Research Design for Causal Inference

High-Level Overview w. Application to Diabetes

Bernie Black
Nicholas Chabraja Professor
Northwestern University, Pritzker School of Law and Kellogg School of Management
bblack@northwestern.edu

(ADA Research Symposium, fall 2018)
[no conflicts]

Causal Inference Workshop(s)

- For (much) more: I co-organize a summer workshop at Northwestern on Research Design for Causal Inference
  - https://northwestern.app.box.com/files/0/f/3437924886/Causal_Inference_Workshops
  - Main workshop w. world-class speakers
  - Advanced workshop: selected topics, vary by year
- A bit about me: Author page on SSRN:
  - http://ssrn.com/author=16042
- Northwestern faculty page:
  - http://www.law.northwestern.edu/faculty/profiles/BernardBlack/

Hierarchy of Research Designs

- Randomized experiments (RE)
  - Simple, block, and pair RE
  - Intent-to-treat designs: One- and two-sided noncompliance
- "Natural" experiment (shock-based) designs
  - Regression discontinuity (RD)
    - Sharp and fuzzy RD
  - Difference-in-differences (DiD)
    - Simple DiD, distributed lag, and leads-and-lags designs
    - Triple difference designs
    - "DiD-continuous" (dose-response) designs
  - combined DiD/RD designs [strengths of both]
  - instrumental variables [will not discuss]
- Pure observational studies [rely on "balancing"]
  - Trimming to common support
  - Matching [many ways]
- Combined DiD/balancing

Others talked about DiD

- Including DiD/balancing
  - Often better than DiD alone
- I will discuss RD (often next best to RCT)
- Confusing terminology: ITS (interrupted time series)
  - **with** a control group **is DiD**
  - For same person (unit) with sharp treatment response to time, and sharp unit response to treatment, **can be RD**
  - Without either of these, is often a weak design

Goal for credible causal inference

- If you don't have an RCT, come as close as you can
- Make your assumptions as weak as you can
- Credible causal inference:
  - comes from clean design; not fancy analysis
- How to look for good research designs
  - And spot them when you bump into them

Toward Stronger Research Design

- Goal is "credible causal inference"
  - No research design is perfect
  - One hopes that a project moves toward that goal
- Often called "identification"
  - loose term, multiple meanings: I will avoid it
- Some projects don't permit causal claims
  - Pure prediction

Regression is (often) evil

- Edward Leamer (1983), Let's Take the Con out of Econometrics, 73 *American Economic Review* 31-43:
  - "Hardly anyone takes data analyses seriously. Or perhaps more accurately, hardly anyone takes anyone else's data analyses seriously."
- Paul Rosenbaum (2017) , *Observation and Experiment* 46:
  - Commonly, statistical hypotheses refer to parameters or aspects of a convenient statistical model, and then a separate argument, not always a particularly clear or compelling argument, is invoked to connect this convenient but rather technical model to the scientific problem at hand [causal inference, say]. . . . [T]hese connectivity arguments are often most compelling to people who do not understand them, and least compelling to people who do.

- IMHO: these skeptical views remain still true today, for "classic" studies, using "regression"

Notation

- "Dependent" or "outcome" variable Y
- Main "Independent" or "predictive" variable X_1
- (Maybe) some "control" variables or "covariates" $\boldsymbol{X}_{-1} = (X_2, X_3, \ldots X_K)$

- **boldface** = vector or matrix
- Sample size N, observations indexed by i
- Often "panel data" over time, indexed by *t*
- Notation convention:
  - CAPITAL LETTERS for random variables (X)
  - Lowercase for *specific realizations* in the sample (x)
  - Exception: bold, capital **X** for a matrix in the sample
  - But I'll sometimes forget my own convention

Design matrix: (cross-section) data looks like . . .

Outcome	Predictor Variable of interest	First Covariate		Last Covariate
y_1	x_{11}	x_{12}	. . .	x_{1K}
y_2	x_{21}	x_{22}	. . .	x_{2K}
y_3	x_{31}	x_{32}	. . .	x_{3K}
. . .	. . .	. . .	. . .	. . .
y_N	x_{N1}	x_{N2}	. . .	x_{NK}

x_{ik} is the ith observation of the kth covariate
Want to know: Will ΔX_1 **cause** ΔY?

