Econometrics

Practical Session 20

Assessing Studies Based on Multiple Regression

Ricardo Gouveia-Mendes
rgouveiamendes@ucp.pt

Spring 2025-26

Católica-Lisbon School of Business and Economics

Theoretical Wrap-up

Internal and External Validity

- We have studied how to **estimate** multiple regression models. Now the harder question: are the estimates **trustworthy**?

Internal Validity

Statistical inferences about causal effects are valid for the **population and setting studied**

Requires: OLS unbiased + consistent + valid standard errors

External Validity

Inferences can be **generalized** to other populations and settings

Threatened by: institutional, legal, physical, or cultural differences

- A study can have **internal** validity but not **external** validity or vice versa

The Five Threats to Internal Validity

Threat	What it is	Solution
1. Omitted variable bias	Omitted factor correlates with x and determines y	Add controls; IV; experiment
2. Functional form	Wrong shape of regression function	Test and use correct form
3. Errors in variables	Regressor x measured with error	Better data; IV
4. Sample selection	Data missing based on y or u	Redesign sample; IV
5. Simultaneous causality	y also causes x	Experiment; IV

All five threats imply $\mathbb{E}[u_i \mid x_{1i}, \dots, x_{ki}] \neq 0$ → **OLS is biased and inconsistent**

1. Omitted Variable Bias

- An omitted variable: **(i)** determines y, and **(ii)** correlates with x

$$\hat{\beta}_j \xrightarrow{p} \beta_j + \frac{\text{cov}(u, x_j)}{\text{var}(x_j)}$$

- **Practical strategy** for deciding whether to **add a control**:
  1. **Identify** the key causal coefficient of interest
  2. **List** the likely determinants of y that also correlate with x
  3. **Augment** the baseline regression with these "questionable" controls
  4. **Assess:** if $\hat{\beta}$ on the variable of interest changes substantially → OVB
- **Trade-off:** adding an irrelevant control increases variance without reducing bias → **bias–variance is always in tension**

2. Functional Form

- **Missing polynomial terms** or interactions → biased estimates
- **Solutions:**
  - Test for nonlinearities
  - Use logarithms
  - Use interactions
  - Use polynomial terms as needed

3. Errors-in-Variables

- **Measurement error in x:** we observe $\tilde{x}_i = x_i + v_i$, then:

$$y_i = \beta_0 + \beta_1 \tilde{x}_i + \underbrace{(u_i - \beta_1 v_i)}_{\tilde{u}_i}$$

- $\tilde{x}_i$ and $\tilde{u}_i$ both contain v_i → $\text{cov}(\tilde{x}_i, \tilde{u}_i) \neq 0$ → **OLS assumption violated**
- **Consequence:** OLS is **biased toward zero** (*attenuation bias*):

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 \cdot \frac{\sigma_x^2}{\sigma_x^2 + \sigma_v^2} = \beta_1 \cdot \underbrace{\frac{1}{1 + \frac{\sigma_v^2}{\sigma_x^2}}}_{<1}$$

3. Errors-in-Variables

- **Measurement error in y:** we observe $\tilde{y}_i = y_i + v_i$, then:

$$\tilde{y}_i = \beta_0 + \beta_1 x_i + \underbrace{(u_i + v_i)}_{\tilde{u}_i}$$

- Two cases emerge:
  1. If $\text{cov}(x_i, v_i) = 0$ → unbiased estimators
  2. If $\text{cov}(x_i, v_i) \neq 0$ → **biased** estimators
- In both cases: $\text{var}(\tilde{u}_i) = \text{var}(u_i + v_i) = \text{var}(u_i) + \text{var}(v_i) + 2 \cdot \underbrace{\text{cov}(u_i, v_i)}_{=0} > \text{var}(u_i)$

  → estimators **lose precision** → higher SE and wider CI
- **Key contrast:** measurement error in y **do not necessarily bias estimators**

