# Instrumental Variables

Ani Katchova

## Instrumental Variables Overview

- Endogeneity examples
- Endogeneity definitions
- Instrumental variables set up
- The two stage least squares (2SLS) estimation procedure
- Identification issues
- Endogeneity tests
- Weak instrumental variables
- Systems of equations (2SLS and 3SLS)

# Instrumental Variables

## Endogeneity examples

- Wages and education jointly depend on ability which is not directly observable. We can use available test results to proxy for ability.
- Consumption and income are both determined by macroeconomic factors. We can use investments to control for endogeneity.

## Causes of endogeneity

- The explanatory variables are measured with errors
- Reverse causality (the explanatory variable is caused by the dependent variable)

## Endogeneity definitions

A regressor is endogenous when it is correlated with the error term.
Example: y is earnings, x is years of schooling, u is error term (including ability), z is proximity to college.

Exogeneity: regressors *x* and the error term *u* are independent causes of the dependent variable *y*.

$$\begin{tikzcd} x \arrow[r] & y \\ u \arrow[ur] & \end{tikzcd}$$

Endogeneity: the error *u* is affecting the regressors *x* and therefore indirectly affecting *y*.

$$\begin{tikzcd} x \arrow[r] & y \\ u \arrow[u] \arrow[ur] & \end{tikzcd}$$

Instrumental variables: instruments *z* are associated with *x* but not with the error term *u*.

$$\begin{tikzcd} z \arrow[r] & x \arrow[r] & y \\ & u \arrow[u] \arrow[ur] & \end{tikzcd}$$

Requirements for instruments $z$:

- $z$ is correlated with the regressors $x$, $\mathrm{E}[z'x] \neq 0$ ($z$ predicts or causes $x$),
- $z$ is uncorrelated with the error term $u$, $\mathrm{E}[z'u] = 0$ ($z$ is not endogenous),
- $z$ is not a direct cause of the dependent variable $y$, $\mathrm{cov}[y, z|x] = 0$ ($z$ is not in the $y$ equation).

**Instrumental variables set up**

- Consider the linear model: $y = x\beta + u$
- Endogeneity is when one or more explanatory variables are correlated with the error term: $\mathrm{E}[x|u] = \mathrm{cov}(x'u) \neq 0$.
- The estimated coefficients from the OLS estimation are biased:

$$b = \beta + (x'x)^{-1}x'u,\ \mathrm{E}[b] \neq \beta.$$

- We re-write the model as the following structural equation:

$$y_1 = y_2^{'}\beta_1 + x_1^{'}\beta_2 + u$$

where $y_1$ is the dependent variable, $y_2$ is the endogenous variable, and $x_1$ are the exogenous variables.

- The structural equation model involves a combined set $x = [y_2, x_1]$ of both endogenous and exogenous variables.
- We need to find a set of instrument $z = [x_1, x_2]$ of only exogenous variables, where $x_1$ is instrument for itself and $x_2$ is instrument for $y_2$.

**The two stage least squares (2SLS) estimation procedure**

- The 2SLS procedure replaces the endogenous variable with predicted values of this endogenous variable when regressed on instruments.

1. Estimate the first stage (reduced form) equation with only exogenous regressors.

$$y_2 = x_1^{'}\gamma_1 + x_2^{'}\gamma_2 + e$$

2. Calculate the predicted values $\hat{y}_2$and substitute them in the structural equation model.

$$y_1 = \hat{y}_2^{'}\beta_1 + x_1^{'}\beta_2 + u$$

**Identification issues**

- Order condition: The number of omitted instrumental variables must be at least as large as the number of endogenous regressor.
- Rank condition: The matrices $z'x$ must have a full rank in order to be inverted.

*Just-identified model*

- An IV model is just identified if there is one instrument $x_2$for each endogenous variable $y_2$.

$$b_{IV} = (z'x)^{-1}z'y = (z'x)^{-1}z'(x\beta + u) = \beta + (z'x)^{-1}z'u$$

- This estimator is unbiased.

*Under-identified model*

- An IV model is under-identified if there are fewer instruments $x_2$than endogenous variables $y_2$.
- The under-identified model has an infinite number of solutions and therefore no consistent estimator exists.

