Causal Inference

A brief introduction to causal reasoning and causal modeling

Hedibert Freitas Lopes

November 2026

Insper ▪ Advanced Econometrics ▪ Professional Master in Economics

These slides were prepared with AI assistance (Claude, Anthropic)
from the author's lecture notes *Causal Inference* (Oct. 2024).

Sources and roadmap

Main sources

- Pearl, Glymour & Jewell (2016) *Causal Inference in Statistics: A Primer*, Ch. 1–2
- Daniels, Linero & Roy (2024) *Bayesian Nonparametrics for Causal Inference and Missing Data*, Ch. 1
- Angrist & Pischke (2015) *Mastering 'Metrics*, Ch. 4

Roadmap

1. Simpson's paradox
2. Why causation? The story behind the data
3. Chains, forks and colliders
4. d-separation
5. Confounding: two examples
6. Potential outcomes and estimands
7. Identification & propensity scores
8. Instrumental variables
9. Difference-in-differences
10. Regression discontinuity

A bit of history: Readings in Statistics & Econometrics (Fall 2015)

A seminar series on causality I organized at Insper, almost a decade ago:

1. Haavelmo (1943) Statistical implications of a system of simultaneous equations. *Econometrica*
2. Rubin (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psych.*
3. Holland (1986) Statistics and causal inference (with discussion). *JASA*
4. Pearl (1995) Causal diagrams for empirical research (with discussion). *Biometrika*
5. Angrist, Imbens & Rubin (1996) Identification of causal effects using IVs (with discussion). *JASA*
6. Dawid (2000) Causal inference without counterfactuals (with discussion). *JASA*
7. Vansteelandt & Goetghebeur (2003) Generalized structural mean models. *JRSS-B*
8. Heckman & Pinto (2015) Causal analysis after Haavelmo. *Econometric Theory*

Annotated bibliography and slides: `https://hedibert.org/previous-teaching/`

Simpson's paradox

Simpson's paradox

> *A statistical association that holds for an entire population is reversed in every subpopulation.*

Named after Edward Simpson (1922–2019), who popularized it.

We will look at three examples ... and at one question:

Which table should we believe?

Example 1: a drug trial (observational)

700 patients (343 women, 357 men) were given access to a drug; 350 chose to take it, 350 did not.

	Drug	No drug
Patients	273/350 = **78%**	289/350 = **83%**

recovered / total

Example 1: a drug trial (observational)

700 patients (343 women, 357 men) were given access to a drug; 350 chose to take it, 350 did not.

	Drug	No drug
Patients	273/350 = **78%**	289/350 = **83%**

recovered / total

Question: should a doctor recommend the drug?

Example 1: split by gender

	Drug	No drug
Men	$81/87 = \mathbf{93\%}$	$234/270 = \mathbf{87\%}$
Women	$192/263 = \mathbf{73\%}$	$55/80 = \mathbf{69\%}$
Combined	$273/350 = \mathbf{78\%}$	$289/350 = \mathbf{83\%}$

- The drug helps men. The drug helps women. The drug hurts *people*?
- Prescribe if we know the gender, but not if we don't? Obviously ridiculous.
- The answer is not in the numbers: we need the story behind the data, the causal mechanism that generated them.

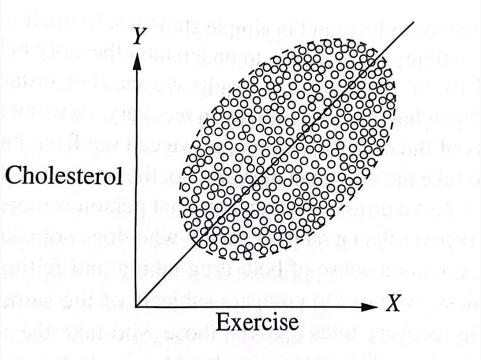
Example 2: same numbers, different variable

Now the rows record blood pressure measured after treatment.

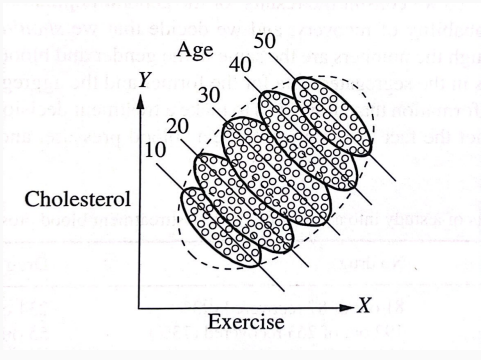
	Drug	No drug
Low BP	$81/87 = \mathbf{93\%}$	$234/270 = \mathbf{87\%}$
High BP	$192/263 = \mathbf{73\%}$	$55/80 = \mathbf{69\%}$
Combined	$273/350 = \mathbf{78\%}$	$289/350 = \mathbf{83\%}$

Question: should a doctor recommend the drug?

Example 3: exercise and cholesterol



Pooled data



Segregated by age

"The more a person exercises, the higher their cholesterol is!" ?

Why causation? The story behind the data

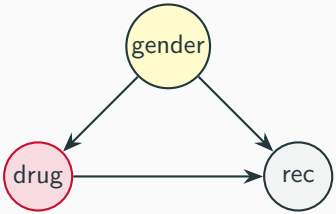
Why study causation

- To make sense of data, to guide actions and policies, and to learn from successes and failures.
- We need to estimate the effect of
  - smoking on lung cancer;
  - education on salaries;
  - carbon emissions on the climate.
- We need to understand how and why causes influence effects.

