Causal Inference I: Correlation, Causation, and Counterfactuals

So far in this course we have built tools that are very good at answering one kind of question: **given what I observe, what should I predict?** Clustering, regression, and classification all learn associations in data.

But many of the questions we actually care about as data scientists are of a different kind. They are questions about **what would happen if we acted**:

Lecture Overview

In this first lecture on causal inference we cover:

A fitted regression is not a causal claim

You just fit regressions. What does a coefficient mean?

In the last two lectures you fit models like

```
1  smf.ols("Lottery ~ Literacy + Wealth + Region", data=df).fit().summary()
```

and read off a coefficient, a confidence interval, and a p-value for each variable.

A regression you would recognize

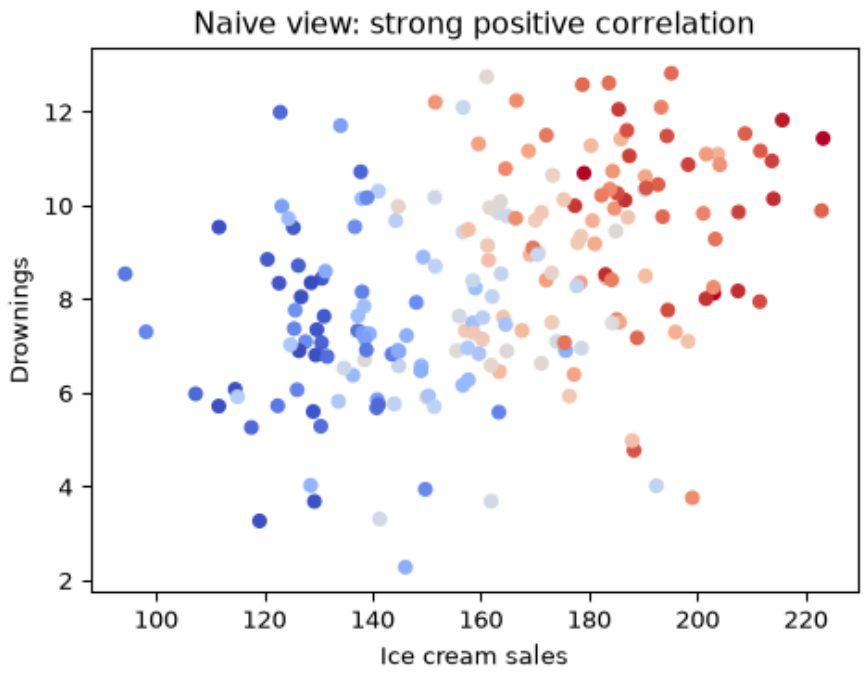
Two variables measured daily across the summer: **ice cream sales** and **drowning deaths**. Let's fit the regression exactly as we did in Lecture 11.

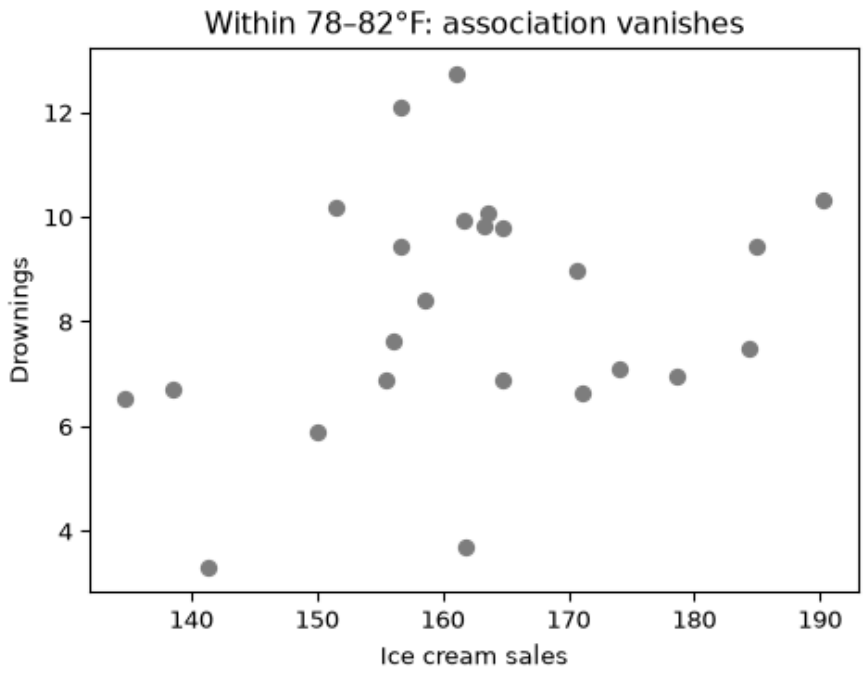
▶ Code

```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      3.1654      0.840      3.770      0.000       1.510       4.821
ice_cream      0.0320      0.005      6.262      0.000       0.022       0.042
==============================================================================
```

