Lecture 7: Instrumental variables estimation II

Prof. dr. Wolter Hassink
Utrecht University School of Economics
w.h.j.hassink@uu.nl

These lecture notes are for your own use. It is not allowed to distribute the notes further by posting them on the Internet or on platforms without explicit and prior permission of the author.
© Utrecht University School of Economics 2024

Contents:

- Simultaneous equations
- 2SLS revisited
- Overidentification
- Test for overidentification: Hansen *J* test (Sargan test)
- Test for endogeneity: Hausman-Wu
- Example of tests for endogeneity and overidentification

Material:

Wooldridge:
Chapter 15: 15.5
Chapter 16: 16.1, 16.2, 16.3

These lecture notes are for your own use. It is not allowed to distribute the notes further by posting them on the Internet or on platforms without explicit and prior permission of the author.
© Utrecht University School of Economics 2024

Motivation: IV can be used to address the following:

- Omitted variable bias (previous lecture)
- Simultaneity bias (first part of this lecture)
- Examples of omitted variables: see previous lecture
- Examples of simultaneity bias: see below

Example 1 of simultaneity bias and 2SLS

IZA DP No. 12479: Open Labor Markets and Firms' Substitution between Training Apprentices and Hiring Workers
Manuel Aepli, Andreas Kuhn
revised version published in: Labour Economics, 2021, 70, 101979

In this paper, we study whether Swiss employers substitute between training apprentices and hiring cross-border workers. Because both training apprentices and hiring skilled workers are costly for firms, we hypothesize that (easier) access to cross-border workers will lead some employers to substitute away from training their own workers. **We account for potential endogeneity issues by instrumenting a firm's share of cross-border workers using a firm's distance to the national border and therefore its possibility to fall back on cross-border workers to satisfy its labor demand.** We find that both OLS and 2SLS estimates are negative across a wide range of alternative specifications, suggesting that firms substitute between training and hiring workers when the supply of skilled workers is higher. Our preferred 2SLS estimate implies that the increase in firms' share of crossborder workers within our observation period, from 1995 to 2008, led to about 3,500 fewer apprenticeship positions (equal to about 2% of the total number of apprentice positions).

Apprentice = F(*Cross-border workers*)

Instrument for Cross-border workers: *Distance to the national border*

Example 2 of simultaneity bias and 2SLS

Yin, Z., Gong, X., Guo, P., & Wu, T. (2019). What drives entrepreneurship in digital economy? Evidence from China. *Economic Modelling*, *82*, 66-73.

Using data collected in the 2017 China Household Finance Survey (CHFS), we study the impact of mobile payment on the likelihood of household entrepreneurship. In the empirical analysis, we use two-stage least squares (2SLS) regression to address the endogeneity of mobile payment. The study finds that mobile payment significantly increases the likelihood of household entrepreneurship. The mechanism could be that the mobile payment: 1) makes users more risk seeking; 2) enriches social networks; 3) provides an additional lending channel.

The key identification requirement for this research is that mobile payment in eq. (1) and eq. (2) may be endogenous due to omitted variables and reverse causality. When households operate industrial or commercial projects, in order to improve the convenience of collection, entrepreneurial households may start to use mobile payment. Therefore, the reverse causal relationship between entrepreneurship and mobile payment cannot be ignored. In addition, the use of mobile payment may be affected by factors such as the ability of individuals to accept new things and local customs, which are unobserved.

Therefore, in this paper, the Instrumental Variable method is adopted to address the endogeneity problem. Referring to previous studies (Yin et al., 2019), in this paper, the ownership of smartphones is used as IV. **The smartphone is an important carrier of mobile payment, and therefore, whether the household uses mobile payment is correlated to whether the household owns a smartphone, but the smartphone does not have a direct impact on the entrepreneurial decision of the household.** Therefore, this IV is valid.

Model:

Entrepreneurship = F(*Mobile Payment*)

Instrument for Mobile Payment: *Ownership of smartphone*

Example 3 of Simultaneity bias and 2SLS

Dogan, E., Madaleno, M., & Taskin, D. (2021). Which households are more energy vulnerable? Energy poverty and financial inclusion in Turkey. *Energy Economics*, *99*, 105306.

This study examines the effects of financial inclusion on energy poverty using the 2018 Turkish Household Budget and Consumption Expenditure Surveys. The study adopts three different measures of energy poverty and then analyzes the impact of financial inclusion proxied by a multidimensional index on energy poverty using different estimation strategies. **After addressing the endogeneity of financial inclusion by instrumenting financial inclusion with access to the nearest bank in a two-stage least squares framework,** the empirical results show that financial inclusion significantly alleviates energy poverty while its impact is higher for female-headed households. These findings are robust to Oster's (2019) bounds estimates that deal with omitted variable bias. The results also suggest that health and income are significant through which financial inclusion influences energy poverty. The findings thus point to the need for policies that promote financial inclusion as a way of alleviating energy poverty.

Model:

Energy poverty = F(*Financial Inclusion*)

Instrument for Financial Inclusion: *Access to nearest bank*

Simultaneous equations

These lecture notes are for your own use. It is not allowed to distribute the notes further by posting them on the Internet or on platforms without explicit and prior permission of the author.
© Utrecht University School of Economics 2024

Simultaneous equations

Aim: to introduce simultaneous equations

- Consider a market (e.g. labour market), which is described by supply and demand equations, as well as an equilibrium condition.
    - Demand equation: there is a negative relationship between price and quantity. **Demand shifters** influence the position of the demand equation.
    - Supply equation: there is a positive relationship between price and quantity. **Supply shifters** influence the position of the supply equation.
- Both the supply and the demand equations describe the relationship between price and quantity.
- The distinction between the structural and reduced-form equations is important.
    - **Structural equation:** relationship between two or more endogenous variables. E.g.

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

    - **Reduced-form equation:** RHS contains exogenous variables only (thus no endogenous variable).
- From an economic perspective, some important questions are:
    - Is it possible to estimate both equations jointly?
    - It is possible to disentangle supply from demand?
    - Are there any exogenous factors which affect supply (but not demand)?
    - Are there any exogenous factors which affect demand (but not supply)?

The structural model of the market

Aim: to introduce a structural model and to consider its implication for OLS.

- The first equation is a labour supply equation in which hours is explained by the wage:
  - $hours_s = \alpha_1 wage + \beta_1 z_1 + u_1$
    - $\alpha_1 > 0$ (the supply curve may be backward bending under some conditions, $\alpha_1 < 0$)
    - Hours $hours_s$ and *wage* are endogenous variables
    - The exogenous variable z_1 is an observed supply shifter.
    - The error term u_1 is an unobserved supply shifter.
- The second equation is a labour demand equation in which hours is explained by the wage:
  - $hours_d = \alpha_2 wage + \beta_2 z_2 + u_2$
    - $\alpha_2 < 0$
    - $hours_d$ and *wage* are endogenous variables
    - The exogenous variable z_2 is an observed demand shifter.
    - The error term u_2 is an unobserved demand shifter.

