbb{E}(\varepsilon\varepsilon') = I\sigma^2 \tag{4}$$

. In particular:
  - the variance of the error term equals σ:

$$var(\varepsilon) = \sigma^2 = var(y). \tag{5}$$

  - the covariance between any pair of ε_i and ε_j is zero"

$$cov(\varepsilon_i, \varepsilon_j) = 0. \tag{6}$$

- **Assumption #4**: **Exogeneity.** The independent variable are **not random** and therefore they are not correlated with the error term.

$$\mathbb{E}(\mathbf{X}\varepsilon) = 0. \tag{7}$$

- **Assumption #5**: the full rank of matrix of explanatory variables (there is no so-called collinearity):

$$rank(\mathbf{X}) = K + 1 \leq N. \tag{8}$$

- **Assumption #6 (optional)**: the normally distributed error term:

$$\varepsilon \sim \mathcal{N}\left(0, \sigma^2\right). \tag{9}$$

Assumptions of the least squares estimators

Under the assumptions A#1-A#5 of the multiple linear regression model, the least squares estimator $\hat{\beta}^{OLS}$ has the smallest variance of all linear and unbiased estimators of β.

$\hat{\beta}^{OLS}$ is the **Best Linear Unbiased Estimators (BLUE)** of β.

- **The least squares estimator**

$$\hat{\beta}^{OLS} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}. \tag{10}$$

- **The variance of the least square estimator**

$$Var(\hat{\beta}^{OLS}) = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} \tag{11}$$

- **If the (optional) assumption about normal distribution of the error term is satisfied** then

$$\beta \sim \mathcal{N}\left(\hat{\beta}^{OLS}, Var(\hat{\beta}^{OLS})\right). \tag{12}$$

Consequences of endogeneity

Random (stochastic) regressors

Are explanatory variables always predetermined?

Endogenous variables

An explanatory variable is said to be **endogenous** when it is correlated with error term, i.e., $\mathbb{E}(\varepsilon|x) \neq 0$.

1. **Bias of the least squares estimator**

$$\mathbb{E}\left(\hat{\beta}^{OLS}\right) = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}(\varepsilon), \tag{13}$$

2. **Inconsistency of the least squares estimator**

$$\text{plim}_{N\to\infty}\hat{\beta}^{OLS} = \beta + \text{plim}_{N\to\infty}\left(\frac{1}{N}\mathbf{X}'\mathbf{X}\right)^{-1}\frac{1}{N}\mathbf{X}'\mathbb{E}(\varepsilon). \tag{14}$$

Examples of endogeneity problems

Standard cases when explanatory variables are endogenous:

1. Measurement error.
2. Omitted variable bias.
3. Simultaneous causality.

In addition,

1. Dynamic panel data models.
2. Autoregression with the serial correlation of the error term.

- Let's assume that **true** DGP (data generating process) for the consumption (c) is as follows:

$$c = \alpha + \beta inc^* + \varepsilon \tag{15}$$

where inc^* is **the permanent income.**

- Usually, we have data on income inc but not **the permanent income.** If so, we can proxy the permanent income by current income:

$$inc^* = inc + \eta, \tag{16}$$

where η stands for the measurement and $\eta \sim \mathcal{N}\left(0, \sigma_\eta^2\right)$.

- The current income (inc) is **proxy variable** for the permanent income (inc^*).
- Substituting the permanent income into (15):

$$c = \alpha + \beta\left(inc + \eta\right) + \varepsilon = \alpha + \beta inc + \beta\eta + \varepsilon = \alpha + \beta inc + \nu, \tag{17}$$

where $\nu = \varepsilon + \beta\eta$.

