

4. MEASUREMENT ERROR AND PROXY VARIABLES

[1] Measurement Errors on Regressors

- OLS estimates are not consistent.
- A simple example:
 - True model: $y_t = \beta x_t^* + \varepsilon_t$.
 - But we only observe $x_t = x_t^* + v_t$ (v_t : measurement error).
[x_t may be a proxy variable for x_t^* .]
 - If we use x_t for x_t^* , $y_t = x_t\beta + [\varepsilon_t - \beta v_t]$ (model we estimate).
 - x_t and $(\varepsilon_t - \beta v_t)$ correlated.
 - Assume that the ε_t are i.i.d. with $N(0, \sigma^2)$; v_t i.i.d. $N(0, \sigma_v^2)$; and ε_t and v_t are stochastically independent.
 - $p \lim_{T \rightarrow \infty} \hat{\beta} = \frac{\beta}{1 + \sigma_v^2 / a}$, where $a = \text{plim } T^{-1} \sum_t (x_t^*)^2$. (Greene)
 - $p \lim_{T \rightarrow \infty} |\hat{\beta}| < |\beta|$ if $\sigma_v^2 > 0$ and $\text{plim } \hat{\beta} = \beta$ only if $\sigma_v^2 = 0$.

- General result:
 - $y = X\beta + \varepsilon$, where some variables in X are correlated with ε .
 - $\text{plim } T^{-1}X'\varepsilon \neq 0_{k \times 1}$.
 - $\hat{\beta} = \beta + (X'X)^{-1}X'\varepsilon$.
 - $\text{plim}_{T \rightarrow \infty} \hat{\beta} = \beta + \text{plim}_{T \rightarrow \infty} \left(\frac{1}{T} X'X \right)^{-1} \left(\frac{1}{T} X'\varepsilon \right)$

$$= \beta + \left(p \lim_{T \rightarrow \infty} \frac{1}{T} X'X \right)^{-1} \left(p \lim_{T \rightarrow \infty} \frac{1}{T} X'\varepsilon \right)$$

$$= \beta + Q_o^{-1} \left(p \lim_{T \rightarrow \infty} \frac{1}{T} X'\varepsilon \right) \neq \beta.$$

[2] The Method of Instrumental Variables (IV)

(1) Assumption:

- $y_t = x_{t\bullet}'\beta + \varepsilon_t = x_{t1}\beta_1 + x_{t2}\beta_2 + \dots + x_{tk}\beta_k + \varepsilon_t,$

or, $y = X\beta + \varepsilon.$

- There exists $z_{t\bullet} = (z_{t1}, \dots, z_{tq})'$ ($q \geq k$) such that:

1) $E(\varepsilon_t | \varepsilon_{t-1}, \dots, \varepsilon_1, z_{t\bullet}, \dots, z_{1\bullet}) = 0;$

$[E(\varepsilon_t) = 0, \text{cov}(\varepsilon_t, \varepsilon_s) = 0, \text{ and } E(z_{t\bullet}\varepsilon_t) = 0;$

$\text{plim } T^{-1}Z'\varepsilon = \text{plim } T^{-1}\sum_t z_{t\bullet}\varepsilon_t = 0.]$

2) The $(z_{t\bullet}', x_{t\bullet}')$ are stationary and ergodic;

$[\text{plim}_{T \rightarrow \infty} \frac{1}{T} Z'Z = \text{plim}_{T \rightarrow \infty} \frac{1}{T} \sum_t z_{t\bullet} z_{t\bullet}' \equiv Q_z \text{ is finite and}$

positive definite; $\text{plim}_{T \rightarrow \infty} \frac{1}{T} \sum_t z_{t\bullet} x_{t\bullet}' \text{ is finite.}]$

3) $\text{var}(\varepsilon_t | \varepsilon_{t-1}, \dots, \varepsilon_1, z_{t\bullet}, \dots, z_{1\bullet}) = \sigma^2$ (Homosked. Assum.);

$[\text{var}(\varepsilon_t) = \sigma^2.]$

4) $\text{plim}_{T \rightarrow \infty} \frac{1}{T} Z'X = \text{plim}_{T \rightarrow \infty} \frac{1}{T} \sum_t z_{t\bullet} x_{t\bullet}'$ is finite, nonzero, with Rank

= k. [All regressors in $x_{t\bullet}$ are correlated with at least one in $z_{t\bullet}$.]

