Reminder

- Education x possibly endogenous (correlated with error u) so need an instrument z for education x that is correlated with education x and not correlated with the error u.
- IV: endogenous regressor / omitted variable problem – use excluded exogenous regressors.
- 2SLS: more than one excluded exogenous regressor.

Errors in Variables

- Today: IV/2SLS also solves Classical Errors in Variables (CEV) problem.

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + u$$

x_1^* not observed. $\quad x_1 = x_1^* + e_1 \quad e_1$ measurement error.

Errors in Variables

- Today: IV/2SLS also solves Classical Errors in Variables (CEV) problem.

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + u$$

x_1^* not observed. $\quad x_1 = x_1^* + e_1 \quad e_1$ measurement error.

- Get second measurement on x_1^*: z_1.
  - Income: ask employers.
  - Education: ask twin.
  - Household savings: ask spouse.

Testing for endogeneity

- 2SLS less efficient than OLS if explanatory variables are exogenous – large standard errors – so should test!
- Hausman (1978) test for endogeneity of y_2.

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1 \tag{1}$$

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_2 \tag{2}$$

Testing for endogeneity

- 2SLS less efficient than OLS if explanatory variables are exogenous – large standard errors – so should test!
- Hausman (1978) test for endogeneity of y_2.

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1 \tag{1}$$

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_2 \tag{2}$$

$Corr(z_j, u_1) = 0$ for $j = 1, 2, 3, 4$ implies $Corr(y_2, u_1) = 0$ if and only if $Corr(v_2, u_1) = 0 \leftarrow$ we want to test this.

$$u_1 = \delta_1 v_2 + e_1 \quad Corr(e_1, v_2) = 0 \quad E(e_1) = 0$$

Test: $\delta_1 = 0$.

Testing for endogeneity

- 2SLS less efficient than OLS if explanatory variables are exogenous – large standard errors – so should test!
- Hausman (1978) test for endogeneity of y_2.

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1 \tag{1}$$

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_2 \tag{2}$$

  $Corr(z_j, u_1) = 0$ for $j = 1, 2, 3, 4$ implies $Corr(y_2, u_1) = 0$ if and only if $Corr(v_2, u_1) = 0 \leftarrow$ we want to test this.

$$u_1 = \delta_1 v_2 + e_1 \quad Corr(e_1, v_2) = 0 \quad E(e_1) = 0$$

  Test: $\delta_1 = 0$.
- Test by putting v_2 as extra regressor in (1) and do t-test.
- Problem: v_2 unobservable.

Testing for endogeneity

- 2SLS less efficient than OLS if explanatory variables are exogenous – large standard errors – so should test!
- Hausman (1978) test for endogeneity of y_2.

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1 \tag{1}$$
$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_2 \tag{2}$$

  $Corr(z_j, u_1) = 0$ for $j = 1, 2, 3, 4$ implies $Corr(y_2, u_1) = 0$ if and only if $Corr(v_2, u_1) = 0 \leftarrow$ we want to test this.

$$u_1 = \delta_1 v_2 + e_1 \quad Corr(e_1, v_2) = 0 \quad E(e_1) = 0$$

  Test: $\delta_1 = 0$.
- Test by putting v_2 as extra regressor in (1) and do t-test.
- Problem: v_2 unobservable.
- Solution: estimate reduced form (2) by OLS to get $\hat{v}_2$.

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \hat{v}_2 + \text{error} \tag{3}$$

  Run OLS on (3) and do a t-test of $H_0 : \delta_1 = 0$.
- Rejection implies endogeneity since $Corr(v_2, u_1) \neq 0$.

Testing for overidentifying restrictions

- Number of overidentifying restrictions = number of extra IVs.

$$\begin{bmatrix} \text{1 endogenous RHS variable} \\ \text{1 IV} \end{bmatrix} \implies \text{None -- nothing can be tested.}$$

$$\begin{bmatrix} \text{1 endogenous RHS variable} \\ \text{2 IVs} \end{bmatrix} \implies \text{One overidentifying restriction.}$$

Testing for overidentifying restrictions

- Number of overidentifying restrictions $=$ number of extra IVs.