- Two statistical problems:
  - Endogeneity – *wage* is an endogenous RHS-variable, which may be correlated with the error term of the demand equation:
    - $Cov(wage, u_1) \neq 0$ and $Cov(wage, u_2) \neq 0$
  - The unobserved demand shifters may be correlated. $Cov(u_1, u_2) \neq 0$.

Simultaneity bias in OLS

Aim: to generalize a structural equation.

- The **structural equations** can be rewritten using two reduced-form equations.
- The explanation is simpler without intercepts in the equations.
- The equations are written as hours as a function of wage (equation (1)) and wage as another function of hours (equation (2)).

The supply equation: **structural equation** of y_1 as a function of y_2

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1 \tag{1}$$

“$hours = \alpha_1 wage + \beta_1 z_1 + u_1$”

The demand equation: **structural equation** of y_2 as a function of y_1

$$y_2 = \alpha_2 y_1 + \beta_2 z_2 + u_2 \tag{2}$$

“$wage = \alpha_2 hours + \beta_2 z_2 + u_2$”

Assumptions:

- z_1 is exogenous in the structural equation (1): $Cov(z_1, u_1) = 0$
- z_2 is exogenous in structural equation (2): $Cov(z_2, u_2) = 0$

Aim: to rewrite a structural equation in a reduced-form equation.

- The reduced-form equation of y_2 can be obtained by substituting equation (1) into the **structural equation** (2):

$$y_2 = \alpha_2 y_1 + \beta_2 z_2 + u_2 \qquad (2)$$

"$wage = \alpha_2 hours + \beta_2 z_2 + u_2$"

$$y_2 = \alpha_2 \underbrace{(\alpha_1 y_2 + \beta_1 z_1 + u_1)}_{=y_1} + \beta_2 z_2 + u_2$$

"$wage = \alpha_2 \underbrace{(\alpha_1 wage + \beta_1 z_1 + u_1)}_{=hours} + \beta_2 z_2 + u_2$"

$$(1-\alpha_1\alpha_2) y_2 = \alpha_2\beta_1 z_1 + \beta_2 z_2 + \alpha_2 u_1 + u_2 \qquad (3)$$

$$y_2 = \frac{\alpha_2\beta_1}{1-\alpha_1\alpha_2} z_1 + \frac{\beta_2}{1-\alpha_1\alpha_2} z_2 + \frac{\alpha_2 u_1 + u_2}{1-\alpha_1\alpha_2}$$

Assumption: $\alpha_1\alpha_2 \neq 1$. This ensures that the denominator is not allowed to be zero.
Note that in the reduced-form equation y_2 ("wage") depends on all exogenous variables (z_1, z_2) as well as the unobserved demand and supply shifters (u_1, u_2).

- Equation (3) can be rewritten as the **reduced-form equation** of y_2 (*wage*)

$$y_2 = \pi_{21} z_1 + \pi_{22} z_2 + v_2 \qquad (4)$$

With $\pi_{21} = \frac{\alpha_2\beta_1}{1-\alpha_1\alpha_2}$; $\pi_{22} = \frac{\beta_2}{1-\alpha_1\alpha_2}$; $v_2 = \frac{\alpha_2 u_1 + u_2}{1-\alpha_1\alpha_2}$

- y_2 (*wage*) is influenced by both z_1 and z_1 (all exogenous variables).

- Since u_1 and u_2 are each uncorrelated with z_1 and z_2, $v_2 (= \alpha_2 u_1 + u_2)$ is uncorrelated with z_1 and z_2. As a consequence, the parameters π_{21} and π_{22} of equation (4) can be estimated consistently with OLS.

Result: simultaneity bias of OLS in structural equation

Aim: to show that OLS on a structural equation leads to inconsistent parameter estimates.

It can be shown that OLS of equation (1) gives **inconsistent parameter estimates** if there is a non-zero correlation between the error term u_1 and y_2 in equation (1). In other words, the covariance between y_2 and u_1 is non-zero. This bias of OLS on equation (1) is referred to as **simultaneity bias**.

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1 \tag{1}$$

Proof:

$$Cov(y_2, u_1) = Cov(\underbrace{\pi_{21} z_1 + \pi_{22} z_2 + v_2}_{\text{equation (4)}}, u_1)$$

$$= \underbrace{Cov(\pi_{21} z_1, u_1)}_{=0} + \underbrace{Cov(\pi_{22} z_2, u_1)}_{=0} + Cov(v_2, u_1) =$$

$$= Cov(\underbrace{\frac{\alpha_2 u_1 + u_2}{1 - \alpha_1 \alpha_2}}_{= v_2\text{ ; see equation (4)}}, u_1)$$

$$= \frac{\alpha_2}{1 - \alpha_1 \alpha_2} Var(u_1) + \frac{1}{1 - \alpha_1 \alpha_2} Cov(u_1, u_2)$$

OLS gives inconsistent parameter estimates.

1) If the correlation between the error terms of both equations is zero ($Cov(u_1, u_2) = 0$) then $Cov(y_2, u_1) \neq 0$

2) If the correlation between the error terms of both equations is nonzero ($Cov(u_1, u_2) \neq 0$) then $Cov(y_2, u_1) \neq 0$

- Note that the motivation for the **simultaneity bias** differs from the motivation provided in lecture 6. In Chapter 15, the motivation for a nonzero correlation between y_2 and u_1 is **omitted variables** (e.g. ability in wage equation).

Identification and estimation of a structural equation

Aim: to introduce exclusion restrictions, which are required to identify and estimate structural equations

Start again with the **structural** equations (1) and (2)

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1 \quad (1)$$

$$y_2 = \alpha_2 y_1 + \beta_2 z_2 + u_2 \quad (2)$$

We have supply equation (1) and demand equation (2) in which:

q : quantity
p : price
z_1: exogenous variable

Supply: $p = \alpha_1 q + \beta_1 z_1 + u_1$ (5)
Demand: $q = \alpha_2 p + u_2$ (6)

Identification: we need to have specific assumptions to identify the parameters of equation (6) In this model, the parameters of the demand equation (6) can be identified. The **identifying assumption** is the following:

- The exogenous variable z_1 has an influence on the supply (5) but not on demand (6). It is assumed that:
  - $\beta_1 \neq 0$
  - equation (6) does NOT contain z_1. **(exclusion restriction)**

