Econometrics II

Course Summary

Lasha Chochua

2026

The Thread of the Course

- One question runs through everything we covered:
  - *When does a regression coefficient measure a **causal effect**, and what do we do when it does not?*

- The logic of the summary:
  - **OLS** delivers a causal effect under one key assumption.
  - **Threats to internal validity** – omitted variables chief among them – break that assumption.
  - **IV, panel, diff-in-diff, RDD** are four ways to restore it.
  - **Binary choice models** extend the toolkit when Y is $0/1$.

- Every remedy slide names the threat it addresses, so the tools stay connected rather than appearing as a list.

OLS and Causal Interpretation

The Multiple Regression Model

- Population model with k regressors:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

- β_1 is the effect on Y of a one-unit change in X_1, **holding the other regressors fixed** (the *ceteris paribus* slope).

- OLS chooses $\hat{\beta}_0, \ldots, \hat{\beta}_k$ to minimize the sum of squared residuals:

$$\min_{b_0,\ldots,b_k} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki})^2$$

- The estimator is mechanical – it always produces numbers. Whether those numbers are **causal** is a separate question.

The Least Squares Assumptions

Least Squares Assumptions for Causal Inference

1. $E[u_i \mid X_{1i}, \ldots, X_{ki}] = 0$ – the error has conditional mean zero.
2. $(X_{1i}, \ldots, X_{ki}, Y_i)$ are i.i.d. draws.
3. Large outliers are unlikely (finite fourth moments).
4. No perfect multicollinearity.

- Assumptions 2–4 are about sampling and the data; they keep OLS well-behaved.
- **Assumption 1 is the one that buys causality.** It says the included regressors carry no information about the omitted determinants of Y collected in u_i.

When OLS Is Causal – Conditional Mean Independence

- If X_1 is **randomly assigned** (an experiment), then $X_1 \perp u$ and Assumption 1 holds automatically.
- With observational data we rarely have that. We instead add control variables W_i and hope for the weaker condition:

Important

Conditional Mean Independence

$$E[u_i \mid X_{1i}, W_i] = E[u_i \mid W_i]$$

- In words: once we control for W, the variable of interest X_1 is *as good as randomly assigned*.
- Under this condition $\widehat{\beta}_1$ consistently estimates the **causal effect** of X_1, even if the coefficients on the controls W are not themselves causal.

What "Causal" Means Here

- The causal effect is the answer to a counterfactual: *if we intervened and raised X_1 by one unit, holding everything else fixed, how would Y change on average?*
- Under conditional mean independence, the OLS slope equals this population causal effect:

$$\hat{\beta}_1 \xrightarrow{p} \frac{\partial\, E[Y \mid X_1, W]}{\partial X_1}$$

- The whole rest of the course is about what happens when Assumption 1 / conditional mean independence **fails**, and how each method restores it.

When OLS Fails – Threats to Internal Validity

Internal vs. External Validity

Two Kinds of Validity

- **Internal validity:** the estimated effect is unbiased and consistent for the **causal effect in the population studied**.
- **External validity:** the findings **generalize** to other populations and settings.

- Everything in this section is about **internal** validity – whether $\widehat{\beta}_1$ is causal at all.
- Internal validity fails through **five threats**, all of which produce the same disease: $E[u_i \mid X_i] \neq 0$.
- We take each in turn – they are **on equal footing**, and any one of them is enough to break the causal interpretation of $\widehat{\beta}_1$.

Threat 1 – Omitted Variable Bias

Omitted Variable Bias

A confounder is **both** a determinant of Y **and** correlated with X_1, so it sits in u_i and correlates with the regressor.

- Consequence – OLS is **biased and inconsistent**:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu} \frac{\sigma_u}{\sigma_X}$$

- The **sign** of the bias is the sign of ρ_{Xu}; the bias does not vanish as $n \to \infty$.
- **Fix:** add controls, or use IV / panel / RDD / diff-in-diff.

Threat 2 – Wrong Functional Form

Functional Form Misspecification

The population regression function is **nonlinear**, but we fit a linear (or otherwise misshaped) model.

- The approximation error is a function of X and falls into u_i, so $E[u_i \mid X_i] \neq 0$.
- Consequence – the estimated marginal effect is **biased**; it is the wrong slope of the wrong curve.
- **Fix:** add polynomials, logs, and interactions; let theory and specification tests guide the form.

