Endogeneity and Instrumental Variables (IV)
Hausman-Taylor Estimator

Jakub Mućk
SGH Warsaw School of Economics

Instrumental variables (IV)

Inconsistency of OLS

Consider standard (general) linear model:

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon \qquad (1)$$

where $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.
The assumptions of OLS (ordinary least squares):

1. **Linearity**: the specification of (1) is correct.
2. **Full rank**: the matrix X, i.e. $X = [x_1, \ldots, x_k]$ has full column rank (not higher than number of observation).
3. **Nonautocorrelation and homoscedasticity of the error term**: $\mathbb{E}(ee') = \sigma_\varepsilon^2 I$.
4. **Independent observations**.
5. **Exogeneity:** $\mathbb{E}(\varepsilon | x_1, \ldots, x_k) = 0$.

It is assumed that all **independent variables are exogenous (assumption #5)**.

Endogenous variables

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

Inconsistency of OLS

An endogeneity problem leads to inconsistency of the OLS estimator.

Endogenous variables

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

Inconsistency of OLS

An endogeneity problem leads to inconsistency of the OLS estimator.

Standard cases when explanatory variables are endogenous:

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

- Let's assume that **true** DGP (data generating process) for the consumption (c) is as follows:
$$c = \alpha + \beta inc^* + \varepsilon \tag{2}$$
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:
$$inc^* = inc + \eta, \tag{3}$$
where η stands for the measurement and $\eta \sim \mathcal{N}\left(0, \sigma^2_\eta\right)$.
- The current income (inc) is **proxy variable** for the permanent income (inc^*).
- Substituting the permanent income into (2):
$$c = \alpha + \beta\left(inc + \eta\right) + \varepsilon = \alpha + \beta inc + \beta\eta + \varepsilon = \alpha + \beta inc + \nu, \tag{4}$$
where $\nu = \varepsilon + \beta\eta$.
- The covariance between inc and error term (ν):
$$\text{cov}(inc, \nu) = \mathbb{E}\left(inc\nu\right) = \mathbb{E}\left((inc^* + \eta)(\varepsilon + \beta\eta)\right) = \mathbb{E}\left(\beta\eta^2\right) = \sigma^2_\eta\beta \neq 0. \tag{5}$$

Omitted variable bias - example I

SGH

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

$$w = \alpha + \rho\mathcal{S} + \beta\mathcal{A} + \varepsilon, \tag{6}$$

  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 (7):

$$w = \alpha + \rho\mathcal{S} + \eta, \tag{7}$$

- 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:

$$\hat{\rho}^{OLS} = \text{cov}(w, \mathcal{S})/Var(\mathcal{S}). \tag{8}$$

- Plugging *true* DGP for wages w:

$$\hat{\rho}^{OLS} = \frac{\text{cov}(\alpha + \rho\mathcal{S} + \beta\mathcal{A} + \varepsilon, \mathcal{S})}{Var(\mathcal{S})}, \tag{9}$$

- After manipulation we get:

$$\hat{\rho}^{OLS} = \frac{1}{Var(\mathcal{S})}\mathbb{E}\left[(\alpha + \rho\mathcal{S} + \varepsilon)\,\mathcal{S} + \beta\mathcal{A}\mathcal{S}\right] = \rho + \underbrace{\beta\frac{\text{cov}(\mathcal{A},\mathcal{S})}{Var(\mathcal{S})}}_{=\text{bias}} \neq \rho. \tag{10}$$

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

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

$$c = \alpha + \beta y + \varepsilon \quad (11)$$
$$y = c + i \quad (12)$$

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

- 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:

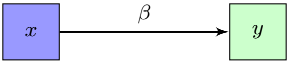
$$\begin{aligned} y &= c + i \\ y &= \alpha + \beta y + \varepsilon + i \\ (1-\beta)y &= \alpha i + \varepsilon \\ y &= \frac{\alpha}{(1-\beta)} + \frac{1}{(1-\beta)} i + \frac{1}{(1-\beta)} \varepsilon. \end{aligned}$$

- The general expression of the OLS estimator of the marginal propensity to consume (β) form equation (11):

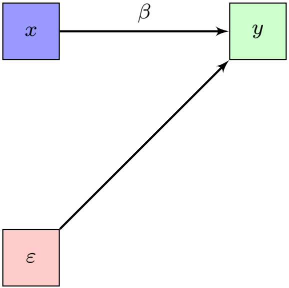
$$\hat{\beta}^{OLS} = \beta + \underbrace{\frac{\sum (y - \bar{y}) \varepsilon}{\sum (y - \bar{y})^2}}_{=0 \quad \text{if} \quad \mathbb{E}(y|\varepsilon)=0} . \tag{13}$$

- 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.

