A Brief Guide to Causal Inference in Social Science Research

Frost Center Colloquium

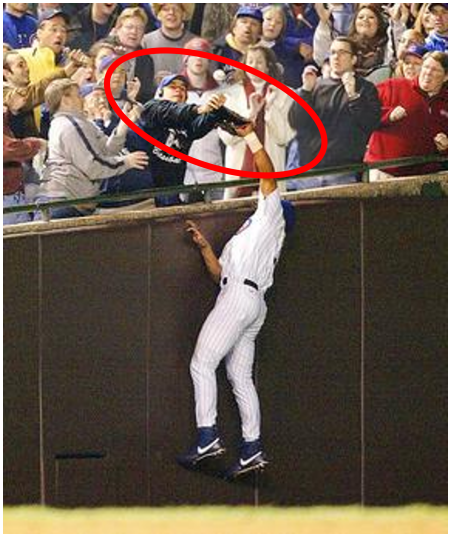
Neil Silveus
Assistant Professor of Economics
Frost Center Fellow

Frost Center for Social Science Research
Hope College

March 18, 2026

"What might have been..."

The essential question of causality

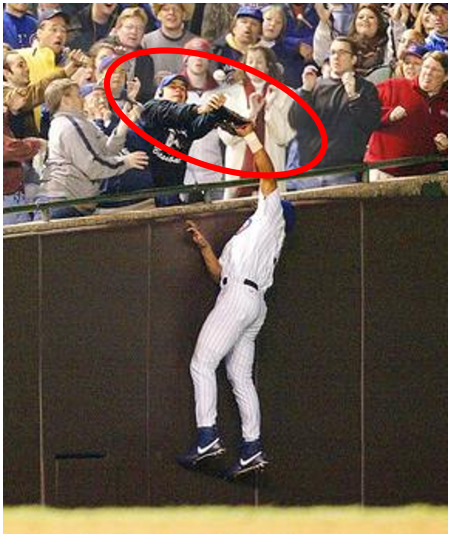


"What might have been..."

The essential question of causality

2003 NLCS, Game 6. Cubs lead 3–0 in the 8th inning.

Fan Steve Bartman deflects a catchable foul ball.
The Cubs surrender 8 runs and lose the series.



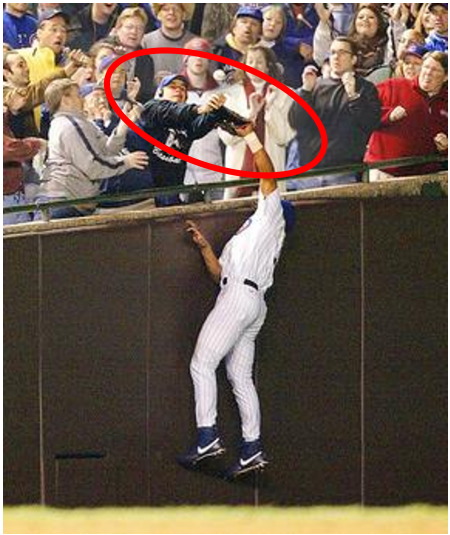
"What might have been..."

The essential question of causality

2003 NLCS, Game 6. Cubs lead 3–0 in the 8th inning.

Fan Steve Bartman deflects a catchable foul ball.
The Cubs surrender 8 runs and lose the series.

What would have happened if he hadn't touched it?



Today's Agenda

Talk ∼30 min

1. Motivation: causation vs. correlation
2. The fundamental problem of causal inference
3. When RCTs are not feasible
4. Three quasi-experimental methods:
   - Difference-in-Differences (DiD)
   - Regression Discontinuity (RD)
   - Synthetic Controls
5. Further reading

Workshop ∼20 min

- Share your research questions
- Identify which method(s) might apply
- Open Q&A

Come ready to talk about your work!

