

ECO447 MACHINE LEARNING FOR ECONOMISTS

Week 2: Causality — Why Correlation Is Not Enough

Assoc. Prof. Dr. Onur Polat

Hacettepe University • Department of Economics

Fall 2026–2027, Lecture 2 / 14

“The first step to doing good economics with ML is
understanding what ML alone cannot do: prove causation.”

Today's Roadmap

Part	Topic	Why It Matters for Economists
I	Correlation $\neq$ Causation	Every policy question is causal, not predictive
II	Confounders and Simpson's Paradox	How aggregate data can deceive policymakers
III	The Potential Outcomes Framework	The formal language of 'what would have happened'
IV	Selection Bias	Why naive comparisons give wrong answers
V	Randomized Experiments (RCTs)	The gold standard — and its limits
VI	Where Does ML Fit?	A preview of the semester ahead

Why Economists Must Think Causally

- A government asks: “Does this job training program increase employment?”
- A central bank asks: “Does raising interest rates reduce inflation?”
- A firm asks: “Did our marketing campaign drive sales, or was it seasonal?”

- ML can predict who will be unemployed next month — but it cannot tell you whether the training program caused fewer people to be unemployed.
- Prediction answers “what will happen?” Causation answers “what happens if we act?”
- Good economic policy requires causal answers. This week we learn why — and what goes wrong when we confuse the two.

Prediction: “What will Y be?” vs. Causation: “What happens to Y if I change X?”



PART I

Correlation $\neq$ Causation

The Classic Trap

Observing that two variables move together tells you NOTHING about whether one causes the other.

- Ice cream sales and drowning deaths are strongly correlated.
→ Does ice cream cause drowning? Of course not. Temperature drives both.
- Countries with more chocolate consumption have more Nobel laureates.
→ Does chocolate cause genius? No. Wealth drives both.
- Per capita cheese consumption correlates with death by bedsheet tangling.
→ Pure coincidence. With enough variables, you WILL find spurious correlations.

The lesson: correlation is a necessary but NOT sufficient condition for causation.

Three Reasons Correlation Misleads

Confounding

A third variable (Z)
causes BOTH X and Y.

Temperature $\rightarrow$ ice cream

Temperature $\rightarrow$ drowning

X and Y move together
but neither causes
the other.

Reverse Causality

You think $X \rightarrow Y$
but actually $Y \rightarrow X$.

More police $\rightarrow$ more crime?

No: more crime $\rightarrow$
more police hired.

The arrow is backwards.

Selection Bias

Your sample is
not random.

“College grads earn more”
— but who goes to college?
People who would earn
more anyway (ability,
family background).

Let's See It in Python

```
# 20 cities: population drives BOTH ice cream sales and drownings
np.random.seed(42)
population = np.random.uniform(10_000, 500_000, 20)

ice_cream = 0.5 * population + np.random.normal(0, 20_000, 20)
drownings = 0.00003 * population + np.random.normal(0, 2, 20)

r = np.corrcoef(ice_cream, drownings)[0, 1]
print(f'Correlation: {r:.3f}')    # Strong positive!

# But when we control for population (the confounder):
# The partial correlation between ice cream and drownings drops to ~0.
# Ice cream does not cause drowning – population size creates
# the illusion of a relationship.
```

This is the code behind your Homework Task 1. You will create these plots and identify the confounder.



PART II

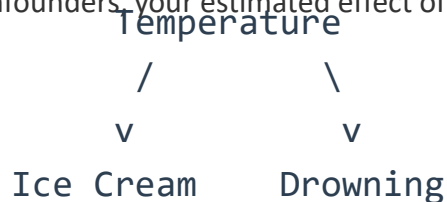
Confounders and Simpson's Paradox

What Is a Confounder?

A confounder is a variable that:

1. Causes (or is associated with) the treatment variable X
2. Causes (or is associated with) the outcome variable Y
3. Is NOT on the causal path from X to Y

If you don't account for confounders, your estimated effect of X on Y will be biased.



Economic example: Ability confounds the education–earnings relationship. Smart people get more education AND earn more, even without the degree.

When Aggregate Data Lies

Simpson's Paradox: a trend that appears in aggregated data REVERSES when you disaggregate by a confounding group.

- UC Berkeley, 1973: Overall, men had higher admission rates than women.

But within EVERY department, women had equal or higher rates!

- How? Women disproportionately applied to competitive departments.

The confounder was department choice, not gender discrimination.

