Econometrics 2

Arieda Muço

Central European University

Spring 2019

Measurement Error

Case 1: Measurement Error in Y_i

- True relationship

$$Y_i^* \;=\; \alpha + X_i'\beta + u_i$$

- But Y_i^* is not observed
- We observe Y_i measured with error

$$Y_i \;=\; Y_i^* + \epsilon_i$$

- Measurement error assumptions
  - $\epsilon_i \sim iid\left(0, \sigma_\epsilon^2\right)$
  - $Cov\left(X_i, \epsilon_i\right) = 0$
- Estimated Model:

$$Y_i \;=\; \alpha + X_i'\beta + \underbrace{\left(u_i + \epsilon_i\right)}_{v_i}$$

Note: $Cov\left(v_i, X_i\right) = 0$

- OLS estimator of β is unbiased
- Error variance: $Var\left(u_i + \epsilon_i\right) = \sigma_u^2 + \sigma_\epsilon^2$

Classical Measurement Error (CME)

Case 2: Measurement Error in regressor

- Consider a bivariate model without constant. We are interested in the effect of education on earnings:

$$\begin{aligned} Y_i &= \rho s_i^* + u_i \\ \text{where} \quad & Cov(s_i^*, u_i) = 0 \end{aligned}$$

- s_i^* true level of schooling
- We observe $s_i = s_i^* + \epsilon_i$
  - $\epsilon_i \sim iid\left(0, \sigma_\epsilon^2\right)$
  - $Cov\left(\epsilon_i, s_i^*\right) = 0$

Classical Measurement Error

- We don't observe s_i^*

$$
\begin{aligned}
Y_i &= \rho s_i^* + u_i \\
&= \rho(s_i - \epsilon_i) + u_i = \rho s_i \underbrace{-\rho\epsilon_i + u_i}_{\tilde{u}_i} \\
Y_i &= \rho s_i + \tilde{u}_i
\end{aligned}
$$

Measurement error bias

Estimating the model with OLS gives

$$
\begin{aligned}
\tilde{\rho} &= \frac{Cov\left(Y_i, s_i\right)}{Var\left(s_i\right)} \\
&= \frac{Cov\left(\rho s_i + \tilde{u}_i, s_i\right)}{Var\left(s_i\right)} \\
&= \rho + \frac{Cov\left(s_i, \tilde{u}_i\right)}{Var\left(s_i\right)} \\
&= \rho - \rho \frac{\sigma_\epsilon^2}{Var\left(s_i\right)}
\end{aligned}
$$

Note that

$$
\begin{aligned}
Cov\left(s_i, \tilde{u}_i\right) &= Cov\left(s_i, u_i - \rho\epsilon_i\right) = Cov\left(s_i, -\rho\epsilon_i\right) \\
&= -\rho Cov\left(s_i, \epsilon_i\right) = -\rho Cov\left(s_i^* + \epsilon_i, \epsilon_i\right) \\
&= -\rho\sigma_\epsilon^2
\end{aligned}
$$

Attenuation Bias

$$\tilde{\rho} = \rho - \rho \frac{\sigma_\epsilon^2}{\sigma_s^2} = (1 - \lambda)\, \rho$$

$$\lambda = \frac{\sigma_\epsilon^2}{\sigma_s^2} = \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_{s*}^2}$$

- λ noise to signal ratio
- 1-λ is the reliability ratio or signal-to-total variance ratio

$$1 - \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_{s*}^2} = \frac{\sigma_{s*}^2}{\sigma_{s*}^2 + \sigma_\epsilon^2} = \frac{Var(s^*)}{Var(s)}$$

Since $0 < \lambda < 1$ the coefficient ρ will be biased towards zero. This bias is therefore called attenuation bias.

Multivariate Model

$$\begin{aligned} Y_i &= \rho s_i^* + \beta X_i + u_i \\ s_i &= s_i^* + \epsilon_i \end{aligned}$$

Classical ME: $Cov\left(s_i^*, \epsilon_i\right) = 0$, $Cov\left(X_i, \epsilon_i\right) = 0$

Attenuation Bias

$$\rho_b = \rho \frac{\sigma_{\tilde{s^*}}^2}{\sigma_{\tilde{s^*}}^2 + \sigma_\epsilon^2}$$

Note that

$$Y_i = \rho s_i^* + \beta X_i + u_i$$

From Regression Anatomy Formula we know that

$$\rho = \frac{Cov(Y_i, \tilde{s_i^*})}{Var(\tilde{s_i^*})}$$

Where $\tilde{s_i^*}$ is the error from a regression of s_i^* on X_i.
Replacing s_i^* with s_i we have

$$\begin{aligned} Y_i &= \rho s_i + \beta X_i + u_i - \rho \epsilon_i \\ \rho_b &= \frac{Cov(Y_i, \tilde{s}_i)}{Var(\tilde{s}_i)} \end{aligned}$$

Where $\tilde{s}_i$ is the error from a regression of s_i on X_i.

