Finance PhD Student Tutorial: Identification and Research Design

Kyle Zimmerschied[1]

[1]University of Arkansas

April 29, 2026

Today's Agenda

Focus: How do I causally answer questions of interest?

1. **Opening Discussion and Reflection (10 min):** How comfortable am I in causal inference?
2. **Overview (20 min):** Integrating causal inference into research process
3. **Methods (30 min):** How causal inference is achieved?

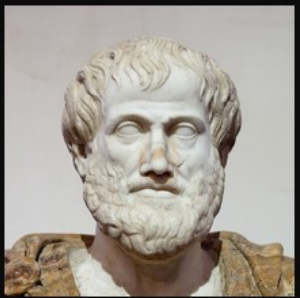
Opening Discussion

Reflect on Your Experience

1. What is my background in causal inference?
2. I have the ability to build research designs and address comments regarding the strength of my design...

Discussion time: 10 minutes

Why Causal Inference Matters?



We do not know a truth without knowing its cause.

~ Aristotle

AZ QUOTES

The Fundamental Problem of Causal Inference

	$Y_i(1)$	$Y_i(0)$
Treated unit ($D_i = 1$)	**Observed**	Counterfactual
Control unit ($D_i = 0$)	Counterfactual	**Observed**

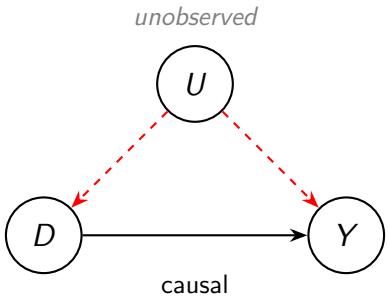
- Each unit reveals **only one** potential outcome — never both
- The missing cell is the counterfactual: what *would have* happened under the other treatment
- The **individual treatment effect** $\tau_i = Y_i(1) - Y_i(0)$ is *never directly observable*
- Every identification strategy is an attempt to **credibly approximate the missing cell**

Causal Inference in Research Development

1. What is the causal relationship of interest?
   - Is my research question a fundamentally identified question?
2. What would the ideal experiment look like to estimate causal effects of interest?
   - How would a randomized control trial answer this question?
3. What's your identification strategy?
   - Identification strategy: Method in which a researcher uses observational data to approximate a real experiment
4. What is the mode of statistical inference
   - Population to be studied, sample of interest, standard errors

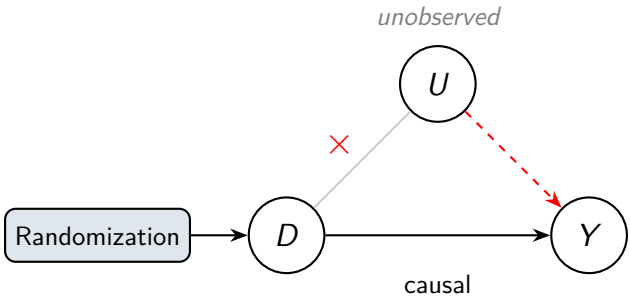
- Key: In research development, should have compelling approach & understand what this approach buys you
  - Crucial: Understanding of design strength and weaknesses comes from well-defined understanding of institutional features

Why Correlation $\neq$ Causation: The DAG



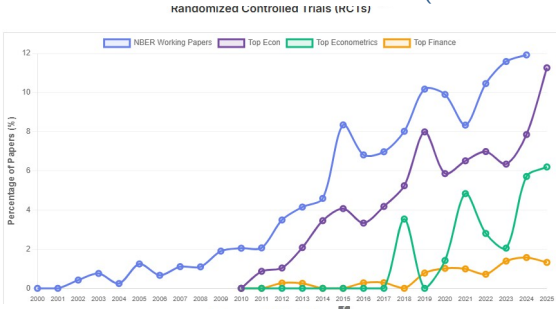
- $D \to Y$: the **causal effect** we want to estimate
- $D \leftarrow U \to Y$: the **backdoor path** — U drives both treatment and outcome
- Naive comparison (OLS regression) picks up **both** paths — this is selection bias
- Example: Sophisticated individuals (U) have higher levels of financial literacy (D) *and* stronger financial health (Y)