A Thought Experiment

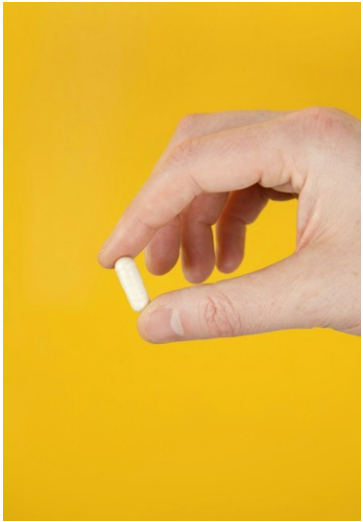
Imagine waking up tomorrow as a member of the **U.S. Food and Drug Administration**.

Your first task: assess the effectiveness of a new blood pressure medication, **Potentia™**, in lowering patient blood pressure.

Your approach: release the drug publicly and compare blood pressure levels of those who *take it* with those who *do not*.

Is this a good approach?

Why or why not? What could go wrong with this comparison?



The Core Question of Causality

"What might have been?"

- Did the policy *cause* the outcome, or would it have happened anyway?
- Does attending college *cause* higher earnings, or do higher-ability individuals both attend college *and* earn more?
- Does moving to a better neighborhood *cause* better child outcomes, or do motivated families both move *and* have children who are observably better off?

The core problem

We observe the world only once. We cannot observe what would have happened without the policy, treatment, or event — the **counterfactual**.

Potential Outcomes and Selection Bias

Notation (Rubin, 1974):

- $Y_i(1)$: outcome for unit i *if treated*
- $Y_i(0)$: outcome for unit i *if not treated*
- We only ever observe **one** of these — the *fundamental problem of causal inference*

Average Treatment Effect on the Treated (ATT):

$$\text{ATT} = \mathbb{E}\big[Y_i(1) - Y_i(0) \;\big|\; D_i = 1\big]$$

Selection bias

A naive comparison $\mathbb{E}[Y \mid D{=}1] - \mathbb{E}[Y \mid D{=}0]$ *conflates* the treatment effect with **selection bias** — treated and untreated units are systematically different.

The Gold Standard: Randomized Controlled Trials

Random assignment eliminates selection bias:

$$\{Y_i(0),\ Y_i(1)\} \perp\!\!\!\perp D_i$$

Treatment is independent of potential outcomes, so a simple difference in means recovers the ATE — no confounding.

Basic requirements for a good RCT:

- Clear, well-defined treatment
- Verified balance across groups
- Pre-registered analysis plan
- IRB approval and informed consent

Moving to Opportunity (HUD, 1994)

Randomly assigned housing vouchers to low-income families.

Chetty, Hendren & Katz (2016) linked MTO data to IRS tax records and found large positive effects on children's long-run earnings — *especially* for those who moved young.

When RCTs Are Not Feasible

Random experiments are the gold standard... but often infeasible:

Cost	Ethics	Real-world policy
Large-scale randomization is expensive and logistically demanding	We cannot randomly assign poverty, incarceration, or environmental exposure	Policies are already enacted; we want to evaluate what actually happened

Solution: Find settings where treatment is *as good as random*
— **quasi-experimental methods**

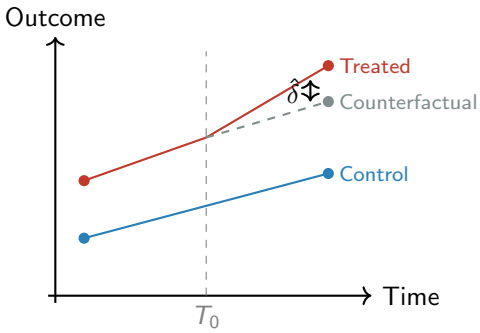
Difference-in-Differences: The Idea

Setting:

- Panel data: units observed before and after an event
- Some units receive treatment at time T_0; others do not
- Compare the *change* in outcomes for treated units to the *change* for untreated units

Classic example: Card & Krueger (1994) — New Jersey raised its minimum wage; neighboring Pennsylvania did not. Did NJ restaurant employment fall?

The control group provides the **counterfactual trend**.