Threat 3 – Errors-in-Variables

Measurement Error in X

We observe a noisy proxy $\tilde{X}_i = X_i + w_i$ instead of the true regressor X_i.

- The error w_i enters u_i and is correlated with $\tilde{X}_i$, so $E[u_i \mid \tilde{X}_i] \neq 0$.
- Consequence – **attenuation bias**: classical noise pulls the slope toward zero,

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 \cdot \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}.$$

- **Fix:** better measurement, or IV using a second independent measure of X.

Threat 4 – Sample Selection

Sample Selection Bias

The sample is selected through a process related to the **outcome** Y (selecting on the dependent variable).

- Selection makes the observed X correlated with u, so $E[u_i \mid X_i] \neq 0$ in the available sample.
- Consequence – **biased** estimates; the direction depends on the selection rule.
- **Fix:** random sampling; model the selection (e.g. a selection-correction approach); understand who is missing and why.

Threat 5 – Simultaneous Causality

Simultaneous (Reverse) Causality

X affects Y, but Y **also** affects X – causation runs both ways.

- The feedback from Y into X makes X correlated with u, so $E[u_i \mid X_i] \neq 0$.
- Consequence – **simultaneous equations bias**; OLS blends both directions of causation.
- **Fix:** a randomized experiment, or IV with an instrument that shifts X but not Y directly.

The Five Threats at a Glance

- All five are **equally fatal** – each one alone makes $E[u_i \mid X_i] \neq 0$.

Threat	Why $E[u \mid X] \neq 0$	Typical fix
Omitted variables	Confounder in u, correlated with X	Controls, IV, panel
Wrong functional form	Misspecification loads into u	Polynomials, logs
Measurement error in X	Noise correlates with $\tilde{X}$	IV, better data
Sample selection	Selecting on Y ties u to X	Selection model
Simultaneous causality	Y feeds back into X	IV, experiment

- The next section develops the fixes – IV, panel, diff-in-diff, RDD.

Remedies

Remedy 1 – Instrumental Variables / 2SLS

- **Threat addressed:** OVB, measurement error, simultaneous causality.
- Idea: find an instrument Z that moves X but is otherwise unrelated to Y.

Two Conditions for a Valid Instrument

1. **Relevance:** $\text{corr}(Z_i, X_i) \neq 0$.
2. **Exogeneity:** $\text{corr}(Z_i, u_i) = 0$ – Z affects Y *only* through X.

- With one instrument and one endogenous regressor:

$$\hat{\beta}_1^{IV} = \frac{s_{ZY}}{s_{ZX}} \qquad \Longrightarrow \qquad \hat{\beta}_1^{IV} \xrightarrow{p} \beta_1$$

Remedy 1 – 2SLS Mechanics

- **Two Stage Least Squares** isolates the exogenous variation in X:

$$\textbf{Stage 1:} \quad X_i = \pi_0 + \pi_1 Z_i + v_i \;\Rightarrow\; \hat{X}_i$$

$$\textbf{Stage 2:} \quad Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i$$

- Stage 1 keeps only the part of X predicted by the (exogenous) instruments; Stage 2 regresses Y on that clean variation.

- **Weak instruments** (low relevance, small π_1) make 2SLS imprecise and biased – check the first-stage F-statistic (rule of thumb $F > 10$).
- Exogeneity cannot be tested with a single instrument; it is argued from economics, not data.

Remedy 2 – Panel Data / Fixed Effects

- **Threat addressed:** OVB from **time-invariant** unobserved heterogeneity.
- Entity fixed-effects model:

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}, \qquad i = 1, \ldots, n; \; t = 1, \ldots, T$$

- α_i absorbs everything about entity i that does not change over time (culture, geography, fixed ability ...) – even if unobserved.

- Within (de-meaning) transformation removes α_i:

$$(Y_{it} - \bar{Y}_i) = \beta_1 (X_{it} - \bar{X}_i) + (u_{it} - \bar{u}_i)$$

- Identification now comes from **within-entity variation over time**. Time-varying confounders are still a threat.

