

Bounding a Class of Parameters in Measurement Error Models under Data Combination

Angda Li

Department of Economics
Texas A&M University

Feb 22, 2026

Outline

- 1 Introduction
- 2 General Framework
- 3 Numerical Illustration
- 4 Empirical Application

Introduction to the Setting

- Measurement error is a pervasive problem in economic studies.
- In *survey data*, we observe
 - X : noisy measurement (e.g, self-reported welfare benefit).
 - Y : outcome variable (e.g, household consumption).

Introduction to the Setting

- Measurement error is a pervasive problem in economic studies.
- In **survey data**, we observe
 - X : noisy measurement (e.g, self-reported welfare benefit).
 - Y : outcome variable (e.g, household consumption).
- Suppose we also have access to **administrative data**:
 - X^* : true variable (e.g, actual welfare benefit).
- However, these two datasets cannot be matched due to:
 - Lack of identifiers.
 - Imperfect linkage or attrition.

Parameters of Interests

- OLS Coefficient: $Y = \alpha + \beta X^* + \epsilon$.
 - Causal effect of X^* on outcome Y .
- Average reporting ratio within specific group defined by $X^* \in \mathcal{X}^*$:

$$\frac{\mathbb{E}[X \mid X^* \in \mathcal{X}^*]}{\mathbb{E}[X^* \mid X^* \in \mathcal{X}^*]}$$

Note: We do not observe (X, Y) and X^* jointly

\Rightarrow These parameters are usually not point identified.

Main Content

This paper proposes a framework for bounding a class of parameters using two unmatched datasets.

- Start from a general decomposition:

$$\text{measurement error} = \text{random error} + \text{systematic error}$$

- Allow for flexible correlation between measurement error and the true variable.

Main Content

This paper proposes a framework for bounding a class of parameters using two unmatched datasets.

- Start from a general decomposition:

$$\text{measurement error} = \text{random error} + \text{systematic error}$$

- Allow for flexible correlation between measurement error and the true variable.
- Develop a linear programming approach to compute bounds.
 - Decision variable: slope function (a measure of systematic error)
 - Accommodate a range of identifying assumptions.
- Our method is illustrated through a numerical example and an empirical application.

Related Literature

Identification and estimation of measurement error models with auxiliary data

- Carroll & Wand (1991); Bound et al. (1994); Lee & Sepanski (1995); Chen et al. (2005, Restud) \Rightarrow require matched data.
- Hu & Ridder (2012, JAE) studied unmatched data but required $u \perp X^*$
 \Rightarrow we allow for flexible specifications.

Data combination in different settings

Pacini (2019a, 2019b); Hwang (2023); D'Haultfoeuille et al., (2021, 2024); Fan & Park (2010); Fan et al. (2014); Russell (2021); Fan et al. (2023)...

Linear programming approach to characterize bounds

Mogstad et al. (2018, ECMA); Torsgovitsky (2019, ECMA); Tabaldi et al. (2022, ECMA); Han & Yang (2024, JoE); Kamat & Norris (2025, Restud) ...

Outline

- 1 Introduction
- 2 General Framework**
- 3 Numerical Illustration
- 4 Empirical Application

Basic Setup

Consider two unmatched datasets:

- **Primary dataset** contains a noisy measurement X .
- **Auxiliary dataset** contains the true variable X^* .
- We also observe covariates Z in both two datasets.

Assumption 1

- (a) The joint distribution of (X, Z) is identified from the primary data.
- (b) The joint distribution of (X^*, Z) is identified from the auxiliary data.
- (c) The joint distribution of (X, X^*, Z) has finite second moments.

Note

- X and X^* can be continuous or discrete.
- We allow for no common covariates.

Decomposition of Measurement Error

Measurement error u can always be decomposed as

$$\begin{aligned} u &= X - X^* \\ &= \underbrace{\left(X - \mathbb{E}[X \mid X^*, Z] \right)}_{\text{random error } \tilde{u}} + \underbrace{\left(\mathbb{E}[X \mid X^*, Z] - X^* \right)}_{\text{systematic error}} \end{aligned}$$

- Random error: $\mathbb{E}[\tilde{u} \mid X^*, Z] = 0$.
- Systematic error can cause the measurement to be consistently higher or lower than the true value.

