Week 8: Instrumental Variables (IVs) and Two Stage Least Squares (2SLS)

Advanced Econometrics 4EK608

Vysoká škola ekonomická v Praze

Outline

1. Introduction & repetition from BSc courses
2. Instrumental variables
3. Two stage least squares
4. IV tests: introduction
   - Durbin-Wu-Hausman (endogeneity in regressors)
   - Weak instruments test
   - Sargan (exogeneity in IVs, over-identification only)
   - IV tests: example

Introduction: endogenous regressors

- CS model: $y_i = \boldsymbol{x}_i \boldsymbol{\beta} + u_i \quad \text{and} \quad E[\boldsymbol{x}_i, u_i] \neq 0.$
  - If important regressors cannot be measured (thus make part of u_i) and are correlated with observed regressors of LRM.
  - Endogeneity can be caused by measurement errors.
  - Always present in simultaneous equations models (SEMs). (SEMs will be discussed in Week 9).
- With endogenous regressors, OLS is biased & inconsistent.

Endogeneity in regressors can sometimes be solved

- By means of proxy variables (if uncorrelated to u_i).
- More detailed (multi-equation) specification, if possible.
- Using panel data methods (data availability permitting).
- Using instrumental variable regression (IVR) (we need "good" instruments, assumptions apply).

Introduction: instrumental variables

Example: $\log(wage_i) = \beta_0 + \beta_1 educ_i + [abil_i + u_i]$

Instrumental variables

1. Not in the main (structural) equation: no effect on the dependent variable after controlling for observed regressors.
2. Correlated (positively or negatively) with the endogenous regressor (this can be tested).
3. Not correlated with the error term (in some cases, this can be tested, see Sargan test discussed next).

- Possible IVs: father's education, mother's education, number of siblings, etc.

  Usually, IQ is not a good IV - it's often correlated with *abil*, i.e. with the error term $[abil_i + u_i]$.

Instrumental variables

- $y_i = \beta_0 + \beta_1 x_i + u_i$ SLRM with exogenous regressor x:

$$\begin{array}{ccc} y & \leftarrow & x \\ & \nwarrow & \\ & & u \end{array} \qquad \text{and} \qquad \frac{\mathrm{d}\, y}{\mathrm{d}\, x} = \beta_1$$

- $y_i = \boldsymbol{x}_i \boldsymbol{\beta} + u_i$ MLRM with exogenous regressor(s):

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \qquad | \text{ subs. for } \boldsymbol{y}$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}) \qquad | \text{ rearr. \& take expects.}$$

$$E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} + E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}] = \boldsymbol{\beta}$$

- With exogenous regressors, OLS is unbiased.

Instrumental variables

- $y_i = \beta_0 + \beta_1 x_i + u_i$ SLRM with endogenous regressor x:

$$\begin{array}{ccc} y & \leftarrow & x \\ & \nwarrow & | \\ & & u \end{array} \qquad \text{and} \qquad \frac{\mathrm{d}\, y}{\mathrm{d}\, x} = \beta_1 + \frac{\mathrm{d}\, u}{\mathrm{d}\, x}$$

- $y_i = \boldsymbol{x}_i \boldsymbol{\beta} + u_i$ MLRM with endogenous regressor(s):

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \qquad | \text{ subs. for } \boldsymbol{y}$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}) \qquad | \text{ rearr. \& take expects.}$$

$$E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} + E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}] \neq \boldsymbol{\beta}$$

- With endogenous regressors, $E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}] \neq \mathbf{0}$. Thus, OLS is biased (and asymptotically biased).

Instrumental variables

- $y_i = \beta_0 + \beta_1 x_i + u_i$ IVR principle (SLRM):

$$\begin{array}{ccccc} y & \leftarrow & x & \leftarrow & z \\ & \nwarrow & | & & \\ & & u & & \end{array} \qquad \text{and} \qquad \frac{\mathrm{d}\,y}{\mathrm{d}\,x} = \frac{\mathrm{d}\,y \;/\; \mathrm{d}\,z}{\mathrm{d}\,x \;/\; \mathrm{d}\,z}$$

- $y_i = \boldsymbol{x}_i \boldsymbol{\beta} + u_i$ IVR in MLRMs:

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$
$$\hat{\boldsymbol{\beta}}_{\text{IV}} = (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{y}$$

where $\boldsymbol{Z}$ is a matrix of instruments, same dimensions as $\boldsymbol{X}$.