(2) How to Choose IV's:

[CASE I] Simple Model:

- $y_t = \beta_1 + \beta_2 x_{t2} + \varepsilon_t$ where $E(x_{t2}\varepsilon_t) \neq 0$.
- Suppose that a single variable z_{t2} is uncorrelated with ε_t ; that is, $E(z_{t2}\varepsilon_t) = 0$.
 - When a variable, say Z , is uncorrelated with the error term ε , the variable Z is called **exogenous**.
 - Conversely, if the variable Z is correlated with ε , it is called **endogenous**.
- Two major conditions for a valid instrumental variable z_{t2} :
 - Instrument Exogeneity: $E(z_{t2}\varepsilon_t) = \text{cov}(z_{t2}, \varepsilon_t) = 0$.
 - Instrument Relevance: $\text{cov}(z_{t2}, x_{t2}) \neq 0$.

[CASE 2] A more general model:

- $y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \dots + \beta_k x_{tk} + \varepsilon_t$,
where $E(x_{tj}\varepsilon_t) = 0$ for $j = 3, \dots, k$ and $E(x_{t2}\varepsilon_t) \neq 0$.
- Suppose that m variable h_{t1}, \dots, h_{tm} are uncorrelated with ε ; that is, $E(h_{tj}\varepsilon_t) = 0$ for $j = 1, \dots, m$.
 - When $m = 1$, the model is said to be exactly identified.
 - When $m > 1$, the model is said to be overidentified.

- Two conditions for valid instrumental variables h_{t1}, \dots, h_{tm} :
 - Instrument Exogeneity: $E(h_{tj}\varepsilon_t) = 0$ for all $j = 1, \dots, m$.
 - Instrument Relevance:
 - $\text{cov}(h_{tj}, x_{t2} | x_{t3}, \dots, x_{tk}) \neq 0$, for some j :

When you regress by OLS the model

$$x_{t2} = \pi + \gamma_1 h_{t1} + \dots + \gamma_m h_{tm} + \delta_3 x_{t3} + \dots + \delta_k x_{tk} + v_{it}$$

you can reject $H_0: \gamma_1 = \dots = \gamma_m = 0$ by the F-test.

(3) IV estimator ($q = k$):

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

Theorem:

$$\text{plim}_{T \rightarrow \infty} \hat{\beta}_{IV} = \beta;$$

$$\sqrt{T}(\hat{\beta}_{IV} - \beta) \rightarrow_d N(0_{k \times 1}, \sigma^2 p \lim_{T \rightarrow \infty} T(Z'X)^{-1}Z'Z(X'Z)^{-1});$$

$$\text{plim}_{T \rightarrow \infty} s_{IV}^2 = \sigma^2, \text{ where } s_{IV}^2 = \frac{(y - X\hat{\beta}_{IV})'(y - X\hat{\beta}_{IV})}{T - k}.$$

Implication:

- IV estimator is consistent.
- $\hat{\beta}_{IV} \sim N(\beta, \hat{\Omega})$, where $\hat{\Omega} = (Z'X)^{-1}s_{IV}^2Z'Z(X'Z)^{-1}$.
- We can use t or Wald statistics to test hypotheses regarding β .

$$[\text{e.g., } t = \frac{(R\hat{\beta}_{IV} - r)}{\sqrt{R\hat{\Omega}R'}}, W_T = (R\hat{\beta}_{IV} - r)'(R\hat{\Omega}R')^{-1}(R\hat{\beta}_{IV} - r).]$$