- Policy danger: if you act on the aggregate data (sue the university), you solve the wrong problem. The real question is: why do women apply to harder departments?

Homework Task 3 asks you to reproduce this paradox with simulated data and discover the reversal yourself.

The Mechanism — Step by Step

	Dept A (easy)	Dept B (hard)	Overall
Men applied	200	500	700
Men admitted	120 (60%)	150 (30%)	270 (38.6%)
Women applied	400	100	500
Women admitted	252 (63%)	28 (28%)	280 (56%)
	Women > Men ✓	Women ≈ Men	Men > Women ?!

The aggregation reverses because the groups have very different sizes across departments.
The confounder (department choice) is not visible until you disaggregate.



PART III

The Potential Outcomes Framework

The Language of ‘What If?’ — Rubin (1974)

For every individual i , there are TWO potential outcomes:

Symbol	Meaning	Example (Job Training)
$Y_i(1)$	Outcome IF treated	Wage if person i gets training
$Y_i(0)$	Outcome IF NOT treated	Wage if person i does NOT get training
$\tau_i = Y_i(1) - Y_i(0)$	Individual treatment effect	How much training changes i 's wage
$ATE = E[\tau_i]$	Average Treatment Effect	Average effect across the population

The Fundamental Problem of Causal Inference:

We can NEVER observe both $Y_i(1)$ and $Y_i(0)$ for the same person. The counterfactual is forever missing.

A Concrete Example

Imagine we could see inside a parallel universe:

Person	Wage WITH Training	Wage WITHOUT	Effect (τ_i)
Ali	\$3,200	\$2,800	+\$400
Ayşe	\$2,900	\$2,500	+\$400
Mehmet	\$3,500	\$2,700	+\$800
Fatma	\$2,600	\$2,400	+\$200
ATE			\$450

**In reality, we observe only ONE column per person — the other is the counterfactual.
The challenge: how do we estimate the ATE without seeing both?**



PART IV

Selection Bias — The Danger of Naive Comparisons

Why 'Just Compare the Groups' Fails

- Suppose people CHOOSE whether to take job training.
- Who signs up? Motivated, ambitious people — the ones who would earn more ANYWAY.
- Naive comparison: trained earn \$3,200, untrained earn \$2,600. Effect = \$600?
- But the true effect is only \$450 — the remaining \$150 is SELECTION BIAS.
- The treated group was already better off before training.

$$\begin{array}{rccccccc} \text{Naive estimate} & = & \text{True effect} & + & \text{Selection bias} \\ \$600 & = & \$450 & + & \$150 \end{array}$$

This is the **CENTRAL** problem of empirical economics. Every identification strategy (IV, RD, DiD, RCT) exists to solve this one problem.

Seeing Selection Bias in Python

```
# Simulation: job training with self selection
np.random.seed(42)
education = np.random.normal(12, 3, 500).clip(6, 20)

# More educated people are more likely to sign up (self-selection)
prob_training = 1 / (1 + np.exp(-(education - 13)))
treated = np.random.binomial(1, prob_training)

# True treatment effect: $500 on average
base_wage = 2000 + 300 * education + np.random.normal(0, 800, 500)
true_effect = 500 + 50 * (education - 12)
wage = np.where(treated == 1, base_wage + true_effect, base_wage)

# Naive estimate vs. truth
naive = wage[treated==1].mean() - wage[treated==0].mean()
print(f'Naive estimate:  ${naive:,.0f}')  # Overestimates!

print(f'True ATE:      $500')  # The real answer
```

Homework Tasks 4–5 walk you through this simulation step by step.



PART V

Randomized Controlled Trials (RCTs)

The Gold Standard — Random Assignment Fixes Everything

If treatment is assigned RANDOMLY, the treated and control groups are comparable on average — eliminating selection bias.

- Random assignment ensures: $E[Y(0) \mid D=1] = E[Y(0) \mid D=0]$
→ The control group is a valid counterfactual for the treated group.
- Simple difference in means estimates the ATE without bias.
- No confounders, no selection bias, no omitted variables.
- Examples: Oregon Health Insurance Experiment, RAND Health, J-PAL development RCTs.
- Nobel 2019: Banerjee, Duflo, Kremer — for experimental approaches to alleviating poverty.

Key intuition: randomization makes the ONLY systematic difference between groups the treatment itself.

