Econometrics 2

Assessing Studies Based on Multiple Regression

Lasha Chochua

2026

Where We Are in the Course

- So far: the **mechanics** of multiple regression – OLS, tests, controls, nonlinearities
- Today: stepping back to ask the **hard question**
  - When does multiple regression actually estimate a **causal effect**?
  - When does it fail – and how would we know?
- This chapter is the **bridge** to panel data, IV, RCTs (Randomized Controlled Trials)

Lecture Roadmap

- The framework: **internal** vs **external validity**
- **Five** threats to internal validity that **bias** $\widehat{\beta}$
- **Two** threats to **standard errors** that invalidate inference
- A different lens: validity for **prediction**, not causation
- Application: replicating the California study in **Massachusetts**

Internal and External Validity

Two Populations, Two Settings

- **Population studied** – the population the sample was drawn from
- **Population of interest** – the population we want to generalize to
- **Setting** – institutional, legal, social, physical, economic environment
- A study can succeed for the **studied** population but fail for the **target** population of interest

A Concrete Example

- **Studied:** California elementary school districts in 1999
- **Of interest:** elementary districts in Massachusetts? Georgia? High schools?
- The closer the match in population *and* setting, the stronger the case for generalization
- Mismatches can be in *who*, in *where*, in *when*, or in *what institutional rules apply*

Definition of External Validity

Definition 1: External Validity

A study is **externally valid** if its inferences and conclusions can be **generalized** from the population and setting studied to other populations and settings of interest.

- External validity is a property of the **mapping** between studied and target populations

Threats to External Validity – Populations

- **Mice vs humans** in toxicology
- **College vs elementary** students for class-size effects
- **Volunteers vs general population** in clinical trials
- **Historical vs current** populations – the study may simply be out of date
- The deeper question: is the **causal mechanism** the same in both populations?

Threats to External Validity – Settings

- **Legal regime** – a binge-drinking study in a state with strict penalties may not generalize
- **Institutional environment** – public vs private universities
- **Physical environment** – tailgating in southern California vs Fairbanks, Alaska
- **Macroeconomic conditions** – a 1995 labor study may not apply post-2020
- **Best diagnostic:** does another study, in a different setting, find similar results?

How to Defend External Validity

- **Compare related studies** – meta-analysis, cross-country replications
- **Argue** for similarity in mechanisms, not just in surface features
- **Design** the study with the target population in mind from the start
- Acknowledge limits **explicitly** – don't oversell

Definition of Internal Validity

Definition 2: Internal Validity

A study is **internally valid** if its statistical inferences about causal effects are valid for the population being studied. Two requirements:

1. The estimator $\hat{\beta}$ is **unbiased** and **consistent** for the true β
2. Standard errors yield CIs with the **correct coverage** and tests with the **correct size**

- Internal validity is a property of the **estimator + data**

Threats to Internal Validity

The Five Threats – Overview

- All five threats induce $\text{Cov}(X_i, u_i) \neq 0$, violating LSA #1:
  1. **Omitted variable bias**
  2. **Functional form misspecification**
  3. **Errors-in-variables** (measurement error in X)
  4. **Sample selection**
  5. **Simultaneous causality**
- Plus a separate threat to **inference**: incorrect standard errors.

Threat 1 – Omitted Variable Bias

- An omitted W that (i) affects Y and (ii) is correlated with included X biases $\hat{\beta}_1$
- The bias **persists** in large samples – OLS is **inconsistent**, not just biased
- Sign of bias $=$ sign of $(\text{Corr}(X, W) \cdot \beta_W)$
- **Solution if W observed:** include it as a regressor
- **Solution if W unobserved but adequate controls exist:** include controls satisfying conditional mean independence

OVB – The Bias-Variance Trade-off

- Adding W when it **belongs** $\Rightarrow$ removes bias
- Adding W when it **doesn't belong** $\Rightarrow$ inflates Var$(\widehat{\beta})$
- Practical question: is the bias reduction worth the precision cost?
- Answer depends on: (i) strength of W's correlation with Y, (ii) collinearity of W with X