Example: (Cross-section data)

$$C_t = \beta_1 + \beta_2 I_t^* + \varepsilon_t; I_t = I_t^* + v_t,$$

where C = consumption expenditure and I_t = income.

$$\rightarrow y_t = C_t; x_{t\bullet} = (1, I_t)'; z_{t\bullet} = (1, ED_t)'$$

Example 2: (Time-series data)

$$INV_t = \beta_1 + \beta_2 GNP_t + \varepsilon_t,$$

where GNP_t is endogenous, and INV = aggregate investment.

$$\rightarrow y_t = INV_t; x_{t\bullet} = (1, GNP_t)'; z_{t\bullet} = (1, GNP_{t-1})'$$

[Proof of the theorem]

$$\hat{\beta}_{IV} = (Z'X)^{-1}Z'y = (Z'X)^{-1}Z'(X\beta + \varepsilon) = \beta + (Z'X)^{-1}Z'\varepsilon$$

$$\rightarrow \hat{\beta}_{IV} = \beta + \left(\frac{1}{T}Z'X\right)^{-1} \left(\frac{1}{T}Z'\varepsilon\right)$$

$$\rightarrow \text{plim}_{T \rightarrow \infty} \hat{\beta}_{IV} = \beta + \left(\text{plim}_{T \rightarrow \infty} \frac{1}{T}Z'X\right)^{-1} \left(\text{plim}_{T \rightarrow \infty} \frac{1}{T}Z'\varepsilon\right).$$

$$\sqrt{T}(\hat{\beta}_{IV} - \beta) = \left(\frac{1}{T} Z'X \right)^{-1} \frac{1}{\sqrt{T}} Z'\varepsilon.$$

→ By the central limit theorem and the given assumptions,

$$\frac{1}{\sqrt{T}} Z'\varepsilon \rightarrow_d N\left(0, \sigma^2 p \lim_{T \rightarrow \infty} \frac{1}{T} Z'Z\right).$$

$$\rightarrow \left(\frac{1}{T} Z'X \right)^{-1} \frac{1}{\sqrt{T}} Z'\varepsilon$$

$$\rightarrow_d N\left(0, \sigma^2 \left(p \lim_{T \rightarrow \infty} \frac{1}{T} Z'X \right)^{-1} \left(p \lim_{T \rightarrow \infty} \frac{1}{T} Z'Z \right) \left(p \lim_{T \rightarrow \infty} \frac{1}{T} X'Z \right)^{-1} \right).$$

(4) Two-Stage Least Squares (2SLS) ($q > k$)

$$\hat{\beta}_{2SLS} = [X'P(Z)X]^{-1} X'P(Z)y, \text{ where } P(Z) = Z(Z'Z)^{-1}Z'.$$

Note 1:

$$\text{If } q = k, \hat{\beta}_{IV} = \hat{\beta}_{2SLS}.$$

Note 2: (Why is 2SLS called 2SLS?)

- $y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + \dots + \beta_k x_{tk} + \varepsilon_t$.
- Suppose x_{t2} contains measurement errors. Then, your instrument set would look like $z_{t\bullet} = (x_{t1}, h_{t1}, h_{t2}, \dots, h_{tm}, x_{t3}, \dots, x_{tk})'$.
- Regress x_{t2} on $z_{t\bullet}$. Then, get fitted values \hat{x}_{t2} .
- Estimate β 's by regressing the model:

$$y_t = \beta_1 x_{t1} + \beta_2 \hat{x}_{t2} + \beta_3 x_{t3} + \dots + \beta_k x_{tk} + \text{error}_t.$$
- Use 2SLS residuals $(y_t - \hat{\beta}_1 x_{t1} - \hat{\beta}_2 x_{t2} - \dots - \hat{\beta}_k x_{tk})$ to estimate σ^2 .