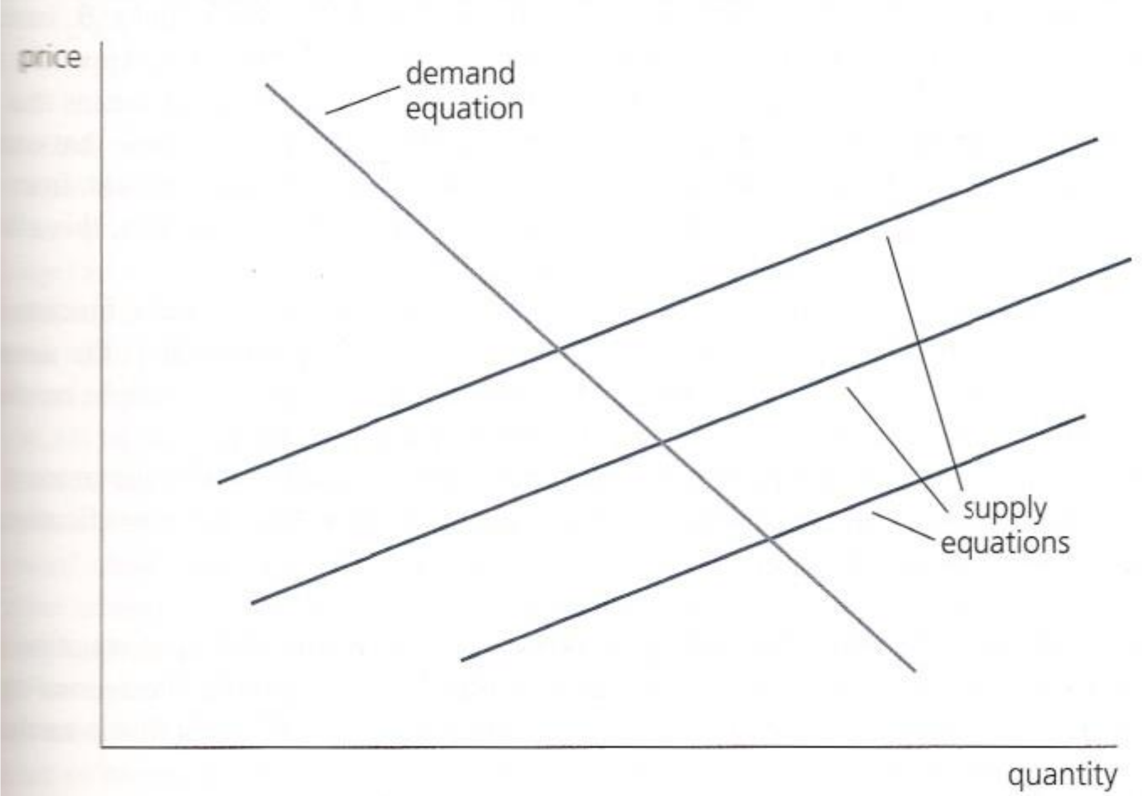
Implications

- The parameters of the supply equation (5) cannot be identified.
- The parameter of the demand equation (6) can be identified. The parameter α_2 gives the **causal** effect of p on q.

Supply: $p = \alpha_1 q + \beta_1 z_1 + u_1$ (5)

Demand: $q = \alpha_2 p + u_2$ (6)

Shifting supply equations trace out the demand equation. Each supply equation is drawn for a different value of the exogenous variable, z_1.



2SLS revisited

These lecture notes are for your own use. It is not allowed to distribute the notes further by posting them on the Internet or on platforms without explicit and prior permission of the author.
© Utrecht University School of Economics 2024

Relationship with 2SLS: Revision from lecture 6

- Consider the case of a **structural equation** with two RHS-variables.
  $$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1 \qquad (9)$$

- We have demonstrated above that because of **simultaneity bias** OLS cannot be applied for equation (9), because $Cov(y_2, u_1) \neq 0$

- We will identify the causal effect of y_2 on y_1 by 2SLS:

- **Assumption 1:** the variable z_2 (not included in equation (9)) has an effect on y_2 ($Cov(y_2, z_2) \neq 0$)
    - This corresponds to the relevance criterion for instrumental variables in the 2SLS-procedure (Chapter 15)

- **Assumption 2:** the instrument variable z_2 has **NO effect on** y_1
    - This is referred to as the **exclusion restriction**, and therefore z_2 is not included in equation (9).
    - This corresponds to the exogeneity criterion for the instrumental variables in the 2SLS-procedure

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1 \tag{9}$$

- The endogenous RHS-variable y_2 in equation (9) may be rewritten in terms of the exogenous variable z_1 and the instrumental variable z_2. This is the reduced-form equation for y_2, which only depends on exogenous variables z_1 and z_2:

$$y_2 = \pi_{20} + \pi_{21} z_1 + \pi_{22} z_2 + v_2 \qquad (10) \qquad \text{(16.14)}$$

with assumption $\pi_{22} \neq 0$ (**assumption 1** of above) and where it is assumed that the explanatory variables z_1 and z_2 are uncorrelated to the error term v_2. $Cov(z_1, v_2) = 0$, and $Cov(z_2, v_2) = 0$.

For the 2SLS-estimator, there are two stages:

- **First stage of 2SLS: the reduced-form equation of** y_2. Regress y_2 on z_1 and z_2 (equation (10)) with OLS and determine the fitted value of y_2, using the estimated parameters:

$$\hat{y}_2 = \hat{\pi}_{20} + \hat{\pi}_{21} z_1 + \hat{\pi}_{22} z_2$$

- **Second stage of 2SLS:** the structural-form equation of y_1. Regress the structural equation (9), in which the fitted value $\hat{y}_2$ is used, instead of its actual value:

$$y_1 \text{ on } \hat{y}_2 \text{ and } z_1 \tag{11}$$

- The second stage does not include z_2 (the **exclusion restriction**; **assumption 2** of above).

- Note that the *t*-values of (11) are wrong, but are corrected using standard 2SLS commands in software packages.

Example 1

Example 1: Application of 2SLS on estimation of structural model

We have 97 daily price and quantity observations of the Fulton Fish market in Manhattan. Suppose that we want to estimate the following demand function for fish:

$$ltotqty_t = \beta_0 + \beta_1 lavgprc_t + u_t$$

$$\text{“} y_1 = \beta_{10} + \alpha_1 y_2 + u_1 \text{”} \qquad (16.17)$$

- $avgprc_t$: average price of fish in period t
- $lavgprc_t$: $\log(avgprc_t)$
- $wave2_t$: measure of wave heights of the sea over the past two days
- $wave3_t$: measure of wave heights of the sea over the past three days
- $totqty_t$: quantity of fish sold.
- $ltotqty_t$: $\log(totqty_t)$
- t : time trend

In the equations we ignore the intercepts (it does not change the answer):

Supply: $$lavgprc_t = \alpha_2 ltotqty_t + \beta_2 wave2_t + u_{2t} \qquad (5')$$

$$\text{“} y_2 = \beta_{20} + \alpha_2 y_1 + \beta_2 z_2 + u_2 \text{”} \qquad (16.18)$$

Demand: $$ltotqty_t = \alpha_1 lavgprc_t + u_{1t} \qquad (6')$$

$$\text{“} y_1 = \beta_{10} + \alpha_1 y_2 + u_1 \text{”} \qquad (16.17)$$

Exclusion restriction for identification of α_1: We assume that $wave2_t$ has an influence on the price of fish ($\beta_2 \neq 0$), AND that there is no influence of $wave2_t$ on $ltotqty_t$ (this is referred to as an exclusion restriction).

