Econometrics

Week 7

Institute of Economic Studies
Faculty of Social Sciences
Charles University in Prague

Fall 2022

Recommended Reading

For today

- Instrumental Variables Estimation and Two Stage Least Squares
- Chapter 15

Next week

- Midterm Exam
- Wednesday, November 23, 2-3:30pm
- Room 109

In two weeks

- Simultaneous Equations Models
- Chapter 16

Today's talk

- We will study the problem of **endogenous explanatory variables**
  - i.e. explanatory variables that are correlated with the disturbance

- Which assumption is violated when an explanatory variable is endogenous?
  (when an explanatory variable is correlated with disturbance)
  - The zero conditional mean assumption

- What happens to the OLS estimator when an explanatory variable is endogenous?
  - It is generally biased and inconsistent
  - You should be able to prove this
    Just work with the expected value of $\widehat{\beta}_{OLS}$

Today's talk

- We will study the properties of OLS with endogenous explanatory variables
- We will discuss what might cause an endogeneity problem
  - Omitted variable(s)
  - Measurement error
  - Simultaneity
- We will study estimation methods dealing with endogeneity
  - Using a **proxy variable**
  - **Instrumental Variables (IV) estimation**.
  - **Two stage least squares (2SLS) estimation**.
  - First differencing and Fixed effects transformation when we have panel data and endogeneity is caused by time-constant unobservables.

OLS with Endogenous Explanatory Variable

Consider a simple regression model:

$$y_i = \beta_0 + \beta_1 x_i + u_i, \text{ where } i = 1, 2, \ldots, n$$

with x and u correlated:

$$Cov(x, u) \neq 0 \Longrightarrow E[u_i | x_i] \neq const.$$

We say that x is an **endogenous** explanatory variable.

OLS with Endogenous Explanatory Variable

$$y_i = \beta_0 + \beta_1 x_i + u_i, \text{ where } i = 1, 2, \ldots, n$$

$$Cov(x, u) \neq 0 \Longrightarrow E[u_i|x_i] \neq 0.$$

$$\hat{\beta}_{1,OLS} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})} = \ldots = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})(u_i - \bar{u_i})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Inconsistency of OLS estimator:

$$plim\hat{\beta}_{1,OLS} = \beta_1 + \underbrace{\frac{Cov(x,u)}{Var(x)}}_{\text{bias}} = \beta_1 + \underbrace{Corr(x,u) \cdot \frac{\sigma_u}{\sigma_x}}_{\text{bias}}$$

where $Var(x) = {\sigma_x}^2$ and $Var(u) = {\sigma_u}^2$

OLS with Endogenous Explanatory Variable

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u},$$

$$Cov(\mathbf{X}, \mathbf{u}) \neq \mathbf{0} \Longrightarrow E[\mathbf{u}|\mathbf{X}] \neq \mathbf{0}.$$

$$\widehat{\beta}_{\mathbf{OLS}} = \left(X^T X\right)^{-1} X^T y = \beta + \left(X^T X\right)^{-1} X^T u$$

Inconsistency of OLS estimator:

$$plim \hat{\beta}_{OLS} = \beta + \left(\frac{1}{n}\mathbf{X^T X}\right)^{-1} plim \left(\frac{1}{n}\mathbf{X^T u}\right)$$

$$= \beta + \underbrace{\left(\mathbf{Var}[\mathbf{X}]\right)^{\mathbf{-1}} \cdot \mathbf{Cov}[\mathbf{X}, \mathbf{u}]}_{\text{bias}}$$

Example: Influence of Immigrants on Local Labor Markets

- Consider a relationship between immigration and unemployment rate among the natives:

$$unempl_i = \beta_0 + \beta_1 share\_immig_i + u_i$$

  where i denotes regions
- Do you expect $share\_immig_i$ to be exogenous or endogenous? Why?
  - We do not observe local labor market conditions across regions
  - Immigrants tend to locate in better-performing regions (e.g. Germany)
  - Unemployment rate is lower in better-performing regions
  - The estimate of β_1 is biased... downwards (derive using slide 6 formula)

Proxy Variable

- Can be used to mitigate the omitted variable bias

$$y_i = \beta_0 + \beta_1 x_{1i} + \underbrace{\beta_2 x_{2i}^* + u_i}_{unobserved}$$

- The proxy variable (x_{2i}) should be closely related to the unobserved omitted variable
  - $x_{2i}^* = \delta_0 + \delta_2 x_{2i} + \nu_i$
  - past unemployment rate measure might be a proxy for local labor market conditions
- Think of the proxy variable as the "second best" option of how to include the unobserved variable in the model.
- When having a proxy variable, we just run a regression of y on x_1, x_2

Instrumental Variable

Consider a simple regression model:

$$y_i = \beta_0 + \beta_1 x_i + u_i, \text{ where } i = 1, 2, \ldots, n$$

with x and u correlated.