The OLS regression model is

- Model: $\mathrm{y_i} = \alpha + \beta x_{i1} + \sum_{k=2}^{K}(x_{ik}\gamma_k) + \varepsilon_i$
- In matrix notation:

$$\boldsymbol{y} = \alpha\mathbf{1}_N + \beta\boldsymbol{x}_1 + \boldsymbol{\gamma}\boldsymbol{X}_{-1} + \boldsymbol{\varepsilon}$$

  - α, β are scalars
  - **y, ε** are $\mathrm{N} \times 1$ "column" vectors
  - $\mathbf{1}_N$ is an $\mathrm{N} \times 1$ column vector of "1's
  - $\boldsymbol{x}_1$ is $\mathrm{N} \times 1$ column vector for principal variable of interest
  - $\boldsymbol{X}_{-1}$ is a $\mathrm{N} \times (k-1)$ matrix of "covariates"
  - γ is a 1 × (k-1) row vector of model parameters
  - $\mathbf{y_i}$, $\mathbf{x_{ik}}$ are elements of the $\mathrm{N} \times (k+1)$ "design matrix"
- Note: different books have different variations of this equation
  - they (should be) equivalent and only look different

OLS Estimation

- Estimation:

$$y_i = \hat{\alpha} + \hat{\beta} x_{i1} + \cdot\, \hat{\boldsymbol{\gamma}} \boldsymbol{x}_{\boldsymbol{i},-\mathbf{1}} + e_i$$

- Two changes:
  - β is an **estimand** (something we want to estimate)
  - OLS provides an **estimator** (one way to estimating the model "parameters", which are the estimands)
  - OLS produces an **estimate** of each parameter
  - Estimated parameters get "hats"
  - OLS replaces the unobserved **error** ε with **residual** e

Causal inference replaces regression with . . .

- What is often called the "Rubin causal model"
- **Major** simplification:

  Replace X_1 with **binary** W (treatment "dummy")
  - Some units are "treated" ($\mathrm{w_i} = 1$)
  - Others are "control" ($\mathrm{w_i} = 0$)
- Multi-valued w = straightforward extension, clunky
  - Continuous = Important in medical research (dose/response), but at research frontier

Major conceptual move: Potential outcomes

- Define: Every unit i has two "potential outcomes"
  - $y_i(w = 1)$:= outcome if treated [shorthand y_{i1}]
  - $y_i(w = 0)$:= outcome if control [shorthand y_{i0}]
- One of these is observed; one is not
  - Missing outcome is often called "counterfactual"
  - I prefer to think of it as "real", just not observed
- **Compare:** $y_i^{obs} := w_i y_i(1) + (1 - w_i) y_i(0)$
- Regression tempts you to treat y_i^{obs} as a real quantity
  - It's not. It's a mixture of y_{i0} and y_{i1} you happen to observe

Causal Inference as Missing Data Problem

- Treatment effect: $\tau_i = (y_{1i} - y_{0i})$
- Rubin's central insight: Causal inference is a missing data problem:
    - Neyman (1923) developed potential outcomes for RCTs
    - Rubin applied this idea to observational studies
  - Must **credibly estimate** the missing potential outcomes
  - "Fundamental problem of causal inference" [Holland, 1986]

Second **major** complication, and conceptual move

- **Heterogeneous** treatment effects
  - Treatment effect: $\tau_i = (y_{1i} - y_{0i})$ depends on characteristics of unit i
  - τ_i depends on (varies with) both $\mathbf{x}_i$ and $\mathbf{u}_i$

Regression uses Y^{obs}

- Regression is really:

$$Y^{obs} = \alpha + \beta W + \boldsymbol{\gamma X_{-1}} + \epsilon$$

- Mixture in; mess out, except special cases
- Regression also assumes homogeneous treatment effects (same β for everyone)

- With two potential outcomes, and missing covariates **u**, the true design matrix is:

The (even more missing) design matrix is. . .