Note that

Under CME ϵ_i is uncorrelated with the covariate, X_i. Then the coefficient from a regression of mismeasured s_i on X_i is the same as the coefficient from a regression of s_i^* on X_i.
Hence

$$\begin{aligned} \tilde{s}_i &= \tilde{s_i^*} + \epsilon \\ s_i - \beta X_i &= s_i^* - \beta X_i + \epsilon \\ Var(\tilde{s}_i) &= Var(\tilde{s_i^*}) + Var(\epsilon) \end{aligned}$$

Attenuation bias in the multivariate case exacerbates the measurement error problem

$$\begin{aligned} \rho_b &= \frac{Cov(Y_i, \tilde{s}_i)}{Var(\tilde{s}_i)} = \rho \frac{Var(\tilde{s_i^*})}{Var(\tilde{s}_i)} \\ &= \rho \frac{Var(\tilde{s_i^*})}{Var(\tilde{s_i^*}) + Var(\epsilon)} = \rho \frac{\sigma^2_{\tilde{s^*}}}{\sigma^2_{\tilde{s^*}} + \sigma^2_\epsilon} \end{aligned}$$

Since $Var(\tilde{s_i^*}) < Var(s_i^*)$ implies that $\frac{\sigma^2_{\tilde{s^*}}}{\sigma^2_{\tilde{s^*}} + \sigma^2_\epsilon} < \frac{\sigma^2_{s^*}}{\sigma^2_{s^*} + \sigma^2_\epsilon}$

Ashenfelter and Krueger (AER, 1994)

"Estimates of the Economic Returns to Schooling from a New Sample of Twins"

- Address problems of ability bias and measurement error in schooling
- Sample of twins
  - identical family
  - identical genes
  - can assume identical A_i
- What happens if we can control for ability bias but there is measurement error in s_i?
- Survey data collected at twins' festival in Ohio

Omitted Variables Bias (recap)

- Short Regression:

$$Y_i = \tilde{\alpha} + \tilde{\rho} s_i + \eta_i$$

- Long Regression (ability, etc.):

$$Y_i = \alpha + \rho s_i + A_i^{'} \gamma + \nu_i$$

- CIA applies given A_i.
- $\tilde{\rho}$ estimated coefficient of the linear causal model when ability is omitted:

$$\tilde{\rho} = \frac{Cov\left(Y_i, s_i\right)}{Var\left(s_i\right)} = \rho + \gamma \frac{Cov\left(A_i, s_i\right)}{Var\left(s_i\right)} = \rho + \gamma \delta_{As}$$

- δ_{As} coefficient from regression of A_i on s_i

Variable	Means (standard deviations in parentheses)		
	Identical twins[a]	Fraternal twins[a]	Population[b]
Self-reported education	14.11 (2.16)	13.72 (2.01)	13.14 (2.73)
Sibling-reported education	14.02 (2.14)	13.41 (2.07)	—
Hourly wage	\$13.31 (11.19)	\$12.07 (5.40)	\$11.10 (7.41)
Age	36.56 (10.36)	35.59 (8.29)	38.91 (12.53)
White	0.94 (0.24)	0.93 (0.25)	0.87 (0.34)
Female	0.54 (0.50)	0.48 (0.50)	0.45 (0.50)
Self-employed	0.15 (0.36)	0.10 (0.30)	0.12 (0.32)
Covered by union	0.24 (0.43)	0.30 (0.46)	—
Married	0.45 (0.50)	0.54 (0.50)	0.62 (0.48)
Age of mother at birth	28.27 (6.37)	29.38 (7.05)	—
Twins report same education	0.49 (0.50)	0.43 (0.50)	—
Twins studied together	0.74 (0.44)	0.38 (0.49)	—
Helped sibling find job	0.43 (0.50)	0.24 (0.43)	—
Sibling helped find job	0.35 (0.48)	0.22 (0.41)	—
Sample size	298	92	164,085

[a]Source: Twinsburg Twins Survey, August 1991.
[b]Source: 1990 Current Population Survey (Outgoing Rotation Groups File). Sample includes workers aged

Is there measurement error?

