Single-Equation GMM: Endogeneity Bias

Lecture for Economics 241B

Douglas G. Steigerwald

UC Santa Barbara

January 2012

Initial Question

How valuable is investment in college education?

- economics - measure value in terms of wage
- How would you determine the return on investment in college education?

Stochastic Model

What are the returns to a college education?

- Random variables of interest
  - W - log of worker wage
  - S - years of schooling
  - A - age
  - M - indicator for male
  - R - indicator for white
  - U - other factors that affect wages

- Stochastic Model

$$W = \beta_0 + \beta_1 S + \beta_2 A + \beta_3 A^2 + \beta_4 M + \beta_5 R + U$$

Estimates

Stochastic Model

$$W = \beta_0 + \beta_1 S + \beta_2 A + \beta_3 A^2 + \beta_4 M + \beta_5 R + U$$

- $\hat{\beta}_1$
  - .084 (and significantly different from zero)
  - each additional year of schooling is worth an additional 8.4% in wages
  - 4 years of college would increase wages by 38% (1.084^4)
- the median full time worker earns about \$550 per week in 2000
  - wage increase of 38% is \$210 per week
- over 30 year work-life, earnings increase by \$170,000 in present value (5% interest), makes public universities a good deal

Potential Endogeneity

$$W = \beta_0 + \beta_1 S + \beta_2 A + \beta_3 A^2 + \beta_4 M + \beta_5 R + U$$

- β_1 may not capture a causal impact on wages
  - workers who obtained more education may have attributes that would have led to higher earnings even without additional education
  - S is endogenous $\Rightarrow Cov\left(S, U\right) \neq 0$
  - $\hat{\beta}_1$ is biased and inconsistent
- What is the direction of bias in $\hat{\beta}_1$?

Potential Endogeneity

$$W = \beta_0 + \beta_1 S + \beta_2 A + \beta_3 A^2 + \beta_4 M + \beta_5 R + U$$

- β_1 may not capture a causal impact on wages
  - workers who obtained more education may have attributes that would have led to higher earnings even without additional education
  - S is endogenous $\Rightarrow Cov\left(S, U\right) \neq 0$
  - $\hat{\beta}_1$ is biased and inconsistent
- What is the direction of bias in $\hat{\beta}_1$?
- $\hat{\beta}_1$ is biased upward, does not provide a helpful bound to argue for benefits of education

Sources of Covariate-Error Correlation

Focus on $Cor(S, U)$

- Endogeneity
  - workers who would otherwise have high wage rates are more likely to obtain higher education
  - $U = \mu + e$ μ - ability e - random shock
  - $S = \alpha\mu + w$
    - describes how schooling is correlated with ability
  - in this application, likely that $Cor(S, U) > 0$
- Measurement Error
  - $S = S^* + v$
    - S^* - actual schooling S - reported v - measurement error
  - in all applications, $Cor(S, U) < 0$

Detail: Measurement Error Correlation

Simplify

$$
\begin{aligned}
\text{population model} \quad W_t &= \beta S_t^* + U_t \qquad S_t = S_t^* + v_t \\
\text{estimated model} \quad W_t &= \beta S_t + (U_t - \beta v_t)
\end{aligned}
$$

- v_t is a component of $S_t \Rightarrow Cor\left[S_t, (U_t - \beta v_t)\right] < 0$
- $$\hat{\beta} = \beta + \frac{\sum_{t=1}^{n} S_t \left[U_t - \beta V_t\right]}{\sum_{t=1}^{n} S_t^2}$$
- in large samples $\hat{\beta}$ tends to
$$\beta\left[1 - \frac{Cov\left(S_t, v_t\right)}{Var\left(S_t\right)}\right] = \beta\left[\frac{Var\left(S_t^*\right)}{Var\left(S_t^*\right) + Var\left(v_t\right)}\right]$$
    - where $Var\left(S_t\right) = Var\left(S_t^*\right) + Var\left(v_t\right)$ $Cov\left(S_t, v_t\right) = Var\left(v_t\right)$
- Iron Law of Econometrics - measurement error leads to attenuation bias

Solutions

- Instrument
  - z is a (valid) instrument if
    - $Cov\,(S, z) \neq 0$ and $Cov\,(U, z) = 0$
  - instruments can address both sources of covariate-error correlation
  - issue - instruments can be difficult to find
- Measurement error assumption
  - $S = S^* + v$ assumptions regard v
    - example: v is symmetric around 0
  - issue - does not address endogeneity

Instrument Solutions

- Standard Instrument Solution
  - implicit model of endogeneity
    - no specified model linking endogenous covariates to error
  - yields classic instrumental variable (IV) estimator
- Model-Based Selection (Endogeneity) Correction
  - explicit model of endogeneity
    - clearly specified model linking endogenous covariates to error
  - yields selection-corrected IV estimator

Standard Instrument Solution : Identification

- $\underset{(K\times 1)}{X}$ covariate vector $\underset{(L\times 1)}{Z}$ instrument vector
- Identification Assumption (Rank Condition)

  The $L \times K$ matrix $\mathbb{E}\left(ZX^{\mathrm{T}}\right)$ has rank K.