- The covariance between inc and error term (ν):

$$cov(inc, \nu) = \mathbb{E}\left(inc\nu\right) = \mathbb{E}\left(\left(inc^* + \eta\right)\left(\varepsilon + \beta\eta\right)\right) = \mathbb{E}\left(\beta\eta^2\right) = \sigma_\eta^2\beta \neq 0. \tag{18}$$

- **Labor economics**: returns to education.
- Let's assume that **true** DGP (data generating process) for the logged wage (w):

$$w = \alpha + \rho\mathcal{S} + \beta\mathcal{A} + \varepsilon, \tag{19}$$

  where $\mathcal{S}$ is the highest grade of schooling completed and $\mathcal{A}$ is a measure of personal ability or(and) motivation.
- **Problem:** data on $\mathcal{A}$ are not unavailable.
- Consider alternative version of (20):

$$w = \alpha + \rho\mathcal{S} + \eta, \tag{20}$$

- where the error term η captures personal abilities $\mathcal{A}$, i.e., $\eta = \varepsilon + \beta\mathcal{A}$.
- The OLS estimator of ρ can be simplified to:

$$\hat{\rho}^{OLS} = cov(w, \mathcal{S})/Var(\mathcal{S}). \tag{21}$$

- Plugging *true* DGP for wages w:

$$\hat{\rho}^{OLS} = \frac{cov(\alpha + \rho\mathcal{S} + \beta\mathcal{A} + \varepsilon, \mathcal{S})}{Var(\mathcal{S})}, \tag{22}$$

- After manipulation we get:

$$\hat{\rho}^{OLS} = \frac{1}{Var(\mathcal{S})}\mathbb{E}\left[(\alpha + \rho\mathcal{S} + \varepsilon)\,\mathcal{S} + \beta\mathcal{A}\mathcal{S}\right] = \rho + \underbrace{\beta\frac{cov(\mathcal{A},\mathcal{S})}{Var(\mathcal{S})}}_{=\text{bias}} \neq \rho. \qquad (23)$$

- The OLS coefficient on schooling would be upward biased if the signs of β and $cov(\mathcal{A},\mathcal{S})/Var(\mathcal{S})$ are the same.

- **Simple (Keynesian) model of consumption**:

$$c = \alpha + \beta y + \varepsilon \tag{24}$$
$$y = c + i \tag{25}$$

where c is the consumption, y is the aggregate product, i stands for the investment and ε is the error term, i.e., $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

- In the above system we have to endogenous variables (c and y) and one exogenous variable (i).
- The reduced form will be defined as model in which endogenous variable(s) is determined by the exogenous variables as well as the stochastic disturbances. In our case:

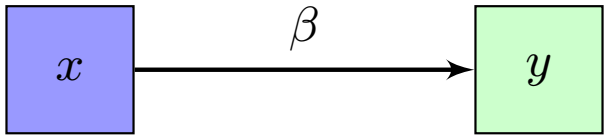
$$\begin{aligned} y &= c + i \\ y &= \alpha + \beta y + \varepsilon + i \\ (1-\beta)y &= \alpha i + \varepsilon \\ y &= \frac{\alpha}{(1-\beta)} + \frac{1}{(1-\beta)} i + \frac{1}{(1-\beta)} \varepsilon. \end{aligned}$$

- The general expression of the OLS estimator of the marginal propensity to consume (β) form equation (24):

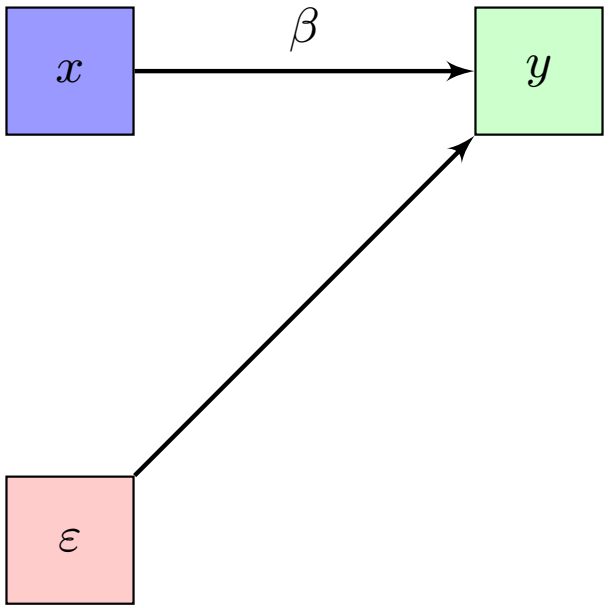
$$\hat{\beta}^{OLS} = \beta + \underbrace{\frac{\sum (y - \bar{y})\, \varepsilon}{\sum (y - \bar{y})^2}}_{=0 \quad \text{if} \quad \mathbb{E}(y|\varepsilon)=0} . \tag{26}$$

