Instrumental Variables and Two-stage Least Squares

Chayanee Chawanote

Semester 2/2013

Concept

- The population model is
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K + u$ (1)
  E(u) = 0, $Cov(x_j, u) = 0$, j = 1,2,..., K-1
- x_K is correlated with u => x_K is endogenous.
- To use the method of instrumental variables (IV), we need an observable variable z_1 not in equation (1) that satisfies the two conditions:
  (i) z_1 must be uncorrelated with u: $Cov(z_1, u) = 0$
  (ii) The relationship between z_1 and the endogenous variable x_K.

Concept

The relationship between z_1 and x_K

- The linear projection of x_K onto all the exogenous variables: $x_K = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + r_K$ (2)
- By definition of a linear projection error, $E(r_K) = 0$ and r_K is uncorrelated with $x_1, x_2, ..., x_{K-1}$,and z_1. And, $\theta_1 \neq 0$.
- $\theta_1 \neq 0$ means that z_1 is partially correlated with x_K once the other exogenous variables have been netted out. If x_K is the only explanatory variable in eq.(1), then the linear projection is $x_K = \delta_0 + \theta_1 z_1 + r_K$, where $\theta_1 = Cov(z_1, x_K)/Var(z_1)$, or we just say that $Cov(z_1, x_K) \neq 0$.
- z_1 solve the identification problem for the β_j in eq.(1), that is, we can write the β_j in terms of population moments in observable variables.

Identification problem and IV estimator

- Write eq.(1) as $y = \mathbf{x}\beta + u$ (3), where x $= (1, x_1, x_2, ..., x_K)$
- Write a vector of all exogenous variables as z $= (1, x_1, x_2, ..., x_{K-1}, z_1)$
- K population orthogonality conditions: $E(z'u) = 0$ (4).
- From (3), we can write (4) as $[E(z'x)]\,\beta = E(z'y)$ (5).
- Eq.(5) represents a system of K linear equations in the K unknowns $\beta_1, \beta_2, ..., \beta_K$. This system has a unique solution if and only if the K$\times$K matrix $E(z'x)$ has full rank; rank $E(z'x)$ = K.
- So, the solution is $\beta = [E(z'x)]^{-1} E(z'y)$ (6). Eq.(6) identifies the vector β.

IV estimator

- Given a random sample $\{(x_i, y_i, z_{i1}) : i = 1, 2, ..., N\}$ from the population, the instrumental variables estimator of β is

$$\hat{\beta^{IV}} = \left(\frac{1}{N} \sum_{i=1} z_i' x_i \right)^{-1} \left(\frac{1}{N} \sum_{i=1} z_i' y_i \right) = (Z'X)^{-1} Z'Y$$

Multiple instruments: Two-Stage Least Squares

- Define the vector of exogenous variables by z = $(1, x_1, x_2, ..., x_{K-1}, z_1, ..., z_M)$, a 1× L vector (L = K + M).
- Out of all possible linear combinations of z that can be used as an instrument for x_K, the method of 2SLS chooses the most highly correlated with x_K.
- If x_K were exogenous, then the best instrument for x_K is simply itself.
- The linear combination of z most highly correlated with x_K is given by the linear projection of x_K on z.
- Write the reduced form for x_K as $x_K = \delta_0 + \delta_1 x_1 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + ... + \theta_M Z_M + r_K$ (7), where r_K has zero mean and is uncorrelated with each RHS variable.

Multiple instruments: Two-Stage Least Squares

- Define the vector $\hat{\mathbf{x}}_i \equiv (1, x_{i1}, ..., x_{i,K-1}, \hat{x}_{iK}), i = 1, 2, ..., N.$ where $\hat{x}_K = \hat{\delta}_0 + \hat{\delta}_1 x_1 + ... + \hat{\delta}_{K-1} x_{K-1} + \hat{\theta}_1 z_1 + ... + \hat{\theta}_M Z_M$ (8)
- Using $\hat{\mathbf{x}}_i$ as the instruments for x_i gives the IV estimator

$$\hat{\beta} = \left(\sum_{i=1} \hat{x}_i' x_i \right)^{-1} \left(\sum_{i=1} \hat{x}_i' y_i \right) = (\hat{X}'X)^{-1} \hat{X}'Y \quad (9)$$

- $\hat{X}$ has N $\times$ (K+1) and $\hat{X} = Z(Z'Z)^{-1}Z'X = P_Z X$
- P_Z is a projection matrix, idempotent and symmetric.
- So, $\hat{X}'X = X'P_z X = (P_Z X)'P_Z X = \hat{X}'\hat{X}$.
- Hence, we have $\hat{\beta}^{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y$.

Two-Stage Least Sqaures

- To summarize, $\hat{\beta}^{2SLS}$ can be obtained from the following steps:
  1. Obtain the fitted values $\hat{x}_K$ from the regression x_K on 1, $x_1, x_2, ..., x_{K-1}, z_1, ..., z_M$ for the first-stage regression.
  2. Run the OLS regression y on $1, x_1, x_2, ..., x_{K-1}, \hat{x}_K$ for the second-stage regression
- In practice, it is best to use a software package with a 2SLS command rather than explicitly carry out the 2-step procedure in which the OLS standard errors reported from the 2nd-stage by hand will be incorrect.