Outcome if treated	Outcome if control	Treatment effect	Treatment dummy	First covariate		Last Covariate	Unobserved covariates
y_{11}	y_{10}	τ_1	w_1	x_{12}	. . .	x_{1K}	$\mathbf{u}_K$
y_{21}	y_{20}	τ_2	w_2	x_{22}	. . .	x_{1K}	$\mathbf{u}_{1K}$
y_{31}	y_{30}	τ_3	w_3	x_{32}	. . .	x_{3K}	$\mathbf{u}_{3K}$
y_{41}	y_{40}	τ_4	w_4	x_{42}	. . .	x_{4K}	$\mathbf{u}_{4K}$
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
y_{N1}	y_{N0}	τ_{N0}	w_N	x_{N2}	. . .	x_{NK}	$\mathbf{u}_{NK}$

red = not observed

Want to know: Is $y_{i1} \neq y_{i0}$? Equivalently, is $\tau_i \neq 0$

This is a hard problem

- Regression, applied to the partial data we observe, won't get us there
  - Except in special cases
- Often not "math hard"
  - Instead "design hard"
- We need research designs that let us:
  - credibly estimate the missing potential outcomes
  - Allow for heterogeneous treatment effects
  - not worry about the omitted covariates
- That's what causal inference is about!

Core assumption 1: manipulation

- w_i is manipulable
- Counterexample: Effect of gender on income
  - Observe y_{i1} = income if male
  - Want to impute y_{i0} = income if female
  - All else about you is the same (*ceteris paribus*)
- Not achievable
  - "no causation without manipulation" [Holland, 1986]
  - If you were dictator, with infinite resources [and no morals], could you design an experiment to answer the question you have in mind? [Dorn, 1953]

Core Assumption 2 (& 3): SUTVA

- "Stable Unit Treatment Value Assumption"
- Really two separate assumptions:
  1. Only one kind of treatment (w = 0 or 1)
     - Can be relaxed (multivalued and continuous treatments)
  2. Responses of different units are **independent:**

$$\tau_i \perp\!\!\!\perp (\tau_j, w_j) \; \forall \, j \neq i$$

Can call this "**SUTVA independence**"

Example: Chronic disease, but not infectious disease

Major concept: "Assignment mechanism"

- Process (perhaps unknown) for determining which units are treated
- For example, is assignment ***random***?

  $w \perp\!\!\!\perp (y_0, y_1, \mathbf{x}_{-1}, \mathbf{u})$
- If yes, then treated and controls are similar on:
  - Observables $\mathbf{x}_{-1}$ **and** unobservables **u**
  - **No omitted variable bias!**
  - Difference in means recovers **average** treatment effect: ATE = $E[y_1 - y_0] = E[y_1|w=1] - E[y_0|w=0]$
  - $\widehat{ATE} = \hat{\tau}_{naive} = \overline{y_1^{obs}} - \overline{y_0^{obs}}$
- So does regression: Stata: `regress y w, robust`

Regression Discontinuity (RD)

- Not really about regression, but can't change the name
  - Units above some sharp (arbitrary) threshold are treated
  - Units below the threshold are controls
- Treated units **above but close** to threshold = very similar to control units **below but close**
  - On observables **and** unobservables
  - Except "running variable" for the threshold
- **(Almost)** "as good as random" assignment to treatment

Some of many medical examples

- Metformin prescribed if HbA1c > 6.5
- Statins prescribed if LDL > [well, its getting complicated]
- Blood pressure meds recommended if systolic pressure > 140 mmHg
- Bariatric surgery recommended if BMI > 40

RD terminology

- "Sharp" RD
  - All units above threshold are treated
  - No units are treated below threshold
- Real world: "fuzzy" RD:
  - More (but not all) units treated above threshold
  - Fewer (but not zero) treated below threshold
- I will discuss only sharp RD (lack of time)
  - Can be seen as "intent to treat"
  - For fuzzy RD, use IV to recover causal estimate for "compliers"instrumental variables
    - Treated only if above threshold

Sharp RD formalism

- "Running variable" r
- [Units treated (w = 1) if above threshold ($r > r_0$)
- Units are control (w = 0) if below threshold ($r < r_0$)
- Within "bandwidth" around r_0: $r \subset [r_0-\pi, r_0+\pi]$
  - units on both sides are similar ➔ w $^{\text{(close to)}} \perp\!\!\!\perp$ (y_0, y_1)
- Use RCT methods within bandwidth around r_0
  - But control for non-random assignment of r

RD can recover RCT estimates

- Across a variety of fields, dual-design studies find similar RD and RCT estimates
  - Buddelmeyer and Hielke (2004)
  - Black, Galdo and Smith (2007)
  - Cook and Wong, (2008)
  - Cook, Shadish and Wong (2008)
  - Green et al., (2009)
  - Berk et al., (2010)
  - Shadish et al. (2011)
  - Gleason, Resch, and Berk (2012)
  - Moss, Yeaton and Lloyd (2014)
- **Not** true for DiD or IV