- But we know that y depends on ε (see the reduced form). If so, then the $\hat{\beta}^{OLS} \neq \beta$ and the OLS estimator is not consistent.

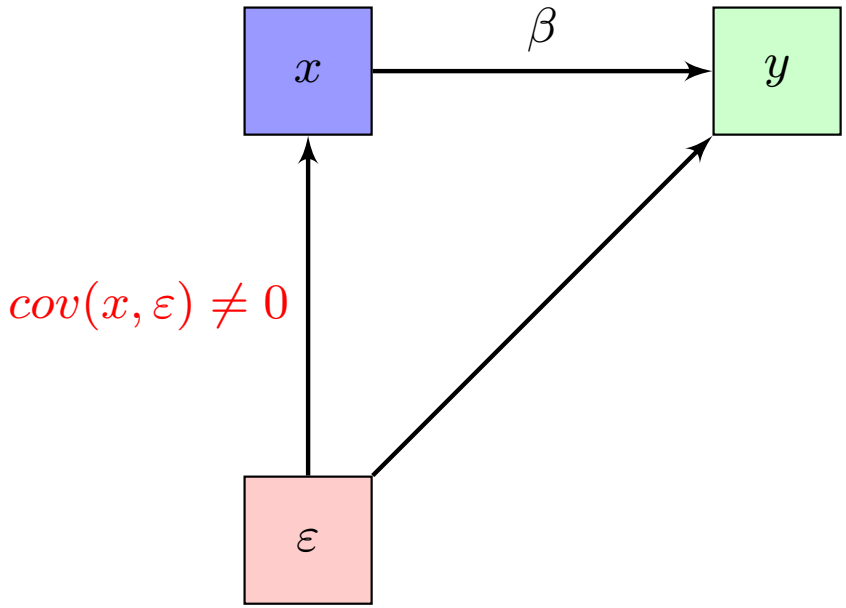
Instrumental variables (IV)



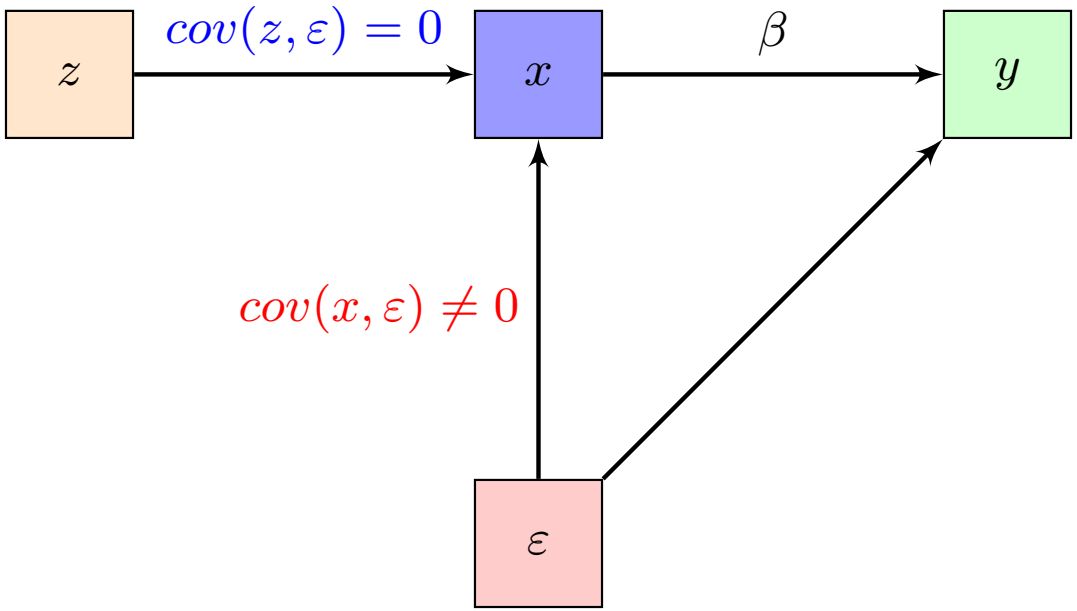
- x – the explanatory variable;
- y – the dependent variable;