4. Sample Selection Bias

Case 1:
Missing at random

Dog ate 20 questionnaires

Equivalent to smaller sample → **no bias**

Case 2:
Missing based on x

Only districts with STR < 20

Cannot generalize outside that x range → **no bias**

Case 3:
Missing based on y or u

Only employed women; only surviving mutual funds

Regressor correlated with error → **biased**

4. Sample Selection Bias

- One of the most common forms: **survivorship bias**
  - Studying only **surviving mutual funds** overstates performance
  - Studying only **employed women** distorts the wage-children relationship
- **Why does it bias OLS?**
  - The selected **sample is not a random draw** from the population
  - Among the selected observations, the **error term is not mean-zero**
  - OLS misattributes the selection effect to the regressors
- **Solutions:**
  - Redesign sample to avoid selection (include the "failures")
  - **Instrumental variables** regression *(covered in sessions 21–24)*

5. Simultaneous Causality Bias

- x **causes** y. But what if y **also causes** x?

$$y_i = \beta_0 + \beta_1 x_i + u_i \qquad \text{(causal: x} \rightarrow \text{y)} \tag{1}$$

$$x_i = \gamma_0 + \gamma_1 y_i + v_i \qquad \text{(causal: y} \rightarrow \text{x)} \tag{2}$$

- If $\gamma_1 \neq 0$:
  - Shock u_i raises y_i through Equation 1
  - Then it feeds back into x_i through Equation 2
  - So x_i itself is partly determined by the same shock u_i → $\text{cov}(x_i, u_i) \neq 0$
    → **OLS is biased**

5. Simultaneous Causality Bias

- **Two examples:**
  1. **Class size example:** suppose poor-performing districts receive extra resources (lower STR) from equalization policy → STR is driven partly by past test scores → $\text{cov}(\text{STR}_i, u_i) \neq 0$
  2. **Estimating demand:** a high price raises supply but also signals high demand → quantity and price are determined jointly → $\text{cov}(p_i, u_i) \neq 0$ in the demand equation
- **Solutions:**
  - Randomized controlled experiment
  - Model both directions simultaneously
  - **Instrumental variables** regression *(Sessions 21–24)*

Threats to External Validity

- Many factors may differ between the **studied population** and the **population of interest**
- The **closer the populations**, the stronger the case for **external validity**
- **How to assess:**
  - Compare results across **two or more studies** on the same topic → similar findings bolster external validity
  - Use expert judgment about **comparability** of populations and settings
- It also matters for **prediction**: high in-sample R^2 is not sufficient → the **prediction sample distribution** must match the training sample

Exercises

Exercise 1 | Measurement Error in y (S&W 9.2)

Consider $y_i = \beta_0 + \beta_1 x_i + u_i$ satisfying the OLS assumptions. y_i is measured with error:

$$\tilde{y}_i = y_i + w_i, \quad w_i \sim \text{i.i.d.}, \quad w_i \perp (y_i, x_i)$$

Using the mismeasured outcome, the composite regression error is $v_i = u_i + w_i$.

a) Does the regression of $\tilde{y}_i$ on x_i satisfy the OLS assumptions? Is OLS consistent?

Exercise 1 | Measurement Error in y (S&W 9.2)

KEY TAKEAWAYS

- The composite error is $v_i = u_i + w_i$, so:

$$\text{cov}(x_i, v_i) = \text{cov}(x_i, u_i + w_i) = \underbrace{\text{cov}(x_i, u_i)}_{=0 \text{ (OLS assumption)}} + \underbrace{\text{cov}(x_i, w_i)}_{=0 \ (w \perp x)} = 0$$

- The OLS assumption holds → **OLS is unbiased and consistent** despite mismeasured y
- **What changes:** $\text{var}(v_i) = \text{var}(u_i) + \text{var}(w_i) > \text{var}(u_i)$ → SER is larger, standard errors wider, confidence intervals less precise → **no directional bias**
- **Intuition:** adding noise to the outcome makes it harder to detect the signal, but does not systematically push the estimate in one direction

Exercise 1 | Measurement Error in y (S&W 9.2)

b) Now suppose instead that x_i is measured with error: $\tilde{x}_i = x_i + v_i$. Using $\tilde{x}_i$ as the regressor, the composite error becomes $\tilde{u}_i = u_i - \beta_1 v_i$. Is $\text{cov}(\tilde{x}_i, \tilde{u}_i) = 0$?

