

Block 4

Instrumental variable regression (IVR)

Two stage least squares (2SLS)

Simultaneous Equation Models

Advanced Econometrics 4EK608

Vysoká škola ekonomická v Praze

# Outline

- ➊ Introduction & repetition from BSc courses
- ➋ Instrumental variables
- ➌ Two stage least squares
- ➍ IVR diagnostic tests
  - Durbin-Wu-Hausman (endogeneity in regressors)
  - Weak instruments test
  - Sargan (exogeneity in IVs, over-identification only)
- ➎ SEM: introduction
- ➏ SEM identification
- ➐ Identification conditions
- ➑ Systems with more than two equations

# Introduction: endogenous regressors

- CS model:  $y_i = \mathbf{x}_i\boldsymbol{\beta} + u_i$  and  $E[\mathbf{x}_i, u_i] \neq 0$ .
  - 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).
- 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

- ❶ Not in the main (structural) equation: no effect on the dependent variable after controlling for observed regressors.
  - ❷ Correlated (positively or negatively) with the endogenous regressor (this can be tested).
  - ❸ 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  $x$ :

$$\begin{array}{ccc} y & \leftarrow & x \\ & \nwarrow & \\ & & u \end{array} \quad \text{and} \quad \frac{dy}{dx} = \beta_1$$

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

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

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

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

- With exogenous regressors, OLS is unbiased.

# Instrumental variables

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

$$\begin{array}{ccc} y & \leftarrow & x \\ & \nwarrow & | \\ & & u \end{array} \quad \text{and} \quad \frac{dy}{dx} = \beta_1 + \frac{du}{dx}$$

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

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

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

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

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

# Instrumental variables

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

$$\begin{array}{ccccc} y & \leftarrow & x & \leftarrow & z \\ & \swarrow & | & & \\ & & u & & \end{array} \quad \text{and} \quad \frac{dy}{dx} = \frac{dy / dz}{dx / dz}$$

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

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

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

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

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

# Instrumental variables: IVR as MM estimator

Exogenous regressors:

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

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

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



# Instrumental variables: IVR as MM estimator

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

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

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

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

...

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

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

# IVR estimator is consistent

$$\hat{\beta}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y} \quad | \text{ subs. for } \mathbf{y}$$

$$\hat{\beta}_{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'(\mathbf{X}\beta + \mathbf{u}) \quad | \text{ rearrange}$$

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

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

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

# Instrumental variables: over-identification

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

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

# Instrumental variables: over-identification

## Projection matrices - repetition

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

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

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

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

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

- Columns of  $\hat{\mathbf{X}}$  are linear combinations of columns in  $\mathbf{Z}$ , i.e. exogenous.
- IV estimator (over-identification):

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

# Instrumental variables: over-identification

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

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

- Exogenous columns (regressors) in  $\mathbf{X}$  appear in  $\mathbf{Z}$  as well. Such columns are perfectly replicated in  $\hat{\mathbf{X}}$ .
- It may be shown that IVR is equivalent to OLS regression  $\mathbf{y} \leftarrow \hat{\mathbf{X}}$ :

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

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

# Instrumental variables: identification conditions

- In  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ , multiple  $\mathbf{x}_j$  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:**  $\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$  has full column rank ( $k$ ) so that  $(\hat{\mathbf{X}}'\mathbf{X})^{-1}$  or  $(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}$  can be calculated in the IV estimator  $\hat{\boldsymbol{\beta}}_{\text{IV}} = (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y}$  (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:**  $y_{i1} = \beta_0 + \beta_1 x_{i1} + u_i \mid x_{i1} \text{ endog., } z_{i1} \text{ exists}$

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

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

# Instrumental variables: statistical properties

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

- 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:**  $y_{i1} = \beta_0 + \beta_1 x_{i1} + u_i \mid x_{i1} \text{ endog.}, z_{i1} \text{ exists}$

- 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:**  $y = X\beta + u$  | valid  $Z$  exists

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

## 2SLS as a special case of IVR

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

### 2SLS:

- Structural equation (as in SEMs)

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

- 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<sup>st</sup> 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

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

### 2SLS:

- Structural equation

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

- 1<sup>st</sup> 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<sup>nd</sup> 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<sup>nd</sup> stage contains all exogenous regressors repeated from  $\mathbf{X}$ , while  $\hat{y}_2$  is  $y_2$  “projected” onto  $\mathbf{Z}$  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<sup>nd</sup> 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<sup>st</sup> 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.