Requirements for running variable

- Ideal: (nearly) continuous around r_0
  - OK if binned, if bin size < plausible π
- Ideal: r $\perp$ other variables
  - small correlation is ok: small change in r ➔ very small predicted change in **x**, **u**
- Testable for **x**: "covariate balance"
  - Similar means on both sides of threshold

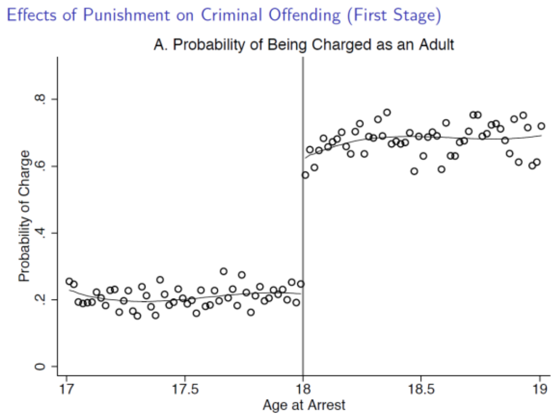
"Almost as good as random"

- Running variable needs special attention
  - For all else, if threshold is truly arbitrary
  - And we're close enough to the threshold
  - Covariates **x** are similar near threshold**:**
    - $E[\mathbf{x} | r_0-\pi < r < r_0] \approx E[\mathbf{x} | r_0 < r < r_0+\pi]$
  - This is also true for unobservables **u**!
- So, if we can control for running variable:
  - We are close to a randomized experiment
  - Can confirm if close enough for observables
  - But must stay near threshold

Some discontinuity examples

- Graphical: Discontinuity in prob. of treatment
  - And in outcome
- [Go to McCrary slides]
  - For each, show discontinuity first
  - Ask if expect an effect
  - Then show effect [or not]
- Discuss local nature of estimate:
  - units near the discontinuity
  - “compliers”: units affected by the discontinuity

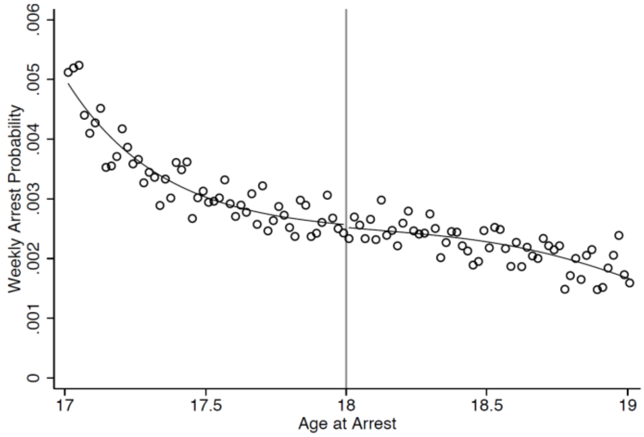
Lee & McCrary (charged as adult)



Practical advice: If can’t see the discontinuity:
It probably isn’t there.