Exercise 1 | Measurement Error in y (S&W 9.2)

KEY TAKEAWAYS

$$y_i = \beta_0 + \beta_1(\tilde{x}_i - v_i) + u_i = \beta_0 + \beta_1\tilde{x}_i + \underbrace{(u_i - \beta_1 v_i)}_{\tilde{u}_i}$$

- Both $\tilde{x}_i = x_i + v_i$ and $\tilde{u}_i = u_i - \beta_1 v_i$ contain the **same noise term** v_i

$$\text{cov}(\tilde{x}_i, \tilde{u}_i) = \text{cov}(x_i + v_i, u_i - \beta_1 v_i) = \underbrace{\text{cov}(x_i, u_i)}_{=0} - \beta_1\underbrace{\text{cov}(x_i, v_i)}_{=0} + \underbrace{\text{cov}(v_i, u_i)}_{=0} - \beta_1\underbrace{\text{var}(v_i)}_{\sigma_v^2} = -\beta_1\sigma_v^2 \neq 0$$

- The OLS assumption is **violated** → OLS is biased and inconsistent (attenuation bias toward zero):

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 \cdot \frac{\sigma_X^2}{\sigma_X^2 + \sigma_v^2} < |\beta_1|$$

- The estimated effect is **smaller in magnitude** than the true effect → systematic understatement

Exercise 1 | Measurement Error in y (S&W 9.2)

c) Evaluate this statement: *"Measurement error in the x's is a serious problem. Measurement error in y is not."*

KEY TAKEAWAYS

- **Broadly correct:** measurement error in x causes attenuation bias; measurement error in y **may not bias** OLS
- **But:**
  - Mismeasurement of y is not free → estimators less precise
  - The statement assumes **classical** measurement error ($w_i \perp x_i$) → if the error in measuring y is systematically related to x (e.g., high-STR districts under-report test scores to attract funding), then $\text{cov}(x_i, w_i) \neq 0$ → **OLS is biased even for mismeasured y**
- **Bottom line:** the statement is a useful rule of thumb, not an absolute law

Exercise 2 | Children Penalty on Earnings (S&W 9.3)

Labor economists regress **wages** on **number of children** and controls (age, education, occupation) using a random sample of **employed** women. They find:

> **Puzzle:** women with **more children** earn **higher wages**, controlling for age, education, and occupation.

Exercise 2 | Children Penalty on Earnings (S&W 9.3)

a) Which of the five threats to internal validity is most likely responsible? Explain the mechanism precisely.

KEY TAKEAWAYS

$$w_i = \beta_0 + \beta_1 \text{children}_i + \gamma_1 \text{age}_i + \gamma_2 \text{educ}_i + \gamma_3 \text{occup}_i + u_i$$

- **Sample selection (Case 3):** the sample excludes women who are **not employed**
- Women with **many children** who **still choose to work** tend to have **unusually** high **unobserved** earnings potential (motivation, ability, strong career preferences) → **systematically higher u_i**
- **Consequence:** $\mathbb{E}[u_i \mid \text{children}_i] > 0$ → OLS conflates the effect of sample selection with the effect of children on wages
- **Historical note:** this puzzle motivated Heckman (1979) → Nobel Prize in Economics 2000

Exercise 2 | Children Penalty on Earnings (S&W 9.3)

b) A panel study compares the same women's wages **before and after** having children and finds a **negative** effect. Why does this design avoid the cross-sectional selection problem?