  - $\boldsymbol{Z}$ follows from $\boldsymbol{X}$, each endogenous regressor (column) is replaced by unique instrument (full column ranks of $\boldsymbol{X}$,$\boldsymbol{Z}$).
  - Exact identification: # endogenous regressors = # IVs
  - In IVR, R^2 has no interpretation (SST $\neq$ SSE + SSR).
  - For IVR, we use specialized robust standard errors
  - **IVR estimator is biased and consistent.**

Instrumental variables: IVR as MM estimator

Exogenous regressors:

- MM: replace $E[\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})] = \boldsymbol{0}$ by $\frac{1}{n}[\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})] = \boldsymbol{0}$ and solve moment equations
- OLS provides identical estimate: $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$

With endogenous regressors (exact identification), moment conditions change:

- MM: replace $E[\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})] = \boldsymbol{0}$ by $\frac{1}{n}[\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})] = \boldsymbol{0}$ and solve moment equations
- IVR provides identical estimate: $\hat{\boldsymbol{\beta}}_{\text{IV}} = (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{y}$

Instrumental variables: IVR as MM estimator

$$y_{i1} = \beta_0 + \beta_1 y_{i2} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i \quad | \; z_1 \text{ is IV for } y_2$$

$$n^{-1} \sum_{i=1}^{n} (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik}) = 0$$

$$n^{-1} \sum_{i=1}^{n} z_{i1} \cdot (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik}) = 0$$

$$n^{-1} \sum_{i=1}^{n} x_{i2} \cdot (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik}) = 0$$

$$\cdots$$

$$n^{-1} \sum_{i=1}^{n} x_{ik} \cdot (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik}) = 0$$

- In moment equations, y_{i2} is replaced by z_{i1}
- Exogenous regressors serve as their own instruments.

IVR estimator is consistent

$$\hat{\beta}_{\text{IV}} = (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{y} \qquad | \text{ subs. for } \boldsymbol{y}$$
$$\hat{\beta}_{\text{IV}} = (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}) \qquad | \text{ rearrange}$$
$$\hat{\beta}_{\text{IV}} = \boldsymbol{\beta} + (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{u}$$

- If consistency condition holds: plim $\left[\frac{1}{n}\boldsymbol{Z}'\boldsymbol{u}\right] = \boldsymbol{0}$, $\hat{\beta}_{\text{IV}}$ is consistent.
- This can be seen from expansion of $[(\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{u}]$:

$$\hat{\beta}_{\text{IV}} = \boldsymbol{\beta} + (n^{-1}\boldsymbol{Z}'\boldsymbol{X})^{-1}\,n^{-1}\boldsymbol{Z}'\boldsymbol{u}$$

Instrumental variables: over-identification

$$y_{i1} = \beta_0 + \beta_1 y_{i2} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i \quad | \, z_1, z_2, z_3 \ \text{ are IVs for } y_2$$

- By choosing any of the z_1, z_2, z_3 IVs (or any linear combination of), we perform IVR
- $\hat{\boldsymbol{\beta}}_{\text{IV}}$ values change, as IV in moment equations changes.
- We cannot "simply" use all three instruments.
  If # columns in $\boldsymbol{Z}$ (l) $>$ # columns in $\boldsymbol{X}$ (k),
  $\boldsymbol{Z}'\boldsymbol{X}$ is $(l \times k)$ with rank k and no inverse:
  $\hat{\boldsymbol{\beta}}_{\text{IV}} = (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{y}$ cannot be calculated
- Solution: Project $\boldsymbol{X}$ to the space column of $\boldsymbol{Z}$ (GMM).
  ($\boldsymbol{X}$ has an endogenous column, $\boldsymbol{Z}$ is purely exogenous).