OVB – Decision Procedure

1. **Identify** the coefficient(s) of interest
2. **A-priori reasoning** – list likely confounders before touching the data
3. **Test** whether questionable controls matter – do they change $\widehat{\beta}$ meaningfully?
4. **Full disclosure** – report multiple specifications side-by-side in a table

OVB – When No Adequate Controls Exist

Three alternative strategies, each exploiting a different data structure:

- **Panel data** – same unit over time; differences out time-invariant confounders (Ch. 10)
- **Instrumental variables** – find Z correlated with X but independent of u (Ch. 12)
- **Randomized controlled experiments** – randomization severs the X-u link (Ch. 13)

Threat 2 – Functional Form Misspecification

Functional Form Bias

If the true PRF (Population Regression Function) is nonlinear but the estimated regression is linear, $\hat{\beta}$ is biased. This is OVB where the omitted variables are the missing nonlinear terms.

- Example: true PRF quadratic; we omit $X^2 \Rightarrow X^2$ is the omitted variable
- **Detection:** scatterplots, residual plots, F-tests on polynomial / interaction terms
- **Solution (continuous Y):** polynomials, logs, interactions (Ch. 8)
- **Solution (binary/discrete Y):** logit, probit (Ch. 11)

Threat 3 – Errors-in-Variables: Setup

- True regressor X_i unobserved; we observe $\widetilde{X}_i$
- Substituting into the population regression:

$$Y_i = \beta_0 + \beta_1 \widetilde{X}_i + \underbrace{[\beta_1 (X_i - \widetilde{X}_i) + u_i]}_{v_i} \tag{1}$$

- The composite error v_i in (1) contains the **measurement error**
- If $\text{Cov}(\widetilde{X}_i, v_i) \neq 0$, OLS is biased and inconsistent

EIV – Sources of Measurement Error

- **Survey misreporting** – last year's earnings are hard to remember
- **Data entry errors** in administrative records
- **Conceptual mismatch** – "STR for the district" vs the class size a particular student actually experienced
- **Stale data** – 1990 Census income used for 1999 outcomes
- **Intentional misreporting** – underreporting taxable income

Classical Measurement Error Model

- Assume $\widetilde{X}_i = X_i + w_i$ with:

$$E[w_i] = 0, \quad \text{Cov}(w_i, X_i) = 0, \quad \text{Cov}(w_i, u_i) = 0$$

Then:

$$\widehat{\beta}_1 \xrightarrow{p} \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2} \beta_1 \tag{2}$$

- The factor in (2) is **strictly less than 1**
- This is **attenuation bias** – $\widehat{\beta}_1$ is biased **toward zero**

Derivation of (2) – Step 1

- From (1), $v_i = \beta_1(X_i - \widetilde{X}_i) + u_i = -\beta_1 w_i + u_i$.
- Compute $\mathrm{Cov}(\widetilde{X}_i, v_i)$:

$$\mathrm{Cov}(\widetilde{X}_i, v_i) = -\beta_1 \,\mathrm{Cov}(\widetilde{X}_i, w_i) + \mathrm{Cov}(\widetilde{X}_i, u_i)$$

- Now $\mathrm{Cov}(\widetilde{X}_i, w_i) = \mathrm{Cov}(X_i + w_i, w_i) = \sigma_w^2$, and $\mathrm{Cov}(\widetilde{X}_i, u_i) = 0$. So:

$$\mathrm{Cov}(\widetilde{X}_i, v_i) = -\beta_1 \sigma_w^2$$

Derivation of (2) – Step 2

- Variance of the observed regressor:

$$\sigma^2_{\widetilde{X}} = \sigma^2_X + \sigma^2_w$$

- Apply the OLS probability limit:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \frac{\mathrm{Cov}(\widetilde{X}_i, v_i)}{\sigma^2_{\widetilde{X}}} = \beta_1 - \frac{\beta_1 \sigma^2_w}{\sigma^2_X + \sigma^2_w}$$