- Note that the parameters α_2 and β_2 from the structural supply equation (5') cannot be identified.

- Parameter α_1 can be estimated by 2SLS on equation (6'), using $Wave2_t$ as an instrument for $lavgprc_t$. The first-stage equation of the 2SLS-procedure:

  $lavgprc_t = \pi_{21} wave2_t + v_{2t}$, where we assume that $\pi_{21} \neq 0$ (note that this is the identifying assumption of instrumental relevance of a 2SLS-procedure).

- In addition, we may substitute equation (5') in to equation (6'):
  $ltotqty_t = \alpha_1(\alpha_2 ltotqty_t + \beta_2 Wave2_t + u_{2t}) + u_{1t}$
  $(1 - \alpha_1\alpha_2) ltotqty_t = \alpha_1\beta_2 wave2_t + \alpha_1 u_{2t} + u_{1t}$

- Note that exclusion restriction for identification, $\beta_2 \neq 0$, is equivalent to the condition of the first-stage equation of 2SLS that $\pi_{21} \neq 0$.

Application: fish.dta

. ivreg ltotqty (lavgprc = wave2) t, first

```
First-stage regressions
-----------------------

      Source |       SS       df       MS              Number of obs =      97
-------------+------------------------------           F(  2,    94) =   15.31
       Model |  3.86072193     2  1.93036097           Prob > F      =  0.0000
    Residual |  11.8523883    94  .126089237           R-squared     =  0.2457
-------------+------------------------------           Adj R-squared =  0.2297
       Total |  15.7131102    96  .163678231           Root MSE      =  .35509

------------------------------------------------------------------------------
     lavgprc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |  -.0021491   .0013472    -1.60   0.114     -.004824    .0005258
       wave2 |    .097487    .021215     4.60   0.000     .0553641      .13961
       _cons |  -.6368629   .1467153    -4.34   0.000    -.9281695   -.3455563
------------------------------------------------------------------------------


Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =      97
-------------+------------------------------           F(  2,    94) =    2.21
       Model |  .830823147     2  .415411573           Prob > F      =  0.1157
    Residual |  55.3016103    94  .588315004           R-squared     =  0.0148
-------------+------------------------------           Adj R-squared = -0.0062
       Total |  56.1324335    96  .584712849           Root MSE      =  .76702

------------------------------------------------------------------------------
     ltotqty |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lavgprc |  -.9714214   .4700701    -2.07   0.042    -1.904757   -.0380861
           t |  -.0028012   .0033493    -0.84   0.405    -.0094513    .0038489
       _cons |   7.984152   .1588031    50.28   0.000     7.668845    8.299459
------------------------------------------------------------------------------
Instrumented:  lavgprc
Instruments:   t wave2
------------------------------------------------------------------------------
```

Overidentification

These lecture notes are for your own use. It is not allowed to distribute the notes further by posting them on the Internet or on platforms without explicit and prior permission of the author.
© Utrecht University School of Economics 2024

Overidentification: omitted variable bias (Section 15.5)

Aim: to show how overidentification may be used to obtain statistically significant t-statistics

The conclusion of the previous slides and of lecture 6 may be that IV (2SLS) is necessary if there is omitted variable bias (chapter 15) or simultaneity bias (chapter 16). In all of the examples, the number of instruments and endogenous RHS variables were equal (in most applications, 1 endogenous variable and 1 instrument). However, there may be more valid instruments than RHS-variables.

- An important feature of IV is that using more instrumental variables for 1 endogenous variable leads to lower standard errors (and higher *t*-value) of:
  - *educ* (example 2 below + estimates)
  - $(y_{it-1} - y_{it-2})$ (example 3 below + estimates)
  - *ltotqty* (example 4 below + estimates)
- Generally, higher *t*-values are more useful in empirical analysis.

This may lead to overidentification.

Next, we explain:

- Overidentification of instrumental variables to address omitted variable bias.
- Overidentification of instrumental variables to address simultaneity bias.

Why do we need to overidentify? Recall lecture 6

Aim: to show why IV leads to a larger standard error (smaller t-statistic) and what happens to this outcome if we add more instruments?

- Recall from Chapter 2 (Wooldridge) that the standard error of the OLS-estimator $\hat{\beta}_1$ is

$$Var(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\hat{\sigma}^2}{SST_x}$$

- It can be shown that the standard error of the IV-estimator $\hat{\beta}_1^{IV}$:

$$Var(\hat{\beta}_1^{IV}) = \frac{\hat{\sigma}^2}{SST_x R_{x,z}^2} = \frac{Var(\hat{\beta}_1)}{R_{x,z}^2}$$

Where $R_{x,z}^2$ is the R^2 from the regression of *x* on *z* (equation (19)), so that

$$Var(\hat{\beta}_1^{IV}) = \frac{Var(\hat{\beta}_1)}{R_{x,z}^2} > Var(\hat{\beta}_1)$$

- If $R_{x,z}^2$ is small, *z* will be a weak instrument and the *t*-value for $\hat{\beta}_1^{IV}$ will be relatively small.
- The instrument may be strengthened by adding a second (or even a third) instrument to the regression of *x* on *z* (and thus on the other instrumental variables).
- Consequence: *t*-values will become larger.
- However, adding more instrumental variables leads to overidentification.
- Below we will motivate and explain the statistical consequences of overidentification.

Examples of Overidentification? Omitted variable bias (I)

Aim: to compare exact identification with overidentification

- **Example 2**: Wage equation with endogenous education.

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

In lecture 6, we saw that *educf* "father's education" may satisfy the criterion of instrument exogeneity. This result relies on the assumption that the father's education is not correlated with their child's ability. The same reasoning may be used to argue that *educm* (education of mother) may satisfy the criterion of instrument exogeneity.

. ivregress 2sls lwage (educ = fatheduc) exper

```
Instrumental variables (2SLS) regression            Number of obs =      2220
                                                    Wald chi2(2)  =    143.35
                                                    Prob > chi2   =    0.0000
                                                    R-squared     =    0.1020
                                                    Root MSE      =    .41657
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1479617   .0129567    11.42   0.000      .122567    .1733564
       exper |   .0688418   .0057758    11.92   0.000     .0575213    .0801622
       _cons |   3.698219    .221646    16.69   0.000     3.263801    4.132637
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   exper fatheduc
```

- Formally, there is no need to add further instrumental variables, since *educ* has large *t*-statistic. However, for the sake of this exercise, let's add another instrumental variable. Consequence of overidentification: the standard error of *educ* becomes somewhat smaller.

ivregress 2sls lwage (educ = fatheduc motheduc) exper

```
Instrumental variables (2SLS) regression            Number of obs =      2220
                                                    Wald chi2(2)  =    178.20
                                                    Prob > chi2   =    0.0000
                                                    R-squared     =    0.0932
                                                    Root MSE      =     .4186
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    .151402   .0117729    12.86   0.000     .1283275    .1744765
       exper |   .0702544   .0053361    13.17   0.000     .0597958     .080713
       _cons |   3.639622   .2015898    18.05   0.000     3.244513     4.03473
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   exper fatheduc motheduc
```

Examples of Overidentification? Omitted variable bias (II)

- **Example 3:** Panel data regression with a lagged dependent variable, applied to first difference estimator.