Correlation is not causation

Confounding, visualized





Confounding

Definition

A **confounder** is a variable that causally affects **both** the treatment (or explanatory variable) X **and** the outcome Y. It creates an association between X and Y that is not due to X causing Y.

Multiple regression can *sometimes* fix this — **if** we measure the confounder and include it:

▶ Code

```
slope on ice_cream, ignoring temp:  +0.032
slope on ice_cream, adjusting temp: -0.007   (truth: 0)
```

Same trap, higher stakes

A hospital analytics team fits a model on patient records and finds a strong, statistically significant relationship:

> Patients who **visit the hospital** are *more* likely to die within the year than patients who do not.

Prediction vs. intervention

Two different questions

Prediction (association)

$$P(Y \mid X = x)$$

"Among people I *observe* to have $X = x$, what is the distribution of Y?"

This is what supervised learning — including every regression you have fit — estimates.

Intervention (causation)

$$P(Y \mid do(X = x))$$

"If I *set* $X = x$ for everyone, what is the distribution of Y?"

This is what a decision-maker needs. (Pearl 2009)

The ladder of causation

Judea Pearl describes three "rungs" of reasoning (Pearl and Mackenzie 2018):

Rung	Question	Example
1. **Association**	What is?	*Seeing* — customers who buy diapers also buy beer
2. **Intervention**	What if I do?	*Doing* — if I discount diapers, will beer sales rise?
3. **Counterfactual**	What if I had?	*Imagining* — would this patient have recovered had they not taken the drug?

Where do spurious associations come from?

A correlation between X and Y can arise for several reasons — only one of which is "X causes Y":

1. See Tyler Vigen's *Spurious Correlations* (https://tylervigen.com/spurious-correlations) for entertaining examples, e.g. US cheese consumption vs. deaths by bedsheet entanglement.

Potential outcomes

The potential outcomes framework

To reason precisely, we use the **Neyman–Rubin potential outcomes** model (Splawa-Neyman et al. 1990; Rubin 1974).

For each unit i (a patient, user, student) and a binary **treatment** $T_i \in \{0, 1\}$:

The fundamental problem of causal inference

We only ever observe the outcome corresponding to the treatment the unit actually received:

$$Y_i^{\text{obs}} = T_i\, Y_i(1) + (1 - T_i)\, Y_i(0).$$

Unit	T_i	$Y_i(0)$	$Y_i(1)$	τ_i
Ann	1	**?**	7	?
Bob	0	4	**?**	?
Cara	1	**?**	9	?
Dan	0	3	**?**	?

The Average Treatment Effect (ATE)

Since individual effects are unknowable, we target **population averages**.

Definition

The **Average Treatment Effect** is the expected difference between a unit's two potential outcomes, averaged over the whole population:

$$\mathrm{ATE} = \mathbb{E}[\,Y(1) - Y(0)\,] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)].$$

Selection bias: why we can’t just compare treated vs. untreated

The tempting estimator is the **naive difference in means**:

$$\underbrace{\mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0]}_{\text{what we can measure}}.$$

Simpson’s paradox

Confounding in numbers: a kidney stone study

A famous real study compared two treatments for kidney stones (Charig et al. 1986). Here are the recovery rates:

	Treatment A (surgery)	Treatment B (less invasive)
Small stones	81/87 = **93%**	234/270 = **87%**
Large stones	192/263 = **73%**	55/80 = **69%**
Overall	273/350 = **78%**	289/350 = **83%**

What is going on?