$$= \beta_1 \cdot \frac{\sigma^2_X}{\sigma^2_X + \sigma^2_w}$$

- This is exactly (2). ■

EIV – Two Limiting Cases

- **No measurement error:** $\sigma_w^2 = 0$ – factor in (2) $= 1$, $\hat{\beta}_1$ consistent
- **Pure noise:** $\sigma_w^2 \to \infty$ – factor in (2) $\to 0$, $\hat{\beta}_1 \xrightarrow{p} 0$
- **Reliability ratio:** $\sigma_X^2/(\sigma_X^2 + \sigma_w^2) \in (0, 1)$ – the share of variation in $\widetilde{X}$ that is "signal"

EIV – The "Best Guess" Model

- Alternative: respondent reports $\widetilde{X}_i = E[X_i \mid \text{info}_i]$
- Because $\widetilde{X}_i$ is the conditional mean, $E[(\widetilde{X}_i - X_i) \mid \widetilde{X}_i] = 0$
- Then $\text{Cov}(\widetilde{X}_i, v_i) = 0$, so $\hat{\beta}_1$ is **consistent**
- But $\text{Var}(v_i) > \text{Var}(u_i)$ – precision still suffers
- **Lesson:** the structure of the error matters as much as its magnitude

Measurement Error in Y

- Suppose $\widetilde{Y}_i = Y_i + w_i$ with w_i independent of X_i and u_i:

$$\widetilde{Y}_i = \beta_0 + \beta_1 X_i + (u_i + w_i) \tag{3}$$

- In (3), the composite error has $E[\cdot \mid X_i] = 0$
- $\hat{\beta}_1$ is **unbiased and consistent**
- Only the **variance** of $\hat{\beta}_1$ is inflated
- **Punchline:** measurement error in X is dangerous; in Y it is merely costly

Solutions to Errors-in-Variables

- **Best:** get a more accurate measure of X
- **Instrumental variables** – find Z correlated with X_i but independent of w_i (Ch. 12)
- **Adjustment:** if σ_w^2/σ_X^2 is known/estimable, invert (2)
- **Sensitivity analysis:** show how $\hat{\beta}$ changes under different assumed reliability ratios

Threat 4 – Sample Selection: Three Cases

- **Case 1: Missing completely at random** – reduces n, no bias
- **Case 2: Missing based on** X – reduces n, narrows support of X, no bias
- **Case 3: Missing based on** Y (beyond depending on X) – induces $\text{Cov}(X_i, u_i) \neq 0$
- Only Case 3 is **sample selection bias**

Sample Selection

Sample Selection Bias

Sample selection bias arises when a selection process influences the availability of data and that process is related to the dependent variable beyond depending on the regressors. The OLS estimator is biased and inconsistent.

Sample Selection – Classic Examples

- **1936 Literary Digest poll** – sampled phone/car owners (Republican-leaning), predicted Landon, Roosevelt won in a landslide
- **Survivorship bias in mutual funds** – defunct funds dropped from databases inflate measured average returns
- **Heckman's wage example** – observed wages only for women who chose to work; self-selection depends on the unobserved wage offer

Sample Selection – Solutions

- **Best defense:** sample design that does **not** condition on Y
- **If unavoidable:** Heckman two-step correction (advanced)
- **Lesson:** sample selection is a *design* problem, fixed before data collection

Threat 5 – Simultaneous Causality

- So far we assumed causality runs $X \to Y$. What if $Y \to X$ as well?