Slope Function: A Measure of Systematic Error

Capture systematic over/under-reporting via a slope function

$$\lambda(x^*, z) = \frac{\mathbb{E}[X \mid X^* = x^*, Z = z]}{x^*}$$

- $\lambda(x^*, z)$ measures the **average reporting rate** at (x^*, z) .
 - > 1 : systematic overreporting
 - < 1 : systematic underreporting
- We impose **prior assumptions**: $\lambda \in \mathcal{L}$.
 - Overreporting only: $\lambda(x^*, z) \geq 1$
 - Bounded overreporting (e.g., at most 20%):

$$1 \leq \lambda(x^*, z) \leq 1.2$$

Step 1: What We Want to Know

- We study the identification of a class of parameters:

$$\gamma = \Gamma^*(\lambda)$$

where Γ^* is an identified (or known) linear functional.

Step 1: What We Want to Know

- We study the identification of a class of parameters:

$$\gamma = \Gamma^*(\lambda)$$

where Γ^* is an identified (or known) linear functional.

- It covers the scalar parameters of the form:

$$\gamma = \mathbb{E}[X \cdot h(X^*, Z)]$$

where h is an identified (or known) function.

Step 1: What We Want to Know

- We study the identification of a class of parameters:

$$\gamma = \Gamma^*(\lambda)$$

where Γ^* is an identified (or known) linear functional.

- It covers the scalar parameters of the form:

$$\gamma = \mathbb{E}[X \cdot h(X^*, Z)]$$

where h is an identified (or known) function.

- Example 1: Covariance

$$\begin{aligned}\text{Cov}(X, X^*) &= \mathbb{E}[XX^*] - \mathbb{E}[X] \cdot \mathbb{E}[X^*] \\ &= \mathbb{E}[X \cdot (X^* - \mathbb{E}[X^*])]\end{aligned}$$

A Quick Detour: Linear Regression with Unmatched Data

Consider a linear regression

$$Y = \alpha + \beta X^* + \epsilon$$

- Two **unmatched** datasets:
 - The primary data contains (Y, X, Z) .
 - The auxiliary data contains (X^*, Z) .
- Assume that $\text{Cov}(X, \epsilon) = 0$, we obtain (use X as IV)

$$\beta = \frac{\text{Cov}(X, Y)}{\text{Cov}(X, X^*)}$$

Bounds of β follow from bounds of $\text{Cov}(X, X^*)$.

Other Examples of Target Parameters

Example 2

The conditional average reporting ratio given $X^* \in \mathcal{X}^*$ and $Z \in \mathcal{Z}$

$$\frac{\mathbb{E}[X \mid X^* \in \mathcal{X}^*, Z \in \mathcal{Z}]}{\mathbb{E}[X^* \mid X^* \in \mathcal{X}^*, Z \in \mathcal{Z}]}$$

Example 3

The expected individual-level reporting ratio given $X^* \in \mathcal{X}^*$ and $Z \in \mathcal{Z}$

$$\mathbb{E} \left[\frac{X}{X^*} \mid X^* \in \mathcal{X}^*, Z \in \mathcal{Z} \right]$$

Note

- Values > 1 indicate overreporting.
- Values < 1 indicate underreporting.

Step 2: What We Know about the Slope Function

From

$$X = \lambda(X^*, Z) \cdot X^* + \tilde{u}, \quad \mathbb{E}[\tilde{u} \mid X^*, Z] = 0$$

- The distribution of X is a **mean-preserving spread** of the distribution of $\lambda(X^*, Z) \cdot X^*$.
- We use a lemma from D'Haultfoeuille et al. (2021, QE).

Lemma 1

Suppose Assumption 1 holds. We have

$$\mathbb{E}[\lambda(X^*, Z) \cdot X^* \mid Z] = \mathbb{E}[X \mid Z],$$

$$\mathbb{E}[(x - \lambda(X^*, Z) \cdot X^*)^+ \mid Z] \leq \mathbb{E}[(x - X)^+ \mid Z], \quad \forall x \in \mathbb{R}.$$

where $a^+ = \max(0, a)$.

