Review: Measurement Error

Qingyuan Fang [1]

[1] Johns Hopkins University

February 9, 2026

Roadmap

- **Part I** Measurement error (ME) in linear regression
    - Classical ME: error in Y vs. error in X
    - Attenuation bias, partial identification
- **Part II** Extensions
    - Non-classical ME
    - Panel data: why differencing may worsen attenuation
- **Part III** Bound et al. (1994)

Setup: true variables vs. observed variables

True (latent) variables: x_i^*, y_i^*, **scalar**
True DGP:

$$y_i^* = \beta x_i^* + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i \mid x_i^*] = 0.$$

Observed variables:

$$x_i = x_i^* + u_i, \qquad y_i = y_i^* + \nu_i.$$

Key question: What happens to OLS when we regress y_i on x_i?

Classical measurement error

We call measurement error **classical** if

$$\mathbb{E}[u_i \mid x_i^*] = 0, \quad \mathbb{C}\mathrm{ov}(u_i, \varepsilon_i) = 0,$$

and similarly, for dependent-variable error,

$$\mathbb{E}[\nu_i \mid y_i^*] = 0, \quad \mathbb{C}\mathrm{ov}(\nu_i, \varepsilon_i) = 0.$$

Interpretation: the error is "noise" *added* on top of the truth, unrelated to truth and the structural disturbance.

Classical ME in the dependent variable (Y)

OLS slope

$$\hat{\beta} = \frac{\sum_i x_i y_i}{\sum_i x_i^2} = \frac{\sum_i x_i(\beta x_i + \varepsilon_i + \nu_i)}{\sum_i x_i^2} = \beta + \frac{\sum_i x_i(\varepsilon_i + \nu_i)}{\sum_i x_i^2}.$$

$\Rightarrow$

$$\mathrm{plim}\hat{\beta} = \beta + \mathbb{C}\mathrm{ov}(x_i, \varepsilon_i) + \mathbb{C}\mathrm{ov}(x_i, \nu_i) = \beta,$$

Unbiased, but may be less efficient (See Ex 4.16 in Hansen's textbook).

Exercise 4.16 (notation is different)

Take the linear homoskedastic CEF

$$Y^* = X'\beta + e, \quad \mathbb{E}[e \mid X] = 0, \quad \mathbb{E}\left[e^2 \mid X\right] = \sigma^2$$

and suppose that instead of Y^*, we observe $Y = Y^* + u$, where u is classical ME with

$$\mathbb{E}[u^2|X] = \sigma_u^2(X)$$

Then,

$$Y = X'\beta + \underbrace{e + u}_{\varepsilon}, \quad \mathbb{V}\mathrm{ar}(\varepsilon|X) = \sigma^2 + \sigma_u^2(X)$$

Classical ME in the independent variable (X)

OLS slope:

$$\hat{\beta} = \frac{\sum_i x_i y_i}{\sum_i x_i^2} = \frac{\sum_i (x_i^* + u_i)(\beta x_i^* + \varepsilon_i)}{\sum_i (x_i^* + u_i)^2}.$$

$\Rightarrow$

$$\text{plim } \hat{\beta} = \frac{\beta \, \mathbb{E}[(x_i^*)^2]}{\mathbb{E}[(x_i^*)^2] + \mathbb{E}[u_i^2]} = \beta \cdot \underbrace{\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_u^2}}_{\text{RR} \in (0,1)}.$$

Attenuation bias: $\text{plim } \hat{\beta} = \text{RR} \cdot \beta$ shrinks toward 0.

Assume $\beta > 0$. Then $\hat{\beta}$ can be viewed as a *lower bound* for β if the sample size is large.

Q: Can you find an upper bound?

Partial-identification

Consider the **reverse regression**:

$$x_i = \gamma y_i + \xi_i.$$

The OLS estimator is

$$\hat{\gamma} = \frac{\sum_i x_i y_i}{\sum_i y_i^2} = \frac{\sum_i (x_i^* + u_i)(\beta x_i^* + \varepsilon_i)}{\sum_i (\beta x_i^* + \varepsilon_i)^2}.$$

Under classical assumptions,

$$\text{plim } \hat{\gamma} = \frac{\beta\,\mathbb{E}[(x_i^*)^2]}{\beta^2\mathbb{E}[(x_i^*)^2] + \mathbb{E}[\varepsilon_i^2]} < \frac{1}{\beta} \quad (\beta > 0).$$

Therefore

$$\text{plim } \hat{\beta} < \beta < \text{plim}\left(\frac{1}{\hat{\gamma}}\right).$$

Non-classical measurement error

In many applications, measurement errors are *not* classical:

- Misreporting correlated with truth (mean reversion, heaping, top-coding, recall bias)
- Constructed variables: e.g. wage = earnings / hours

Matters a lot in survey data!