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{4}$$

$$X_i = \gamma_0 + \gamma_1 Y_i + v_i \tag{5}$$

  - Equation (4) is the causal effect we want
  - Equation (5) is the **reverse channel**
  - Both operate **simultaneously** in the data

Test Score Example

- We want: effect of STR on test scores – equation (4)
- Suppose government **funds extra teachers** in low-scoring districts
- Then low scores $\rightarrow$ low STR – equation (5) with $\gamma_1 < 0$
- OLS sees the **mixture** of both effects $\Rightarrow$ biased estimate of β_1

Why (5) Breaks OLS – 1

- Take $\text{Cov}(\cdot, u_i)$ of equation (5):

$$\text{Cov}(X_i, u_i) = \gamma_1 \text{Cov}(Y_i, u_i) + \text{Cov}(v_i, u_i)$$

- Assume $\text{Cov}(v_i, u_i) = 0$. Then:

$$\text{Cov}(X_i, u_i) = \gamma_1 \text{Cov}(Y_i, u_i)$$

Why (5) Breaks OLS – 2

- From (4), $\text{Cov}(Y_i, u_i) = \beta_1 \text{Cov}(X_i, u_i) + \sigma_u^2$. Substituting:

$$\text{Cov}(X_i, u_i) = \gamma_1 \beta_1 \text{Cov}(X_i, u_i) + \gamma_1 \sigma_u^2$$

- Solving for $\text{Cov}(X_i, u_i)$:

$$\text{Cov}(X_i, u_i) = \frac{\gamma_1 \sigma_u^2}{1 - \gamma_1 \beta_1} \tag{6}$$

  - The covariance in (6) is **non-zero** whenever $\gamma_1 \neq 0$
  - This biases OLS in (4)

Simultaneous Causality – Solutions

- **Instrumental variables** – find Z that shifts X but does not enter (4) (Ch. 12)
- **Randomized experiments** – randomization severs the reverse channel (Ch. 13)
- **Natural experiments** – exogenous shocks that shift X for reasons unrelated to Y
- Also called **simultaneous equations bias**

Inconsistent Standard Errors

- Two sources, both fixable:
  - **Heteroskedasticity** with homoskedasticity-only SEs $\Rightarrow$ use **robust** SEs
  - **Correlation across observations** (serial / spatial / cluster) $\Rightarrow$ use HAC / cluster-robust SEs (Chs. 10, 16)
  - **Important:** these problems do **not** bias $\widehat{\beta}$ – they only invalidate t-stats and CIs

Summary

Threat	Consequence / Remedy
Omitted variables	Cov$(X, u) \neq 0$; controls, panel, IV, RCT
Functional form	Bias in partial effects; nonlinear specifications
Errors in variables	Attenuation bias (2); IV, better data
Sample selection	Bias when selection depends on Y; redesign
Simultaneous causality	Reverse channel (6); IV, RCT
Bad standard errors	Wrong inference; robust / HAC / cluster SEs

Validity for Prediction

Two Different Goals

- **Causal goal:** the superintendent asks "if I cut STR, what happens to scores?"
- **Predictive goal:** the parent asks "what scores can I expect in district X?"
- These are **fundamentally different** statistical problems
- Confusing them is one of the most common applied mistakes

Three Requirements for Reliable Prediction

1. **Same distribution** for training and prediction observations (external validity in the prediction context)
2. **Predictors that explain** Y – causal status is irrelevant
3. **An estimator suited to the dimension** – when p is large, ridge / lasso may dominate OLS (Ch. 14)

Application: California vs Massachusetts

Why Compare Two States?

- California results (Chs. 4–8): small but significant negative effect of STR on test scores
- **External validity check:** replicate on Massachusetts (220 districts, 1998)
- **Similar findings** $\Rightarrow$ supports external validity
- **Divergent findings** $\Rightarrow$ raises questions about internal validity of one or both

Summary Statistics – Two States

Variable	CA mean	CA SD	MA mean	MA SD
Test scores	654.1	19.1	709.8	15.1
STR	19.6	1.9	17.3	2.3
% English learners	15.8	18.3	1.1	2.9
% subsidized lunch	44.7	27.1	15.3	15.1
District income ($)	15,317	7,226	18,747	5,808
n	420		220	