Back to Example 1: gender is a common cause

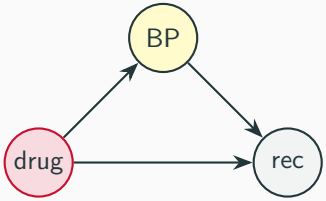
- Estrogen lowers recovery: women recover less, with or without the drug.
- Women are much more likely to take the drug.
- Gender affects both treatment and outcome: a confounder.

Use the segregated data: compare like with like, within gender. The drug helps.



Back to Example 2: blood pressure is a mediator

- The drug improves recovery *by lowering BP*, but it also has a toxic effect.
- Post-treatment BP lies on the causal path drug $\rightarrow$ BP $\rightarrow$ rec.
- Within BP strata we only see the toxic effect; we block the beneficial channel.

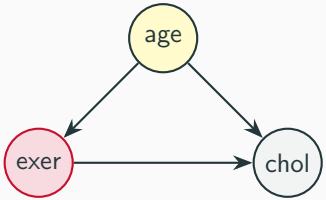


Use the aggregated data.

Same numbers, opposite answers: segregated data for gender, aggregated data for BP.

Back to Example 3: age is a common cause

- Within each age group: more exercise, lower cholesterol.
- Pooled: more exercise, higher cholesterol.
- Older people exercise more *and* have higher cholesterol regardless of exercise.
- Age is a common cause of treatment (exercise) and outcome (cholesterol) $\Rightarrow$ segregate by age.



Chains, forks and colliders

The three building blocks



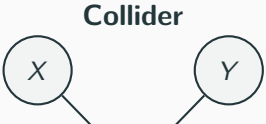
$X \not\perp\!\!\!\perp Z$

$X \perp\!\!\!\perp Z \mid Y$



$Y \not\perp\!\!\!\perp Z$

$Y \perp\!\!\!\perp Z \mid X$



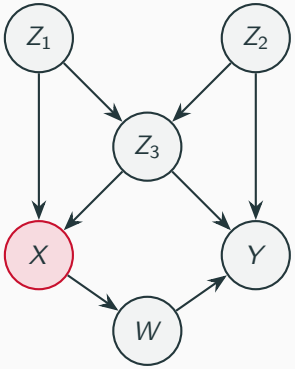
$X \perp\!\!\!\perp Y$

$X \not\perp\!\!\!\perp Y \mid Z$ (!)

Rule 1 (chains): $X \perp\!\!\!\perp Z \mid Y$ if there is only one unidirectional path between X and Z and Y intercepts it.
Rule 2 (forks): if X is a common cause of Y and Z and there is only one path between them, $Y \perp\!\!\!\perp Z \mid X$.
Rule 3 (colliders): if Z is a collision node between X and Y and there is only one path, then $X \perp\!\!\!\perp Y$ unconditionally, but they become dependent given Z *or any descendant of* Z.

A DAG with all three junctions

(G. Harrison, *Towards Data Science*, 2024)



Paths from treatment X to outcome Y:

1. $X \to W \to Y$ (the causal path)
2. $X \leftarrow Z_1 \to Z_3 \to Y$
3. $X \leftarrow Z_1 \to Z_3 \leftarrow Z_2 \to Y$
4. $X \leftarrow Z_3 \to Y$
5. $X \leftarrow Z_3 \leftarrow Z_2 \to Y$

A **junction** has exactly 3 nodes and 2 edges. The DAG has 15 junctions: 7 chains, 3 forks, 5 colliders.

Classifying the 15 junctions

Chains

$X \to W \to Y$
$Z_1 \to X \to W$
$Z_3 \to X \to W$
$Z_1 \to Z_3 \to X$
$Z_1 \to Z_3 \to Y$
$Z_2 \to Z_3 \to X$
$Z_2 \to Z_3 \to Y$

Forks

$X \leftarrow Z_1 \to Z_3$
$X \leftarrow Z_3 \to Y$
$Y \leftarrow Z_2 \to Z_3$

Colliders

$W \to Y \leftarrow Z_2$
$W \to Y \leftarrow Z_3$
$Z_2 \to Y \leftarrow Z_3$
$Z_1 \to X \leftarrow Z_3$
$Z_1 \to Z_3 \leftarrow Z_2$

Causal discovery vs. causal validation

Discovery: learn a DAG from data. *Validation*: test a proposed DAG against data. Usually more than one DAG passes the tests, so both are hard and uncertain.

Simulating data from the DAG

$$Z_1 \sim N(4.75,\ 1.72^2)$$
$$Z_2 \sim N(3.29,\ 1.89^2)$$
$$Z_3 = 3Z_1 - 1.5Z_2 + \epsilon_{z_3}$$
$$X = 2Z_1 + 2.5Z_3 + \epsilon_x$$
$$W = 3X + \epsilon_w$$
$$Y = 2W + 2Z_2 - 3Z_3 + \epsilon_y$$

Each structural equation: a node and its parents.

```
set.seed(4321); n = 100
z1 = rnorm(n,4.75,1.72)
z2 = rnorm(n,3.29,1.89)
z3 = 3*z1 - 1.5*z2
z3 = z3 + rnorm(n,0,sd(z3))
x  = 2*z1 + 2.5*z3
x  = x  + rnorm(n,0,sd(x))
w  = 3*x
w  = w  + rnorm(n,0,sd(w))
y  = 2*w + 2*z2 - 3*z3
y  = y  + rnorm(n,0,sd(y))
```

(noise added sequentially, so each child sees its parents' noisy values)

d-separation

d-separation ("directional" separation)

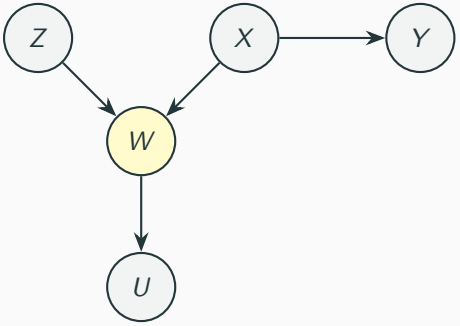
- **d-connected**: there is an open (connecting) path between the two nodes.
- **d-separated**: there is no such path.
- With no conditioning, only colliders block a path.