Difference-in-Differences: The Math

DiD estimator (2×2 case):

$$\hat{\delta}_{\text{DiD}} = \underbrace{\left(\bar{Y}_{T,\text{post}} - \bar{Y}_{T,\text{pre}}\right)}_{\text{treated group change}} - \underbrace{\left(\bar{Y}_{C,\text{post}} - \bar{Y}_{C,\text{pre}}\right)}_{\text{control group change}}$$

Equivalent regression:

$$Y_{it} = \alpha + \beta\,\text{Treated}_i + \gamma\,\text{Post}_t + \delta\,(\text{Treated}_i \times \text{Post}_t) + \varepsilon_{it}$$

$\hat{\delta}$ is the ATT estimate — the coefficient on the interaction term.

Two-way fixed effects (TWFE) — multiple periods and units:

$$Y_{it} = \alpha_i + \lambda_t + \delta\,D_{it} + \varepsilon_{it}$$

α_i absorbs unit-level differences; λ_t absorbs common features across units within time unit.

DiD: The Key Assumption

Parallel Trends Assumption

In the *absence of treatment*, the treated and control groups would have followed the **same trend** over time:

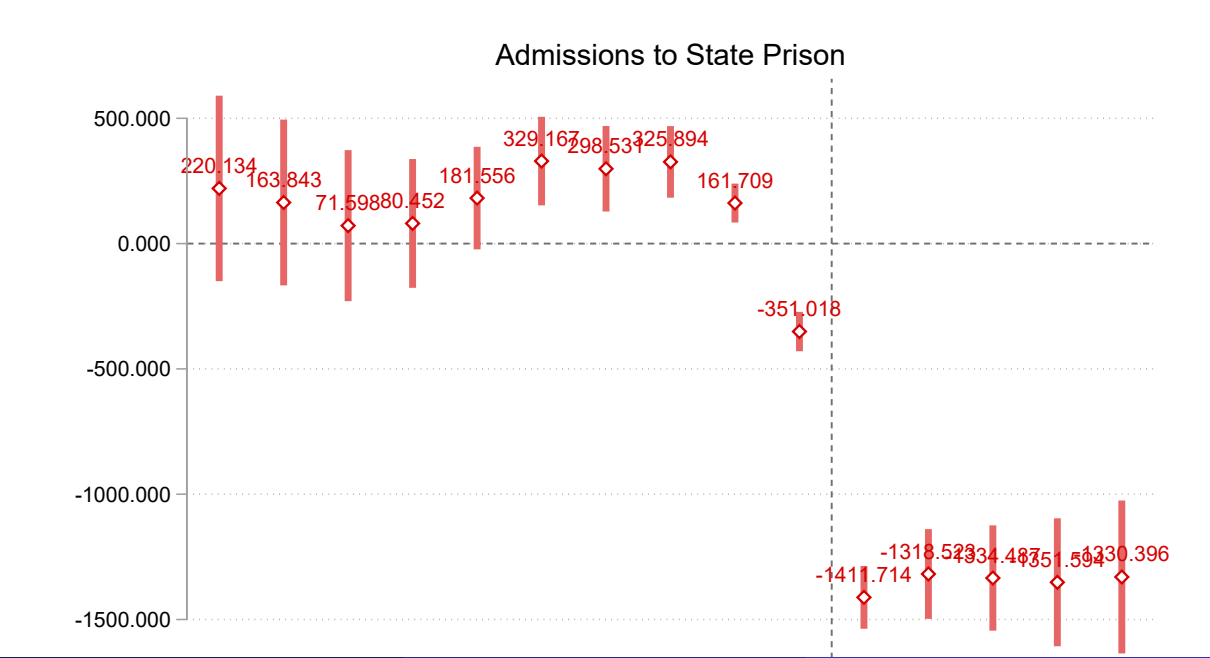
$$\mathbb{E}\big[Y_{it}(0) - Y_{i,t-1}(0) \mid \text{Treated}_i = 1\big] = \mathbb{E}\big[Y_{it}(0) - Y_{i,t-1}(0) \mid \text{Treated}_i = 0\big]$$

How to support the assumption:

- Plot pre-treatment outcome trends
- Event study: test whether pre-period "effects" are near zero
- Choose a plausible, similar comparison group

Common threats:

- Treated units are fundamentally different
- Simultaneous policy changes in treated units
- Anticipation effects before T_0
- Differential time trends


Admissions to State Prison
500.000
0.000
-500.000
-1000.000
-1500.000
220.134
163.843
71.598
80.452
181.556
329.167
298.531
325.894
161.709
-351.018
-1411.714
-1318.523
-1434.487
-1351.594
-1230.396

Difference-in-Differences: R Code

Simple 2×2 OLS:

```r
# treated:post interaction = delta_DiD
model <- lm(
  outcome ~ treated * post,
  data = panel_data
)
summary(model)
```

Two-way fixed effects (TWFE):

```r
library(fixest)

# Unit + time fixed effects
twfe <- feols(
  outcome ~ i(year, treated, ref = 0) |
            unit_id + year,
  data    = panel_data,
  cluster = ~unit_id
)

# Event study plot
iplot(twfe, main = "Event Study")
```

`fixest` (Bergé, 2018) is fast and handles clustering, multiple fixed effects, and event-study plots in one workflow.

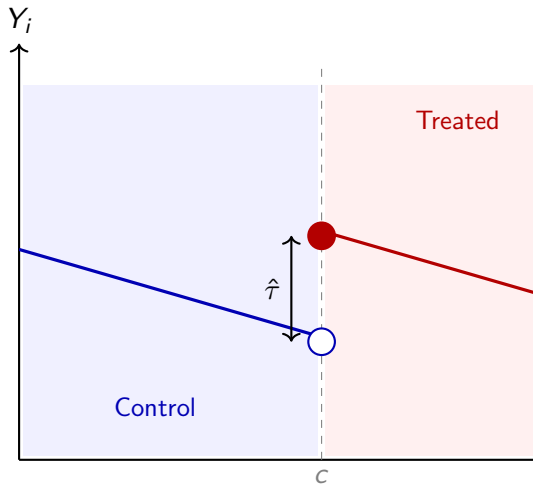
Regression Discontinuity: The Idea

Setting: Treatment is determined by whether a **running variable** X_i crosses a known **cutoff** c.

- Units just above and just below c are *very similar*
- Any discontinuous jump in the outcome at c is the causal effect of treatment

Classic examples:

- Test score $\geq$ cutoff $\Rightarrow$ scholarship eligibility
- Vote share $\geq 50\%$ $\Rightarrow$ election winner
- Age ≥ 65 $\Rightarrow$ Medicare eligibility



Regression Discontinuity: The Math

Sharp RD: Treatment is a deterministic function of X_i:

$$D_i = \mathbf{1}[X_i \geq c]$$

RD estimand — Local Average Treatment Effect (LATE) at the cutoff:

$$\hat{\tau}_{\text{RD}} = \lim_{x \downarrow c} \mathbb{E}[Y_i \mid X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i \mid X_i = x]$$

Estimated with local polynomial regression on either side of c.

Key assumption: Continuity

Potential outcomes are continuous at the cutoff.

Units cannot *precisely* manipulate X_i to sort just above or just below c.

Fuzzy RD

When crossing c raises the *probability* of treatment (not a certainty), use 2SLS with $\mathbf{1}[X_i \geq c]$ as an instrument for D_i.

Regression Discontinuity: R Code

```r
library(rdrobust)

# Sharp RD estimate at cutoff c
rd <- rdrobust(
  y = data$outcome,
  x = data$running_var,
  c = cutoff_value      # e.g., c = 0 or c = 50
)
summary(rd)
# Reports: local polynomial estimate, MSE-optimal bandwidth,
#          bias-corrected confidence interval

# Bin-scatter plot around the cutoff
rdplot(
  y     = data$outcome,
  x     = data$running_var,
  c     = cutoff_value,
  title = "Regression Discontinuity"
)
```

`rdrobust` (Calonico, Cattaneo & Titiunik, 2014) selects the MSE-optimal bandwidth and applies bias correction automatically.