- x – the explanatory variable;
- y – the dependent variable;
- ε – the error term;



- x – the explanatory variable;
- y – the dependent variable;
- ε – the error term;



- x – the explanatory variable;
- y – the dependent variable;
- ε – the error term;
- z – the instrumental variable.

- Consider the linear model with single explanatory variable:

$$y = \alpha + \beta x + \varepsilon \quad \text{and} \quad cov(\varepsilon|x) \neq 0. \tag{27}$$

- The OLS estimates of β will be inconsistent.
- **Instrumental variable regression (IV)** divides variation of the endogenous variable (x) in two parts:
  1. a part that might be **not** correlated with the error term (ε),
  2. a part that might be correlated with the error term (ε).
- It is possible due to using **instrumental variable (instrument, z)** which is not correlated with ε.
- The instrument (z) allows to identify the variation in endogenous variable that is not correlated with ε and, therefore, can be used to estimate β.

- More generally, the IV regression is:

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_k x_k + \beta_{k+1} w_1 + \ldots + \beta_{k+r} w_r + \varepsilon, \tag{28}$$

where

- y is the dependent variable;
- ε is the error term. In the context of the endogeneity, it might capture omitted factors as well as measurement error;
- $x_1, \ldots, x_k$ are k **endogenous** variables that can be correlated with the error term ε;
- $w_1, \ldots, w_r$ are r **exogenous** variables that are potentially not correlated with the error term ε;
- $z_1, \ldots, z_m$ are m **instrumental variables**.

Identification

The coefficients $\beta_1, \ldots, \beta_{k+r}$ are said to be:

- **exactly identified** if $m = k$;
- **underidentified** if $m < k$;
- **overidentified** if $m > k$.

The coefficients have to be **exactly identified** or **overidentified** if we want to apply IV regression.

Two conditions for valid instruments

1. **Instrument Relevance**
   A set of instrumental variables $(z_1, \ldots, z_m)$ must be related to the endogenous explanatory variables $(x_1, \ldots, x_k)$. Formally,

$$cov(z_i, x_j) \neq 0.$$

2. **Instrument Exogeneity**
   A set of instrumental variables $(z_1, \ldots, z_m)$ cannot be correlated with the error term ε. Formally,

$$cov(\varepsilon, z_i) = 0.$$

examples of instruments (Angrist, Krueger, 2001)

Dependent variable	Endogenous x	Source of Instrumental variable	Reference
Earnings	Years of schooling	Region and time variation in school construction	Duflo (2001)
Earnings	Years of schooling	Proximity to college	Card (1995)
Earnings	Years of schooling	Quarter of birth	Angrist and Krueger (1991)
Earnings	Veteran status	Cohort dummies	Imbens and van der Klaauw (1995)
Birth weight	Maternal smoking	State cigarette taxes	Evans and Ringel (1999)
Health	Heart attack surgery	Proximity to cardiac care centers	McClellan, McNeil and Newhouse (1994)
College enrollment	Financial aid	Discontinuities in financial aid formula	van der Klaauw (1996)
Crime	Police	Electoral cycles	Levitt (1997)