Definition
A path p is **blocked** by a set of nodes $\mathcal{Z}$ iff

1. p contains a chain $A \to B \to C$ or a fork $A \leftarrow B \to C$ with $B \in \mathcal{Z}$; or
2. p contains a collider $A \to B \leftarrow C$ with $B \notin \mathcal{Z}$ and no descendant of B in $\mathcal{Z}$.

If $\mathcal{Z}$ blocks *every* path between X and Y: X and Y are d-separated given $\mathcal{Z}$, hence $X \perp\!\!\!\perp Y \mid \mathcal{Z}$.

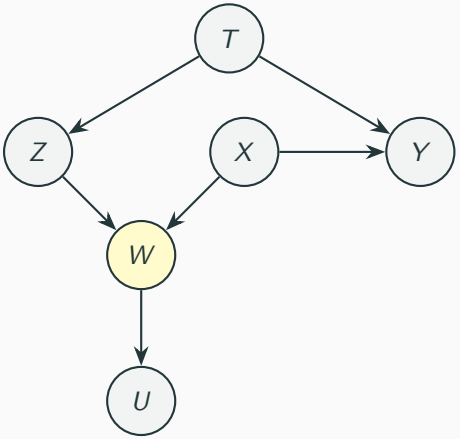
Example: a collider with a child, plus a fork



Path: $Z \to W \leftarrow X \to Y$

- $Z \perp\!\!\!\perp Y$: the collider W blocks the path.
- Given W: path opens $\Rightarrow Z \not\perp\!\!\!\perp Y \mid W$.
- Given U (descendant of the collider): also opens $\Rightarrow Z \not\perp\!\!\!\perp Y \mid U$.
- Given $\{W, X\}$: the fork at X blocks again $\Rightarrow Z \perp\!\!\!\perp Y \mid W, X$.

Adding a second (forked) path through T



- $Z \not\perp\!\!\!\perp Y$: the path $Z \leftarrow T \rightarrow Y$ has no collider.
- $Z \perp\!\!\!\perp Y \mid T$
- $Z \not\perp\!\!\!\perp Y \mid T, W$
- $Z \perp\!\!\!\perp Y \mid T, W, X$

d-connected given: W, U, $\{W, U\}$, $\{W, T\}$, $\{U, T\}$, $\{W, U, T\}$, $\{W, X\}$, $\{U, X\}$, $\{W, U, X\}$.

d-separated given: T, $\{X, T\}$, $\{W, X, T\}$, $\{U, X, T\}$, $\{W, U, X, T\}$.

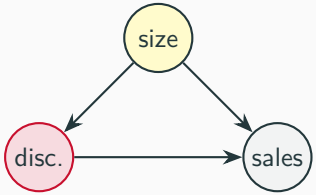
Exercise in class: verify one line of each list by checking both paths.

Confounding: two examples

Estimating the effect of a marketing campaign

(H. Kögel, *Causal Machine Learning in Marketing*, Medium, 2023)

- A firm sells computer equipment to other businesses; some customers got a discount.
- Managers: should we keep the discounts?
- Naive approach: compare average sales of firms with and without the discount.
- But larger firms were more likely to get the discount *and* buy more.



Treatment: discount. Outcome: sales. Confounder: firm size. Ignoring the back-door path disc. $\leftarrow$ size $\rightarrow$ sales biases the estimate; we must control for size.

Two classes of hypertension medication

(Daniels, Linero & Roy, 2024, Sec. 1.1; see also Imbens & Rubin, 2015; Rubin, 1974)

- Observational study comparing two drug classes for hypertension:
  - ACEIs (angiotensin-converting enzyme inhibitors) $\Rightarrow A = 1$
  - ARBs (angiotensin II receptor blockers) $\Rightarrow A = 0$
- Outcome Y: systolic blood pressure (SBP) some months later.
- Physicians do not assign drugs at random: patient characteristics drive both the choice of drug and SBP.
- What we *want*: what would average SBP be if everyone got ACEIs vs. if everyone got ARBs?

Causal inference requires assumptions that cannot be checked from the data.

Potential outcomes

Potential outcomes and causal estimands

Binary treatment $A \in \{0, 1\}$. **Potential outcome** $Y(a)$: outcome that *would* be observed under treatment a.

Average treatment effect (ATE)	$E\{Y(1)\} - E\{Y(0)\}$
ATE on the treated (ATT)	$E\{Y(1) \mid A = 1\} - E\{Y(0) \mid A = 1\}$
Quantile treatment effect (QTE)	$F_1^{-1}(p) - F_0^{-1}(p), \;\; F_a(y) = \Pr(Y(a) \leq y)$
Causal relative risk (binary Y)	$E\{Y(1)\}/E\{Y(0)\}$
Conditional ATE (CATE)	$E\{Y(1) - Y(0) \mid V = v\}$

- ATE: how much higher would average SBP be if *everyone* took ACEIs rather than ARBs?
- ATT: for the treated, $Y(0)$ is the counterfactual, the outcome had they, contrary to fact, not been treated.
- QTE: useful when Y is skewed or the effect is concentrated in a small subgroup.