Instrumental variables: over-identification

Projection matrices - repetition

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y}$$

$$\boldsymbol{y} = \hat{\boldsymbol{y}} + \hat{\boldsymbol{u}} = \boldsymbol{P}\boldsymbol{y} + \boldsymbol{M}\boldsymbol{y}, \text{ where}$$

$$\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{I} - \boldsymbol{P}$$

- Projection of columns of $\boldsymbol{X}$ in the column space of $\boldsymbol{Z}$:

$$\hat{\boldsymbol{X}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X},$$

- Columns of $\hat{\boldsymbol{X}}$ are linear combinations of columns in $\boldsymbol{Z}$, i.e. exogenous.

- IV estimator (over-identification):

$$\hat{\boldsymbol{\beta}}_{\text{IV}} = (\hat{\boldsymbol{X}}'\boldsymbol{X})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y}$$

Instrumental variables: over-identification

- Projection of columns of $\boldsymbol{X}$ in the column space of $\boldsymbol{Z}$:

$$\hat{\boldsymbol{X}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X},$$

- Exogenous columns (regressors) in $\boldsymbol{X}$ appear in $\boldsymbol{Z}$ as well. Such columns are perfectly replicated in $\hat{\boldsymbol{X}}$.

- It may be shown that IVR is equivalent to OLS regression $\boldsymbol{y} \leftarrow \hat{\boldsymbol{X}}$:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\mathrm{IV}} &= (\hat{\boldsymbol{X}}'\boldsymbol{X})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y} \\ &= (\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{M}_Z)\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{M}_Z)\boldsymbol{y} \\ &= (\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y}\end{aligned}$$

- $\boldsymbol{y} \leftarrow \hat{\boldsymbol{X}}$ is part of a two-stage LS (2SLS) method, (discussed next).

Instrumental variables: identification conditions

- In $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$, multiple $\boldsymbol{x}_j$ regressors may be endogenous.
- Identification (estimability) conditions:
  - **Order condition:** We need at least as many IVs (excluded exogenous variables) as there are included endogenous regressors in the main (structural) equation.

    This is a necessary condition for identification.
  - **Rank condition:** $\hat{\boldsymbol{X}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}$ has full column rank (k) so that $(\hat{\boldsymbol{X}}'\boldsymbol{X})^{-1}$ or $(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}$ can be calculated in the IV estimator $\hat{\boldsymbol{\beta}}_{\text{IV}} = (\hat{\boldsymbol{X}}'\boldsymbol{X})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y}$ (will be discussed in detail with respect to 2SLS method and for SEM models).

    This is a necessary and sufficient condition for identification.

Instrumental variables: statistical properties

SLRM: $y_{i1} = \beta_0 + \beta_1 x_{i1} + u_i \quad | \; x_{i1}$ endog., z_{i1} exists

- In large samples, IV estimator has approximately normal distribution (MM/GMM properties).
- For calculation of standard errors, we usually need assumption of homoskedasticity conditional on IV(s). Alternatively, we calculate robust errors.
- Asymptotic variance of the IV estimator is always higher than of the OLS estimator.

$$\mathrm{var}(\hat{\beta}_{1,IV}) = \frac{\hat{\sigma}^2}{SST_x \cdot R^2_{x,z}} \quad > \quad \mathrm{var}(\hat{\beta}_{1,OLS}) = \frac{\hat{\sigma}^2}{SST_x}$$

Instrumental variables: statistical properties

SLRM: $y_{i1} = \beta_0 + \beta_1 x_{i1} + u_i \quad | \; x_{i1}$ endog., z_{i1} exists

- Asymptotic variance of the IV estimator decreases with increasing correlation between z and x.
- IV-related routines & tests are implemented in R, ...
- Both endogenous explanatory variables and IVs can be binary variables.
- R^2 can be negative and has no interpretation nor relevance if IVR is used.

Instrumental variables: statistical properties

SLRM: $y_{i1} = \beta_0 + \beta_1 x_{i1} + u_i \quad | \; x_{i1}$ endog., z_{i1} exists

- If (small) correlation between u and instrument z is possible, inconsistency in the IV estimator can be much higher than in the OLS estimator:

$$\text{plim}\hat{\beta}_{1,OLS} = \beta_1 + \text{corr}(x,u) \cdot \frac{\sigma_u}{\sigma_x}$$

$$\text{plim}\hat{\beta}_{1,IV} = \beta_1 + \frac{\text{corr}(z,u)}{\text{corr}(z,x)} \cdot \frac{\sigma_u}{\sigma_x}$$

- Weak instrument: if correlation between z and x is small.