The Two Stages Least Squares (TSLS) estimator

Two Stage Least Squares (TSLS):

$$y = \beta_0 + \beta_1 x + \varepsilon. \tag{29}$$

1. **First-Stage Regression (reduced form)**:
   Regress the endogenous variable (x) on the instrument (z):

$$x = \pi_0 + \pi_1 z + \eta, \tag{30}$$

   Based on the OLS estimates calculate predicted values, i.e., $\hat{x}$.
2. **Second -Stage Regression (structural form/equation)**:
   Using OLS regress dependent variable y on the predicted values $\hat{x}$:

$$y = \beta_0 + \beta_1 \hat{x}_1 \varepsilon. \tag{31}$$

   The TSLS estimator $\hat{\beta}_1^{TSLS}$ stands for the estimates obtained in the second-stage regression.

Two Stage Least Squares (TSLS):

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \beta_{k+1} w_1 + \ldots + \beta_{k+r} w_r + \varepsilon. \tag{32}$$

1. **First-Stage Regression(s) (reduced form)**:
   Regress each of the endogenous variable (x_i) on the instruments ($z_1, \ldots, z_m$) as well as the exogenous variables ($w_1, \ldots, w_r$):

$$\forall_{i \in 1,\ldots,k} \quad x_i = \pi_0 + \pi_1 z_1 + \ldots + \pi_m z_m + \pi_{m+1} w_1 + \ldots + \pi_{m+r} w_r + \eta, \tag{33}$$

   Based on the OLS estimates calculate predicted values, i.e., $\hat{x}_i$.
2. **Second -Stage Regression (structural form/equation)**:
   Using OLS regress dependent variable y on the predicted values $\hat{x}_1$, ..., $\hat{x}_k$ as well as the exogenous variables ($w_1, \ldots, w_r$):

$$y = \beta_0 + \beta_1 \hat{x}_1 + \ldots + \beta_k \hat{x}_k + \beta_{k+1} w_1 + \ldots + \beta_{k+r} w_r + \varepsilon. \tag{34}$$

   The TSLS estimator $\hat{\beta}_1^{TSLS}$, ..., $\hat{\beta}_k^{TSLS}$, ..., $\hat{\beta}_{k+r}^{TSLS}$ stands for the estimates obtained in the second-stage regression.

Specification Tests

- Consider case when there is one endogenous variable (x) and one instrumental variable (z).
- Then the asymptotic properties of the OLS and TSLS estimators:

$$\text{plim}\left(\hat{\beta}^{OLS}\right) = \beta + cor(x,e)\frac{\sigma_e}{\sigma_x}, \tag{35}$$

$$\text{plim}\left(\hat{\beta}^{IV}\right) = \beta + \frac{cor(z,e)}{cor(z,x)}\frac{\sigma_e}{\sigma_z}. \tag{36}$$

- Then, the two stages least squares estimator is consistent if

$$\frac{cor(z,e)}{cor(z,x)} = 0, \tag{37}$$

  so when the instrumental variable is **strong** and **exogenous**.
- In the IV estimation the TSLS estimator is less efficient than the OLS estimator.

- To test the strength of instruments it is useful to analyze the first step regression:

$$x_i = \pi_0 + \pi_1 z_1 + \ldots + \pi_m z_m + \pi_{m+1} w_1 + \ldots + \pi_{m+r} w_r + \eta, \tag{38}$$

the intuitive null hypothesis:

$$\mathcal{H}_0 \quad \pi_1 = \pi_2 = \ldots = \pi_m = 0, \tag{39}$$

is related to the weak instruments case.

- It can be tested with $\mathcal{F}$ test.
- The rule of the thumb: if the $\mathcal{F}$ statistic is larger than 10 then the null can be rejected.