- To obtain consistent estimates of β_0 and β_1, we can use **a new exogenous variable**.

Instrumental Variable z

- This variable has to satisfy the following properties:
  - (1) z is uncorrelated with u, $Cov(z, u) = 0$.
  - (2) z is correlated with x, $Cov(z, x) \neq 0$.

- (1) $\Rightarrow$ z is exogenous in the regression equation.
- (2) $\Rightarrow$ z must be related to the endogenous variable x.

IV Estimation - intuition

$$y_i = \beta_0 + \beta_1 x_i + u_i, \text{ with } Cov(x_i, u_i) \neq 0.$$

- We use z to consistently estimate regression parameters.
  - z is exogenous in the regression equation
  - z is related to the endogenous variable x
  - unlike a proxy variable, z does not affect y on its own, just through x
- z is called an **instrumental variable (IV)**
- Proper IV *identifies* the β_1 parameter by filtering the data
- For estimation we use only the variation in the data (i.e., among others, in the endogenous explanatory variable) "allowed" by z
- Think of the IV as of a **polarizing filter**

IV Estimation - summation notation

$$y_i = \beta_0 + \beta_1 x_i + u_i, \qquad \text{with } Cov(x,u) \neq 0.$$

- Let us use an instrumental variable z satisfying:
  - $Cov(z,x) \neq 0 \Rightarrow z$ is relevant.
  - $Cov(z,u) = 0 \Rightarrow z$ is exogenous.
- From the regression equation we have: $Cov(z,y) = \beta_1 Cov(z,x) + Cov(z,u)$.
- Because $Cov(z,u) = 0$ and $Cov(z,x) \neq 0$, we get:

$$\beta_1 = \frac{Cov(z,y)}{Cov(z,x)}.$$

- Under random sampling (applying the LLN):

Instrumental Variable (IV) estimator

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

IV Estimation - summation notation

$$y = \beta_0 + \beta_1 x + u, \qquad \text{with } Cov(x, u) \neq 0.$$

- The slope coefficient is estimated as:

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

- Intercept can be estimated as:

$$\hat{\beta}_{0,IV} = \bar{y} - \hat{\beta}_{1,IV}\bar{x}.$$

NOTE

When $z = x$, we have OLS estimator of β_1.

- In other words, when x is exogenous, IV estimator is identical to OLS estimator.

- IV estimator is consistent $plim(\hat{\beta}_{1,IV}) = \beta_1 + \frac{cov(z,u)}{cov(z,x)}$.

IV Estimation - matrix notation

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \qquad \text{with } Cov(\mathbf{X}, \mathbf{u}) \neq \mathbf{0}.$$

- $Cov(\mathbf{Z}, \mathbf{y}) = Cov(\mathbf{Z}, \mathbf{X})\beta + Cov(\mathbf{Z}, \mathbf{u})$.
- Because $Cov(\mathbf{Z}, \mathbf{u}) = \mathbf{0}$ and $Cov(\mathbf{Z}, \mathbf{X}) \neq \mathbf{0}$, we get:

$$\beta = \mathbf{Cov}(\mathbf{Z}, \mathbf{X})^{-1}\mathbf{Cov}(\mathbf{Z}, \mathbf{y}).$$

- Under random sampling (applying the LLN):

Instrumental Variable (IV) estimator

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

Valid Instruments

A valid instrumental variable has to satisfy two conditions:

- $Cov(z, u) = 0$ can never be tested (u is unobservable error), so we have to rely on economic theory and intuition to decide about exogeneity of z. // Can residuals from this regression $y = \beta_0 + \beta_1 x + u$ be used for testing when we expect x to be endogenous?

- $Cov(z, x) \neq 0$ can easily be tested by running a simple regression:

$x_i = \pi_0 + \pi_1 z_i + \nu_i$

$Cov(z, x) \neq 0$ holds if and only if $\pi_1 \neq 0$.