Identification: linking data to potential outcomes

We observe (A_i, Y_i), $i = 1, \ldots, N$, never both $Y_i(0)$ and $Y_i(1)$. We need causal assumptions:

SUTVA	Y_i does not depend on other units' treatments: $Y_i(\mathbf{a}) = Y_i(\mathbf{a}')$ whenever $a_i = a_i'$.
Consistency	If $A_i = a$, then $Y_i = Y_i(a)$.
Ignorability	$\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid L$, L = observed pre-treatment confounders (a fork!).
Positivity	$\Pr(A = a \mid L = l) > 0$ for all a and l.

Randomized trials: the gold standard
Randomization gives $\{Y(0), Y(1)\} \perp\!\!\!\perp A$, so
$E\{Y(a)\} = E\{Y(a) \mid A = a\} = E(Y \mid A = a)$.

Ignorability ("no unmeasured confounders") cannot be checked from data, whatever the sample size.

Propensity scores

Propensity score: $e(l) = \Pr(A = 1 \mid L = l)$

Proposition
Under SUTVA, consistency, ignorability and positivity, with binary A,

$$\text{ATE} = E\left[\frac{AY}{e(L)} - \frac{(1-A)Y}{1-e(L)}\right] = E\left[Y\,\frac{A - e(L)}{e(L)\{1-e(L)\}}\right].$$

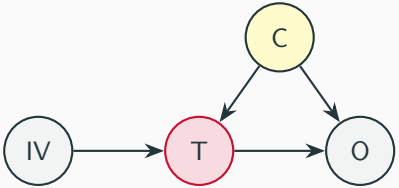
If $e(\cdot)$ is estimated well, the plug-in (inverse probability weighting) estimator

$$\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n} Y_i\,\frac{A_i - \hat{e}(L_i)}{\hat{e}(L_i)\{1-\hat{e}(L_i)\}}$$

is consistent. The propensity score is a balancing score: a one-dimensional summary of L.

Instrumental variables

Instrumental variables: the idea



IV: instrument T: treatment O: outcome
C: (un)measured confounders

The instrument moves the treatment, and affects the outcome *only through* the treatment.

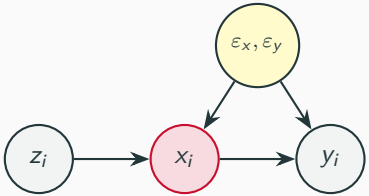
	Confounders (C)	Treatment (T)	Outcome (O)	IV
i	Maternal characteristics	Smoking in pregnancy	Low birth weight	Cigarette taxes
ii	Smoking, caffeine, alcohol	Body mass index	Parkinson disease	FTO gene variant
iii	Prognostic factors	Catheter use	Mortality	Catheter use at facility
iv	Ability	Education	Earnings	Mother's education
v	Proximity	Tutoring program	GPA	Library hours

The linear Gaussian IV model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_{yi}$$
$$x_i = \delta_0 + \delta_1 z_i + \varepsilon_{xi}$$

$(\varepsilon_{xi}, \varepsilon_{yi})$ bivariate normal, variances σ_x^2, σ_y^2, correlation ρ.

- $\rho = 0 \Rightarrow$ no endogeneity: $z \to x \to y$ is a *chain*, OLS of y on x is causal.
- $\delta_1 = 0 \Rightarrow$ no instrument.



Two-stage logic

Substitute the treatment equation into the outcome equation:

$$\begin{aligned} y_i &= \beta_0 + \beta_1[\delta_0 + \delta_1 z_i + \varepsilon_{xi}] + \varepsilon_{yi} \\ &= \underbrace{(\beta_0 + \beta_1\delta_0)}_{\theta_0} + \underbrace{(\beta_1\delta_1)}_{\theta_1} z_i + u_i, \qquad u_i \sim N(0, \tau^2), \end{aligned}$$

with $\tau^2 = \beta_1^2\sigma_x^2 + \sigma_y^2 + 2\beta_1\rho\sigma_x\sigma_y$.

1. First stage: regress x on $z \Rightarrow \hat{\delta}_1$
2. Reduced form: regress y on $z \Rightarrow \hat{\theta}_1$
3. Since $\theta_1 = \beta_1\delta_1$:

$$\hat{\beta}_1^{IV} = \frac{\hat{\theta}_1}{\hat{\delta}_1} = \frac{\widehat{\text{cov}}(z, y)}{\widehat{\text{cov}}(z, x)} = \frac{\text{effect of } z \text{ on } y}{\text{effect of } z \text{ on } x}$$

Why not just add z to the outcome regression?

Bivariate normality gives $\varepsilon_{yi} = \dfrac{\sigma_{xy}}{\sigma_x^2}\varepsilon_{xi} + \omega_i, \;\; \omega_i \sim N(0, \sigma_y^2(1-\rho^2))$.