Theorem:

$$\text{plim}_{T \rightarrow \infty} \hat{\beta}_{2SLS} = \beta;$$

$$\sqrt{T}(\hat{\beta}_{2SLS} - \beta) \rightarrow_d N(0_{k \times 1}, \sigma^2 p \lim_{T \rightarrow \infty} T(X'P(Z)X)^{-1});$$

$$\text{plim}_{T \rightarrow \infty} s_{2SLS}^2 = \sigma^2, \text{ where } s_{2SLS}^2 = \frac{(y - X \hat{\beta}_{2SLS})'(y - X \hat{\beta}_{2SLS})}{T - k}.$$

[3] Hausman Test for Measurement Error

[Hausman, 1978, ECONOMETRICA]

H_0 : No measurement error in regressors.

(1) Intuition

Under H_0 , both $\hat{\beta}$ and $\hat{\beta}_{IV}$ are consistent. And $\hat{\beta}$ is more efficient.

Under H_a , only $\hat{\beta}_{IV}$ is consistent.

(2) Hausman Statistic

- Under H_0 ,

$$H_T = (\hat{\beta}_{IV} - \hat{\beta})' [Cov(\hat{\beta}_{IV}) - Cov(\hat{\beta})]^+ (\hat{\beta}_{IV} - \hat{\beta}) \rightarrow_d \chi^2(df),$$

where $df = \#$ of regressors with measurement errors.

- You can use $\hat{\beta}_{2SLS}$ instead.

DIGRESSION TO GENERALIZED G-INVERSE

1) g-inverse

A: a square matrix. Then, $A^-AA^- = A^-$.

2) Moore-Penrose g-inverse

$A^+AA^+ = A^+$; $AA^+A = A$; AA^+ and A^+A are symmetric.

END OF DIGRESSION

(3) Alternative Hausman test.

1) Case in which only one regressor (say g) has measurement error. [$Z = [X^g, h_1, \dots, h_m]$, where X^g includes all regressors except g .]

STEP 1: Do OLS on $g = Z\gamma + \text{error}$, and get the residual vector \hat{u}

STEP 2: Do OLS on $y = X\beta + \hat{u}\xi + \text{error}$ and do t-test for $H_0: \xi = 0$.

→ This alternative test is asymptotically identical to H_T .

- Another alternative procedure (Ahn's test, Oxford, 1997):

STEP 1: Do OLS on $y = X\beta + H\lambda + \text{error}$ and do F-test (χ^2 test) for

$H_0: \lambda = 0_{m \times 1}$, where $H = (h_1, \dots, h_m)$.

→ If $m = 1$, this test is asymptotically identical to H_T .

2) Case in which multiple regressors (say G) have measurement errors.

[$Z = [X^G, h_1, \dots, h_m]$, where X^G includes regressors other than $G = (g_1, \dots, g_r)$ ($T \times r$).]

STEP 1: Do OLS on $g_1 = Z\gamma + \text{error}$, and get the residual vector, \hat{u}_1 .

Do OLS on $g_2 = Z\gamma + \text{error}$, and get the residual vector, \hat{u}_2 .

Do OLS on $g_r = Z\gamma + \text{error}$, and get the residual vector, \hat{u}_r .

[Let $\hat{U} = (\hat{u}_1, \dots, \hat{u}_r)$.]

STEP 2: Do OLS on $y = X\beta + \hat{U}\xi + \text{error}$ and do F or χ^2 tests for H_0 :
 $\xi = 0_{r \times 1}$.

→ This alternative test is asymptotically identical to H_T .

- Another alternative procedure (Ahn's test, Oxford, 1997):

STEP 1: Do OLS on $y = X\beta + H\lambda + \text{error}$ and do F or χ^2 tests for H_0 :
 $\lambda = 0_{m \times 1}$.

→ If $m = r$, this test is asymptotically identical to H_T .

