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

Advanced Econometrics 4EK608

Vysoká škola ekonomická v Praze

Outline

1. Instrumental variables
2. Two stage least squares
3. IV Tests: introduction
   - Durbin-Wu-Hausman (endogeneity in regressors)
   - Weak instruments Test
   - Sargan (exogeneity in IVs, over-identification only)
   - IV Tests: example

Instrumental variables

- IVs help to solve the endogeneity problem.
- Endogeneity exists in social sciences and economics everywhere.
    - Many important variables cannot be measured and often are correlated with observed explanatory variables.
    - Endogeneity can be caused by measurement errors.
    - It is always present in simultaneous equations models.
- In this case, estimators are biased and inconsistent.

Instrumental variables

- Endogeneity can sometimes be ignored, e.g. if the estimates are coupled with the direction of the biases for key parameters and if we can draw some useful conclusion. (job training effect on wages: attenuated by self-selection)
- Endogeneity can sometimes be solved
  - with the use of proxy variables
  - with panel data methods for models with time-invariant unobserved effects (FD, FE, RE estimators)
  - with instrumental variables

Instrumental variables

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

Definition of instrumental variables

- They are not in the regression (do not have partial effect on the dependent variable after controlling for $\boldsymbol{x}$ regressors and omitted variables.
- They are correlated (positively or negatively) with the endogenous variable – instrument is relevant; can be tested
- They are not correlated with the error term (that is why IQ is not a good IV); often, this cannot be tested
- Possible instrumental variables: father's education, mother's education, number of siblings, school proximity, month of birth (children born after August usually start school one year later)

Instrumental variables

Consistency of the OLS estimator in simple regression

$$y_i = \beta_0 + \beta_1 x_i + u_i$$
$$\mathrm{cov}(x_i, u_i) = 0 \quad \text{exogeneity assumption}$$

$$\mathrm{cov}(x_i, y_i - \beta_0 - \beta_1 x_i) = 0 \Leftrightarrow \mathrm{cov}(x_i, y_i) - \beta_1 \mathrm{var}(x_i) = 0,$$

$$\beta_1 = \mathrm{cov}(x_i, y_i)/\mathrm{var}(x_i) \Rightarrow \hat{\beta}_1 = \widehat{\mathrm{cov}}(x_i, y_i)/\widehat{\mathrm{var}}(x_i)$$

This holds as long as the data are such that sample variances and covariances converge to their theoretical counterparts as n grows. OLS will be consistent if, and only if, exogeneity holds.

Instrumental variables

IV estimator in simple regression

We assume instrumental variable z exists:

$$\text{cov}(y_i|\boldsymbol{x}_i, z_i) = 0 \qquad \text{cov}(z_i, u_i) = 0 \qquad \text{cov}(x_i, u_i) \neq 0$$

$$\text{cov}(z_i, y_i - \beta_0 - \beta_1 x_i) = 0 \Leftrightarrow \text{cov}(z_i, y_i) - \beta_1 \text{cov}(z_i, x_i) = 0$$

$$\beta_1 = \text{cov}(z_i, y_i)/\text{cov}(z_i, x_i) \Rightarrow \hat{\beta}_{1,IV} = \widehat{\text{cov}}(z_i, y_i)/\widehat{\text{cov}}(z_i, x_i)$$

$$\hat{\beta}_{1,OLS} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \; ; \qquad \hat{\beta}_{1,IV} = \frac{\sum_{i=1}^{n}(z_i - \overline{z})(y_i - \overline{y})}{\sum_{i=1}^{n}(z_i - \overline{z})(x_i - \overline{x})}$$

Instrumental variables

Statistical inference with IV estimator (SLRM)

- In large samples, IV estimator has approximately normal distribution.
- For calculation of standard errors, we usually need assumption of homoskedasticity conditional on the instrumental variable.
- Asymptotic variance of the IV estimator is always higher than of the OLS estimator.

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

Instrumental variables

Statistical inference with IV estimator (SLRM)

- 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.

Instrumental variables

Statistical inference with IV estimator (SLRM)

- If (small) correlation between u and instrument is possible, the 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

Coefficient of determination after IV estimation

- R^2 can be negative; SSR can be higher than SST.
- It does not have natural interpretation and any importance/relevance when IV method is used.
- IV method is for estimation of the ceteris paribus effect, not for maximization of the coefficient of determination (for forecasting needs).