$$\begin{bmatrix} \text{1 endogenous RHS variable} \\ \text{1 IV} \end{bmatrix} \Longrightarrow \text{None -- nothing can be tested.}$$

$$\begin{bmatrix} \text{1 endogenous RHS variable} \\ \text{2 IVs} \end{bmatrix} \Longrightarrow \text{One overidentifying restriction.}$$

- Testing overidentifying restrictions:

Testing for overidentifying restrictions

- Number of overidentifying restrictions = number of extra IVs.

$$\begin{bmatrix} \text{1 endogenous RHS variable} \\ \text{1 IV} \end{bmatrix} \implies \text{None – nothing can be tested.}$$

$$\begin{bmatrix} \text{1 endogenous RHS variable} \\ \text{2 IVs} \end{bmatrix} \implies \text{One overidentifying restriction.}$$

- Testing overidentifying restrictions:
  1. Estimate structural equation (e.g. (1)) by 2SLS $\rightarrow \hat{u}_1$.

Testing for overidentifying restrictions

- Number of overidentifying restrictions = number of extra IVs.

$$\begin{bmatrix} \text{1 endogenous RHS variable} \\ \text{1 IV} \end{bmatrix} \Longrightarrow \text{None – nothing can be tested.}$$

$$\begin{bmatrix} \text{1 endogenous RHS variable} \\ \text{2 IVs} \end{bmatrix} \Longrightarrow \text{One overidentifying restriction.}$$

- Testing overidentifying restrictions:
  1. Estimate structural equation (e.g. (1)) by 2SLS $\rightarrow \hat{u}_1$.
  2. Regress $\hat{u}_1$ on all exogenous variables $\rightarrow R_1^2$.

Testing for overidentifying restrictions

- Number of overidentifying restrictions $=$ number of extra IVs.

$$\begin{bmatrix} \text{1 endogenous RHS variable} \\ \text{1 IV} \end{bmatrix} \Longrightarrow \text{None – nothing can be tested.}$$

$$\begin{bmatrix} \text{1 endogenous RHS variable} \\ \text{2 IVs} \end{bmatrix} \Longrightarrow \text{One overidentifying restriction.}$$

- Testing overidentifying restrictions:
  1. Estimate structural equation (e.g. (1)) by 2SLS $\rightarrow \hat{u}_1$.
  2. Regress $\hat{u}_1$ on all exogenous variables $\rightarrow R_1^2$.
  3. Under H_0, all IVs uncorrelated with u_1, $nR_1^2 \overset{a}{\sim} \chi_q^2$.
     $q =$ number of overidentifying restrictions.
     $nR_1^2 > \chi^2_{q,0.05} \longrightarrow$ reject H_0, so at least some IVs are not exogenous.

Testing for overidentifying restrictions

- Number of overidentifying restrictions = number of extra IVs.

$$\begin{bmatrix} \text{1 endogenous RHS variable} \\ \text{1 IV} \end{bmatrix} \Longrightarrow \text{None -- nothing can be tested.}$$

$$\begin{bmatrix} \text{1 endogenous RHS variable} \\ \text{2 IVs} \end{bmatrix} \Longrightarrow \text{One overidentifying restriction.}$$

- Testing overidentifying restrictions:
  1. Estimate structural equation (e.g. (1)) by 2SLS $\rightarrow \hat{u}_1$.
  2. Regress $\hat{u}_1$ on all exogenous variables $\rightarrow R_1^2$.
  3. Under H_0, all IVs uncorrelated with u_1, $nR_1^2 \overset{a}{\sim} \chi_q^2$.
     q = number of overidentifying restrictions.
     $nR_1^2 > \chi^2_{q,0.05} \longrightarrow$ reject H_0, so at least some IVs are not exogenous.
- Can test only when have more instruments than we need – just enough $\rightarrow$ **just** identified – can also make tests robust to heterogeneity of arbitrary form.

2SLS with Heteroscedasticity

- Similar issues as with OLS; metrics packages do this routinely.
- Breusch-Pagan test:
  - 2SLS $\rightarrow \hat{u}$ and let $z_1, \ldots, z_m$ be all exogenous variables including IVs.
  - Under assumptions, use F test for joint significance of z's in a regression of $\hat{u}^2$ on $z_1, \ldots, z_m$.
  - H_0 of homogeneity rejected if z_j jointly significant.
- Can use W2SLS if know how error variance depends on exogenous variables.