[EXAMPLE]

- Data: (WAGE2.WF1 or WAGE2.TXT – from Wooldridge's website)

| | |
|------------------------|-------------------------------|
| # of observations (T): | 935 |
| 1. wage | monthly earnings |
| 2. hours | average weekly hours |
| 3. IQ | IQ score |
| 4. KWW | knowledge of world work score |
| 5. educ | years of education |
| 6. exper | years of work experience |
| 7. tenure | years with current employer |
| 8. age | age in years |
| 9. married | =1 if married |
| 10. black | =1 if black |
| 11. south | =1 if live in south |
| 12. urban | =1 if live in SMSA |
| 13. sibs | number of siblings |
| 14. brthord | birth order |
| 15. meduc | mother's education |
| 16. feduc | father's education |
| 17. lwage | natural log of wage |

OLS Results:

Dependent Variable: LOG(HOURS)

Method: Least Squares

Sample: 1 935

Included observations: 935

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|--------------------|-------------|-----------------------|-------------|--------|
| C | 3.825292 | 0.090201 | 42.40863 | 0.0000 |
| AGE | 0.001574 | 0.001628 | 0.966675 | 0.3340 |
| LWAGE | -0.015767 | 0.012012 | -1.312559 | 0.1897 |
| R-squared | 0.002469 | Mean dependent var | 3.770463 | |
| Adjusted R-squared | 0.000329 | S.D. dependent var | 0.152592 | |
| S.E. of regression | 0.152567 | Akaike info criterion | -0.91922 | |
| Sum squared resid | 21.69393 | Schwarz criterion | -0.90369 | |
| Log likelihood | 432.7353 | F-statistic | 1.153519 | |
| Durbin-Watson stat | 1.978186 | Prob(F-statistic) | 0.315974 | |

Measurement Error in LWAGE?

Or LWAGE correlated with ε ?

IV ESTIMATION (USING TSLS in EViews)

Dependent Variable: LOG(HOURS)

Method: Two-Stage Least Squares

Sample: 1 935

Included observations: 935

Instrument list: C AGE EDUC

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|--------------------|-------------|--------------------|-------------|----------|
| C | 3.034888 | 0.249017 | 12.18748 | 0.0000 |
| AGE | -0.001290 | 0.001918 | -0.672376 | 0.5015 |
| LWAGE | 0.114801 | 0.040057 | 2.865968 | 0.0043 |
| R-squared | -0.123988 | Mean dependent var | | 3.770463 |
| Adjusted R-squared | -0.126400 | S.D. dependent var | | 0.152592 |
| S.E. of regression | 0.161949 | Sum squared resid | | 24.44407 |
| F-statistic | 4.366135 | Durbin-Watson stat | | 2.015547 |
| Prob(F-statistic) | 0.012961 | | | |

<HAUSMAN TEST>

Dependent Variable: LWAGE

Method: Least Squares

Sample: 1 935

Included observations: 935

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|--------------------|-------------|-----------------------|-------------|--------|
| C | 5.225153 | 0.159896 | 32.67855 | 0.0000 |
| AGE | 0.022450 | 0.004153 | 5.406114 | 0.0000 |
| EDUC | 0.060228 | 0.005875 | 10.25107 | 0.0000 |
| R-squared | 0.124860 | Mean dependent var | 6.779004 | |
| Adjusted R-squared | 0.122982 | S.D. dependent var | 0.421144 | |
| S.E. of regression | 0.394398 | Akaike info criterion | 0.980292 | |
| Sum squared resid | 144.9725 | Schwarz criterion | 0.995823 | |
| Log likelihood | -455.2863 | F-statistic | 66.48613 | |
| Durbin-Watson stat | 1.805028 | Prob(F-statistic) | 0.000000 | |

Get RES (residuals) using the Genr command.