Classical ME assumptions:

$$\begin{aligned} s_k^j &= s_k^* + \epsilon_k^j, \quad j,k = 1,2 \\ Cov(s_k^*, \epsilon_k^j) &= Cov(s_k^*, \epsilon_j^k) = Cov(\epsilon_k^j, \epsilon_j^k) = 0 \end{aligned}$$

To check correlations

$$\begin{aligned} Corr(s_1^1, s_1^2) &= \frac{Var(s_1^*)}{\sqrt{Var(s_1^1)Var(s_1^2)}} \\ &= 1 - \frac{\sigma_\epsilon^2}{\sigma_s^2} = \underbrace{1-\lambda}_{\text{reliability ratio}} \end{aligned}$$

- $Corr(s_1^1, s_1^2) = 0.92$, $Corr(s_2^2, s_2^1) = 0.88$ in Table 2
- 8-12% in the measured variance in schooling levels is error
- Measurement error in parental schooling levels
  $EF = 0.86$, father, $EM = 0.84$

TABLE 2—CORRELATION MATRICES

A. *Identical Twins*

Variable	Y_1	Y_2	S_1^1	S_1^2	S_2^2	S_2^1	E_F^1	E_F^2	E_M^1	E_M^2
Y_1	1.000									
Y_2	0.563	1.000								
S_1^1	0.382	0.168	1.000							
S_1^2	0.375	0.140	0.920	1.000						
S_2^2	0.267	0.272	0.658	0.697	1.000					
S_2^1	0.248	0.247	0.700	0.643	0.877	1.000				
Father's education (E_F^1)	0.155	0.088	0.345	0.266	0.361	0.416	1.000			
Father's education (E_F^2)	0.159	0.091	0.357	0.278	0.320	0.389	0.857	1.000		
Mother's education (E_M^1)	0.102	0.088	0.348	0.343	0.392	0.410	0.614	0.644	1.000	
Mother's education (E_M^2)	0.126	0.087	0.316	0.321	0.322	0.337	0.503	0.579	0.837	1.000

First difference model

- Y_{1i}, log wage twin 1
- Y_{2i}, log wage twin 2
- X_i, variables that vary by family
- s_{1i}, s_{2i} schooling

$$\begin{aligned} Y_{1i} &= \alpha + \rho s_{1i} + X_i^{'}\beta + u_{1i} + A_i^{'}\gamma \\ Y_{2i} &= \alpha + \rho s_{2i} + X_i^{'}\beta + u_{2i} + A_i^{'}\gamma \end{aligned}$$

- Difference:

$$Y_{1i} - Y_{2i} = \rho\left(s_{1i} - s_{2i}\right) + u_{1i} - u_{2i}$$

- Fixed effects estimator, A_i eliminated by differencing

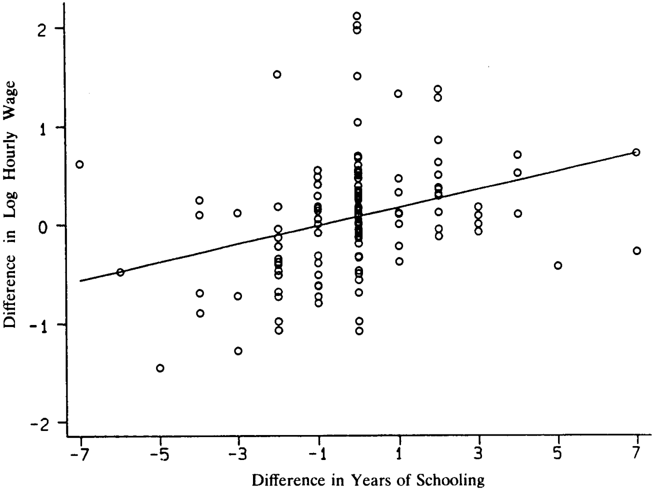

FIGURE 1. INTRAPAIR RETURNS TO SCHOOLING, IDENTICAL TWINS

Measurement error

- Measurement error in s

$$\begin{aligned} s_{1i} &= s_{1i}^{*}+\epsilon_{1i} \\ s_{2i} &= s_{2i}^{*}+\epsilon_{2i} \end{aligned}$$

- Classical ME:

$$\begin{aligned} Cov\left(s_{1i}^{*},\epsilon_{1i}\right) &= 0 \\ Cov\left(s_{2i}^{*},\epsilon_{2i}\right) &= 0 \\ Cov\left(s_{1i}^{*},\epsilon_{2i}\right) &= Cov\left(s_{2i}^{*},\epsilon_{1i}\right)=0 \\ Cov\left(\epsilon_{1i},\epsilon_{2i}\right) &= 0 \end{aligned}$$