$(y_{it} - y_{it-1}) = \gamma(y_{it-1} - y_{it-2}) + (u_{it} - u_{it-1}) \quad i = 1,..., N; t = 3,..., T$

In lecture 6, we saw that y_{it-2} might be a valid instrument. One could also argue that additional valid instruments are $y_{it-3}, y_{it-4},..., y_{i1}$ as well as $y_{it-2} - y_{it-3}, y_{it-3} - y_{it-4},..., y_{i2} - y_{i1}$. All of these instruments may be correlated with $(y_{it-1} - y_{it-2})$ but they remain uncorrelated with $(u_{it} - u_{it-1})$.

- Consequence of overidentification: the standard errors of estimated regression parameters will become smaller by allowing for overidentification.

Examples of overidentification? simultaneity bias

- Next, we consider overidentification if IV is needed to correct for simultaneity bias.
- It may be the case that there are many variables that are suitable for the exclusion restriction. E.g. there are two variables z_1 and z_2, that do affect y_1 but that do not affect y_2. Thus,

Structural equation of y_1:

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + \beta_2 z_2 + u_1 \tag{13}$$

Structural equation of y_2:

$$y_2 = \alpha_2 y_1 + \beta_3 z_3 + u_2 \tag{14}$$

- The reduced-form equation of y_2 can be obtained by substituting equation (13) in (14):

$$y_2 = \pi_{21} z_1 + \pi_{22} z_2 + \pi_{23} z_3 + v_2$$

Example 4:

Supply: $lavgprc_t = \alpha_2 ltotqty_t + \beta_2 Wave2_t + \beta_3 Wave3_t + u_{2t}$ (5'')

Demand: $ltotqty_t = \alpha_1 lavgprc_t + u_{1t}$ (6'')

- There are two instruments (*Wave*2 and *Wave*3) in the first equation and one endogenous variable *ltotqty* in the second equation, so that α_1 can be identified by both (only if overidentification does not turn out to be a problem - see tests for overidentification on slides below). Note again that the structural parameter α_2 cannot be identified.

How to estimate with overidentification?
2SLS (Section 15.5)

2SLS: No problem at all to calculate $\hat{\beta}_0^{IV}$ and $\hat{\beta}_1^{IV}$. E.g. Structural equation of y_1:

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + \beta_2 z_2 + u_1. \tag{15}$$

Structural equation of y_2:

$$y_2 = \alpha_2 y_1 + \beta_3 z_3 + u_2 \tag{16}$$

Overidentification: there are **two instrumental variables** z_1 and z_2), although there is **one endogenous variable** only (y_1).

The 2SLS estimate of α_2 in case of overidentification:

Stage 1: $y_1 = \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2$ gives $\hat{y}_1 = \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 z_3$
Stage 2: regression of y_2 on $\hat{y}_1$ and z_3.

Why is overidentification a statistical problem? (I)

We make a comparison with OLS. OLS leads to a zero correlation between residual and explanatory variables

Example 5 (data: Card.dta):

First, we dropped the cases with missing variables in *fatheduc* or *motheduc*

. drop if fatheduc == .

(690 observations deleted)

. drop if motheduc == .

(100 observations deleted)

. reg lwage educ exper

```
      Source |       SS       df       MS              Number of obs =    2220
-------------+------------------------------           F(  2,  2217) =  234.24
       Model |  74.8394528     2  37.4197264           Prob > F      =  0.0000
    Residual |  354.160051  2217  .159747429           R-squared     =  0.1745
-------------+------------------------------           Adj R-squared =  0.1737
       Total |  428.999503  2219  .193330105           Root MSE      =  .39968
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0894429   .0041958    21.32   0.000     .0812149     .097671
       exper |   .0448139   .0027611    16.23   0.000     .0393993    .0502286
       _cons |   4.694929   .0741562    63.31   0.000     4.549506    4.840352
------------------------------------------------------------------------------
```

. predict uhat, resid

. reg uhat educ exper

```
      Source |       SS       df       MS              Number of obs =    2220
-------------+------------------------------           F(  2,  2217) =    0.00
       Model |  6.8212e-13     2  3.4106e-13           Prob > F      =  1.0000
    Residual |  354.160052  2217   .15974743           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared = -0.0009
       Total |  354.160052  2219  .159603448           Root MSE      =  .39968
------------------------------------------------------------------------------
        uhat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.45e-10   .0041958     0.00   1.000     -.008228     .008228
       exper |   1.05e-10   .0027611     0.00   1.000    -.0054147    .0054147
       _cons |  -3.45e-09   .0741562    -0.00   1.000    -.1454229    .1454229
------------------------------------------------------------------------------
```

. corr uhat educ exper

```
(obs=2220)
             |     uhat     educ    exper
-------------+---------------------------
        uhat |   1.0000
        educ |   0.0000   1.0000
       exper |   0.0000  -0.6239   1.0000
```

- Conclusion with OLS: residuals are uncorrelated with all of the explanatory variables.

Why is overidentification a statistical problem? (II)

Example 6:

- Now we estimate a 2SLS regression, for which there are as many instruments as endogenous variables. So there is no overidentification.

. ivregress 2sls lwage (educ = fatheduc) exper

```
Instrumental variables (2SLS) regression               Number of obs =    2220
                                                       Wald chi2(2)  =  143.35
                                                       Prob > chi2   =  0.0000
                                                       R-squared     =  0.1020
                                                       Root MSE      =  .41657
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1479617   .0129567    11.42   0.000      .122567    .1733564
       exper |   .0688418   .0057758    11.92   0.000     .0575213    .0801622
       _cons |   3.698219    .221646    16.69   0.000     3.263801    4.132637
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   exper fatheduc
```

. predict uhat, resid
. reg uhat fatheduc exper

```
      Source |       SS       df       MS              Number of obs =    2220
-------------+------------------------------           F(  2,  2217) =    0.00
       Model |           0     2           0           Prob > F      =  1.0000
    Residual |  385.234439  2217  .173763843           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared = -0.0009
       Total |  385.234439  2219  .173607228           Root MSE      =  .41685
------------------------------------------------------------------------------
        uhat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    fatheduc |   1.57e-11   .0025619     0.00   1.000    -.0050241    .0050241
       exper |   1.02e-10   .0024093     0.00   1.000    -.0047246    .0047246
       _cons |  -1.06e-09    .038892    -0.00   1.000    -.0762685    .0762685
------------------------------------------------------------------------------
```

. corr uhat educ fatheduc exper

```
 (obs=2220)
             |     uhat     educ fatheduc    exper
-------------+------------------------------------
        uhat |   1.0000
        educ |  -0.2219   1.0000
    fatheduc |  -0.0000   0.4692   1.0000
       exper |   0.0000  -0.6239  -0.3571   1.0000
```

- Conclusion: the residuals of 2SLS are uncorrelated with instruments (*fatheduc*) and the exogenous variable (*exper*) in case of exact identification.