Replace $\varepsilon_{xi} = x_i - \delta_0 - \delta_1 z_i$:

$$y_i = \Big(\beta_0 - \frac{\sigma_{xy}}{\sigma_x^2}\delta_0\Big) + \Big(\beta_1 + \frac{\sigma_{xy}}{\sigma_x^2}\Big)x_i + \Big(-\frac{\sigma_{xy}}{\sigma_x^2}\delta_1\Big)z_i + \omega_i$$

The coefficient on x is $\beta_1 + \sigma_{xy}/\sigma_x^2$, not β_1. Simply including the instrument as a regressor does *not* recover the causal effect.

Worked examples

A Simulated exercise: the omitted variable problem.

B Return to education, married women ($n = 753$). y: log wage; x: education; z: father's education.
OLS: 11% per year 2SLS: 5.9% per year. OLS suffers from omitted ability bias.

C Smoking and birth weight ($n = 1388$). z: average cigarette price in the state.
The IV fails the one IV requirement we can always test (relevance).

D College proximity, Card (1995) ($n = 3010$). z: grew up near a 4-year college.

	$\hat{\beta}$	s.e.	95% CI
OLS	0.075	0.003	(0.069, 0.081)
IV	0.132	0.055	(0.022, 0.242)

Wider intervals: the price paid for consistency under endogeneity.

A few of my own papers on Bayesian IV / treatment effects

- **Hahn, He & Lopes (2018)** Bayesian factor model shrinkage for linear IV regression with many instruments. *JBES* 36(2), 278–287.

  Slice sampler from a Bayesian analogue of 2SLS; factor-based shrinkage prior for many instruments.

- **Lopes & Polson (2014)** Bayesian instrumental variables: likelihoods and priors. *Econometric Reviews* 33, 100–121.

  Likelihood pathologies, Jeffreys priors, a Cholesky-based alternative to the inverted Wishart; weak instruments.

- **Heckman, Lopes & Piatek (2014)** Treatment effects: a Bayesian perspective. *Econometric Reviews* 33, 36–67.

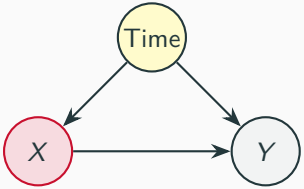
  Potential outcomes with a choice equation; dependence driven by low-dimensional latent factors.

Difference-in-differences

Difference-in-differences: motivation

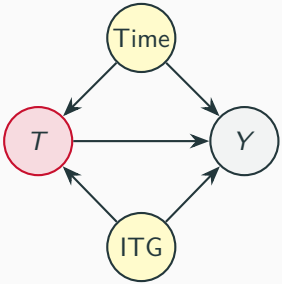
(partially based on N. Huntington-Klein, `nickchk.com/causalgraphs.html`)

- An office switches from cubicles to open plan on Jan 1, 2017. Did productivity go up?
- Before (2016) vs. after (2017) mixes two things:
  - the effect of treatment on Y (want);
  - changes in Y over time unrelated to treatment, e.g. the economy improved (don't want).
- Time opens a back door: Treatment $\leftarrow$ Time $\rightarrow Y$.



Adding a control group

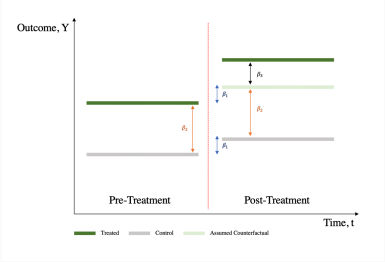
- Add a Control office that keeps its cubicles in 2016 and 2017.
- Treatment = in the Treated group **and** after the policy.
- New back door: Treatment $\leftarrow$ In Treated Group (ITG) $\rightarrow Y$, since the groups may differ.
- DiD closes both back doors: compare the before/after change in the treated group with the before/after change in the control group.



DiD regression

$$y_{ti} = \beta_0 + \beta_1 P_t + \beta_2 T_i + \beta_3 P_t \times T_i + x_{ti}'\gamma + \varepsilon_{ti}$$

$P_t = 1$ post-period; $T_i = 1$ treated group; x_{ti} extra controls.



	Treatment Group ($T_i = 1$) (1)	Control Group ($T_i = 0$) (2)	Difference (1) – (2)
Post-Treatment Period ($P_t = 1$) (a)	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_0 + \beta_1$	$\beta_2 + \beta_3$
Pre-Treatment Period ($P_t = 0$) (b)	$\beta_0 + \beta_2$	β_0	β_2
Difference (a) – (b)	$\beta_1 + \beta_3$	β_1	$\boldsymbol{\beta_3}$

Regression vs. means by hand: standard errors, multiple periods/groups, covariates. (figures: `bookdown.org/cuborican/RE_STAT`)

Two equivalent ways to take the two differences

There are two equivalent strategies to think about the two "differences":

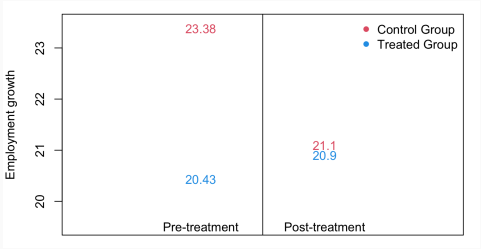
	Strategy #1	Strategy #2
Difference 1	**Average change of treated over time**	**Average change between treated and control in post-treatment period**
	$E(Y_{it}\|T_i = 1, P_t = 1) - E(Y_{it}\|T_i = 1, P_t = 0)$	$E(Y_{it}\|T_i = 1, P_t = 1) - E(Y_{it}\|T_i = 0, P_t = 1)$
Difference 2	**Average change of control over time**	**Average change between treated and control in pre-treatment period**
	$E(Y_{it}\|T_i = 0, P_t = 1) - E(Y_{it}\|T_i = 0, P_t = 0)$	$E(Y_{it}\|T_i = 1, P_t = 0) - E(Y_{it}\|T_i = 0, P_t = 0)$
Difference in Differences	**Difference 1 – Difference 2**	