Measurement error bias

- Differenced equation:

$$Y_{1i} - Y_{2i} = \rho\left(s^*_{1i} - s^*_{2i}\right) + u_{1i} - u_{2i}$$

- Estimated equation

$$Y_{1i} - Y_{2i} \quad = \quad \rho\left(s_{1i} - s_{2i}\right) + u_{1i} - u_{2i} + \rho\left(\epsilon_{1i} - \epsilon_{2i}\right)$$

- Bias

$$\begin{aligned} \tilde{\rho} &= \rho\left(1 - \frac{\sigma^2_\epsilon}{\sigma^2_{s^*} + \sigma^2_\epsilon}\frac{1}{1 - r_s}\right) \\ &= \rho\left(1 - \lambda\frac{1}{1 - r_s}\right) \end{aligned}$$

- r_s within family correlation in schooling levels
- Differencing takes out a lot of the signal, but not the noise

Approximation of attenuation bias

$$\tilde{\rho} \;=\; \rho\left(1-\lambda\frac{1}{1-r_s}\right)$$

From Table 2:

- $\lambda = 0.1$
- $r_s = 0.66$.
- Bias is $\frac{0.1}{1-0.66} \approx 30\%$

Simple OLS procedure to deal with measurement error

- Use average over multiple education reports $(\frac{s_1^1+s_1^2}{2})-(\frac{s_2^1+s_2^1}{2})$
- Averaging decreases measurement error as a fraction of total variance.

$$\tilde{\rho} = \rho\left(1 - \frac{\lambda}{1-r_s} - \frac{2Var(s_1^* - s_2^*)}{2}\right)$$

Instrumental variables strategy

- Get another measure on s_i^* from an independent source
- Ask *twin 2* about *twin 1* schooling and vice versa
- s_j^k, $j = 1, 2$ and $k = 1, 2$
- s_1^1, s_2^2 self reports
- s_1^2, s_2^1 cross reports
- All measures are highly correlated (see Table 2)

IV Procedure

- Independent measure of schooling as instrument

$$Y_1 - Y_2 = \rho\left(s_1^1 - s_2^2\right) + \underbrace{(u_1 - u_2) + \left(\epsilon_1^1 - \epsilon_2^2\right)}_{v_1 - v_2}$$

- Use independent measures of s_j^* to construct an instrument for $(s_1^1 - s_2^2)$

$$z = (s_1^2 - s_2^1)$$

- Exclusion restriction: z_i uncorrelated with $(v_{1i} - v_{2i})$.
- Relevance: z_i correlated with $(s_{1i}^* - s_{2i}^*)$

TABLE 3—ORDINARY LEAST-SQUARES (OLS), GENERALIZED LEAST-SQUARES (GLS), INSTRUMENTAL-VARIABLES (IV), AND FIXED-EFFECTS ESTIMATES OF LOG WAGE EQUATIONS FOR IDENTICAL TWINS[a]

Variable	OLS (i)	GLS (ii)	GLS (iii)	IV[a] (iv)	First difference (v)	First difference by IV (vi)
Own education	0.084 (0.014)	0.087 (0.015)	0.088 (0.015)	0.116 (0.030)	0.092 (0.024)	0.167 (0.043)
Sibling's education	—	—	−0.007 (0.015)	−0.037 (0.029)	—	—
Age	0.088 (0.019)	0.090 (0.023)	0.090 (0.023)	0.088 (0.019)	—	—
Age squared (÷100)	−0.087 (0.023)	−0.089 (0.028)	−0.090 (0.029)	−0.087 (0.024)	—	—
Male	0.204 (0.063)	0.204 (0.077)	0.206 (0.077)	0.206 (0.064)	—	—
White	−0.410 (0.127)	−0.417 (0.143)	−0.424 (0.144)	−0.428 (0.128)	—	—
Sample size:	298	298	298	298	149	149
R^2:	0.260	0.219	0.219	—	0.092	—

Notes: Each equation also includes an intercept term. Numbers in parentheses are estimated standard errors.

[a]Own education and sibling's education are instrumented for using each sibling's report of the other sibling's education as instruments.

TABLE 4—ESTIMATES USING AVERAGE OF SCHOOLING REPORTS, LOG WAGE EQUATIONS FOR IDENTICAL TWINS

Variable	OLS (i)	GLS (ii)	GLS (iii)	First difference (iv)
Average own education[a]	0.087 (0.015)	0.094 (0.016)	0.098 (0.016)	0.117 (0.026)
Average sibling's education[b]	—	—	−0.017 (0.016)	
Age	0.089 (0.019)	0.091 (0.023)	0.091 (0.023)	—
Age squared (÷100)	−0.088 (0.023)	−0.091 (0.029)	−0.091 (0.029)	—
Male	0.203 (0.063)	0.202 (0.077)	0.208 (0.077)	—
White	−0.406 (0.127)	−0.382 (0.144)	−0.385 (0.144)	—
Sample size:	298	298	298	149
R^2:	0.272	0.223	0.225	0.122

Notes: Each equation also includes an intercept term. Numbers in parentheses are estimated standard errors.