## 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.

## 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 heteroscedasticity/serial correlation analogous to OLS
- 2SLS/IV estimation easily extends to time series and panel data situations



# IVR diagnostic tests: introduction

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

IV regression advantages for endogenous  $y_2$ :

- $\hat{\beta}_{1,OLS}$  is a **biased and inconsistent estimator**  
(asymptotic errors)
- $\hat{\beta}_{1,IV}$  is a **biased and consistent estimator** (increased sample size ( $n$ ) lowers estimator bias and s.e.)

IVR disadvantages (price for the IV regression):

- $\text{s.e.}(\hat{\beta}_{1,IV}) > \text{s.e.}(\hat{\beta}_{1,OLS})$
- $\hat{\beta}_{1,IV}$  is biased, even if  $y_2$  is actually exogenous  
 $\hat{\beta}_{1,OLS}$  is unbiased for exogenous regressors  
(potentially, pending other G-M conditions).

# IVR diagnostic tests: introduction

LRM:  $y_{i1} = \beta_0 + \beta_1 y_{i2} + \beta_2 x_{i1} + u_i$ ;  $z$  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 ( $\hat{\beta}_{\text{IV}} - \hat{\beta}_{\text{OLS}}$ ) under the corresponding  $H_0$ .

# Durbin-Wu-Hausman endogeneity test

$$y_{i1} = \beta_0 + \beta_1 y_{i2} + \beta_2 x_{i1} + u_i \quad | \quad z_{i1}, \quad (1)$$

## 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  $\varepsilon$  (error from reduced form equation) is correlated with  $u$ .

- $y_2$  in (1) is endogenous  $\Leftrightarrow \text{corr}(y_2, u) \neq 0$
- Reduced form:  $y_2 = l.f.(x_1, z_1) + \varepsilon \Rightarrow y_2 = \hat{y}_2 + \hat{\varepsilon}$
- $\text{corr}(y_2, u) \neq 0 \wedge \text{corr}(\hat{y}_2, u) = 0 \Rightarrow \text{corr}(\hat{\varepsilon}, u) \neq 0$
- $y_1$  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:

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

Reduced form for  $y_2$ :

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

$H_0$ :  $y_2$  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).

## 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:

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

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 (heteroscedasticity-robust) or the  $LM$  test ( $\chi^2$ ) to test for the joint null hypothesis.

# Sargan test (over-identification only)

Structural equation:

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

$H_0$ : 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:  $\hat{u} \leftarrow f(\mathbf{x}, \mathbf{z})$  and save the  $R_a^2$
- 3 Under  $H_0$ :  $nR_a^2 \sim \chi_q^2$  where  
 $q = (\text{number of IVs}) - (\text{number of endogenous regressors})$   
i.e.  $q$  is the number of over-identifying variables.
- 4 If the observed test statistic exceeds its critical value (at a given significance level), we reject  $H_0$ .

# IVR diagnostic 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  
(IVR adequate)

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



# Simultaneous equation model (SEM)

- SEM: outline
- SEM: identification
- Identification conditions
- SEMs with more than two equations

## Simultaneity is another important form of endogeneity

Simultaneity occurs if at least two variables are jointly determined. A typical case is when observed outcomes are the result of separate behavioral mechanisms that are coordinated in an equilibrium.

Prototypical case: a system of demand and supply equations:

- $D(p)$  how high *would* demand be if the price was set to  $p$ ?
- $S(p)$  how high *would* supply be if the price was set to  $p$ ?
- Both mechanisms have a ceteris paribus interpretation.
- Observed quantity and price will be determined in equilibrium, where  $D(p) = S(p)$ .

Simultaneous equations systems can be estimated by 2SLS/IVR  
... Identification conditions apply.