Key Differences Between the Samples

- Massachusetts has **lower** STR (17.3 vs 19.6)
- Massachusetts is **richer** on average but with **less** spread
- California has **far more** English learners (15.8% vs 1.1%)
- California has **far more** subsidized-lunch students (44.7% vs 15.3%)
- Different tests $\Rightarrow$ raw coefficients are **not directly comparable**

Comparable Estimates – Standardized Units

Specification	$\hat{\beta}_{STR}$	**Effect of** $\Delta STR = -2$	**In SDs**
CA linear	-0.73 (0.26)	1.46	0.076 (0.027)
CA cubic, STR: $20 \to 18$	–	2.93	0.153 (0.037)
MA linear	-0.64 (0.27)	1.28	0.085 (0.036)

- Both states: roughly 0.08 SD per 2-student cut – 95% CIs overlap

Why Divide by the Standard Deviation?

- Raw estimated effects of $\Delta STR = -2$:
  - California: 1.46 points on the Stanford 9 test
  - Massachusetts: 1.28 points on the MCAS test
- **Cannot compare directly** – different tests, different scales
- Analogous to comparing 1.46 dollars to 1.28 euros without an exchange rate
- **Solution:** convert both effects into a common, scale-free unit
- The natural "exchange rate" is the **standard deviation** of test scores in each state
- Dividing by the SD rescales the effect into *fractions of a typical district-to-district gap*

Standardization in Action – CA vs MA

- Apply $\dfrac{\text{effect in test points}}{\text{SD of test scores}}$ in each state:

$$\text{CA:}\quad \frac{1.46}{19.1} \approx 0.076 \text{ SDs} \qquad \text{MA:}\quad \frac{1.28}{15.1} \approx 0.085 \text{ SDs}$$

- Units of "test points" cancel – result is **dimensionless**
- Interpretation: a 2-student cut moves a district about 8% of one SD up the distribution
- Both states now on the **same yardstick** – direct comparison is meaningful
- The two estimates differ by less than 0.01 SDs $\Rightarrow$ **strong external validity**
- Why the SD and not the mean? The mean is an arbitrary anchor; the SD measures *real variation across districts*
- Same logic as a z-score – the SD is the natural ruler of the data

Interpreting the Magnitude

- 0.08 SD is a **small** effect
- For perspective: median-to-75th-percentile gap in CA $= 0.64$ SD
- A 2-student cut moves a district about **one-tenth** of the way from median to 75th
- Statistically significant, but **economically modest**

Internal Validity Check – Omitted Variables

- Controls used: % English learners, % free lunch, district income
- **Residual concerns:** teacher quality, parental investment
- One way to settle this: a **randomized experiment** (Tennessee STAR, Ch. 13)

Internal Validity Check – Other Threats

- **Functional form:** explored polynomials, logs, interactions $\rightarrow$ robust
- **Errors in variables:** STR is a noisy proxy $\rightarrow$ via (2), true effect may be slightly larger
- **Sample selection:** all districts included $\rightarrow$ no concern
- **Simultaneity:** no achievement-based funding mechanism $\rightarrow$ unlikely
- **Standard errors:** robust SEs used; spatial correlation a residual concern

Bottom Line for the Superintendent

- A 2-student STR cut raises scores by ≈ 0.08 SD
- The effect is **real but small**
- The two-state replication strengthens **external validity** for U.S. elementary districts
- The leading internal-validity concern is **residual omitted-variable bias**

Looking Ahead

- **Chapter 10:** panel data $\rightarrow$ time-invariant unobservables
- **Chapter 11:** binary dependent variables (logit, probit)
- **Chapter 12:** instrumental variables
- **Chapter 13:** randomized experiments
- **Chapter 14:** prediction with many regressors

Required Reading

- Stock, J. H. and Watson, M. W. *Introduction to Econometrics*, 4th Global Edition, **Chapter 9** (pp. 330–361)