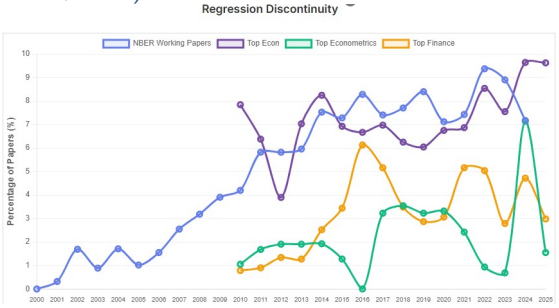
How RCTs Close the Backdoor Path

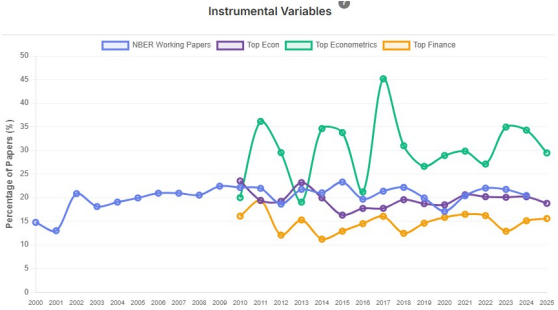


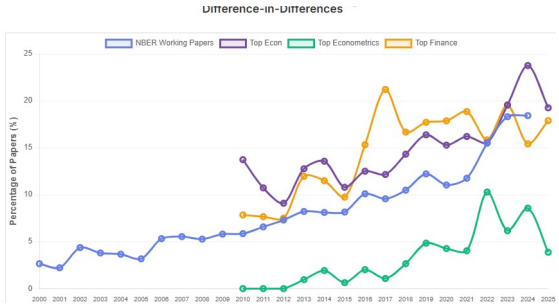
- Randomisation assigns D **independently** of U: $D \perp U$
- The backdoor path $D \leftarrow U \rightarrow Y$ is **severed by design**
- Now the only path from D to Y is the causal arrow
- $E[Y \mid D = 1] - E[Y \mid D = 0]$ recovers the **ATE** — no assumptions needed
- Identification strategies **close backdoor paths through assumptions** (Regression Discontinuity Design, Instrumental Variables, Differences-in-Differences. . .)

The Rise of Causal Inference Methods (Goldsmith-Pinkham, 2025)









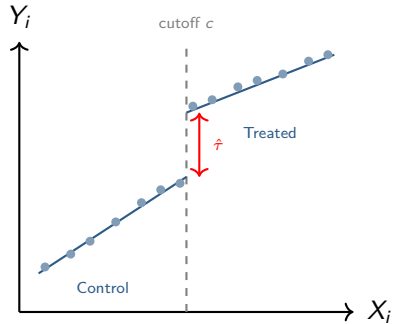
Regression Discontinuity Design

Identifying Variation

- Treatment assigned by a **cutoff rule** in a running variable X_i
- Units just above and below the threshold are **as-good-as randomly assigned**
- Causal effect estimated at the threshold: $\hat{\tau}_{RDD} = \lim_{x \downarrow c} E[Y \mid X = x] - \lim_{x \uparrow c} E[Y \mid X = x]$

Identifying Assumptions

- **Continuity:** $E[Y_i(0) \mid X_i]$ and $E[Y_i(1) \mid X_i]$ are continuous at the cutoff
- **No manipulation:** Units cannot precisely sort around the threshold
  - Bunching estimator actually relies upon this for causal inference



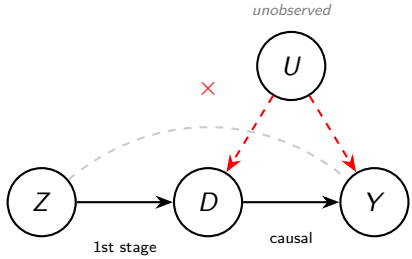
Practical examples:

1. Using market capitalization for delisting or index inclusion
2. Access to credit given credit scores
3. Age cut-offs for program access
4. Voting outcomes in policy or elections

Instrumental Variables

Identifying Variation

- An instrument Z_i provides **exogenous variation** in treatment D_i
- IV isolates only the variation in D_i driven by Z_i — purging the endogenous part
- Estimates the **LATE**: causal effect for *compliers* — units whose treatment status is changed by Z_i
- $\hat{\tau}_{IV} = \dfrac{E[Y_i \mid Z_i = 1] - E[Y_i \mid Z_i = 0]}{E[D_i \mid Z_i = 1] - E[D_i \mid Z_i = 0]}$