Step 2: What We Know about the Slope Function

We use a trick to generate linear moment conditions:

$$a^+ = \max(0, a) \geq a \cdot d, \quad \text{for } d \in \{0, 1\}$$

Step 2: What We Know about the Slope Function

We use a trick to generate linear moment conditions:

$$a^+ = \max(0, a) \geq a \cdot d, \quad \text{for } d \in \{0, 1\}$$

Proposition 1: Unconditional Linear Moment Conditions

Suppose that $s : \mathbb{R}^{dz} \rightarrow \mathbb{R}^+ \cup \{0\}$ is an identified (or known) function and $t \in \mathbb{R}$. Define $d_t(u) = \mathbf{1}\{u \leq t\}$. Then,

$$\mathbb{E} \left[\lambda(X^*, Z) \cdot s(Z) X^* \right] = \mu_s \quad (1)$$

$$\mathbb{E} \left[\lambda(X^*, Z) \cdot s(Z) X^* \cdot d_t(X^*) \right] \geq \delta_{s,t} \quad (2)$$

where $\mu_s = \mathbb{E}[s(Z)X]$ and

$$\delta_{s,t} = \sup_{x \in \mathbb{R}} \mathbb{E} \left[s(Z) \left(x \cdot d_t(X^*) - (x - X)^+ \right) \right]$$

Step 3: Bounds of the Target Parameter

Recall that objective function is $\gamma = \Gamma^*(\lambda)$.

Proposition 2

Suppose $\lambda \in \mathcal{L}$. Let \mathcal{S} denote a collection of function s and $\mathcal{T} \subseteq \mathbb{R}$. Bounds of γ are given by

$$\underline{\gamma}^* = \inf_{\lambda \in \mathcal{L}} \Gamma^*(\lambda), \quad \text{s.t. (1) and (2) hold for all } s \in \mathcal{S} \text{ and } t \in \mathcal{T}$$

$$\bar{\gamma}^* = \sup_{\lambda \in \mathcal{L}} \Gamma^*(\lambda), \quad \text{s.t. (1) and (2) hold for all } s \in \mathcal{S} \text{ and } t \in \mathcal{T}$$

Note

- Two linear programming problems.
- Constraints may be infeasible \Rightarrow **model is misspecified**.
- For appropriate \mathcal{S} and \mathcal{T} , we may exhaust all the information in two datasets under mild assumptions.

Additional Restrictions on the Slope Function

Shape Restrictions

- (a) (Bounds of the slope function) $\lambda_l \leq \lambda(x^*, z) \leq \lambda_u$.
- (b) (Monotonicity) $\lambda(x^*, z)$ is weakly increasing in x^* .
- (c) (Concavity) $\lambda(x^*, z)$ is weakly concave in x^* .

Functional Form Restrictions

- (a) (Conditional mean independence: $\mathbb{E}[X | X^*, Z] = \mathbb{E}[X | X^*]$)

$$\lambda(x^*, z) = \tilde{\lambda}(x^*).$$

- (b) (Separability between X^* and Z)

$$\lambda(x^*, z) = \lambda_1(x^*) + \lambda_2(z).$$

Note: we can easily incorporate them into our framework.

Outline

- 1 Introduction
- 2 General Framework
- 3 Numerical Illustration**
- 4 Empirical Application

Numerical Illustration

$$X = \underbrace{\{-0.1(3 - X^*)^2 + 1.1\}}_{\lambda(X^*)} \cdot X^* + \tilde{u}$$

where $X^* \sim \text{Unif}[2, 3]$ and $\tilde{u} | X^* \sim \mathcal{N}(0, 0.3^2)$.