Instrumental variables

IV estimation in multiple regression:

- Structural equation (as in SEMs)
  $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \cdots + \beta_k z_{k-1} + u$
- Reduced form for y_2 – endogenous variable as function of all exogenous variables (including IVs)
  $y_2 = \pi_0 + \pi_1 z_1 + \cdots + \pi_{k-1} z_{k-1} + \pi_k z_k + v$

z_k is some exogenous variable, excluded from structural equation (order condition for identification of the structural equation),

z_k is an instrumental variable for y_2, its coefficient must not be zero (rank condition for identification for this model) in the reduced form equation.

Instrumental variables

Calculating IV estimates in multiple regression

Exogeneity conditions:

$$\text{cov}(z_j, u) = 0, \quad j = 1, \ldots, k \quad \mathrm{E}(u) = 0$$

We use their sample analogs:

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

$$n^{-1} \sum_{i=1}^{n} z_{ij} \cdot \hat{u}_i = \widehat{\text{cov}}(z_j, \hat{u}) = 0, \quad j = 1, \ldots, k$$

$k + 1$ equations for $k + 1$ estimated parameters

Instrumental variables

Conditions for z_k

- It is excluded from the estimated structural equation
- It must be correlated with y_2
- It must not be correlated with u

We can have more instrumental variables, all must fulfill the conditions above (exclusion restrictions).

The best IV is some linear combination of the vector $\boldsymbol{z}_i$; one that is most correlated with y_2. It is given by the reduced form.

Instrumental variables

- In the reduced form, there are both original exogenous variables (from the structural equation) and excluded exogenous variables.

- In the reduced form, at least some coefficient for the excluded variables must be different from zero, otherwise we would get perfect collinearity between the instrumental variable given by the reduced form and original exogenous variables.
  - In other words, rank condition for identification would not be fulfilled.

Two stage least squares

IV estimator is equivalent to the following procedure:

$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + ... + \beta_k z_{k-1} + u$

1^{st} stage: reduced form regression

$$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \cdots + \hat{\pi}_{k-1} z_{k-1} + \hat{\pi}_k z_k$$

2^{nd} stage:

$$y_1 = \beta_0 + \beta_1 \hat{y}_2 + \beta_2 z_1 + \cdots + \beta_k z_{k-1} + \varepsilon$$

In the 2^{nd} stage, all variables are exogenous, because y_2 was replaced by its prediction which is dependent on exogenous information only.
Alternative description: In the second stage of 2SLS, y_2 is cleared from its endogenous part (which is correlated with error).

Two stage least squares

2SLS properties

- The standard errors from the OLS second stage regression are wrong. However, it is not difficult to compute correct standard errors (and software gives it 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.
- The 2SLS estimation can also be used if there is more than one endogenous variable. Conditions for identification must be fulfilled.

Two stage least squares

Conditions for identification:

- **Order condition:** We need at least as many excluded exogenous variables as there are included endogenous explanatory variables in the structural equation.

  This is a necessary condition for identification.

- **Rank condition:** We touched it before;

  This is a necessary and sufficient condition for identification.

Two stage least squares

Using 2SLS/IV as a solution to errors-in-variables:

- If a second measurement of the mis-measured variable is available, this can be used as an instrumental variable for the mis-measured variable

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

Two stage least squares

Statistical properties of the 2SLS/IV estimator

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

IV Tests: introduction

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

- s.e.$(\hat{\beta}_{1,IV})$ > s.e.$(\hat{\beta}_{1,OLS})$
- $\hat{\beta}_{1,IV}$ is always 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
  (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 tests & specifications for IV-tests exist, often focusing on the distribution of the difference between IVR and OLS estimators $(\hat{\boldsymbol{\beta}}_{\boldsymbol{IV}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{OLS}})$ under the corresponding H_0.

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).

Durbin-Wu-Hausman endogeneity test

$$y_{i1} = \beta_0 + \beta_1 y_{i2} + \beta_2 x_{i1} + \beta_3 \hat{\varepsilon} + u_i,$$

DWH test explanation:

If z_j are proper instruments (uncorrelated with u), then y_2 is endogenous (correlated with u) if and only if ε is correlated with u.

- y_2 in (1) is endogenous $\Leftrightarrow$ $\text{corr}(y_2, u) \neq 0$
- From (2), $y_2 = l.f.(instruments\ \boldsymbol{z}) + \varepsilon \;\Rightarrow\; y_2 = \hat{y}_2 + \hat{\varepsilon}$
- $\text{corr}(y_2, u) \neq 0 \;\wedge\; \text{corr}(\boldsymbol{z}, u) = 0 \;\Rightarrow\; \text{corr}(\varepsilon, u) \neq 0$
- y_1 is always correlated with u in (1).
- Hence, $\hat{\varepsilon}$ is significant in the regression, if y_2 is endogenous.
- $\forall z_j$ uncorrelated with u is essential for DWH to "work".

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

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!)