Why is overidentification a statistical problem? (III)

Example 7:

For the test for overidentification, compute the residual after 2SLS.

```
. ivregress 2sls lwage (educ =  fatheduc  motheduc) exper

Instrumental variables (2SLS) regression              Number of obs =    2220
                                                      Wald chi2(2)  =  178.20
                                                      Prob > chi2   =  0.0000
                                                      R-squared     =  0.0932
                                                      Root MSE      =   .4186

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    .151402   .0117729    12.86   0.000     .1283275    .1744765
       exper |   .0702544   .0053361    13.17   0.000     .0597958     .080713
       _cons |   3.639622   .2015898    18.05   0.000     3.244513     4.03473
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   exper fatheduc motheduc
```

. predict uhat, resid

. corr uhat educ fatheduc motheduc exper

```
(obs=2220)

             |     uhat     educ fatheduc motheduc    exper
-------------+---------------------------------------------
        uhat |   1.0000
        educ |  -0.2339   1.0000
    fatheduc |  -0.0052   0.4692   1.0000
    motheduc |   0.0060   0.4396   0.6315   1.0000
       exper |  -0.0000  -0.6239  -0.3571  -0.3163   1.0000
```

- Conclusion: the residual (*uhat*) is correlated with instrumental variables *fatheduc* and *motheduc* (correlations are -0.0052 and 0.0060, respectively.

Is overidentification a serious statistical problem? (IV)

Aim: To explain when overidentification is no problem for inconsistency of parameter estimates.

- Consider 3 linear equations in X-Y which must solved simultaneously.

  e.g.: Y+X = 10
  -Y + 3X = 5
  2Y +3X = 2

- Two equations will intersect, but the third equation does not intersect at the same point

- If the third line is “close” to the intersection of the other two lines, overidentification will not be a problem. “How close is close?” This is where the statistical test is about.

- Overidentification will be a problem if the third line is not “close” to the intersection of the other two lines.

Test for overidentification: Hansen *J* test (Sargan test)

These lecture notes are for your own use. It is not allowed to distribute the notes further by posting them on the Internet or on platforms without explicit and prior permission of the author.
© Utrecht University School of Economics 2024

Test for overidentification: Hansen *J* test (Sargan test)

Aim: to outline the Hansen J (or Sargan) test of overidentification

- Please remember that after OLS, the residual $\hat{u}$ must be uncorrelated with all of the explanatory variables. And $\hat{u}$ is on average zero. In addition, the regression of $\hat{u}$ on all explanatory variables gives an R^2 of zero.

- The R^2 will be zero for 2SLS if the there is no overidentification (# endogenous RHS variables = # instruments)

- It is possible to construct the following hypothesis test
    - H_0: no overidentification
    - H_1: overidentification.

- On average, if there is no overidentification, the correlation between residual of the second stage equation and the explanatory variables should be small after 2SLS (example 7). Without overidentification (# instruments = # endogenous variables), the correlation would be zero (example 6). This can be tested in the following way.

- The test on overidentification is illustrated using the following model:

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + u\,. \tag{17}$$

In equation (17), the instruments are z_2 and z_3 for y_2 (endogenous RHS-variable).

Step 1: determine $\hat{u}$ after 2SLS.

Step 2: regress $\hat{u}$ on all exogenous variables z_1 + instrumental variables, z_2 and z_3 (the exogenous regressors of first-stage equation).

Step 3: calculate R^2 of regression of Step 2. The R^2 will be small (though not exactly equal to zero as in the case of exact identification). However, if H_0 is true, $n \cdot R^2$ follows a Chi-square distribution with q degrees of freedom. Thus, $n \cdot R^2 \sim \chi^2_{(q)}$.

q: number of instruments – number of endogenous variables.

Thus, for equation (15), we have $q = 1$.

We have indication of overidentification if $n \cdot R^2 > \chi^2_{(q;\alpha)}$, where $\chi^2_{(q;\alpha)}$ is the α percent critical value of the Chi-square distribution.

- What conclusion can be drawn if H_0 (no overidentification) is rejected? We cannot trust the 2SLS or GMM-estimates. They do not solve the bias problems we already knew existed when estimating by OLS.

- Solution? Omit one of the instruments, but this results in lower *t*-statistics.

- If we do not reject H_0 (no overidentification) there may be overidentification (more instruments than endogenous variables), however it is not a statistical problem. The *t*-values of the endogenous variables will be larger.

Sargan-test for overidentification

We continu with the residual of the 2SLS estimator of Example 7. We run a regression of residual on all exogenous variables; instruments included)

. reg uhat fatheduc motheduc exper

```
      Source |       SS       df       MS              Number of obs =    2220
-------------+------------------------------           F(  3,  2216) =    0.13
       Model |  .067087737     3  .022362579           Prob > F      =  0.9439
    Residual |  388.928517  2216  .175509259           R-squared     =  0.0002
-------------+------------------------------           Adj R-squared = -0.0012
       Total |  388.995605  2219   .17530221           Root MSE      =  .41894
------------------------------------------------------------------------------
        uhat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    fatheduc |  -.0017186   .0031752    -0.54   0.588    -.0079453    .0045081
    motheduc |    .002119   .0037904     0.56   0.576    -.0053141    .0095521
       exper |  -.0000573   .0024406    -0.02   0.981    -.0048433    .0047288
       _cons |  -.0046828   .0458766    -0.10   0.919    -.0946485    .0852828
------------------------------------------------------------------------------
```

```
. display e(r2)       (Stata command to show R2)
```

.00017246

$2220 \cdot 0.00017246 = 0.38286$ (computed manually)

Under H_0, the statistic is Chi-squared distributed with one degree of freedom

Critical value of $\chi^2_{(1;0.05)}$: 3.84. (see Table G.4 of Wooldridge; 1 overidentifying restriction).

Alternative Stata command

. estat overid

```
  Tests of overidentifying restrictions:

  Sargan (score) chi2(1) =   .38287  (p = 0.5361)
  Basmann chi2(1)        =  .382246  (p = 0.5364)
```

- Conclusion: do not reject the null hypothesis (no overidentification), because *p*-value (0.38) is above 0.05. Thus, overidentification is not a statistical problem.