*Over-identified model*

- An IV model is over-identified if there are more instruments than endogenous variables.
- There are two efficient estimators that can be used:

- The two stage least squares (2SLS) (best if the error term is iid and homoskedastic):

$$b_{2sls} = [x'z(z'z)^{-1}z'x]^{-1}x'z(z'z)^{-1}z'y$$

- The generalized method of moments (GMM):

$$b_{GMM} = (x'zwz'x)^{-1}x'zwz'y$$

If $w = (z'z)^{-1}$, then this is the 2SLS estimate.
Usually $w = \hat{S}^{-1}$, where $\hat{S}$ is the estimated variance of *z'u.*
This estimator is optimal in presence of heteroscedasticity.

## Instrumental variables tests

*Hausman test for endogeneity*

- The Hausman test checks if a regressor is exogenous or endogenous.
- The Hausman test compares the OLS and IV estimates to check for significant differences.
  - If there are significant differences, then the regressor is endogenous.
  - If there are no significant differences, then the regressor is exogenous.

*Durbin-Wu-Hausman test for exogenous regressors*

- The Durbin-Wu-Hausman test is a procedure that checks whether $\mathrm{E}[\mathrm{x}|e] = \mathrm{cov}(\mathrm{x}e) \neq 0$.
- Estimate the first-stage model: $y_2 = x_1^{'}\gamma_1 + x_2^{'}\gamma_2 + u$
- Include the residuals ($\hat{u}$) from the first-stage regression in the structural equation regression: $y_1 = y_2^{'}\beta_1 + x_1^{'}\beta_2 + \hat{u}\rho + e$
  - If the coefficient on the residuals from the first-stage regression $\rho$ is not significantly different from zero then the regressors are exogenous.
  - If the coefficient $\rho$ is significantly different from zero then the regressors are endogenous.

*Tests for overidentifying restrictions*

- Estimate model using GMM and form a test statistic:

$$Q(\beta) = (1/N)(y - x\beta)'z(S^{-1})(1/N)z'(y - x\beta)$$

- It is distributed as chi-square with degrees of freedom of the number of overidentifying restrictions.
- Rejection of null hypothesis – at least one instrument is not valid.

## Weak Instrumental Variables

A weak instrument has a low correlation with the endogenous variable.

**Tests for weak instruments**

- In a case of one endogenous regressor and one instrument, a low correlation between instrument and the endogenous variable would indicate a weak instrument.
- When several instruments are used for one endogenous variable, the weakness of the instruments can be measured by the partial $R^2$ and partial F-statistic from the first stage regression.
  - The instrument is weak if the partial F-statistic testing the joint significance of the coefficients of the instruments ($\gamma_2 = 0$) is less than 10.

**Consequences of weak instruments**

- A weak instrument will undermine the precision of the estimator.

$$V\left(\hat{\beta}_{IV}\right) = V\left(\hat{\beta}_{OLS}\right)/r_{xz}^2$$

- The IV estimator is asymptotically consistent but biased toward OLS estimator in finite-sample. The size of the bias is positively related to the weakness of the instrument(s) and inversely related with the size of the sample.

## Instrumental Variables and Simultaneous Systems of Equations

Simultaneous systems of equations with two endogenous variables

- The system of structural equations is:

$$y_1 = y_2'\beta_1 + z_1'\gamma_1 + u_1$$
$$y_2 = y_1'\beta_2 + z_2'\gamma_2 + u_2$$

- There are endogenous variables as independent variables in both equations.
- The reduced form equation is:

$$y = z'\Gamma + u$$

The two stage least squares (2SLS) or three stage least squares (3SLS) procedure:

1. Estimate the reduced form equation by OLS regression and obtain $\hat{y}$.
2. Use the estimates $\hat{y}$ from the first stage to estimate the structural equations:

$$y_1 = \hat{y}_2'\beta_1 + z_1'\gamma_1 + u_1$$
$$y_2 = \hat{y}_1'\beta_2 + z_2'\gamma_2 + u_2$$

These estimates are the 2SLS estimates.

3. Use the 2SLS estimates to compute the 3SLS using the following estimator:

$$\hat{\beta}_{3SLS} = \{X'(\Sigma^{-1} \otimes I_N)X\}^{-1}\{X'(\Sigma^{-1} \otimes I_N)y\}$$

**2SLS and 3SLS comparison**

- 3SLS is more efficient than 2SLS because it uses cross-equation information.
- 3SLS is inconsistent if the error term is heteroscedastic.