2SLS assumptions

- Assumption 2SLS.1: For some $1\times$ L vector z, $E(\mathbf{z}'\mathbf{u}) = 0$
  - Also the zero conditional mean assumption: $E(u|\mathbf{z}) = 0$
- Assumption 2SLS.2: (a) rank $E(\mathbf{z}'\mathbf{z}) = L$; (b) rank $E(\mathbf{z}'\mathbf{x}) = K$
  - Exogenous variables are linearly independent in the population
  - z is sufficiently linearly related to x so that rank $E(\mathbf{z}'\mathbf{x})$ has full column rank.
  - Order condition: $L \geq K$, we must have at least as many instruments as we have explanatory variables, or β is not identified. We also need that at least one element of z not in x is significant.

2SLS assumptions

- Assumption 2SLS.3: $E(u^2\mathbf{z}'\mathbf{z}) = \sigma^2 E(\mathbf{z}'\mathbf{z})$, where $\sigma^2 = E(u^2)$
- With conditional on z, $E(u^2|\mathbf{z}) = \sigma^2$
  - Under 2SLS.1-3, $\sqrt{N}(\hat{\beta} - \beta)$ is asymptotically normally distributed with mean zero and variance matrix $\sigma^2 \left\{E(\mathbf{x}'\mathbf{z})[E(\mathbf{z}'\mathbf{z})]^{-1}E(\mathbf{z}'\mathbf{x})\right\}^{-1}$
  - We need to estimate $\hat{\sigma}^2$. Define 2SLS residuals: $\hat{u}_i = y_i - \mathbf{x}_i\hat{\beta}$. So, $\hat{\sigma}^2 \equiv (N-K)^{-1}\sum^N \hat{u}_i^2$
  - Therefore an estimator of the asymptotic variance of $\hat{\beta}$ is $\hat{\sigma}^2 \left(\sum^N \hat{x}_i'\hat{x}_i\right)^{-1} = \hat{\sigma}^2(\hat{X}'\hat{X})^{-1}$

Testing for Endogeneity

- 2SLS estimator is less efficient than OLS when the explanatory variables are exogenous.
- 2SLS estimates may provide very large standard errors.
- Hausman: directly comparing OLS and 2SLS estimated to see whether the differences are statistically significant or not.
- Both OLS and 2SLS are consistent if all variables are exogenous (H_0)
- Suppose we have: $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1$ (10)
  - z_1 and z_2 are exogenous.
- If 2SLS and OLS differ significantly, then y_2 must be endogenous (H_a).

Testing for Endogeneity

- Estimating the reduced form for y_2:
  $y_2 = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_3 + \gamma_4 z_4 + v_2$ (11)
- Since each z_j is uncorrelated with u_1, y_2 is uncorrelated with u_1 iff v_2 is uncorrelated with u_1.
- $u_1 = \delta_1 v_2 + e_1$, $cov(e_1, v_2) = 0$ and $E(e_1) = 0$
  - u_1 and v_2 are uncorrelated iff $\delta_1 = 0$
  - Include v_2 as an additional regressor in (10), and do t test.

Testing for Endogeneity

- (i) Estimate the reduced form for y_2 by regressing eq (11), then obtain the residuals $\hat{v}_2$
- (ii) Add $\hat{v}_2$ to the structural equation (including y_2) and test for significance of $\hat{v}_2$ using OLS.
  $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \hat{v}_2 + error$
- test $H_0 : \delta_1 = 0$ using t statistic
  - If reject H_0, then y_2 is endogenous because u_1 and v_2 are correlated.
- If multiply endogenous explanatory variables, for each suspected endogenous variable, obtain the reduced form residuals. Then, test for joint significance of these residuals in the structural equation, using F test.
  - # of exclustion restriction = # of suspected endogenous explanatory variables

Testing Overidentification Restrictions

- We have more instruments than we need. If all instruments are exogenous, the 2SLS residuals should be uncorrelated with the instruments, up to sampling error.
- The test checks whether the 2SLS residuals are correlated with q linear functions of the instruments.
- (i) Estimate the structural equation by 2SLS and obtain the 2SLS residuals $\hat{u}_1$
- (ii) Regress $\hat{u}_1$ on all exogenous variables. Obtain the R-squared: R_1^2
- (iii) H_0 : all IVs are uncorrelated with u_1, $nR_1^2 \sim \chi_q^2$,
  - q = # IVs from outside the model - # endogenous explanatory variables
  - If nR_1^2 exceeds 5% (or 10%) critical value in the χ_q^2 distribution, we reject H_0.
  - reject H_0 = at least some of the IVs are not exogenous.

IV Solutions to Errors-in-Variables Problems

- Consider the model: $y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + u$ (12)
  - y and $x_2$are observed, but x_1^* is not.
- Let $x_1 = x_1^* + e_1$, where e_1 is the measurement error.
- The model becomes: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + (u - \beta_1 e_1)$ (13)
- The classical errors-in-variables (CEV) assumption: e_1 is uncorrelated with x_1^* and x_2.
  - x_2 is exogenous, but x_1 is correlated with e_1.
  - Hence, we need an IV for x_1: z_1 is correlated with x_1, but uncorrelated with u and e_1

IV Solutions to Errors-in-Variables Problems

- We can obtain a second measurement on x_1^*: $z_1 = x_1^* + a_1$ as an IV for x_1.
- How can we get two measurements on a variable?
  - For married couples, each spouse can independently report information
  - For twins, ask each of them the other's years of education, an IV for self-reported education in a wage equation.
- Althernative: use other exogenous variables as IVs for a potentially mismeasured variable.
  - Use *motheduc* and *fatheduc* as IVs for education
- Use proxy to control for unobserved characteristics.
  - e.g., when use IQ as a proxy variable for unobserved ability, we might need another test score for ability.