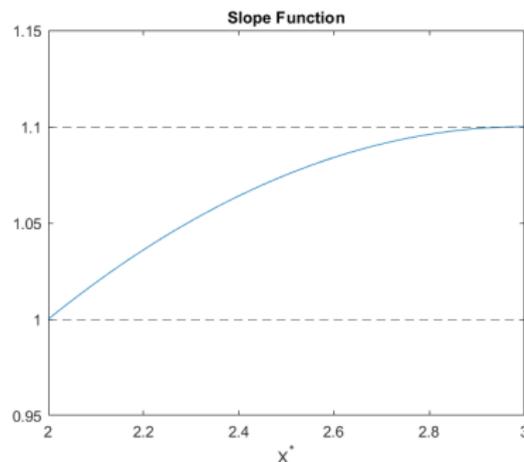


Figure: $\lambda(\cdot)$ ranges from 1 to 1.1 and is both strictly increasing and concave.

Numerical Illustration

- Target parameter $\gamma = \text{Corr}(X, X^*)$.
- Two independent samples with 10^5 observations.

Table: Bounds of γ under various assumptions (True value is 0.78)

Assumptions	$(\lambda_l, \lambda_u) = (1, 1.1)$	$(\lambda_l, \lambda_u) = (1, 1.2)$
No assumption	$[-0.99, 0.99]$	$[-0.99, 0.99]$
Only bounds of $\lambda(\cdot)$	$[0.46, 0.82]$	$[0.25, 0.94]$
$\lambda(\cdot)$ increasing	$[0.64, 0.82]$	$[0.64, 0.94]$
$\lambda(\cdot)$ increasing + concave	$[0.64, 0.80]$	$[0.64, 0.83]$

Notes: We assume the slope function is bounded between λ_l and λ_u and approximate $\lambda(\cdot)$ using Bernstein polynomials of order $K = 20$.

Outline

- 1 Introduction
- 2 General Framework
- 3 Numerical Illustration
- 4 Empirical Application**

Two Unmatched Datasets in Hu & Ridder (2012, JAE)

- X : reported benefit level from SIPP.
- X^* : actual benefit level from AFDC QC.

	X	X^*	Ratio
Sample size	520	3318	0.157
Mean	304.2	327.7	0.928
Median	280	294	0.952
SD	180.9	169.4	1.068

Table: Comparison of reported (X) and actual (X^*) benefit levels

Reported vs. Actual Benefit Levels

	X	X^*	Ratio
Mean in $[0, 200)$	127.3	139.1	0.915
Mean in $[200, 400)$	288.0	289.6	0.994
Mean in $[400, 600)$	473.5	472.0	1.003
Mean in $[600, 800)$	672.8	678.1	0.992
Mean in $[800, \infty)$	938.6	965.8	0.972

Table: Comparison of reported (X) and actual (X^*) benefit levels

Average Degree of Underreporting

- Assume that individuals tend to underreport $\Rightarrow \lambda(x^*) \leq 1$.
- We focus on the average degree of underreporting given $X^* \in \mathcal{X}^*$:

$$1 - \frac{\mathbb{E}[X \mid X^* \in \mathcal{X}^*]}{\mathbb{E}[X^* \mid X^* \in \mathcal{X}^*]}$$

- In particular, we consider:
 - Low-benefit group: X^* below the 25th percentile.
 - High-benefit group: X^* above the 75th percentile.

Empirical Results

	(1)	(2)
	Only bounds on $\lambda(\cdot)$	$\lambda(\cdot)$ increasing
<i>Panel A: $(\lambda_l, \lambda_u) = (0.8, 1)$</i>		
Low-benefit group	[0.9%, 15.3%]	[7.1%, 15.3%]
High-benefit group	[2.9%, 12.0%]	[2.9%, 7.2%]
<i>Panel B: $(\lambda_l, \lambda_u) = (0.9, 1)$</i>		
Low-benefit group	[3.5%, 9.7%]	[7.1%, 9.7%]
High-benefit group	[5.0%, 8.9%]	[5.0%, 7.2%]

Table: Bounds for the average degree of underreporting in two groups

Conclusion

- A general framework for bounding a class of parameters in measurement error models under data combination.
 - Fast to compute using linear programming.
 - Flexible: accommodates several assumptions on measurement error.