Y_{it}: *Outcome of unit i at time t*
T_i: *Dummy variable indicating 1 if treatment is assigned and 0 otherwise*
P_t: *Dummy variable indicating 1 if post – treatment period and 0 otherwise*

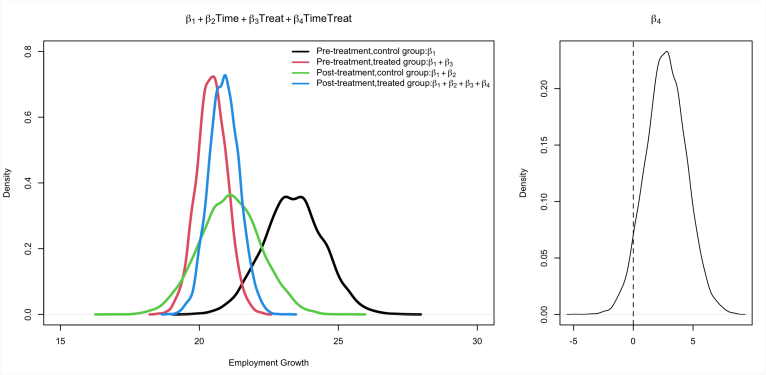
Card & Krueger (1994): minimum wage and employment

On April 1, 1992, New Jersey raised the minimum wage from \$4.25 to \$5.05; Pennsylvania stayed at \$4.25. Employment in 384 fast-food restaurants, Feb. 1992 (time = 0) and Nov. 1992 (time = 1). Treatment = 1: NJ.

```
A = post.treated   - pre.treated      #  0.467
B = post.untreated - pre.untreated    # -2.283
effect = A - B                        #  2.750
lm(outcome ~ time + treatment + time:treatment)
#              Estimate Std.Err t value
# (Intercept)   23.380   1.098  21.29
# time          -2.283   1.553  -1.47
# treatment     -2.949   1.224  -2.41
# interaction    2.750   1.731   1.59
```



Card & Krueger: a Bayesian look



Posterior draws of β under a flat prior (normal approximation). Left: the four cell means. Right: posterior of the DiD effect β_4; the dashed line marks zero.

Regression discontinuity

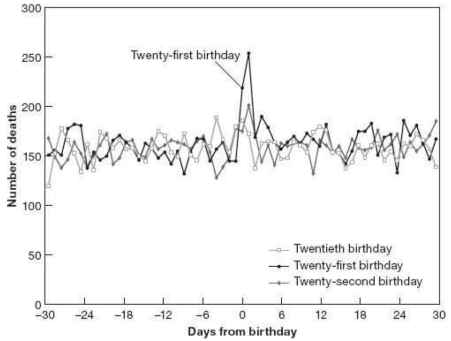
Human behavior is constrained by rules

- California caps elementary school classes at 32 students; 33 is one too many.
- Social Security pays no retirement benefits before age 62.
- Recruits with low test scores are ineligible for military service.

Such rules generate natural experiments. RD does not answer every causal question, but when it works the results have almost the causal force of a randomized trial.

Birthdays and funerals

FIGURE 4.1
Birthdays and funerals

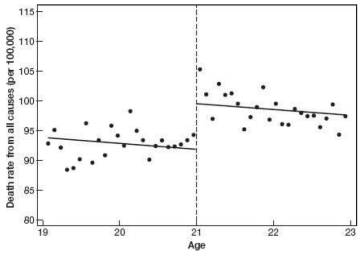


- US minimum legal drinking age (MLDA): 21.
- A tiny change in age (days) $\Rightarrow$ a big change in legal access.
- Deaths among Americans aged 20–22, 1997–2003.
- Mortality spikes on and right after the 21st birthday: about 100 extra deaths on a baseline of about 150 per day.

Sharp RD

FIGURE 4.2

A sharp RD estimate of MLDA mortality effects



Notes: This figure plots death rates from all causes against age in months. The lines in the figure show fitted values from a regression of death rates on an over-21 dummy and age in months (the vertical dashed line indicates the minimum legal drinking age (MLDA) cutoff).

$$D_a = \begin{cases} 1 & a \geq 21 \\ 0 & a < 21 \end{cases}$$

- Treatment is a *deterministic* and *discontinuous* function of a.
- a is the running variable.
- Sharp RD: treatment switches cleanly at the cutoff.

$$y_a = \alpha + \rho D_a + \gamma a + e_a$$

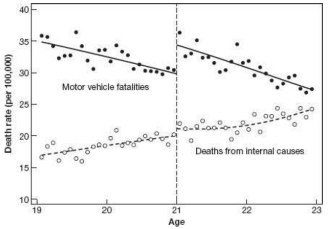
y_a: death rate in month a (30-day bins from the 21st birthday).

$\gamma < 0$: smoothly declining death rates with age.

$\hat{\rho} = 7.7$ deaths per 100,000: a large jump at the cutoff.