KEY TAKEAWAYS

- We can include the same controls plus an individual fixed effect a_i:

$$y_{it} = a_i + \beta_1 \text{children}_{it} + \gamma_1 \text{age}_{it} + \gamma_2 \text{educ}_{it} + \gamma_3 \text{occup}_{it} + u_{it}$$

- First-differencing eliminates a_i:

$$\Delta y_i = \beta_1 \Delta \text{children}_i + \gamma_1 \Delta \text{age}_i + \gamma_2 \Delta \text{educ}_i + \gamma_3 \Delta \text{occup}_i + \Delta u_i$$

  The selection source drops out entirely → $\text{cov}(\Delta \text{children}_i, \Delta u_i) = 0$

- **Why the sign reverses:** the cross-section captured selection into employment (positive bias); the within-person panel captures the actual "child penalty" on earnings

Exercise 2 | Children Penalty on Earnings (S&W 9.3)

(Preview — covered in more detail in sessions 21–24)

c) The textbook proposes instrumental variables (IV) as a solution to the selection problem here. Without knowing IV yet, think about what properties a variable z would need to solve the problem: what should z be correlated with, and what should it be independent of?

Exercise 2 | Children Penalty on Earnings (S&W 9.3)

CATÓLICA LISBON BUSINESS & ECONOMICS

KEY TAKEAWAYS

- A valid instrument z must satisfy two conditions:
  - **Relevance:** $\text{cov}(z_i, \text{children}_i) \neq 0$ → z must predict the number of children
  - **Exogeneity:** $\text{cov}(z_i, u_i) = 0$ → z must affect wages **only through** the number of children
- **IV 1: sex composition of first two children (Angrist & Evans, 1998):** couples with two same-sex children are more likely to have a third; the sex of the first child is essentially random and unrelated to parental wages → valid instrument for having ≥ 3 children
- **IV 2: unexpected twin births:** an exogenous shock to family size; correlated with having more children but unrelated to parental wages
- Both strategies exploit **quasi-random variation in family size** that is unrelated to unobserved productivity → they identify the causal effect, not the selection effect
- **Limitation:** IV estimates the effect for **compliers** — those induced to change family size by the instrument — which may not equal the average treatment effect for all women

Exercise 3 | The California Test Score Study

The table shows the OLS coefficient on the student–teacher ratio (STR) across specifications estimated on California data (n = 420):

Specification	$\hat{\beta}_{STR}$	SE	Controls added
(1) Simple bivariate	−2.28	(0.52)	None
(2) + student and poverty	−1.10	(0.43)	(1) + PctEL, LunchPct
(3) + income (cubic)	−0.73	(0.26)	(2) + Income, $Income^2$, $Income^3$
(4) + nonlinear STR	−0.60	(0.28)	(3) + STR^2, STR^3

Exercise 3 | The California Test Score Study

CATÓLICA LISBON BUSINESS & ECONOMICS

a) The STR coefficient falls from –2.28 to –0.73 as controls are added. What does this reveal about specification (1)? Is OVB fully resolved in specification (3)?

Exercise 3 | The California Test Score Study

KEY TAKEAWAYS

- **Specification (1) suffers from severe OVB:** the coefficient falls by 68%
  - Districts with **lower STR tend** to be **wealthier** and have **fewer English learners**
  - Both factors independently **raise test scores**
  - The simple regression **attributes to class size** what partly reflects socioeconomic and demographic advantage
- **The controls are proxies:** PctEL proxies for immigrant language background; LunchPct proxies for family poverty; income captures household resources and parental investment
- **OVB is reduced, not eliminated:** unobservable factors — like teacher quality, school culture, parental involvement — may still correlate with STR; the **stability** of the coefficient across (3) and (4) is reassuring but does not prove the problem is gone

Exercise 3 | The California Test Score Study

b) Assess the functional form and errors-in-variables threats for the California study.

