Instrumental Variables

Kakhramon Yusupov
IFPRI/WIUT Summer School 2015
June 8, 2015

Example of Endogenous Independent Variable

- Suppose price of fuel must be raised in order cut consumption by 20%
- Sample: Price P and quantity Q at various times and markets
- We want the demand function $E(Q|P = p) = q(p)$, i.e. what mean consumption would have been if price is set to p
- The interaction between demand and supply means that Q and P are determined simultaneously – P is endogenous, and is correlated with Q both through the demand and the supply curves

Example of Endogenous Independent Variable

- In our market causality goes both ways Q $\leftrightarrow$P
- Way out is finding an instrument variable
- Instrumental variable must be correlated with P but not the residual of Q–q(p)
  - Instrumental variable must affect Q only through price P and not in any other way
- Example of instrumental variable in this case
  - Number of terrorist attacks in the Middle East
  - Number of American troops present in the area

Instrument Variable

- If all assumptions of OLS are met
  $E\,(Y \mid X = x) = \alpha + \beta x$
- X is endogenous when causality goes both ways
  $\mathrm{cov}\,(X, Y - E\,(Y \mid X)) = \mathrm{cov}\,(X, u) \neq 0$
  $u = Y - (\alpha + \beta x)$
- Z is a useful instrument variable when
  - Z is exogenous: $\mathrm{cov}\,(Z, u) = 0$
  - Z is only correlated with Y through X
  - Z is exogenous to the Y and X interaction
  - Z is relevant: $\mathrm{cov}\,(Z,X) \neq 0$
  - Z is a predictor for X

Estimation: Two Stage Least Squares (TSLS)

- The linear regression of X on Z:
  $E(X|Z) = a + bZ$, $b = \mathrm{cov}(X,Z)/\mathrm{var}(Z)$
- The linear regression of Y on Z:
  $E(Y|Z) = E(E(Y|X, Z)|Z) = E(\alpha + \beta X|Z) = \alpha + \beta(X|Z) = \alpha + \beta a + \beta b Z$
- $\beta b = \mathrm{cov}(Y,Z)/\mathrm{var}(Z)$
- Then $\beta = (\beta b)/b = \mathrm{cov}(Y,Z)/\mathrm{cov}(X,Z)$
- This is IV estimator

Identification

- The regression model is
  - Exactly identified when the number of instruments is equal to the number of endogenous variables
  - Over identified when the number of instruments is greater than the number of endogenous variables
  - Under identified when the number of instruments is less than the number of endogenous variables

General Two Stage Least Squares

- Regress each endogenous regressor on all instruments and exogenous variables

$$X_{ji} = \gamma_0 + \gamma_1 Z_{1i} + \ldots + \gamma_n Z_{ni} + \delta_1 W_{1i} + \ldots + \delta_k W_{ki} + v_{ji}$$

- Regress the dependent variable on predicted values of endogenous variable

$$Y_{ji} = \beta_0 + \beta_1 \hat{Z}_{1i} + \ldots + \beta_n \hat{Z}_{ni} + \delta_1 W_{1i} + \ldots + \delta_k W_{ki} + \varepsilon_{ji}$$

Instrument Validity

- Invalid instruments produce meaningless results
- Instrument are checked by regressing endogenous variables on instruments
  - Instrument is weak if correlation between instrument and endogenous variable is close to zero
  - Weak instruments results in poor results
    - Biased
    - High fluctuation when specification or sample changes

Instrument Validity

- What if the number of instruments is greater than the number of endogenous variables
  - Run TSLS estimator and obtain the predicted values of Y
  - Generate the residuals $\hat{u}$
  - Regress $\hat{u}$ against all instruments Z and all exogenous variables W (unrestricted) and test whether all instruments are jointly insignificant (restricted) in a standard F-test

Testing for Endogeniety (Wu-Hausman test)

- Run the first stage regression

$$X_{ji} = \gamma_0 + \gamma_1 Z_{1i} + \ldots + \gamma_n Z_{ni} + \delta_1 W_{1i} + \ldots + \delta_k W_{ki} + v_{ji}$$

- Obtain the predicted residuals from the first stage $\hat{v}_{ji}$
- Include the predicted residuals $\hat{v}_{ji}$ into the second stage regression

$$Y_{ji} = \beta_0 + \beta_1 \hat{Z}_{1i} + \ldots + \beta_n \hat{Z}_{ni} + \delta_1 W_{1i} + \ldots + \delta_k W_{ki} + \gamma \hat{v}_{ji} + \varepsilon_{ji}$$

- Use F-test to test if the coefficient of $\hat{v}_{ji}$ is zero

Testing for Overidentifying Instruments

- Use 2SLS estimator to run the following regression

$$Y_{ji} = \beta_0 + \beta_1 \hat{Z}_{2i} + \ldots + \beta_n \hat{Z}_{ni} + \delta_1 W_{1i} + \ldots + \delta_k W_{ki} + \varepsilon_{ji}$$

- Note that we omitted Z_1
- Obtain the predicted residuals $\hat{\varepsilon}_{ji}$
- If Z_1 and $\hat{\varepsilon}_{ji}$ are correlated then Z_1 is not a valid instrument

Key points

- Omitted variables, measurement errors in variables and the regressor *X* being causally influenced by *Y* are sources of endogeniety
- Instrumental variable regression can provide valid inference for $E(Y \mid X = x)$
- Applying instrumental variable technique is a challenging task
  - Hard to come up with a convincing instrument
  - Own judgement is needed
- Weak instruments are dangerous
  - Increasing sample size or the number of weak instruments does not help
- IV estimates are less efficient than OLS

Empirical Literature

- Acemoglu et al (2001) wanted to measure the effect of institution quality on economic performance in former colonial countries
- Good economy allows good institutions, and good institution helps the economy, thus causality goes both ways
- In colonial time, better institutions were set up in colonies where European settlers wanted to settle
- They wanted to settle if they did not die in the early days due to deceases
- Early mortality of settlers for soldiers, missionaries and sailors is exogenous to later institutions and economic conditions, and is correlated with settlement and thus institution quality. Very clever!
- We will replicate this paper today in STATA!