## Example 1: Labor supply and demand in agriculture

$$h_s = \alpha_1 w + \beta_1 z_1 + u_1$$

$$h_d = \alpha_2 w + \beta_2 z_2 + u_2$$

- Endogenous variables, exogenous variables, observed and unobserved supply shifter, observed and unobserved demand shifter
- We have  $n$  regions, market sets equilibrium price and quantity in each. We observe the equilibrium values only

$$h_{is} = h_{id} \Rightarrow (h_i, w_i)$$

Example 1: Labor supply and demand in agriculture contnd.

$$h_i = \alpha_1 w_i + \beta_1 z_{i1} + u_{i1}$$

$$h_i = \alpha_2 w_i + \beta_2 z_{i2} + u_{i2}$$

- If we have the same exogenous variables in each equation, we cannot identify (distinguish) equations.
- We assume independence between errors in structural equations & exogenous regressors.

**Example 1:** Labor supply and demand in agriculture contnd.

If we estimate the structural equation with OLS method, estimators will be biased – so called “simultaneity bias”.

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1$$

$$y_2 = \alpha_2 y_1 + \beta_2 z_2 + u_2$$

$y_2$  is dependent on  $u_1$

(substitute RHS of the 1<sup>st</sup> equation for  $y_1$  in the 2<sup>nd</sup> eq.)

$$\Rightarrow y_2 = \left[ \frac{\alpha_2 \beta_1}{1 - \alpha_2 \alpha_1} \right] z_1 + \left[ \frac{\beta_2}{1 - \alpha_2 \alpha_1} \right] z_2 + \left[ \frac{\alpha_2 u_1 + u_2}{1 - \alpha_2 \alpha_1} \right]$$

# Structural and reduced form equations, 2SLS method

## Structural equations (example)

$$y_1 = \beta_{10} + \beta_{11}y_2 + \beta_{12}z_1 + u_1$$

$$y_2 = \beta_{20} + \beta_{21}y_1 + \beta_{22}z_2 + u_2$$

## Reduced form equations

$$y_1 = \pi_{10} + \pi_{11}z_1 + \pi_{12}z_2 + \varepsilon_1 \quad \Rightarrow \quad \hat{y}_1 \text{ by OLS}$$

$$y_2 = \pi_{20} + \pi_{21}z_1 + \pi_{22}z_2 + \varepsilon_2 \quad \Rightarrow \quad \hat{y}_2 \text{ by OLS}$$

## 2SLS (a special case of IVR)

- 1<sup>st</sup> stage: Estimate reduced forms, get  $\hat{y}_1$  and  $\hat{y}_2$ .
- 2<sup>nd</sup> stage: Replace endogenous regressors in structural equations by fitted values from 1<sup>st</sup> stage, estimate by OLS.

Estimation assumptions and “problems” involved:

- ... Identification of structural equations,
- ... Statistical inference in structural equations (2<sup>nd</sup> stage).

## Example 2: (Structural equations)

Estimation of murder rates

$$murdpc = \beta_{10} + \alpha_1 polpc + \beta_{11} incpc + u_1$$

$$polpc = \beta_{20} + \alpha_2 murdpc + \beta(other\ factors) + u_2$$

- 1<sup>st</sup> equation describes the behaviour of murderers, 2<sup>nd</sup> one the behaviour of municipalities.  
Each one has its ceteris paribus interpretation.
- For the municipality policy, the 1<sup>st</sup> equation is interesting: what is the impact of exogenous increase of police force on the murder rate?
- However, the number of police officers is not exogenous (simultaneity problem).