Controlling for smooth variation in death rates

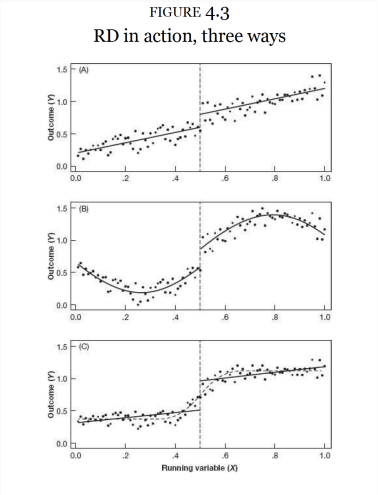
FIGURE 4.5
RD estimates of MLDA effects on mortality by cause of death



Notes: This figure plots death rates from motor vehicle accidents and internal causes against age in months. Lines in the figure plot fitted values from regressions of mortality by cause on an over-21 dummy and a quadratic function of age in months, interacted with the dummy (the vertical dashed line indicates the minimum legal drinking age [MLDA] cutoff).

- Mortality changes with age for reasons unrelated to the MLDA.
- Internal causes (e.g. cancer): low but rising in the late teens / early 20s.
- External causes (car accidents, homicides, suicides): falling.
- RD separates this smooth trend from the MLDA effect by regression control, hence the name.

Nonlinearity mistaken for a discontinuity



A RD with a linear model for $E(y_i \mid X_i)$.

B Adds curvature.

C Nonlinearity mistaken for a discontinuity.

Linear:

$$y_a = \alpha + \rho D_a + \gamma a + e_a$$

Quadratic with interactions:

$$y_a = \alpha + \rho D_a + \gamma_1 (a - a_0) + \gamma_2 (a - a_0)^2 + \delta_1 (a - a_0) D_a + \delta_2 (a - a_0)^2 D_a + e_a$$

Sharp RD estimates of MLDA effects

TABLE 4.1

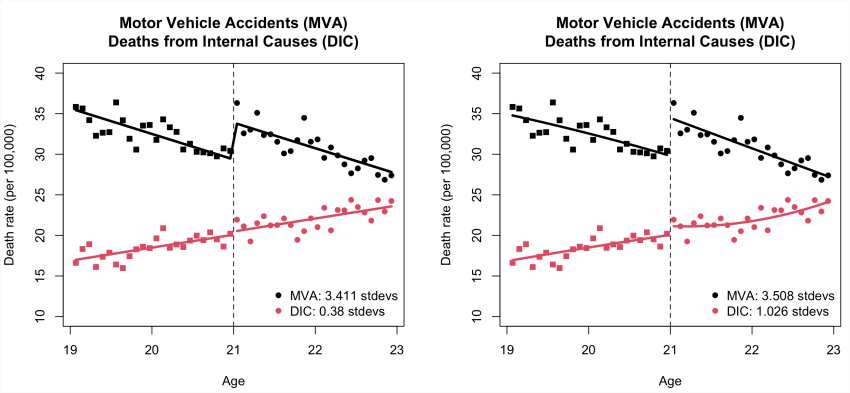
Sharp RD estimates of MLDA effects on mortality

Dependent variable	Ages 19–22 (1)	Ages 19–22 (2)	Ages 20–21 (3)	Ages 20–21 (4)
All deaths	7.66 (1.51)	9.55 (1.83)	9.75 (2.06)	9.61 (2.29)
Motor vehicle accidents	4.53 (.72)	4.66 (1.09)	4.76 (1.08)	5.89 (1.33)
Suicide	1.79 (.50)	1.81 (.78)	1.72 (.73)	1.30 (1.14)
Homicide	.10 (.45)	.20 (.50)	.16 (.59)	−.45 (.93)
Other external causes	.84 (.42)	1.80 (.56)	1.41 (.59)	1.63 (.75)
All internal causes	.39 (.54)	1.07 (.80)	1.69 (.74)	1.25 (1.01)
Alcohol-related causes	.44 (.21)	.80 (.32)	.74 (.33)	1.03 (.41)
Controls	age	age, age^2, interacted with over-21	age	age, age^2, interacted with over-21
Sample size	48	48	24	24

Notes: This table reports coefficients on an over-21 dummy from regressions of month-of-age-specific death rates by cause on an over-21 dummy and linear or interacted quadratic age controls. Standard errors are reported in parentheses.

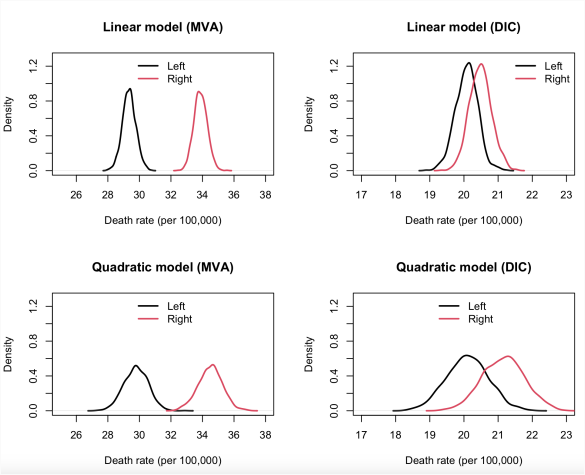
- Columns: ages 19–22 vs. 20–21; linear vs. quadratic with interactions.
- Large and robust jumps for all deaths, motor vehicle accidents and alcohol-related causes.
- Internal causes: no meaningful jump, a built-in placebo check.

Replicating Figure 4.5 in R



Jump at 21 in units of residual s.d.: linear (left) and quadratic with interactions (right). Motor vehicle accidents jump; internal causes barely move.