[a]Average own education is equal to $(S_1^1 + S_1^2)/2$.

[b]Average sibling's education is equal to $(S_2^2 + S_2^1)/2$.

TABLE 5—GLS, IV, AND FIXED-EFFECTS ESTIMATES OF AUGMENTED LOG-WAGE EQUATIONS FOR IDENTICAL TWINS

Variable	GLS (i)	GLS (ii)	IV[a] (iii)	First difference (iv)	First difference by IV (v)
Own education	0.105 (0.016)	0.105 (0.016)	0.147 (0.034)	0.091 (0.022)	0.179 (0.041)
Sibling's education	—	−0.008	−0.062 (0.016)	— (0.035)	—
Age	0.082 (0.023)	0.082 (0.023)	0.082 (0.019)	—	—
Age squared (÷100)	−0.094 (0.029)	−0.094 (0.029)	−0.092 (0.024)	—	—
Male	0.147 (0.080)	0.149 (0.081)	0.139 (0.066)	—	—
White	−0.472 (0.143)	−0.482 (0.144)	−0.506 (0.130)	—	—
Covered by union	0.115 (0.072)	0.118 (0.072)	0.153 (0.081)	0.063 (0.090)	0.095 (0.095)
Married	0.089 (0.065)	0.086 (0.065)	0.051 (0.073)	0.142 (0.081)	0.140 (0.086)
Years of tenure	0.025 (0.005)	0.024 (0.005)	0.020 (0.005)	0.028 (0.006)	0.028 (0.006)
Father's education	0.001 (0.014)	0.001 (0.014)	0.006 (0.013)	—	—
Mother's education	0.013 (0.017)	0.015 (0.018)	0.019 (0.017)	—	—
Sample size:	284	284	284	147	147
R^2:	0.320	0.320	—	0.257	—

Notes: Each equation also includes an intercept term. Numbers in parentheses are estimated standard errors.

[a]Own education and sibling's education are instrumented using sibling's report of the other sibling's education as instruments.

Correlated measurement error

Individuals who report upward biased measure of own education may be more likely to report upward biased education for their sibling

$$\rho_v = corr(\epsilon_1^1, \epsilon_2^1) = corr(\epsilon_1^2, \epsilon_2^2) > 0$$

- this implies that s_1^1 and s_2^1 are more strongly correlated than s_1^1 and s_2^2, see Table 2
- the previous IV strategy fails, because

$$cov(s_1^* - s_2^* + (\epsilon_1^1 - \epsilon_2^2), (u_1 - u_2) + (\epsilon_1^2 - \epsilon_2^1)) \neq 0$$

rewrite the model

$$Y_1 - Y_2 = \delta(s_1^1 - s_2^1) + \tilde{\varepsilon}$$

and use $z = s_1^2 - s_2^2$ as an instrument

TABLE 6—OLS AND IV FIRST-DIFFERENCE ESTIMATES OF LOG-WAGE EQUATIONS FOR IDENTICAL TWINS, ASSUMING CORRELATED MEASUREMENT ERRORS

Variable	OLS (i)	IV (ii)	OLS (iii)	IV (iv)
ΔS^*	0.107	0.129	0.112	0.132
	(0.025)	(0.030)	(0.023)	(0.028)
Δ Covered by union	—	—	0.089	0.099
			(0.088)	(0.089)
Δ Married	—	—	0.157	0.160
			(0.080)	(0.080)
Δ Years of tenure	—	—	0.028	0.028
			(0.006)	(0.006)
Sample size:	149	149	147	147
R^2:	0.105	—	0.286	—

Notes: ΔS^* is the difference between sibling 1's report of her (his) own education and her (his) report of sibling 2's education. The instrument used for ΔS^* is ΔS^{**}, the difference between sibling 2's report of sibling 1's education and sibling 2's report of sibling 2's own education. Numbers in parentheses are estimated standard errors.

Book References

- Angrist, Joshua D., and Jörn-Steffen Pischke. Mastering' metrics: The path from cause to effect. Princeton University Press, 2014.
- Wooldridge, Jeffrey M. Introductory econometrics: A modern approach (pages 318-324)