- Example $X^{\mathrm{T}} = (1, S)$ $Z^{\mathrm{T}} = (1, z)$

$$\mathbb{E}\left(ZX^{\mathrm{T}}\right) = \begin{bmatrix} 1 & \mathbb{E}(S) \\ \mathbb{E}(z) & \mathbb{E}(Sz) \end{bmatrix}$$

- Rank is K if determinant is not zero $\Leftrightarrow$ $Cov(S, z) \neq 0$

Identification

- Identification Assumption (Order Condition)

  There are at least as many instruments as endogenous covariates: $L \geq K$.

- Over identification
  - rank condition satisfied and $L > K$
- Exact identification
  - rank condition satisfied and $L = K$
- No identification
  - $L < K$ (rank condition cannot hold)

Selection (Endogeneity) Correction

- Key - construct $\mathbb{E}[U|X, Z]$
  - add to regression, remaining error uncorrelated with covariates

Wage Regression Application

- data on twins (indexed by i) who share family characteristics
- Selection (Endogeneity) Model

$$U_i = \mu + \varepsilon_i \qquad \mu = \gamma S_1 + \gamma S_2 + \omega$$

  - μ - latent family characteristics, correlated with S
  - could relax assumption that γ is constant (use equation for twin 1 to identify γ_2)
    - γ - selection effect : $\gamma > 0 \Rightarrow$ families that would otherwise have high wages are more likely to educate their children

Selection Correction

- wage regression (twin 1 $C_1' = (A_1, A_1^2, M_1, R_1)$)

$$\begin{aligned} W_1 &= \beta_0 + \beta_1 S_1 + C_1'\delta + (\mu + \varepsilon_1) \\ &= \beta_0 + \beta_1 S_1 + C_1'\delta + (\gamma S_1 + \gamma S_2 + \omega + \varepsilon_1) \end{aligned}$$

- identification assumption
  - $\mathbb{E}[U_1 | X, Z] = \gamma S_1 + \gamma S_2$
- selection-corrected regression

$$W_1 = \beta_0 + (\beta_1 + \gamma) S_1 + \gamma S_2 + C_1'\delta + (\varepsilon_1 + \omega)$$

-

Variable	OLS	Include S_2
Own education	0.084 (0.014)	0.088 (0.015)
Sibling's education	-	-0.007 (0.015)

- Twins data - endogeneity bias is negative!

Review

Stochastic Model

$$W_i = \beta_0 + \beta_1 S_i + U_i$$

- What two issues lead to correlation between S_i and U_i?

Review

Stochastic Model

$$W_i = \beta_0 + \beta_1 S_i + U_i$$

- What two issues lead to correlation between S_i and U_i?
  - 1) endogeneity - latent ability that impacts both S_i and U_i

Review

Stochastic Model

$$W_i = \beta_0 + \beta_1 S_i + U_i$$

- What two issues lead to correlation between S_i and U_i?
  - 1) endogeneity - latent ability that impacts both S_i and U_i
  - 2) measurement error in S_i

Review

Stochastic Model

$$W_i = \beta_0 + \beta_1 S_i + U_i$$

- What two issues lead to correlation between S_i and U_i?
  - 1) endogeneity - latent ability that impacts both S_i and U_i
  - 2) measurement error in S_i
- What is the impact of the correlation on B_{OLS}?

Review

Stochastic Model

$$W_i = \beta_0 + \beta_1 S_i + U_i$$

- What two issues lead to correlation between S_i and U_i?
  - 1) endogeneity - latent ability that impacts both S_i and U_i
  - 2) measurement error in S_i
- What is the impact of the correlation on B_{OLS}?
  - biased and inconsistent

Review

Stochastic Model

$$W_i = \beta_0 + \beta_1 S_i + U_i$$

- What two issues lead to correlation between S_i and U_i?
  - 1) endogeneity - latent ability that impacts both S_i and U_i
  - 2) measurement error in S_i
- What is the impact of the correlation on B_{OLS}?
  - biased and inconsistent
- What is needed to construct a consistent estimator?

Review

Stochastic Model

$$W_i = \beta_0 + \beta_1 S_i + U_i$$

- What two issues lead to correlation between S_i and U_i?
  - 1) endogeneity - latent ability that impacts both S_i and U_i
  - 2) measurement error in S_i
- What is the impact of the correlation on B_{OLS}?
  - biased and inconsistent
- What is needed to construct a consistent estimator?
  - 1) instrument Z_i $\quad Cor\left(Z_i, S_i\right) \neq 0 \quad Cor\left(Z_i, U_i\right) = 0$

Review

Stochastic Model

$$W_i = \beta_0 + \beta_1 S_i + U_i$$

- What two issues lead to correlation between S_i and U_i?
  - 1) endogeneity - latent ability that impacts both S_i and U_i
  - 2) measurement error in S_i
- What is the impact of the correlation on B_{OLS}?
  - biased and inconsistent
- What is needed to construct a consistent estimator?
  - 1) instrument Z_i $\quad Cor(Z_i, S_i) \neq 0 \quad Cor(Z_i, U_i) = 0$
  - 2) assumption about measurement error