Instrumental variables: statistical properties

MLRM: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \quad | \text{ valid } \boldsymbol{Z} \text{ exists}$

- IVR method is a "trick" for consistent estimation of the ceteris paribus effects, i.e. $\hat{\beta}_{j,\mathrm{IV}}$.
- Fitted values are generated as $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$ (NOT from $\hat{\boldsymbol{y}} = \hat{\boldsymbol{X}}\hat{\boldsymbol{\beta}}_{\mathrm{IV}}$).
- Similarly: $\mathrm{var}(\hat{u}_i) = \hat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_{\mathrm{IV}})^2$ d.f. correction is superfluous (asymptotic use only).
- $\mathrm{Asy.Var}(\hat{\boldsymbol{\beta}}_{\mathrm{IV}}) = \hat{\sigma}^2(\boldsymbol{Z}'\boldsymbol{X})^{-1}(\boldsymbol{Z}'\boldsymbol{Z})(\boldsymbol{X}'\boldsymbol{Z})^{-1}$ for the exactly identified & homoskedastic case.
- With heteroskedasticity and/or over-identification, the $\mathrm{Asy.Var}(\hat{\boldsymbol{\beta}}_{\mathrm{IV}})$ formula is complex and built into all SW packages.

2SLS as a special case of IVR

$$\hat{\boldsymbol{\beta}}_{\text{IV}} = (\hat{\boldsymbol{X}}'\boldsymbol{X})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y} = (\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y}$$

2SLS:

- Structural equation (as in SEMs)

  $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x_2 + \cdots + \beta_k x_k + u \quad | \; z_1$ exists

- Reduced form for y_2 – endogenous variable as function of all exogenous variables (including IVs)

  $y_2 = \pi_0 + \pi_1 z_1 + \pi_2 x_2 + \cdots + \pi_k x_k + \varepsilon$

- 1^{st} stage of 2SLS: Estimate reduced form by OLS
  - Order condition for identification of the structural equation: at least one instrument for each endogenous regressor).
  - If z_1 is an IV for y_2, its coefficient must not be zero (rank condition for identification) in the reduced form equation - see stage 2 of 2SLS.

2SLS as a special case of IVR

$$\hat{\boldsymbol{\beta}}_{\text{IV}} = (\hat{\boldsymbol{X}}'\boldsymbol{X})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y} = (\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y}$$

2SLS:

- Structural equation

  $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x_2 + \cdots + \beta_k x_k + u \quad | \; z_1$ exists

- 1^{st} stage of 2SLS: estimate reduced form for y_2:

  $\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 x_2 + \cdots + \hat{\pi}_k x_k$

- 2^{nd} stage of 2SLS: Use $\hat{y}_2$ to estimate structural equation:

  $y_1 = \beta_0 + \beta_1 \hat{y}_2 + \beta_2 x_2 + \cdots + \beta_k x_k + u$

- Note that RHS in the 2^{nd} stage contains all exogenous regressors repeated from $\boldsymbol{X}$, while $\hat{y}_2$ is y_2 "projected" onto $\boldsymbol{Z}$ and thus uncorrelated with u.
- Order condition explained: if $\pi_1 = 0$, $\hat{y}_2$ is a perfect linear combination of the remaining RHS regressors in 2^{nd} stage.

Instrumental variables

Instrumental variables: summary

- Excluded from the main / structural equation
- Must be correlated with endogenous regressor(s)
- Must not be correlated with u

All IVs used in IVR / 2SLS estimation must fulfill the conditions above.

In 2SLS, 1$^{\text{st}}$ stage is used to generate the “best” IV.
With multiple endogenous regressors, reduced forms for each endogenous regressor must be constructed and estimated, rank and order conditions apply.

Two stage least squares

2SLS properties

- The standard errors from the OLS second stage regression are biased and inconsistent estimators with respect to the original structural equation (SW handles this problem automatically).
- If there is one endogenous variable and one instrument then 2SLS = IV
- With multiple endogenous variables and/or multiple instruments, 2SLS is a special case of IVR.