Synthetic Controls: The Quantitative Case Study

Setting:

- *One* (or very few) treated unit(s) — a state, country, or city
- Treatment is a major policy or event
- DiD is hard: no obvious comparison group; one treated region

Key idea: Construct a **synthetic** version of the treated unit from a weighted combination of *donor pool* units that best matches its pre-treatment trajectory.

Classic example: Abadie, Diamond & Hainmueller (2010) — Effect of California's 1988 tobacco tax (Prop 99) on cigarette consumption.

Why this is powerful

- Transparent: weights are explicit
- No global parallel-trends assumption needed
- Rich pre-period fit *validates* the counterfactual
- Designed for the "$n = 1$ treated unit" problem

Synthetic Controls: Setup

Notation:

- $J+1$ units, T time periods; unit 1 is treated starting at T_0
- Units $2, \ldots, J+1$ form the **donor pool** (never treated)
- $\mathbf{X}_1$: vector of pre-treatment predictors for the treated unit
- $\mathbf{X}_0$: matrix of pre-treatment predictors for donor units

The synthetic control is a convex combination of donor units:

$$\hat{Y}_{1t}(0) = \sum_{j=2}^{J+1} w_j^* \, Y_{jt}, \qquad t > T_0$$

with weights $W^* = (w_2^*, \ldots, w_{J+1}^*)$ satisfying $w_j \geq 0$ and $\sum_j w_j = 1$.

Estimated treatment effect:

$$\hat{\alpha}_{1t} = Y_{1t} - \hat{Y}_{1t}(0), \qquad t > T_0$$

Synthetic Controls: Finding the Weights

Choose W^* to minimize pre-treatment discrepancy:

$$W^* = \arg\min_W \left\| \mathbf{X}_1 - \mathbf{X}_0 \, W \right\|_{\mathbf{V}}^2 = \arg\min_W \sum_{k=1}^{K} v_k \left(X_{1k} - \sum_{j=2}^{J+1} w_j \, X_{jk} \right)^2$$

subject to: $w_j \geq 0 \;\; \forall j$ and $\sum_j w_j = 1$

What goes in **X**?

- Pre-treatment outcome lags (crucial)
- Economic / demographic predictors
- More pre-periods $\Rightarrow$ more credibility

Diagnosing fit

Compare Y_{1t} to $\hat{Y}_{1t}(0)$ in the pre-period. Close tracking before T_0 validates the post-period counterfactual.

Synthetic Controls: Inference via Placebo Tests

Traditional standard errors do not apply when $n = 1$. Instead: **permutation / placebo inference**.

1. Apply the synthetic control method to *each donor unit* as if it were the treated unit
2. Compute a placebo treatment effect for each
3. Compare the treated unit's estimated effect to the distribution of placebo effects

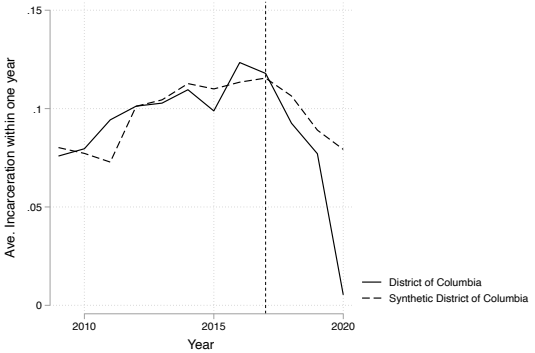
Inference rule

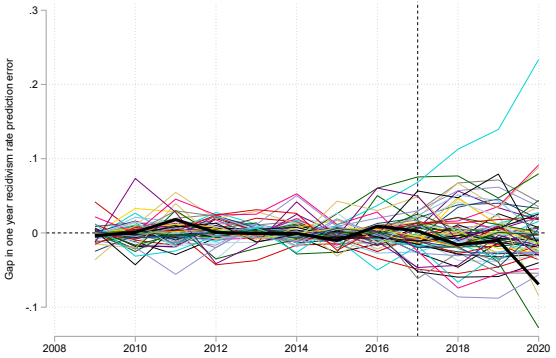
If the treated unit's gap is *extreme* relative to the placebo distribution, we have evidence of a real effect.

$$\text{Pseudo } p\text{-value} = \frac{\text{rank of treated unit's effect}}{\text{total number of units}}$$

Exposition follows Cunningham (2021), Causal Inference: The Mixtape — `mixtape.scunning.com`

Example: Fair Chance Housing in Washington DC





Suggestive, but not significant evidence that FCH in DC led to reduction in recidivism probabilities.