RCT vs. Naive Comparison in Python

```
# Random assignment: treatment is INDEPENDENT of education
random_treated = np.random.binomial(1, 0.5, 500) # Fair coin flip
wage_rct = np.where(random_treated == 1,
                    base_wage + true_effect, base_wage)

rct_estimate = wage_rct[random_treated==1].mean() - wage_rct[random_treated==0].mean()
print(f'RCT estimate:  ${rct_estimate:,.0f}') # Close to $500!
print(f'Naive estimate: ${naive:,.0f}')       # Biased
print(f'True ATE:      $500')
```

The RCT estimate is close to the truth. The naive estimate is inflated by selection bias.

This is the fundamental reason why economists care so much about identification strategy.

When RCTs Are Not Possible

RCTs are the gold standard — but they are often impractical or unethical:

Limitation	Example	Alternative
Ethical	Can't randomly deny healthcare	Natural experiments
Compliance	People don't follow assignment	Intent-to-treat analysis
Cost	Randomizing tax policy is expensive	Quasi-experiments
External validity	Lab results may not generalize	Field experiments
Duration	Some effects take decades to emerge	Longitudinal data

When RCTs aren't possible, economists use quasi-experimental methods: IV, RD, DiD, synthetic control. We will learn these in Weeks 8–10.



PART VI

Where Does ML Fit? — A Preview

Prediction vs. Causation — The Two Worlds

	Prediction (ML)	Causation (Econometrics)
Goal	Forecast Y as accurately as possible	Estimate the effect of X on Y
Correlation	A feature — use anything that predicts	A problem — must separate from causation
Confounders	Useful! They improve prediction	Dangerous! They bias causal estimates
Overfitting	The main enemy	Less of a concern (simple models)
Evaluation	RMSE, AUC, accuracy on test set	Bias of the causal estimate
Question	“What will Y be?”	“What happens if we change X?”

This Semester's Journey: ML × Causality

The power of this course is learning BOTH worlds and how they combine:

Weeks	ML Skills	Causal Thinking
2–3	Data exploration, visualization	Correlation $\neq$ causation, confounders
4–6	Regression, regularization (LASSO)	OVB, control strategies, IV selection
7–8	Classification, trees, ensembles	Propensity scores, matching
9–10	Cross-validation, model selection	RD, DiD, synthetic control
11–12	Neural networks, deep learning	Causal forests, heterogeneous effects
13–14	Student presentations	Double ML, policy evaluation

Double ML and causal forests (Weeks 13–14) combine ML prediction power with causal econometric rigor. But first: the foundations.

Week 2 — Five Things to Remember

The Big Ideas

- ✓ Correlation $\neq$ causation: confounders, reverse causality, selection bias
 - ✓ Confounders create spurious relationships; Simpson's Paradox shows how aggregation can hide or reverse real effects
 - ✓ Potential outcomes: $Y(1)$ and $Y(0)$ define the treatment effect; the counterfactual is never observed
 - ✓ Selection bias inflates naive estimates; RCTs fix this via random assignment
 - ✓ ML predicts; econometrics identifies causes.
- This course teaches you to do both.

Practical Skills

- ✓ Simulate a spurious correlation and identify the confounder in Python
- ✓ Reproduce Simpson's Paradox with `groupby` and disaggregation
- ✓ Simulate self-selection and compute naive vs. true treatment effects
- ✓ Run an RCT simulation and see how random assignment eliminates bias
- ✓ Compare estimates in a summary table and write a causal conclusion

Week 2 Assignment: Causality & Data

- T1. Simulate a spurious correlation (ice cream vs drownings) and identify the confounder
- T2. Reveal the confounder with side-by-side scatterplots
- T3. Reproduce Simpson's Paradox: compute aggregate vs disaggregate admission rates
- T4. Simulate a job training program with self-selection and compute the naive estimate
- T5. Visualize selection bias: show that education differs between treated/control groups
- T6. Run an RCT simulation and compare the unbiased estimate with the naive one
- T7. Summary table + 3–4 sentence conclusion: what did you learn about correlation vs causation?

Due: Before Week 3 class • Moodle → ECO447 → Week 2 → Assignment

Grading: Code 30%, Interpretation 40%, Completeness 20%, Presentation 10%

Week 3: Exploratory Data Analysis and Summary Statistics

Preparation

- Complete and submit the Week 2 assignment
- Install/verify: numpy, pandas, matplotlib, seaborn in your Colab
- Read: Bruce et al. (2026), Chapter 1 — Exploratory Data Analysis

Next week we shift from “why causality matters” to “how to look at data” — the practical EDA skills that every analysis (causal or predictive) begins with.

Contact: onurpolat@hacettepe.edu.tr