Further Reading: Course, Textbook

Case 1: u_i correlated with x_i^*

One can show:

$$\text{plim } \hat{\beta} = \underbrace{\frac{\left(\mathbb{E}[(x_i^*)^2] + \mathbb{E}[u_i x_i^*]\right)}{\mathbb{E}[(x_i^*)^2] + 2\mathbb{E}[u_i x_i^*] + \sigma_u^2}}_{1-b_{u\tilde{X}}} \cdot \beta.$$

$$Bias = \text{plim } \hat{\beta} - \beta = -b_{u\tilde{X}} \cdot \beta$$

Implication: if $\mathbb{E}[u_i x_i^*] < 0$, it may alleviate the classical attenuation bias.

Case 2: ν_i correlated with y_i^*

Suppose x_i is measured without error, but $y_i = y_i^* + \nu_i$ and

$$\nu_i = \delta y_i^* + \tilde{\nu}_i, \qquad \tilde{\nu}_i \perp (x_i, y_i^*).$$

Then

$$y_i = (1+\delta)y_i^* + \tilde{\nu}_i = (1+\delta)\beta x_i + (1+\delta)\varepsilon_i + \tilde{\nu}_i.$$

OLS of y_i on x_i yields

$$\text{plim } \hat{\beta} = (1+\delta)\beta.$$

$$Bias = \text{plim } \hat{\beta} - \beta = \delta\beta = b_{v\tilde{x}}$$

General form

$$\hat{\beta} = \left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'(\overbrace{\underbrace{\tilde{X}\beta - u\beta}_{X^*\beta} + \nu}^{Y^*} + \varepsilon)$$

$$= \beta + \left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'(-u\beta + \nu + \varepsilon)$$

$$Bias = \operatorname{plim}\ \hat{\beta} - \beta = -b_{u\tilde{X}}\beta + b_{\nu\tilde{X}}$$

Case 3: Panel data

Suppose $x_{it} = x^*_{it} + \varepsilon_{it}$ with (approximately) i.i.d. measurement errors over time:

$$\Delta x_{it} = (x^*_{it} - x^*_{i,t-1}) + (\varepsilon_{it} - \varepsilon_{i,t-1}), \qquad \mathbb{V}\mathrm{ar}(\Delta\varepsilon) = 2\,\mathbb{V}\mathrm{ar}(\varepsilon).$$

If the true x^* is persistent, $\mathbb{V}\mathrm{ar}(\Delta x^*)$ may be small, so the **signal-to-noise ratio deteriorates**.

Heuristic: differencing increases noise variance but often reduces true variance ⇒ stronger attenuation in Δ regressions.

Bound et al. (1994)

It studies measurement error in labor-market survey variables using a validation design:

- Validation data allow direct study of whether errors correlate with truth and other regressors.

Findings (see Abstract):

- “Individuals’ reports of annual earnings are fairly accurate. Errors are negatively related to true earnings, reducing bias due to measurement error when earnings are used as an independent variable.” See Case 1 above
- “Biases are moderately larger for changes in earnings.” See Case 3 above
- “Earnings per hour are less reliably reported than annual earnings.”
- “Biases in estimating earnings functions are relatively small, but those in labor supply functions may be important.”

Data: PSID Validation Study

- Two-wave panel survey of workers in a single large manufacturing firm.
- 1983 wave: 418 interviews out of 534 potential respondents (78.3% response).
- 1987 wave: reinterviews with 341 of the remaining 1983 sample; 275 in both waves.
- Additional hourly-worker sample in 1987; total 1987 interviews = 492 (79.9% response).
- Validation sources: detailed company payroll records and (for hourly workers) day-level activity records ⇒ treated as the "true" value.

Section III.

For log(earnings), they find moderate measurement noise but also *negative correlation* with true earnings:

- Std dev of error is large (about 2/3 of the std dev of the true variable).
- Variance ratio (= 1 - RR) reported: 0.302.
- Bias when ln(earnings) is an explanatory variable: $b_{u\tilde{X}} = 0.239 < 0.302$ because error is negatively correlated with truth.
- When ln(earnings) is the dependent variable, this induces bias ($b_{vY} = -0.172 < 0$).

Takeaway: the *direction and magnitude* of bias depend on non-classical covariance patterns.

Section V.

- **Earnings functions:** ln(annual earnings) on education, tenure, experience
  - Coefficient biases tend to be relatively small (in their setting).
- **Hours on wage:** ln(hours) on ln(earnings/hour)
  - Because wage is constructed from earnings and hours, measurement error can be severe.
  - They caution against naive structural interpretation in a one-firm sample, but the econometric warning generalizes.

$$Bias = -b_{u\tilde{X}}\beta + b_{v\tilde{X}}$$

References I

Bound, J., Brown, C., Duncan, G. J., & Rodgers, W. L. (1994).
Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data.
Journal of Labor Economics, 12(3), 345–368.