Identifying Assumptions

- **Relevance:** Z_i has a strong first stage — $\text{Cov}(Z_i, D_i) \neq 0$
- **Exclusion:** Z_i affects Y_i *only through* D_i
- **Independence:** $Z_i \perp \{Y_i(0), Y_i(1)\}$ — instrument is as-good-as random

Practical examples:

1. Effects of distance to instrument for treatment at private-equity backed nursing home
2. Effects of weather to instrument for agricultural production and bank size

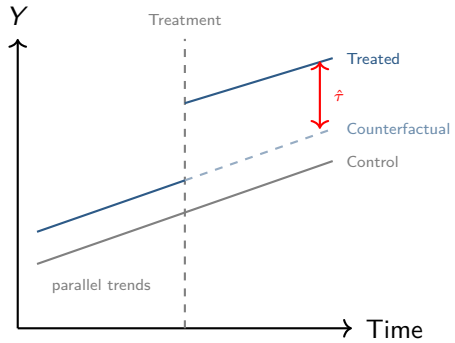
Differences-in-Differences

Identifying Variation

- Compares **change over time** in treated vs. control units
- Differences out time-invariant unobservables and common time trends
- $\hat{\tau}_{DiD} = (E[Y_{post} - Y_{pre} \mid D = 1]) - (E[Y_{post} - Y_{pre} \mid D = 0])$

Identifying Assumptions

- **Parallel trends:** Absent treatment, treated and control units would have followed the same trend
- **No anticipation:** Treatment effect begins only after treatment is received



Practical examples:

1. Effects of policy increasing cost of filing for bankruptcy on consumer credit terms
2. Effects of investment banking consolidation on underwriting spreads in municipal finance

Before Next Week (1 Hour)

- Explore (30 minutes):
  1. Causal Inference: *The Mixtape*
  2. Mostly Harmless Econometrics: An Empiricist's Companion
  3. Paul Goldsmith-Pinkham's causal modeling series
  4. Northwestern University Causal Inference Workshop 2026
- Reflect (30 minutes):
  1. How compelling is my current research design?
  2. What assumptions am I relying upon and what arguments are they sufficient or insufficient to address?
  3. Can I provide a compelling design to answer my research question of interest?

References

Angrist, Joshua D, & Pischke, Jörn-Steffen. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Cunningham, Scott. 2021. *Causal inference: The mixtape*. Yale university press.

Goldsmith-Pinkham, Paul. 2025. *Economics Literature Search: Empirical Methods Trends Dashboard*. `https://paulgp.com/econlit-pipeline/dashboard.html`. Accessed: 2026-04-29.

Extra: Correlation $\neq$ Causation: The Selection Bias Problem

What we *want*	*ATE*	$= E[Y_i(1) - Y_i(0)]$
What we *observe*	Naive comparison	$= E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0]$

It can be shown that the naive comparison equals:

$$\underbrace{E[Y_i(1) - Y_i(0)]}_{\text{ATE}} + \underbrace{E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0]}_{\text{Selection Bias}}$$

- **Selection bias:** Treated and control units differ in their *baseline outcomes* $Y_i(0)$ — they were never comparable to begin with
- Example: Firms that adopt ESG policies may already be higher performers — comparing them to non-adopters confounds the ESG effect with pre-existing differences
- Correlation picks up **both** the treatment effect **and** the selection bias

Extra: How RCTs Solve the Selection Bias Problem

	Treated ($D_i = 1$)	Control ($D_i = 0$)
$E[Y_i(1)]$	**Observed**	Counterfactual
$E[Y_i(0)]$	Counterfactual	**Observed**

Randomisation ensures: $\{Y_i(1), Y_i(0)\} \perp D_i$

- Independence kills selection bias: $E[Y_i(0) \mid D_i = 1] = E[Y_i(0) \mid D_i = 0]$
- So the naive comparison **recovers the ATE**:

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = ATE$$

- The control group becomes a **valid counterfactual** for the treated group — by design, not assumption
- All observational strategies try to **recreate this independence** through assumptions