Test for endogeneity: Hausman-Wu

These lecture notes are for your own use. It is not allowed to distribute the notes further by posting them on the Internet or on platforms without explicit and prior permission of the author.
© Utrecht University School of Economics 2024

Introduction to a test of endogeneity

- Consider a structural wage equation (with experience as exogenous variable) and education as an endogenous RHS-variable. There is omitted variable bias (ability bias), which is addressed by 2SLS, when *ability* is instrumented using father's education.

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + u_1$$

- After the first-stage regression of the RHS-endogenous variable on the instrument and the exogenous variable,
  $educ = \pi_0 + \pi_1 exper + \pi_2 feduc + v_2$
  we may keep the residuals:
  . predict uhat, resid
- We add the residual $\hat{v}_2$ to the structural equation

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \delta_1 \hat{v}_2 + u_1 \qquad (18)$$

- and estimate equation (18) with OLS.
- Interestingly, the $\hat{\beta}_1$ and $\hat{\beta}_2$ are equal to those of the 2SLS (ivreg) procedure (see below) but only the *t*-values of ivregress 2SLS are correct.

Example of test of endogeneity: Card.dta

Example 8:

- We run the first-stage equation of 2SLS procedure:

Regression of endogenous RHS variable on instruments and exogenous variables:

. reg educ fatheduc motheduc exper

```
      Source |       SS       df       MS              Number of obs =    2220
-------------+------------------------------           F(  3,  2216) =  666.67
       Model |  7048.85948     3  2349.61983           Prob > F      =  0.0000
    Residual |  7810.06439  2216  3.52439729           R-squared     =  0.4744
-------------+------------------------------           Adj R-squared =  0.4737
       Total |  14858.9239  2219  6.69622527           Root MSE      =  1.8773

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    fatheduc |   .1302532   .0142287     9.15   0.000     .1023502    .1581563
    motheduc |   .1373783   .0169854     8.09   0.000     .1040692    .1706873
       exper |  -.3331567   .0109367   -30.46   0.000    -.3546039   -.3117096
       _cons |      13.62   .2055813    66.25   0.000     13.21685    14.02316
------------------------------------------------------------------------------
```

. predict vhat, resid

- **Regression of structural model, including the residual of first stage:**

. reg lwage educ exper vhat

```
      Source |       SS       df       MS              Number of obs =    2220
-------------+------------------------------           F(  3,  2216) =  170.57
       Model |  80.4783086     3  26.8261029           Prob > F      =  0.0000
    Residual |  348.521195  2216  .157274907           R-squared     =  0.1876
-------------+------------------------------           Adj R-squared =  0.1865
       Total |  428.999503  2219  .193330105           Root MSE      =  .39658
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    .151402   .0111537    13.57   0.000     .1295293    .1732748
       exper |   .0702544   .0050554    13.90   0.000     .0603405    .0801683
        vhat |  -.0719885   .0120246    -5.99   0.000    -.0955651   -.0484118
       _cons |   3.639622   .1909864    19.06   0.000     3.265091    4.014153
------------------------------------------------------------------------------
```

- Compare this with the 2SLS outcome - the estimated parameters are similar, but the standard errors are slightly different.
- *vhat* is statistically significant (-5.99)

. ivregress 2sls lwage (educ = fatheduc motheduc) exper

```
Instrumental variables (2SLS) regression               Number of obs =    2220
                                                       Wald chi2(2)  =  178.20
                                                       Prob > chi2   =  0.0000
                                                       R-squared     =  0.0932
                                                       Root MSE      =   .4186

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    .151402   .0117729    12.86   0.000     .1283275    .1744765
       exper |   .0702544   .0053361    13.17   0.000     .0597958     .080713
       _cons |   3.639622   .2015898    18.05   0.000     3.244513     4.03473
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   exper fatheduc motheduc
```

Alternative command in Stata:

. estat endogenous

```
  Tests of endogeneity
  Ho: variables are exogenous

  Durbin (score) chi2(1)          =  35.3463  (p = 0.0000)
  Wu-Hausman F(1,2216)            =  35.8535  (p = 0.0000)
```

Test for endogeneity (Section 15.5; (Hausman-Wu))

Model $y_1 = \alpha_1 y_2 + \beta_1 z_1 + u$ (19)

- This test is based on insights from the previous slide.
- It is possible to test for endogeneity of y_2 in structural model.
- Hypotheses: $H_0 : Cov(y_2, u) = 0$ (no endogeneity)

  $H_1 : Cov(y_2, u) \neq 0$ (endogeneity)

- Under H_0, OLS may be safely applied on equation (19) and gives $\hat{\alpha}_1^{OLS}$ and $\hat{\beta}_1^{OLS}$.
- Under H_1, 2SLS must be applied on equation (19), which is very similar to OLS on:

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + \delta_1 \hat{v}_2 + u, \quad (20)$$

Where $\hat{v}_2$ is the residual of the first-stage equation of the 2SLS-procedure:
$y_2 = \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2$ and $\hat{v}_2 = y_2 - \hat{y}_2 = y_2 - \hat{\pi}_1 z_1 - \hat{\pi}_2 z_2 - \hat{\pi}_3 z_3$

- If H_0 is true: $\delta_1 = 0$; If H_0 is not true: $\delta_1 \neq 0$. δ_1 refers to the parameter on $\hat{v}_2$ in equation (20). After OLS on equation (20), a *t*-test on δ_1 to test for endogeneity may be applied. There is an indication of endogeneity of y_2 if δ_1 in equation (20) is statistically different from zero.
- Hypotheses: $H_0 : \delta_1 = 0$ (no endogeneity)

  $H_1 : \delta_1 \neq 0$ (endogeneity)

Test for endogeneity in Stata (Hausman-Wu):

Example 9:

. reg educ feduc meduc exper **(first stage of 2SLS procedure)**

. predict uhat, resid **(residual after first-stage regression)**

. reg lwage educ exper uhat **(structural model with residual)**

```
      Source |       SS       df       MS              Number of obs =     722
-------------+------------------------------           F(  3,   718) =   44.87
       Model |   20.022858     3  6.67428601           Prob > F      =  0.0000
    Residual |  106.789058   718  .148731278           R-squared     =  0.1579
-------------+------------------------------           Adj R-squared =  0.1544
       Total |  126.811916   721  .175883378           Root MSE      =  .38566

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    .142298    .018011     7.90   0.000     .1069375    .1776585
       exper |   .0376059    .005374     7.00   0.000     .0270552    .0481565
        uhat |  -.0769213   .0196251    -3.92   0.000    -.1154509   -.0383918
       _cons |   4.429426   .2962819    14.95   0.000     3.847744    5.011108
------------------------------------------------------------------------------
```

Observation:

- The IV-estimate is 0.142, which is close to 0.151 (parameter estimate using OLS)

Conclusion:

- |*t*-value| on *uhat* > 1.96. Reject the null hypothesis and conclude that there is indication of endogeneity.