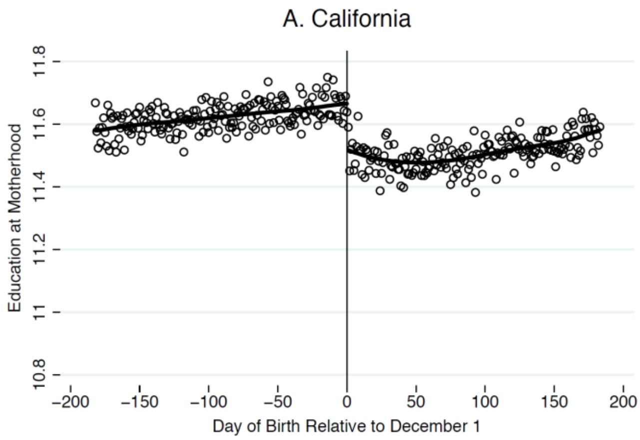
Second stage

Small Deterrence Effects for 18-Year-Olds

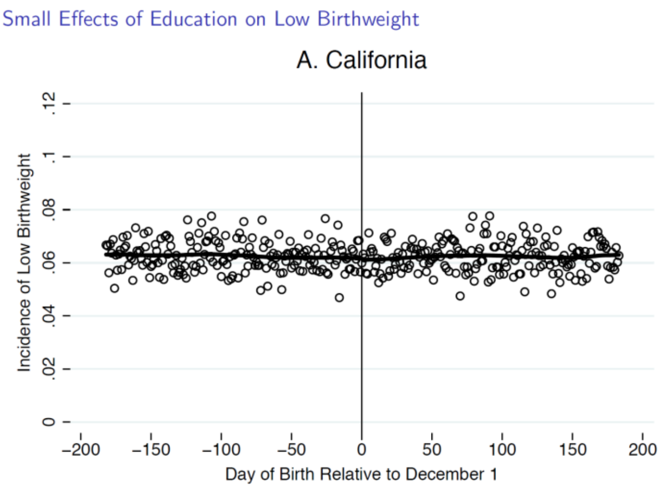


Mother birthdate and education

Effects of Education on Infant Health: California (First Stage)



Mother birthdate and low-weight birth



Two (apparently) Cleaner Health Care Examples

- Newborn birthweight ("very low" < 1,500 grams)
  - Almond, Doyle, Kowalski and Williams (QJE 2010)
    - More intense treatment
  - 18% lower 1-year mortality just below threshold!
    - Apparently clean . . .
    - But Barecca, Guldi, Lindo and Waddell (2011) (donut holes)
- Mother length of stay (two midnights)
  - Almond and Doyle (2011)
  - No benefit of longer stay [readmissions, mortality]

RD and value of graphing

- If you can't **easily** see the treatment discontinuity
  - Hard to find results
  - Hard for them to be convincing, if you find them
- If you can't see the outcome discontinuity . . .
  - It probably isn't there

A result that isn't there

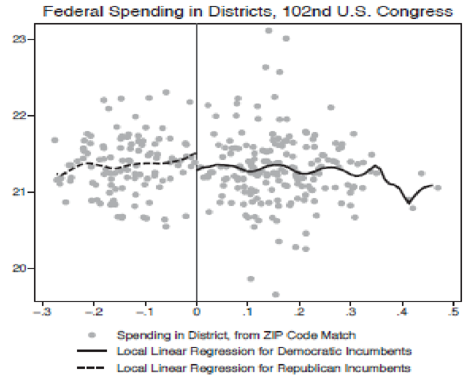


Figure 2: RD example

Even if the author thinks it is. Source:
Austin Nichols (2007), Causal Inference with Observational Data, 7 *Stata Journal* 507-541 2007)

Manipulation risk

- Can units manipulate which side of the threshold they are on?
- Careful check for covariate balance
    - below vs. above threshold
    - for fuzzy RD, actual treated vs. actual controls
  - Distribution of (**x**|r) smooth for broader bandwidths
- If units can choose whether to be treated:
  - **Similar densities** below and above threshold
  - Density **continuous and smooth** at r_0 [McCrary (2008)]

Placebo tests

- Placebo tests:
  - Placebo discontinuity at different thresholds
    - pick lots of them: Compute jumps at threshold for each.
    - Randomization inference can be useful
      - Is observed jump in upper tail of distribution of jumps
  - Placebo outcomes: other covariates
  - If threshold introduced at time T
    - Should be no effect before that

Control for running variable: options

- None (if bandwidth is narrow enough)
  - With unit fixed effects, covariates, one time period:
  - $y_i = \alpha + \delta_{RD}*w_i + \mathbf{x_i}\boldsymbol{\beta} + \varepsilon_i$ [With $w_i = 1$ if $r_i > r_0$]
- Linear plus jump at threshold
  - $y_i = \alpha + \gamma*r_i + \delta_{RD}*w_i + \mathbf{x_i}\boldsymbol{\beta} + \varepsilon_i$
- Linear (different slopes) plus jump
  - $y_i = \alpha + \gamma_{below}*r_i + \gamma_{above}*r_i *w_i + \delta_{RD}*w_i + \mathbf{x_i}\boldsymbol{\beta} + \varepsilon_i$
- Quadratic (or higher polynomial) in running variable, plus jump
- Local linear regression on each side of jump
  - How flexible?
  - Is regression line a plausible model of the world?
    - You assume it is, when estimating jump at threshold
- Try various approaches, assess robustness!

I tend to prefer

- Start simple:
  - Linear with a jump, maybe different slopes
  - Quadratic with a jump
  - Maybe higher order polynomial with a jump
- Advantage:
  - Your model of the world is (continuous plus jump)
    - At least for first derivative
    - Often for second derivative too
      - Can't get a plot like Lieber's
      - Or Card, Dobkin and Maestas for that matter