- The Hausman test allows to investigate endogeneity of explanatory variable.
- **Key assumption:** the IV estimates are unbiased, i.e., instrumental variables are strong and exogenous.
- The null hypothesis:

$$\mathcal{H}_0 : \quad cov(x, \varepsilon) = 0, \tag{40}$$

refers to exogeneity of explanatory variable while in the alternative hypothesis

$$\mathcal{H}_0 : \quad cov(x, \varepsilon) \neq 0, \tag{41}$$

the explanatory variable is endogenous and, therefore, the OLS estimates are inconsistent.

- There are several version of the Hausman test.

- The Hausman-Wu test statistic:

$$\mathcal{H} = \left(\hat{\beta}^{OLS} - \hat{\beta}^{TSLS}\right)' \left(Var(\hat{\beta}^{TSLS}) - Var(\hat{\beta}^{OLS})\right)^{-1} \left(\hat{\beta}^{OLS} - \hat{\beta}^{TSLS}\right) \tag{42}$$

  is χ^2 distributed with K degrees of freedom.
- In general, the above version of Hausman test allows to check whether the IV and OLS estimates are statistically significant.
- It is assumed that the IV estimator is consistent under both null and alternative.
- The null is about a lack of differences between the OLS and TSLS estimators. In other words, the null is about consistency of the least squares.
- The alternative hypothesis postulates that the difference between the OLS and TSLS estimator is systematic/significant. In other words, the alternative states that the OLS estimator is inconsistent.

- The other version of Hausman test bases on including the residuals from the first step regression into the structural equation. Let's assume simply regression model:

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{43}$$

where x is being tested for endogeneity.

- First step is the regression of x on the instrumental variables (e.g. z_1 and z_2):

$$x = \theta_0 + \theta_1 z_1 + \theta_2 z_2 + \eta, \tag{44}$$

and obtaining the residuals $\hat{\eta}$.

- In the second step, the structural equation is extended by residuals from the first step, i.e., $\hat{\eta}$

$$y = \beta_0 + \beta_1 x + \delta\hat{\eta} + \varepsilon, \tag{45}$$

The null hypothesis is related to exogeneity:

$$\mathcal{H}: \quad \delta = 0, \tag{46}$$

or no correlation between x and ε. It can be tested with a standard t-test.

- When there are more explanatory variables that are tested to be endogenous then:

- regression in the first step is repeated for all variables. The residuals from each equation are collected,
- the auxilary regression in the second step is extended by residuals from each first step regression,
- the $\mathcal{F}$ statistic is used to test joint significance of the coefficients on the included residuals.

When the number of instruments is larger than number of endogenous variables (overidentification) we can test their validity.

1. **Informal strategy**
   - Try different combinations of instrumental variables and compare estimates.
2. **Sargan test**
   - In the first step, we perform IV estimation using all instrumental variables in order to get the residuals $\hat{\varepsilon}$.
   - Then, the residuals $\hat{\varepsilon}$ are regressed on all available instruments.
   - The surplus instruments are tested with the statistics NR^2, where N is the number of observations and R^2 is the coefficient of determination.
   - NR^2 is χ^2 distributed with $m-k$ degrees of freedom. $m-k$ is the surplus of the instruments.
   - The null refers to validity of instruments.

- **Standards errors** are little bit more complicated than in the OLS estimator, the variation of endogenous variables is exploited together with consistent estimates.
- **Serial correlation and heteroskedasticty** should also be taken into account.
- **Weak instruments** explain little of variation of the endogenous variables. If the instruments are weak then the TSLS estimates are not reliable.
  - It can be tested with standard $\mathcal{F}$ statistics (testing the hypothesis that the coefficients on the all instruments are zero) in the first stage.
- **Endogeneity of instruments**
  - There is no formal statistical test allowing for testing whether instruments are correlated with the error term.