Remedy 3 – Differences-in-Differences

- **Threat addressed:** OVB in program/policy evaluation when treatment is not randomized.
- Compare the **change** in a treated group to the **change** in a control group:

Important

The DiD Estimator

$$\hat{\beta}^{DiD} = (\bar{Y}^{\text{after}}_{\text{treat}} - \bar{Y}^{\text{before}}_{\text{treat}}) - (\bar{Y}^{\text{after}}_{\text{control}} - \bar{Y}^{\text{before}}_{\text{control}})$$

- Equivalent regression with a treatment–post interaction:

$$Y_{it} = \beta_0 + \beta_1 \text{Treat}_i + \beta_2 \text{Post}_t + \beta_3 (\text{Treat}_i \times \text{Post}_t) + u_{it}$$

- The causal effect is β_3, **valid under parallel trends**: absent treatment, both groups would have moved together.

Remedy 4 – Regression Discontinuity

- **Threat addressed:** OVB when treatment is assigned by a cutoff in a running variable W.
- **Sharp RDD:** treatment is a deterministic step at the threshold w_0:

$$X_i = \mathbf{1}\{W_i \geq w_0\}$$

- Units just below vs. just above the cutoff are **comparable**; the jump in Y at w_0 is the causal effect.

- **Identifying assumption:** all other determinants of Y are **continuous** at w_0 – only treatment status jumps.
- **Fuzzy RDD:** crossing w_0 changes the *probability* of treatment; combine with IV using $\mathbf{1}\{W_i \geq w_0\}$ as the instrument.

Remedies – One-Slide Map

Method	Key assumption	Threat fixed
IV / 2SLS	Relevance + exogeneity of Z	OVB, meas. error, simultaneity
Panel / FE	No time-varying confounders	Time-invariant OVB
Diff-in-diff	Parallel trends	OVB in policy evaluation
RDD	Continuity at the cutoff	OVB via cutoff assignment

- All four share one goal: recover an as-good-as-random comparison so that $E[u \mid X] = 0$ holds.

Binary Dependent Variables

When Y Is $0/1$

- Now the outcome is a **binary** event: $Y_i \in \{0, 1\}$ (e.g. loan approved / denied).
- The conditional expectation is a **probability**:

$$E[Y_i \mid X_i] = \Pr(Y_i = 1 \mid X_i)$$

- Three workhorse models, all modeling $\Pr(Y_i = 1 \mid X_i)$:
  - **Linear Probability Model** (LPM)
  - **Probit**
  - **Logit**

The Linear Probability Model

- Just OLS with a binary Y:

$$\Pr(Y_i = 1 \mid X_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$$

- β_1 is the change in **probability** of $Y = 1$ for a one-unit change in X_1 – easy to interpret.

- Drawbacks:
  - Predicted probabilities can fall **below 0 or above 1**.
  - Errors are inherently **heteroskedastic** – use robust standard errors.
- Still useful as a fast, transparent benchmark; the slope is a genuine marginal effect.

Probit and Logit

- Force fitted probabilities into $[0, 1]$ by passing a linear index through a CDF:

Probit and Logit

$$\Pr(Y_i = 1 \mid X_i) = F(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})$$

- **Probit:** $F = \Phi$, the standard normal CDF.
- **Logit:** $F = \dfrac{1}{1 + e^{-z}}$, the logistic CDF.

- Nonlinear $\Rightarrow$ estimated by **maximum likelihood**, not OLS.
- Coefficients are **not** marginal effects: the effect of X_1 depends on where you are on the curve. Report **marginal effects** or predicted probabilities for interpretation.

Binary Choice – Takeaways

- LPM: linear, interpretable slopes, but unbounded fits and heteroskedasticity.
- Probit / logit: bounded, principled, but require ML and marginal-effect interpretation.

- The causal-inference logic is **unchanged**: β_1 is causal only if X_1 is (conditionally) as good as randomly assigned. The threats and remedies from earlier sections all still apply – the binary models only change the *functional form* of $\Pr(Y = 1 \mid X)$.

Required Reading

- Stock, J. H. and Watson, M. W., *Introduction to Econometrics*.
  - Ch. 6–7: Multiple regression and the least squares assumptions.
  - Ch. 9: Assessing studies based on multiple regression (threats to internal validity).
  - Ch. 10: Regression with panel data.
  - Ch. 11: Regression with a binary dependent variable.
  - Ch. 12: Instrumental variables regression.
  - Ch. 13: Experiments and quasi-experiments (diff-in-diff, RDD).