▶ Code

```
Treatment A: small=93%, large=73%, overall=78%
Treatment B: small=87%, large=69%, overall=83%
```

Simpson’s paradox

Definition

Simpson’s paradox occurs when an association between X and Y that holds in **every subgroup** of the data **reverses** (or disappears) when the subgroups are combined. It is caused by a confounder whose distribution differs across the levels of X.

A second real example: Berkeley admissions

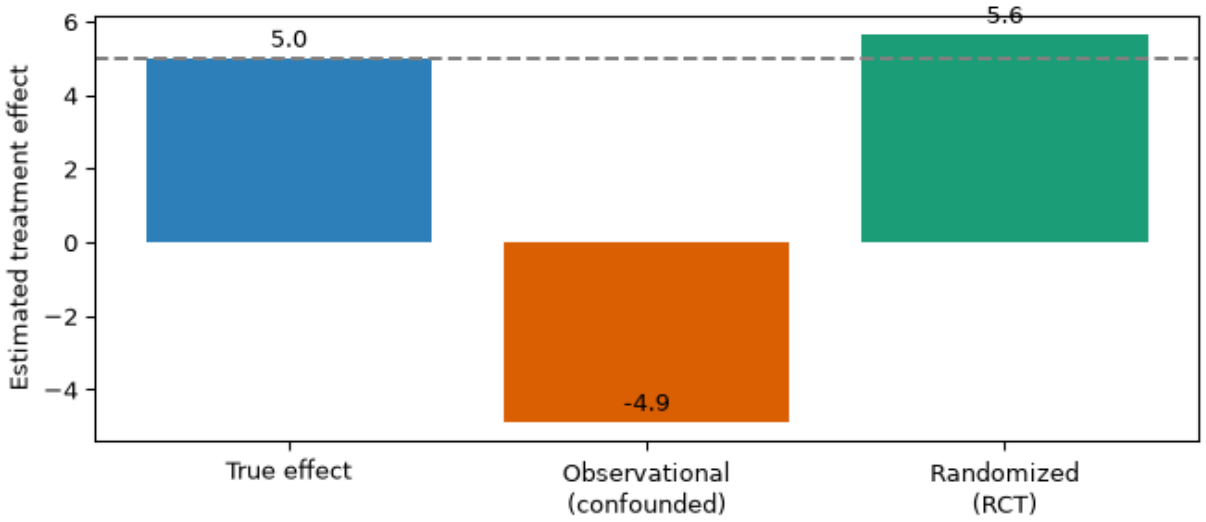
In 1973, UC Berkeley graduate admissions data showed (Bickel et al. 1975):

Randomization

The gold standard: randomized experiments

How do we break confounding by design? **Randomly assign** the treatment.

Randomization removes confounding



The true effect is $+5$. The **observational** estimate is badly biased (sick patients both got the drug *and* had worse outcomes). The **randomized** estimate recovers the truth.

RCTs in the wild

Randomized experiments are everywhere once you look:

🎯 Quick check: think–pair–share (5 min)

For each headline, name the most likely **confounder** and state the correct causal caveat.

Turn to a neighbor, pick the trickiest one, and be ready to share your reasoning.

Summary

Summary

References

- *Causal Inference: What If* — Hernán & Robins (free PDF)
- *The Book of Why* — Pearl & Mackenzie (popular intro)
- *Causal Inference: The Mixtape* — Cunningham (free online)
- *Mastering 'Metrics* — Angrist & Pischke

Bibliography

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Bickel, P. J., E. A. Hammel, and J. W. O'Connell. 1975. "Sex Bias in Graduate Admissions: Data from Berkeley." *Science* 187 (4175): 398–404. https://doi.org/10.1126/science.187.4175.398.

Charig, C. R., D. R. Webb, S. R. Payne, and J. E. Wickham. 1986. "Comparison of Treatment of Renal Calculi by Open Surgery, Percutaneous Nephrolithotomy, and Extracorporeal Shockwave Lithotripsy." *British Medical Journal (Clinical Research Ed.)* 292 (6524): 879–82. https://doi.org/10.1136/bmj.292.6524.879.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (5): 688–701.