Example of tests for endogeneity and overidentification

These lecture notes are for your own use. It is not allowed to distribute the notes further by posting them on the Internet or on platforms without explicit and prior permission of the author.
© Utrecht University School of Economics 2024

Application of method below

Dutta, N., Sobel, R. S., & Roy, S. (2013). Entrepreneurship and political risk. *Journal of Entrepreneurship and Public Policy*.

Purpose
Previous literature has clearly demonstrated the need for sound government policies or "institutions" to promote and support entrepreneurship in a country. The purpose of this paper is to explore the role of one such institution – political stability – in boosting entrepreneurial endeavors. A politically stable nation will have lower risk and transaction/contracting costs, and higher levels of government transparency, predictability, and accountability. Thus, the paper should expect that with greater political stability there should be a greater degree of entrepreneurial activity.

Design/methodology/approach
Using dynamic panel estimators (System GMM estimators) and considering multiple proxies of political risk, our results confirm this hypothesis. Such estimators handle challenges associated with panel data efficiently.

Findings
The paper's results show that greater political stability for a country does indeed lead to an increased rate of entrepreneurship and wealth creation.

Originality/value
Entrepreneurship is critical to the process of economic growth and development. To prosper, countries must unleash the creative talents of their citizens through the decentralized process of formal private sector entrepreneurship. New legal businesses create jobs, opportunities, wealth, and goods and services that make a nation grow. Sadly in many nations, this process is stifled and poverty is the result. While previous research has examined which types of specific policies matter for promoting entrepreneurship, the paper considers the different question of how the stability of political institutions impacts the rate of entrepreneurship.

Example of tests for endogeneity and overidentification (II): panel data with lagged dependent variable.

Data: wagepan.dta

Example 10:

Next, we apply tests for endogeneity and overidentification to a panel data equation, in which there is a lagged dependent variable.

```
. ivreg d.lwage (d.l.lwage=l2.lwage l3.lwage), cluster(nr) first

First-stage regressions
-----------------------

      Source |       SS       df       MS              Number of obs =    2725
-------------+------------------------------           F(  2,  2722) =  424.36
       Model |  116.838912     2  58.4194561           Prob > F      =  0.0000
    Residual |   374.72058  2722  .137663696           R-squared     =  0.2377
-------------+------------------------------           Adj R-squared =  0.2371
       Total |  491.559492  2724  .180455027           Root MSE      =  .37103

------------------------------------------------------------------------------
    LD.lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lwage |
         L2. |   -.507284   .0175622   -28.88   0.000    -.5417206   -.4728473
         L3. |   .2586314   .0172092    15.03   0.000      .224887    .2923758
       _cons |   .4797474   .0251195    19.10   0.000     .4304922    .5290026
------------------------------------------------------------------------------


Instrumental variables (2SLS) regression               Number of obs =    2725
                                                       F(  1,   544) =    3.45
                                                       Prob > F      =  0.0638
                                                       R-squared     =       .
                                                       Root MSE      =  .42679

                                   (Std. Err. adjusted for 545 clusters in nr)
------------------------------------------------------------------------------
             |               Robust
     D.lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lwage |
         LD. |   .0784685   .0422525     1.86   0.064    -.0045296    .1614666
       _cons |   .0544607   .0046531    11.70   0.000     .0453205    .0636008
------------------------------------------------------------------------------
Instrumented:  LD.lwage
Instruments:   L2.lwage L3.lwage
------------------------------------------------------------------------------
```

Test for endogeneity (Hausman-Wu):

. reg d.l.lwage l2.lwage l3.lwage **(first-stage regression of 2SLS)**

```
      Source |       SS       df       MS              Number of obs =    2725
-------------+------------------------------           F(  2,  2722) =  424.36
       Model |  116.838912     2  58.4194561           Prob > F      =  0.0000
    Residual |   374.72058  2722  .137663696           R-squared     =  0.2377
-------------+------------------------------           Adj R-squared =  0.2371
       Total |  491.559492  2724  .180455027           Root MSE      =  .37103

------------------------------------------------------------------------------
    LD.lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lwage |
         L2. |   -.507284   .0175622   -28.88   0.000    -.5417206   -.4728473
         L3. |   .2586314   .0172092    15.03   0.000      .224887    .2923758
       _cons |   .4797474   .0251195    19.10   0.000     .4304922    .5290026
------------------------------------------------------------------------------
```

. predict uhat, resid **(residual after first-stage regression)**
(1635 missing values generated)

. reg d.lwage d.l.lwage uhat **(OLS on structural equation, included residual of first-stage regression)**

```
      Source |       SS       df       MS              Number of obs =    2725
-------------+------------------------------           F(  2,  2722) =  489.97
       Model |  122.017973     2  61.0089864           Prob > F      =  0.0000
    Residual |  338.931188  2722  .124515499           R-squared     =  0.2647
-------------+------------------------------           Adj R-squared =  0.2642
       Total |  460.949161  2724  .169217754           Root MSE      =  .35287

------------------------------------------------------------------------------
     D.lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lwage |
         LD. |   .0784685   .0326451     2.40   0.016     .0144568    .1424802
        uhat |  -.6474184   .0373897   -17.32   0.000    -.7207336   -.5741033
       _cons |   .0544607   .0070144     7.76   0.000     .0407067    .0682147
------------------------------------------------------------------------------
```

Conclusion: |*t*-value| for *uhat* > 1.96. Reject the null hypothesis and conclude that there is indication of endogeneity

Test for overidentification (Sargan):

. ivreg d.lwage (l.d.lwage = l2.lwage l3.lwage) **(2SLS)**
. predict uhat, resid **(residual after 2SLS)**
(1090 missing values generated)

. reg uhat l2.lwage l3.lwage **(regression of residual on all instruments and exogenous variables)**

```
      Source |       SS       df       MS              Number of obs =    2725
-------------+------------------------------           F(  2,  2722) =    2.75
       Model |  1.00198268     2  .500991339           Prob > F      =  0.0638
    Residual |  494.993579  2722  .181849221           R-squared     =  0.0020
-------------+------------------------------           Adj R-squared =  0.0013
       Total |  495.995562  2724   .18208354           Root MSE      =  .42644

------------------------------------------------------------------------------
        uhat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lwage |
         L2. |   .0061685   .0201848     0.31   0.760    -.0334106    .0457477
         L3. |  -.0397735   .0197791    -2.01   0.044     -.078557   -.0009899
       _cons |   .0519134   .0288706     1.80   0.072    -.0046972     .108524
------------------------------------------------------------------------------
```

. display e(r2) (**R2 in STATA)**
.00202014
.00202014* 2725=5.5045. Critical value of Chi-square (1): 3.84.

Conclusion: Reject H_0. There is indication of overidentification. Parameter estimates using 2SLS are inconsistent.

Wrapping up

- Simultaneity of supply and demand equation
- Simultaneity bias due to endogenous variables
- Instrumental variables and exclusion restriction
- Solving simultaneity bias by 2SLS
- Testing overidentifying restrictions in 2SLS (Sargan)
- Testing for endogeneity after 2SLS (Hausman-Wu)