KEY TAKEAWAYS

- **Functional form:** adding STR^2 and STR^3 barely changes the STR coefficient (from –0.73 to –0.60), and the F-test on the nonlinear STR terms fails to reject linearity → **not a serious threat in STR**; nonlinearity matters for income, hence the cubic specification
- **Errors-in-variables in STR:** STR is a district-wide average; students taking the test may not have experienced the district-average STR (students move, class composition varies) → **attenuation bias** (the true effect may be **more negative** than −0.73)
- **Errors-in-variables in income:** district income is from the 1990 Census for a 1999 test → if income shifted differentially across districts over the decade, this measure is imprecise and may attenuate the income coefficient

Exercise 3 | The California Test Score Study

c) Assess the sample selection and simultaneous causality threats.

KEY TAKEAWAYS

- **Sample selection:** the data cover **all** California public elementary school districts satisfying minimum size requirements, not a selected subset → **no sample selection bias**
- **Simultaneous causality:** for bias to exist, low test scores would have to cause lower STR via a political equalization mechanism; California had equalization based on **budgets**, not **test scores**, during this period → **simultaneous causality is unlikely**
- **Remaining subtlety:** if parents self-select into districts based on STR (choosing low-STR districts for other reasons), the composition of each district's student body may be endogenous — a subtler form of omitted variable bias, harder to rule out

Exercise 3 | The California Test Score Study

The table below compares the estimated effect of reducing STR by 2 students per teacher for California and Massachusetts (S&W Table 9.3):

State	Specification	Effect (test score points)	Effect (standard deviations)
California	Linear	1.46 (0.52) 95% CI: 0.46–2.48	0.076 (0.027)
California	Cubic	2.93 (0.70)	0.153 (0.037)
Massachusetts	Linear	1.28 (0.54) 95% CI: 0.22–2.34	0.085 (0.036)

Exercise 3 | The California Test Score Study

d) Are the California and Massachusetts results consistent? What does this tell us about external validity?

KEY TAKEAWAYS

- **Estimates are very similar:** CA linear = 1.46 points (0.076 SDs); MA linear = 1.28 points (0.085 SDs); the 95% confidence intervals overlap substantially
- **External validity is supported:**
  - Different states, different standardized tests (STAR vs. MCAS), different school districts, yet the effect size is essentially identical
  - Both studies are US elementary schools in the late 1990s, so the populations and settings are closely comparable

Discussion

CATÓLICA LISBON BUSINESS & ECONOMICS

1. A criminologist regresses per-capita crime rates on per-capita police presence across US cities. Which of the five threats to internal validity are most serious?

KEY TAKEAWAYS

- **Simultaneous causality (most fundamental):** cities with more crime hire more police → $\text{cov}(\text{police}, u) \neq 0$; OLS will likely yield a **positive** coefficient even if the true effect is negative
- **OVB:** poverty, inequality, unemployment all affect crime and correlate with police presence; controlling for all of them is nearly impossible
- **Measurement error in y:** crime is systematically underreported; if under-reporting **correlates** with police presence, this creates bias even though the mismeasured variable is y, not x
- **Sample selection:** minor if all cities are included; but cities that voluntarily share crime data may be a selected sample

Discussion

2. A policy-maker says: *"The California study shows class size reduction significantly increases test scores. Therefore, Portugal should reduce class sizes in 2026."* Evaluate this recommendation using the internal and external validity framework.

Discussion

KEY TAKEAWAYS

- **Internal validity:** residual OVB (teacher quality, parental involvement) likely overstates the effect; measurement error in STR attenuates it; the two biases partially offset, but neither disappears → the true effect is genuinely uncertain
- **External validity:**
  - **Against:** California 1999 ≠ Portugal 2026 (educational systems, teacher contracts and incentives, family culture, socioeconomic composition, and testing methodology)
  - **In favor:** the CA/MA replication gives credible evidence that reducing class size has a positive effect on test performance **within US elementary schools** → reasonable prior for the direction in Portugal
- **Policy caution:** the effect is **small** (≈ 0.08 standard deviations) relative to the **large cost** of hiring more teachers and building classrooms → needed **Portugal-specific study** to weigh the marginal benefit against the fiscal cost