Dependent Variable: LOG(HOURS)
 Method: Least Squares
 Sample: 1 935
 Included observations: 935

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|--------------------|-------------|-----------------------|-------------|--------|
| C | 3.034888 | 0.233033 | 13.02340 | 0.0000 |
| AGE | -0.001290 | 0.001795 | -0.718494 | 0.4726 |
| LWAGE | 0.114801 | 0.037486 | 3.062542 | 0.0023 |
| RES | -0.145289 | 0.039542 | -3.674268 | 0.0003 |
| R-squared | 0.016727 | Mean dependent var | 3.770463 | |
| Adjusted R-squared | 0.013559 | S.D. dependent var | 0.152592 | |
| S.E. of regression | 0.151554 | Akaike info criterion | -0.93148 | |
| Sum squared resid | 21.38385 | Schwarz criterion | -0.91077 | |
| Log likelihood | 439.4658 | F-statistic | 5.279410 | |
| Durbin-Watson stat | 1.982304 | Prob(F-statistic) | 0.001298 | |

Reject H_0 : No measurement error nor endogeneity of wage.

AHN's TEST

Dependent Variable: LOG(HOURS)

Method: Least Squares

Sample: 1 935

Included observations: 935

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|--------------------|-------------|-----------------------|-------------|--------|
| C | 3.794048 | 0.090005 | 42.15397 | 0.0000 |
| AGE | 0.001972 | 0.001621 | 1.216923 | 0.2239 |
| LWAGE | -0.030488 | 0.012587 | -2.422189 | 0.0156 |
| EDUC | 0.008751 | 0.002382 | 3.674268 | 0.0003 |
| R-squared | 0.016727 | Mean dependent var | 3.770463 | |
| Adjusted R-squared | 0.013559 | S.D. dependent var | 0.152592 | |
| S.E. of regression | 0.151554 | Akaike info criterion | -0.93148 | |
| Sum squared resid | 21.38385 | Schwarz criterion | -0.91077 | |
| Log likelihood | 439.4658 | F-statistic | 5.279410 | |
| Durbin-Watson stat | 1.982304 | Prob(F-statistic) | 0.001298 | |

Reject H_0 : No measurement error nor endogeneity of wage.

How can we check the quality of instruments?

(1) Regress the regressor on the instrumental variables.

Dependent Variable: LWAGE
 Method: Least Squares
 Sample: 1 935
 Included observations: 935

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|--------------------|-------------|-----------------------|-------------|----------|
| C | 5.225153 | 0.159896 | 32.67855 | 0.0000 |
| AGE | 0.022450 | 0.004153 | 5.406114 | 0.0000 |
| EDUC | 0.060228 | 0.005875 | 10.25107 | 0.0000 |
| R-squared | 0.124860 | Mean dependent var | | 6.779004 |
| Adjusted R-squared | 0.122982 | S.D. dependent var | | 0.421144 |
| S.E. of regression | 0.394398 | Akaike info criterion | | 0.980292 |
| Sum squared resid | 144.9725 | Schwarz criterion | | 0.995823 |
| Log likelihood | -455.2863 | F-statistic | | 66.48613 |
| Durbin-Watson stat | 1.805028 | Prob(F-statistic) | | 0.000000 |

(2) Then, test H_0 : The coefficients on the instruments that are not regressors in the original model equal zeros.

Wald Test:
 Equation: Untitled

| Null Hypothesis: C(3)=0 | | | |
|-------------------------|----------|-------------|----------|
| F-statistic | 105.0844 | Probability | 0.000000 |
| Chi-square | 105.0844 | Probability | 0.000000 |

(3) If the F test rejects the null hypothesis and the value of the statistic is smaller than 10, the instrumental variables may not be a high-quality instrument. (Staiger and Stock, *Econometrica*, 1997).

Testing Instrument Exogeneity.

- Are instrumental variables exogenous?
- How can we check whether instruments are exogenous?
 - If $q = k$, it is not possible to test.
 - If $q > k$, it is possible.

- Test Procedure:

STEP 1: Do TSLS and get residuals, $\hat{\varepsilon} = y - X\hat{\beta}_{2SLS}$.

STEP 2: Do OLS on

$$\hat{\varepsilon} = Z\gamma + error,$$

and get R^2 .

STEP 3: J-statistic = $T \times R^2$. If J-stat $> c$ from $\chi^2(q-k)$, we can conclude that some of instruments are endogenous. If J-stat $< c$, we can conclude that no instrument is endogenous.