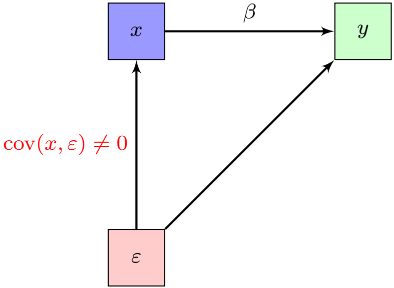
Instrumental variables (IV)– general idea



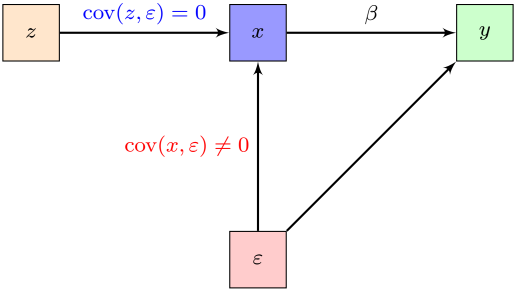
- x – the explanatory variable;
- y – the dependent variable;



- 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:

$$y = \alpha + \beta x + \varepsilon \quad \text{and} \quad \text{cov}(\varepsilon|x) \neq 0. \tag{14}$$

- 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:

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_k x_k + \beta_{k+1} w_1 + \ldots + \beta_{k+r} w_r + \varepsilon, \quad (15)$$

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,
   $$\text{cov}(z_i, x_j) \neq 0.$$
2. **Instrument Exogeneity**
   A set of instrumental variables $(z_1, \ldots, z_m)$ cannot be correlated with the error term ε. Formally,
   $$\text{cov}(\varepsilon, z_i) = 0.$$

Two Stage Least Squares (TSLS):

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_k x_k + \beta_{k+1} w_1 + \ldots + \beta_{k+r} w_r + \varepsilon. \quad (16)$$

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

   $$\forall_{i \in 1,\ldots,k} \quad x_i = \pi_0 + \pi_1 z_1 + \ldots + \pi_m z_m + \pi_{m+1} w_1 + \ldots + \pi_{m+r} w_r + \eta, \quad (17)$$

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

   $$y = \alpha + \beta_1 \hat{x}_1 + \ldots + \beta_k \hat{x}_k + \beta_{k+1} w_1 + \ldots + \beta_{k+r} w_r + \varepsilon. \quad (18)$$

   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.

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)

Instrumental variables (IV)– general remarks

- **Standards errors** are little bit more complicated than in the OLS estimator.
- **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.

Hausman-Taylor estimator

- Let's consider the following one-way RE model:

$$y_{it} = x_{1it}\beta_1 + x_{2it}\beta_2 + z_{1i}\gamma_1 + z_{2i}\gamma_2 + \mu_i + u_{it} \quad (19)$$

where:
x_{1it} are **time-varying** variables; **not correlated with** μ_i
x_{2it} are **time-varying** variables; **correlated with** μ_i
z_{1i} are **time-invariant** variables; **not correlated with** μ_i
z_{2i} are **time-invariant** variables; **correlated with** μ_i

- The RE model estimates on γ_2 are inconsistent.
- The estimator proposed by Hausman and Taylor (1981) takes into account the above correlation.

- **First step**: Within regression for the model including only time-variable regressors, both x_{1it} and x_{2it}. Here, the usual differences from the *temporal* mean are used:

$$(y_{it} - \bar{y}_i) = \beta_1(x_{1i,} - \bar{x}_{1i}) + \beta_2(x_{2it} - \bar{x}_{2i}) + (u_{it} - \bar{u}_i) \tag{20}$$

- Based on the expression above we can estimate variance of the idiosyncratic error, i.e., $\hat{\sigma}^2_{\varepsilon}$.

- **Second step**: construct the *intra-temporal* mean of the residuals from (20):

$$\bar{e} = [\underbrace{(\bar{e}_1, \bar{e}_1, \dots, \bar{e}_1)}_{T}, \dots, \underbrace{(\bar{e}_N, \bar{e}_N, \dots, \bar{e}_N)}_{T}]' \tag{21}$$

- Then make TSLS for $\bar{e}_i$ using:
  variables: z_{1it} (time invariant, not correlated with μ_i), z_{2it} (time invariant, correlated with μ_i)
  instruments: z_{1it}, x_{1it} (time invariant, not correlated with μ_i)
  Specifically,
  1. Regress z_{2it} on z_{1it} as well as x_{1it}.
  2. Use the predicted value from the above regression and create new matrix, i.e., $Z = [z_{1it}, \hat{z}_{21t}]$.
  3. Regress $\bar{e}_i$ on Z to get estimates of γ_1 and γ_2.
  4. Calculate $\sigma^2_{TSLS,\bar{e}}$ the variance of the error components from the above regression.
- Now, we can calculate the variation of the individual-specific error component:

$$\sigma^2_\mu = \sigma^2_{TSLS,\bar{e}} - \frac{\sigma^2_\varepsilon}{T}. \tag{22}$$

- Based on the estimates of σ_μ^2 and σ_ε^2 calculate the conventional in the FGLS regression scale parameter θ:

$$\theta = \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T^{-1}\sigma_\mu^2}} \tag{23}$$

- Finally, do a TSLS regression of y^* on X^* with instruments described by V:

$$y^* = y_{it} - \theta y_{it}, \tag{24}$$
$$X^* = [x_{1it}, x_{2it}, z_{1i}, z_{2i}] - \theta[x_{1it}, x_{2it}, z_{1i}, z_{2i}], \tag{25}$$
$$V = [(x_{1it} - \bar{x}_{1i}), (x_{2it} - \bar{x}_{2i}), z_{1i}, \bar{x}_{1i}], \tag{26}$$

  more specifically:

  1. Regress X^* on the instruments (V) and obtain fitted values, i.e., $\hat{X}^*$,
  2. Regress y^* on the predicted values from the previous step, i.e, $\hat{X}^*$, in order to get the estimates of $[\beta, \gamma]$.

- The estimates of the variance-covariance of the structural parameters are a little bit more complicated.