- Thus, for a valid instrument we should be able to reject the null hypothesis:
- $H_0 : \pi_1 = 0$

against the two-sided alternative that $H_A : \pi_1 \neq 0$

Example: Influence of immigrants on local labor markets

- Consider an equation for unemployment rate among the low-skilled natives:

$$unempl_i = \beta_0 + \beta_1 share\_immig_i + u_i$$

- variable $share\_immig_i$ is endogenous
- When we expect that $Cov(share\_immig_i, u) \neq 0$, we need an instrumental variable which:
  - influences the decision about where to immigrate (relevance).
  - does not affect local unemployment rate directly (exogeneity).

Example: Influence of immigrants on local labor markets

Good instrumental variables in this case might be:

Historical location of immigrants

- Immigrants tend to move to regions where their fellow citizens reside (relevance)
- Immigrants moving to Europe in 20th century faced different economic conditions (exogeneity).

Local policies towards immigrants

- Immigrants move to regions where it is easier to get asylum (relevance).
- immigration policies were designed earlier and are not affected by current economic conditions (exogeneity).

Statistical Inference under IV Estimation

- IV estimates are asymptotically normal

We need to assume homoskedasticity:

$$Var(u) = E(u^2|z) = \sigma^2$$

The asymptotic variance of $\hat{\beta}_{1,IV}$

Under homoskedasticity, the asymptotic variance of $\hat{\beta}_{1,IV}$ is:

$$Var(\hat{\beta}_{1,IV}) = \frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2}.$$

- where $\rho_{x,z}^2$ is the square of the correlation between x and z
- compare with the variance of the OLS slope estimator!

Statistical Inference under IV Estimation

- Thus we can estimate the standard error of IV estimator

Standard errors of $\hat{\beta}_{1,IV}$

The (asymptotic) variance of $\hat{\beta}_{1,IV}$ can be estimated as:

$$\widehat{Var(\hat{\beta}_{1,IV})} = \frac{\hat{\sigma}^2}{SST_x R^2_{x,z}},$$

where $\hat{\sigma}^2$ can be estimated from the IV residuals, SST_x is total sum of squares of the x and $R^2_{x,z}$ is simple R^2 from the regression of x on z

- Resulting standard errors allow us to construct t statistics for testing the hypotheses about β_1 and to form confidence intervals of β_1.

IV versus OLS Estimation

$$
\begin{array}{cc}
OLS & IV \\
\hat{\beta}_{1,OLS} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})} & \hat{\beta}_{1,IV} = \frac{\sum_{i=1}^{n}(z_i-\bar{z})(y_i-\bar{y})}{\sum_{i=1}^{n}(z_i-\bar{z})(x_i-\bar{x})} \\
Var(\hat{\beta}_{1,OLS}) = \frac{{\sigma_u}^2}{SST_x}, & Var(\hat{\beta}_{1,IV}) = \frac{{\sigma_u}^2}{SST_x \rho_{x,z}^2}, \\
\hat{\beta}_{OLS} = (\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T y}, & \hat{\beta}_{IV} = (\mathbf{Z^T X})^{-\mathbf{1}}\mathbf{Z^T y}, \\
Var(\hat{\beta}_{OLS}) = \sigma^2(\mathbf{X^T X})^{-\mathbf{1}}, & Var(\hat{\beta}_{IV}) = \sigma^2(\mathbf{Z^T X})^{-\mathbf{1}}\mathbf{Z^T Z}(\mathbf{Z^T X})^{-\mathbf{1}},
\end{array}
$$

- IV standard errors differ from OLS only by the $\rho_{x,z}^2$.
- Since $\rho_{x,z}^2 < 1$, IV standard errors are always larger than OLS standard errors.
- The stronger the correlation between z and x, the smaller the IV standard errors (in case of $\rho_{x,z}^2 = 1$, it is equivalent to OLS).

The Quality of Instruments

- What happens if $Cov(z, u) \neq 0$?
- IV estimator will be inconsistent.
- However, it can still be better than OLS under certain conditions!

Asymptotic bias of IV and OLS estimators

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

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

- Asymptotic bias in IV will be smaller than asymptotic bias in OLS if:

$$\frac{Corr(z, u)}{Corr(z, x)} < Corr(x, u)$$

IV estimation in Multiple Regression Case

- We can extend the IV estimation to multiple regression.
- Let's start with the case, where only one of the explanatory variables is correlated with the error:
  $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$.
- We use the name **structural equation** for such models, where we distinguish between endogenous and exogenous variables.
  - y_1 is obviously endogenous, as it is by definition correlated with u_1
  - z_1 is assumed to be exogenous (uncorrelated with u_1, $Cov(z_1, u_1) = 0$).
  - y_2 is suspected of being endogenous (correlated with u_1, $Cov(y_2, u_1) \neq 0$).