Synthetic Controls: California Prop 99

Abadie, Diamond & Hainmueller (2010): Did California's 1988 tobacco tax reduce cigarette consumption?

- **Treated unit:** California
- **Donor pool:** 38 other U.S. states with no major tobacco legislation
- **Outcome:** Per-capita cigarette sales (packs/year)
- **Synthetic California:** weighted combination of Colorado, Connecticut, Montana, Nevada, Utah, and others
- **Finding:** By 1997, California consumed ≈26 fewer packs/person/year than its synthetic counterfactual

Pre-treatment fit

Synthetic California closely tracks actual California from 1970–1988, lending credibility to the post-1988 counterfactual.

Placebo test

California's post-1988 gap is among the largest of any state in the donor pool — a compelling permutation test result.

Synthetic Controls: R Code

```r
library(tidysynth)   # modern tidy interface; install.packages("tidysynth")

synth_out <- smoking_data |>
  synthetic_control(
    outcome  = cigsale,         # per-capita cigarette sales
    unit     = state,           # unit identifier
    time     = year,            # time identifier
    i_unit   = "California",    # treated unit
    i_time   = 1988,            # treatment year
    generate_placebos = TRUE    # enable permutation inference
  ) |>
  generate_predictor(time_window = 1980:1988,
    avg_sales = mean(cigsale),
    avg_beer  = mean(beer)) |>
  generate_weights(optimization_window = 1970:1988) |>
  generate_control()

synth_out |> plot_trends()         # treated vs. synthetic control
synth_out |> plot_placebos()       # permutation / placebo inference
synth_out |> grab_unit_weights()   # inspect donor pool weights
```

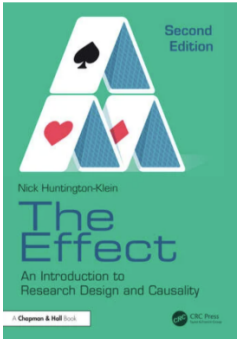
Package: `tidysynth` (Dunford, 2021). Data: `smoking` dataset shipped with the `Synth` package.

Which Method Should You Use?

Method	What you need	Key assumption
Diff-in-Diff (DiD)	Panel data; pre/post periods; treated and control group	Parallel trends: groups would have trended together absent treatment
Reg. Discontinuity (RD)	Running variable with a known cutoff that determines treatment	Continuity at cutoff; units cannot precisely manipulate the running variable
Synthetic Controls	Aggregate units; few treated; long pre-treatment history	Pre-treatment fit is good; no interference across units

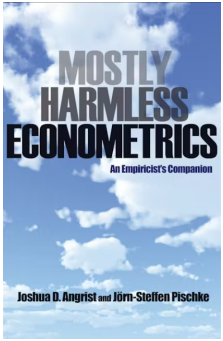
The best method is the one that most credibly mimics random assignment for your specific research setting.

Further Reading



The Effect
Nick Huntington-Klein
Introductory; undergrad
Free online:
theeffectbook.net

Causal Inference: The Mixtape
Scott Cunningham
Adv. UG / Master's
Free online:
mixtape.scunning.com



Mostly Harmless Econometrics
Angrist & Pischke
Graduate / practitioner
Classic reference text

Workshop: Let's Talk About Your Research

$\sim$20 minutes — bring your research questions!

Questions to think about:

- What is your core research question?
- What is the *treatment* or *event* you want to study?
- What is your outcome variable?
- Do you have data on multiple units over time?
- Is there a natural threshold or cutoff?
- Did a policy change affect only some units or regions?

Quick method-matching guide

DiD — before/after period + treated/control group

RD — a score or threshold determines who gets treatment

Synthetic Controls — one region / country, long history, no clean comparison group

No question too basic — ask anything about methods, data, or interpretation!