Two stage least squares

Statistical properties of the 2SLS/IV estimator

- Under assumptions completely analogous to OLS, but conditioning on z_i rather than on x_i , 2SLS/IV is consistent and asymptotically normal.
- 2SLS/IV estimator is typically much less efficient than the OLS estimator because there is more multicollinearity and less explanatory variation in the second stage regression
- Problem of multicollinearity is much more serious with 2SLS than with OLS

Statistical properties of the 2SLS/IV estimator

- Corrections for heteroskedasticity/serial correlation analogous to OLS
- 2SLS/IV estimamtion easily extends to time series and panel data situations

IV tests: introduction

LRM: $y_{i1} = \beta_0 + \beta_1 y_{i2} + \beta_2 x_{i1} + u_i$; $\boldsymbol{z}$ instruments exist

IV regression advantages for endogenous y_2:

$\rightarrow$ $\hat{\beta}_{1,OLS}$ is a biased and inconsistent estimator (asymptotic errors)

$\rightarrow$ $\hat{\beta}_{1,IV}$ is a biased and consistent estimator (increased sample size (n) lowers estimator bias and s.e.)

IVR disadvantages (price for the IV regression):

- s.e.($\hat{\beta}_{1,IV}$) > s.e.($\hat{\beta}_{1,OLS}$)
- $\hat{\beta}_{1,IV}$ is biased, even if y_2 is actually exogenous

  $\hat{\beta}_{1,OLS}$ is unbiased for exogenous regressors (potentially, pending other G-M conditions).

IV tests: introduction

LRM: $y_{i1} = \beta_0 + \beta_1 y_{i2} + \beta_2 x_{i1} + u_i$; $\boldsymbol{z}$ instruments exist

- Is the regressor y_2 endogenous / $\text{corr}(y_2, u) \neq 0$ / ?
  Is it meaningful to use IVR (considering IVRs "price")?
  Durbin-Wu-Hausman endogeneity test
- Are the instruments actually helpful
  (weakly or strongly correlated with endogenous regressors)?
  Weak instruments test
- Are the instruments really exogenous / $\text{corr}(z_j, u) = 0$ / ?
  Sargan test (only applicable in case of over-identification)

Different types & specifications for IV-tests exist, often focusing on the distribution of the difference between IVR and OLS estimators $(\hat{\boldsymbol{\beta}}_{\text{IV}} - \hat{\boldsymbol{\beta}}_{\text{OLS}})$ under the corresponding H_0.

Durbin-Wu-Hausman endogeneity test

$$y_{i1} = \beta_0 + \beta_1 y_{i2} + \beta_2 x_{i1} + u_i \quad \mid z_{i1}, \tag{1}$$

DWH test motivation:

If z_1 is a proper instrument (uncorrelated with u), then y_2 is endogenous (correlated with u) if and only if ε (error from reduced form equation) is correlated with u.

- y_2 in (1) is endogenous $\quad \Leftrightarrow \quad \text{corr}(y_2, u) \neq 0$
- Reduced form: $y_2 = l.f.(x_1, z_1) + \varepsilon \;\Rightarrow\; y_2 = \hat{y}_2 + \hat{\varepsilon}$
- $\text{corr}(y_2, u) \neq 0 \;\wedge\; \text{corr}(\boldsymbol{z}, u) = 0 \;\Rightarrow\; \text{corr}(\varepsilon, u) \neq 0$
- y_1 is always correlated with u in (1).
- Hence, $\hat{\varepsilon}$ is significant in the regression, if y_2 is endogenous.
- IV/IVs uncorrelated with u is essential for DWH to “work”.

Note: other versions of the DWH test exist...

Durbin-Wu-Hausman endogeneity test

Structural equation:

$$y_{i1} = \beta_0 + \beta_1 y_{i2} + \beta_2 x_{i1} + u_i; \quad \text{IVs: } z_1 \text{ and } z_2 \tag{1}$$

Reduced form for y_2:

$$y_{i2} = \pi_0 + \pi_1 z_{i1} + \pi_2 z_{i2} + \pi_3 x_{i1} + \varepsilon_i \tag{2}$$

H_0: y_2 is exogenous $\leftrightarrow$ $\hat{\varepsilon}$ is not significant when added to equation (1)

H_1: y_2 is endogenous $\rightarrow$ OLS is not consistent for (1) estimation, use IVR (2SLS).