IV estimation in Multiple Regression Case

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$$

- With endogenous regressor the OLS estimator is biased and inconsistent $\Rightarrow$ we need to find a proper instrument for y_2, let's call it z_2.
- z_2 needs to be correlated with y_2 **after partialing out** the effect of exogenous variable(s) included in the structural model:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \nu_2$$

  - The key identification condition is $\pi_2 \neq 0$.
  - This **reduced form equation** regresses the endogenous variable on all exogenous variables.
- z_2 needs to be exogenous in the structural model: $Cov(z_2, u_1) = 0$

Two Stage Least Squares

- We may have multiple instruments for the endogenous variable y_2, say z_2 and z_3
- In this case we would have more than one IV estimator.
- BUT: None of the IV estimators would be efficient. Why?
- Since z_1, z_2 and z_3 are all uncorrelated with u_1, any linear combination of exogenous variables would be a valid IV.
- Thus, we choose the linear combination that is most highly correlated with y_2.
- The IV estimator using such instrument is known as the **two stage least squares (2SLS)** estimator.

Two Stage Least Squares

- Consider the following model:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1,$$

  with two potential instruments for y_2: z_2 and z_3
- 2SLS estimate is obtained in two stages:

Two-stage least squares (2SLS)

- (1): Obtain OLS fitted values of endogenous variable: (run the reduced-form equation)
  $\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 z_3$
- (2): Use the fitted values in the structural regression instead of the endogenous variable:
  $y_1 = \beta_0 + \beta_1 \hat{y}_2 + \beta_2 z_1 + u_1$

- But let R do the estimation for you to get the correct (robust) standard errors.

Two Stage Least Squares

- The 2SLS approach can be extended to multiple endogenous variables.
- $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 y_3 + \beta_3 z_1 + \beta_4 z_2 + u_1$,
- BUT, we need **at least** as many instruments as there are endogenous variables!
- In this case we need two IV's, let's call them z_3 and z_4
  - $cov(z_3, u_1) = 0$ and $cov(z_4, u_1) = 0$
  - $cov(z_3, y_2) \neq 0$ and $cov(z_4, y_3) \neq 0$
- (proper conditions statement in Advanced Econometrics).

Testing for Endogeneity

- When all explanatory variables are exogenous, both OLS and 2SLS are consistent estimators.
- BUT: 2SLS is less efficient than OLS $\Rightarrow$ OLS is preferred.
- If we have endogeneity problem, only 2SLS(or IV) is consistent.
- Thus it is good to have a test for endogeneity (to see if the 2SLS is necessary).

Hausman test for endogeneity

H_0 : OLS and IV are consistent.
H_A : OLS is inconsistent and IV is consistent.

- We simply compute both estimates and use Hausman test for comparison.
- (more about this test in the Advanced Econometrics course.)

Testing for Endogeneity

- Another alternative is to use a **regression-based test**.
- If y_2 is endogenous, then ν_2 from the reduced model and u_1 from the structural model are correlated.

Regression-based test for endogeneity

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1$$

- (1): Estimate the reduced-form equation for y_2 and obtain residuals $\hat{\nu}_2$:
  $y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + \nu_2.$
- (2): Run the structural model including endogenous variable and residual $\hat{\nu}_2$:
  $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \hat{\nu}_2 + \widetilde{u}_1$
- (3): If $H_0 : \delta_1 = 0$ is rejected against $H_A : \delta_1 \neq 0$ on small significance level $\Rightarrow Cov(\nu_2, u_1) \neq 0 \Rightarrow y_2$ is endogenous.

Testing Overidentification Restrictions

- If we have only one instrument for our endogenous variable, we can not test whether the instrument is uncorrelated with the error.
- We say that model is just identified.
- In case of multiple instruments for each endogenous variable, it is possible to test whether some of the instruments are correlated with the error.
- We call this testing for **overidentifying restrictions**

Testing Overidentification Restrictions

- (1): Estimate the structural model by 2SLS and obtain residuals, $\hat{u}_1$.
- (2): Regress $\hat{u}_1$ on all *exogenous variables* and obtain R^2

Test the H_0 : all IVs are uncorrelated with u_1

$$LM = nR^2 \overset{a}{\sim} \chi^2_q$$

- where q is the number of instrumental variables from outside minus the total number of endogenous explanatory variables.
- If we reject the H_0, at least some of the IV are not exogenous.

Thank you

Thank you very much for your attention!