## Example 3: (Identification)

### Identification problem in a SEM

- Example: Supply and demand for milk

Supply of milk:  $q = \alpha_1 p + \beta_1 z_1 + u_1$

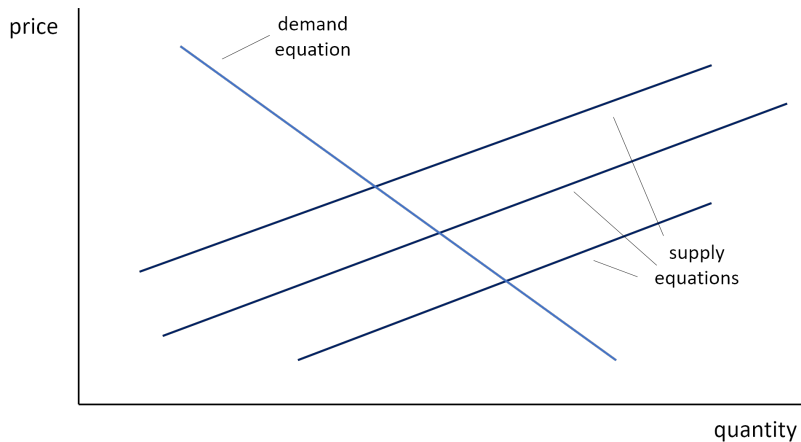
Demand for milk:  $q = \alpha_2 p + u_2$

- Supply of milk cannot be consistently estimated because we do not have (at least) one exogenous variable “available” to be used as instrument for  $p$  in the supply equation.
- Demand for milk can be consistently estimated because we can use exogenous variable  $z_1$  as instrument for  $p$  in the demand equation.



# SEM identification

- Illustration



# Identification conditions

Identification conditions for a sample 2-equation SEM  
(individual  $i$  subscripts omitted)

$$y_1 = \beta_{10} + \alpha_1 y_2 + \beta_{11} z_{11} + \beta_{12} z_{12} + \cdots + \beta_{1k} z_{1k} + u_1$$

$$y_2 = \beta_{20} + \alpha_2 y_1 + \beta_{21} z_{21} + \beta_{22} z_{22} + \cdots + \beta_{2k} z_{2k} + u_2$$

- Order condition (necessary): 1<sup>st</sup> equation is identified if at least one exogenous variable  $z$  is excluded from 1<sup>st</sup> equation (yet in the SEM).
- Rank condition (necessary and sufficient): 1<sup>st</sup> equation is identified if and only if the second equation includes at least one exogenous variable excluded from the first equation with a nonzero coefficient, so that it actually appears in the reduced form.
- For the second equation, the conditions are analogous.
- Some estimation approaches allow for identification through IVs not explicitly included in the SEM.

# Examples

## Example 4: (Identification)

Labor supply of married working women

Supply (workers):

$$\begin{aligned} hours = \alpha_1 \log(wage) + \beta_{10} + \beta_{11} educ + \beta_{12} age + \beta_{13} kidslt6 \\ + \beta_{14} nwifeinc + u_1 \end{aligned}$$

Demand (enterprises):

$$\log(wage) = \alpha_2 hours + \beta_{20} + \beta_{21} educ + \beta_{22} exper + \beta_{23} exper^2 + u_2$$

Order condition is fulfilled in both equations.

## Example 4: (Identification)

Labor supply of married working women contnd.

- Identification of the first equation (Supply). For the rank condition, either  $\beta_{22}$  or  $\beta_{23}$  non-zero population coefficient (in the second equation) is required – so that *exper*, *exper*<sup>2</sup> (or both) can be used in the reduced form.
- To evaluate the rank condition for supply equation, we estimate the reduced form for  $\log(\textit{wage})$  and test if we can reject the null hypothesis that coefficients for both coefficients for *exper* and *exper*<sup>2</sup> are zero.  
If  $H_0$  is rejected, the rank condition is fulfilled.
- We would do the evaluation of the rank condition for the demand equation analogically.

- We can consistently estimate identified equations with the 2SLS method.
- In the 1<sup>st</sup> stage, we regress each endogenous variable on all exogenous variables (“reduced forms”).
- In the 2<sup>nd</sup> stage we put into the structural equations instead of endogenous variables their predictions from the 1<sup>st</sup> stage and estimate with the OLS method.
- The reduced form can be always estimated (by OLS).
- In the 2<sup>nd</sup> stage, we cannot estimate unidentified structural equations.
- With some additional assumptions, we can use a more efficient estimation method than 2SLS: 3SLS.