The same comparison, Bayesian style



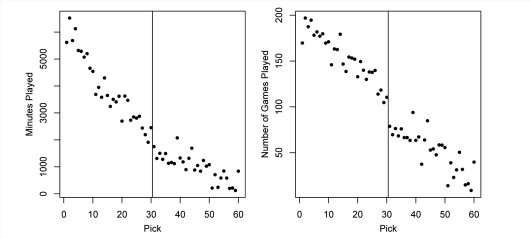
Posterior of the regression function just to the **left** and just to the **right** of age 21.

- MVA: the two posteriors are well separated.
- Internal causes: they largely overlap.

- **Local behavior:** within $[a_0 - b,\ a_0 + b]$ for small b, nonlinear trends matter little.
- **Strategy:** compare averages just to the left and just to the right of the cutoff.
- **High variance:** very narrow window $\Rightarrow$ few observations $\Rightarrow$ imprecise estimates.
- **Trade-off:** bias reduction near the boundary vs. increased variance (bandwidth choice).
- **Local linear regression:** fit $y_a = \alpha + \rho D_a + \gamma a + e_a$ only for $a \in [a_0 - b, a_0 + b]$.

Example: the NBA draft

Branson, Rischard, Bornn & Miratrix (2019) A nonparametric Bayesian methodology for RDDs. *JSPI* 202, 14–30.



- 1,238 players drafted 1995–2016; two rounds of 30 picks.
- Running variable: pick number; cutoff between pick 30 and 31.
- Question: are first-round picks given more playing time beyond what skill explains?

NBA draft: Gaussian process RD

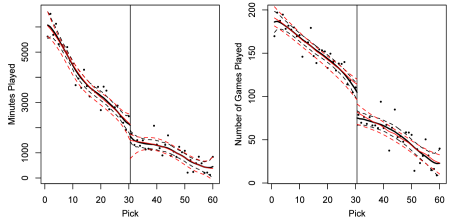


Figure 5: The estimated mean functions (solid lines) and corresponding confidence intervals (dashed lines) for LLR (black lines) and GPR (red lines). The lines for LLR were produced by the `rdd` R package (Dimmery, 2013), but using the bandwidth estimated by the `rdrobust` R package. The two treatment groups are the first round of picks (picks 30 and below) and second round of picks (picks 31 and above). We set the boundary to be $b = 30.5$ to minimize the amount of extrapolation that needs to be conducted on both sides of the boundary to estimate the treatment effect.

Table 1: Treatment effect estimation for LLR, robust LLR, and GPR on NBA data

	LLR		**Robust LLR**		**GPR**	
Outcome	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Box Plus-Minus	**-3.06**	[-5.20, -0.92]	**-3.37**	[-5.86, -0.88]	-2.42	[-5.02, 0.01]
Win Shares	-1.58	[-4.09, 0.93]	-1.73	[-4.76, 1.29]	-1.08	[-3.04, 0.88]
Minutes Played	-446.29	[-1042.81, 150.23]	-406.63	[-1123.99, 310.72]	-640.75	[-1259.65, 5.51]
Games Played	**-29.71**	[-40.20, -19.22]	**-28.16**	[-40.81, -15.51]	**-32.00**	[-45.69, -18.14]

Point estimates and 95% confidence intervals for the treatment effect on each of the four outcomes: box plus-minus, win shares played, number of minutes played, and number of games played. Statistically significant point estimates are in bold.

Being a *second-round* pick significantly reduces games played and marginally reduces minutes played (GPR vs. local linear and robust local linear).

RDD references

- Alcantara, Wang, Hahn & Lopes (2024) Modified BART for learning heterogeneous effects in RDDs. arXiv:2407.14365
- Chib, Greenberg & Simoni (2023) Nonparametric Bayes analysis of the sharp and fuzzy RDDs. *Econometric Theory*
- Cattaneo, Idrobo & Titiunik (2020) A practical introduction to RDDs: foundations. arXiv:1911.09511
- Gelman & Imbens (2019) Why high-order polynomials should not be used in RDDs. *JBES*
- Gelman (2019) Another regression discontinuity disaster... (blog)
- Cattaneo, Titiunik & Vazquez-Bare (2017) Comparing inference approaches for RDDs. *JPAM*
- Gelman & Zelizer (2015) Deleterious impact of polynomial regression on causal inference. *Research & Politics*
- Cattaneo, Frandsen & Titiunik (2015) Randomization inference in the RDD. *J. Causal Inference*
- Calonico, Cattaneo & Titiunik (2014) Robust nonparametric CIs for RDDs. *Econometrica*
- Imbens & Kalyanaraman (2012) Optimal bandwidth choice for the RD estimator. *REStud*
- Imbens & Lemieux (2008) RDDs: a guide to practice. *J. Econometrics*
- Hahn, Todd & Van der Klaauw (2001) Identification and estimation of treatment effects with a RDD. *Econometrica*
- Thistlethwaite & Campbell (1960) Regression-discontinuity analysis. *J. Educ. Psych.*

Take-home messages

1. The same numbers can call for opposite conclusions: the causal story decides, not the table.
2. Chains and forks are blocked by conditioning; colliders are *opened* by it.
3. Control for confounders (forks), not for mediators or colliders.
4. Potential outcomes make the estimand explicit: ATE, ATT, QTE, CATE.
5. Identification needs untestable assumptions: SUTVA, consistency, ignorability, positivity.
6. When ignorability fails, use design: **IV**, **DiD**, **RDD**.