Testing algorithm:

1. Estimate equation (2) and save residuals $\hat{\varepsilon}$.
2. Add residuals $\hat{\varepsilon}$ into equation (1) and estimate using OLS (use HC inference).
3. H_0 is rejected if $\hat{\varepsilon}$ in the modified equation (1) is statistically significant (t-test).

Weak instruments

Motivation for Weak instruments and Sargan tests:

LRM: $y_{i1} = \beta_0 + \beta_1 y_{i2} + \beta_2 x_{i1} + u_i;\;$ z instrument exists

- IVR is consistent if $\text{cov}(z, y_2) \neq 0$ and $\text{cov}(z, u) = 0$
- If we allow for (weak) correlation between z and u, the asymptotic error of IV estimator is:

$$\text{plim}(\hat{\beta}_{1,IV}) = \beta_1 + \frac{\text{corr}(z, u)}{\text{corr}(z, y_2)} \cdot \frac{\sigma_u}{\sigma_{y_2}}$$

- If $\text{corr}(z, y_2)$ is too weak (too close to zero in absolute value), OLS may be better than IV. The asymptotic bias for OLS (LRM with endogenous y_2):

$$\text{plim}(\hat{\beta}_{1,OLS}) = \beta_1 + \text{corr}(y_2, u) \cdot \frac{\sigma_u}{\sigma_{y_2}}$$

Rule of thumb: IF $|\text{corr}(z, y_2)| < |\text{corr}(y_2, u)|$, do not use IVR.

Weak instruments

Structural equation:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x_1 + \cdots + \beta_{k+1} x_k + u; \quad \text{IVs: } z_1, z_2, \ldots, z_m$$

The reduced form for y_2:

$$y_2 = \pi_0 + \pi_1 x_1 + \pi_2 x_2 + \cdots + \pi_k x_k + \theta_1 z_1 + \cdots + \theta_m z_m + \varepsilon$$

H_0: $\theta_1 = \theta_2 = \cdots = \theta_m = 0$
interpretation: "instruments are weak".

H_1: $\neg H_0$

Testing for weak instruments:

Use F-test (heteroskedasticity-robust) or the LM test (χ^2) to test for the joint null hypothesis.

Sargan test (over-identification only)

Structural equation:

$$y_{i1} = \beta_0 + \beta_1 y_{i2} + \beta_2 x_{i1} + u_i; \quad \text{IVs: } z_1, z_2, \ldots \tag{3}$$

H_0: all IVs are uncorrelated with u

H_1: at least one instrument is endogenous

Testing algorithm:

1. Estimate equation (3) using IVR and save the $\hat{u}$ residuals.
2. Use OLS to estimate auxiliary regression: $\hat{u} \leftarrow f(\boldsymbol{x}, \boldsymbol{z})$ and save the R_a^2
3. Under H_0: $nR_a^2 \sim \chi_q^2$ where $q =$ (number of IVs) - (number of endogenous regressors) i.e. q is the number of over-identifying variables.
4. If the observed test statistics exceeds its critical value (at a given significance level), we reject H_0.

IV tests: example

Wooldridge, bwght dataset
R code, {AER} package

```
Call:
ivreg(formula = lbwght ~  packs + male  |  faminc + motheduc + male,
    data = bwght)

Residuals:
     Min        1Q    Median        3Q       Max
-1.66291  -0.09793   0.01717   0.11616   0.82793

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.77419    0.01099 434.478  < 2e-16 ***
packs       -0.25584    0.07613  -3.361 0.000798 ***
male         0.02422    0.01048   2.311 0.021003 *

Diagnostic tests:
                  df1  df2 statistic p-value
Weak instruments    2 1383    38.732  <2e-16 ***
Wu-Hausman          1 1383     5.385  0.0205 *
Sargan              1   NA     4.476  0.0344 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

Residual std. error: 0.195 on 1384 d.f.
Multiple R-Squared: -0.04371,   Adj R-sqr: -0.04522
Wald test: 8.342 on 2 and 1384 DF,  p-value: 0.0002504
```

IVs

Regressors explicitly included in equation

✓Reject H_0: IVs are weak

✓Reject H_0: pack are exogenous

!! Reject H_0: all IVs are uncorrelated with u (!DWH assumptions!)