# Systems with more than two equations

## Example 5: Keynesian macroeconomic model

$$C_t = \beta_0 + \beta_1(Y_t - T_t) + \beta_2 r_t + u_{t1}$$

$$I_t = \gamma_0 + \gamma_1 r_t + u_{t2}$$

$$Y_t \equiv C_t + I_t + G_t$$

Endogenous:  $C_t, I_t, Y_t$

Exogenous:  $T_t, G_t, r_t$

- Order condition for identification is the same as for two equations systems, rank condition is more complicated.
- There exist complicated models based on macroeconomic time series. There is a lot of problems with these models: series are usually not weakly dependent, it is difficult to find enough exogenous variables as instruments. Question is, if any macroeconomic variables are exogenous at all.

# Identification in SEMs with more than two equations

$y_i = X_i\beta + u_i$  is the  $i$ -th equation of a SEM.

$K$  - # of exogenous/predetermined variables in the SEM,

$K_i$  - # of  $K$  in the  $i$ -th equation,

$G_i$  - # of endogenous variables in the  $i$ -th equation.

**Order condition** for the  $i$ -th equation:

necessary, not sufficient condition for identification

$$K - K_i \geq G_i - 1$$

Condition evaluates as:

- = Equation  $i$  is just-identified,
- > Equation  $i$  is over-identified,
- < Equation  $i$  is not identified,  
structural equation  $i$  cannot be estimated by 2SLS/IVR.

# Identification in SEMs with more than two equations

Rank condition: based on matrix algebra & IV estimator

Consider IVR for an identified  $i$ -th equation of SEM

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i$$

$\mathbf{X}_i$  is a  $(n \times k)$  matrix, includes the intercept column and all endogenous regressors of the  $i$ -th equation,

$\hat{\mathbf{X}}_i$  is a  $(n \times k)$  matrix, includes the intercept column.

Exogenous regressors are repeated from  $\mathbf{X}_i$ , endogenous are projected to the column space of  $\mathbf{Z}$ : a  $(n \times l)$  matrix of all exogenous variables in the SEM.

Single equation (limited information) estimator for each  $i$ -th equation:

- $\hat{\boldsymbol{\beta}}_{IVR} = \hat{\boldsymbol{\beta}}_{2SLS,i} = \left( \hat{\mathbf{X}}_i' \mathbf{X}_i \right)^{-1} \hat{\mathbf{X}}_i' \mathbf{y}$
- $\hat{\mathbf{X}}_i' (\mathbf{y} - \mathbf{X}_i \hat{\boldsymbol{\beta}}) = \mathbf{0}$  (GMM moment conditions)



# Identification in SEMs with more than two equations

Rank condition: based on matrix algebra & IV estimator (cont.)

$$\hat{\beta}_{IVR} = \left( \hat{\mathbf{X}}_i' \mathbf{X}_i \right)^{-1} \hat{\mathbf{X}}_i' \mathbf{y}$$

- **Order condition:** The necessary condition for the  $i$ -th equation to be identified is that the number of columns (exogenous variables of SEM) in  $\mathbf{Z}$  should be no less than the number of columns (explanatory variables) in  $\mathbf{X}_i$ .
- **Rank condition:** The necessary and sufficient condition for identification of the  $i$ -th equation is that  $\hat{\mathbf{X}}_i'$  has full column rank of  $\mathbf{X}_i$ .  
...ensures the existence of  $\left( \hat{\mathbf{X}}_i' \mathbf{X}_i \right)^{-1}$ .

# Identification in SEMs with more than two equations

## Identification: recap & final remarks

- Reduced form equations can always be estimated.
- Structural equations can be estimated (IV/2SLS) only if identified: i.e. if rank condition is met.
- With SW, checking rank condition for  $\left(\hat{\mathbf{X}}_i' \mathbf{X}_i\right)^{-1}$  is easy for finite datasets.
- Asymptotic identification may be “tricky”:  
because some columns in  $\mathbf{X}_i$  are endogenous,  
 $\text{plim } n^{-1} \hat{\mathbf{X}}_i' \mathbf{X}_i$   
depends on the parameters of the DGP.  
...see Davidson-MacKinnon (2009) Econometric theory and methods