Introduction to causal inference for data scientists

Michael Gill

2018-01-23

Center for Data Science | New York University

Preliminaries

Preliminaries

Lecture: Tues 1:00pm–2:40pm. 60 Fifth Avenue, C10.

Lab: Wed 8:35pm–9:25pm. 60 Fifth Avenue, C12. Leader: Lei Xu.

Office hours: (Gill) Tue, 3pm-4:30pm, 60 Fifth Avenue, 620. (Xu) Wed, 1:30pm-3pm, 60 Fifth Avenue, 665.

Course webpage: on NYU classes. The lecture slides are posted after each lecture.

Exams: Tuesday, March 6, 1:00pm–2:40pm (Midterm) and Tuesday, May 15, 2:00pm-3:50pm (Final). Note: date and timing of final subject to change, from NYU Registrar.

Preliminaries

Texts: There is NO required textbook. Recommended books:

- [IR] Imbens, G. and Rubin, D. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences. An Introduction*. Cambridge University Press.
- [MW] Morgan, S., and Winship, C. (2014). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, 2nd Edition. Cambridge.
- [PGJ] Pearl, J., Glymour, M., and Jewell, N. (2016). *Causal Inference in Statistics: A Primer*. Wiley.

Others listed on syllabus.

This course

The course is on **causal inference**, mostly with applications from business, economics, and politics. The course will discuss controlled experiments as well as observational studies.[1]

Our focus will be both on theory and application. Where appropriate, I will attempt to tailor applications towards issues you may face in your own research or work.

I will consider the course a success if you grasp fundamental issues in causal inference, and can apply those concepts to everyday problems you face in application.

[1] Major credit due to Alfred Galichon, who first taught this course last year at CDS, and from whom much of content and structure of this course is derived.

This course

Programming: The official language of the course will be R, but you are welcome to use the language of your choice — e.g. Python, Julia... A short tutorial on R will be provided by Lei in this week's lab.

Final grade: Class attendance/participation (10%), Homework (35%), Midterm (25%), Final Exam (30%).

Homework: roughly every other week. Will frequently involve analyzing a dataset using the methods seen in class; sometimes it may involve critiquing a research paper. Homework will be managed through NYU Classes. Full homework policies are listed on syllabus.

Questions?

Course outline

Outline: parts I and II

	Part I. Introduction	
L1	Basic questions	Jan 23
L2	The potential outcome model	Jan 30
	Part II. Randomized experiments	
L3	RCTs, AB testing, business experiments (1)	Feb 6
L4	RCTs, AB testing, business experiments (2)	Feb 13
L5	Noncompliance, instrumental variables (1)	Feb 20

Outline: part III

	Part III: observational studies	
L6	IV (2) and observational studies	Feb 27
Midterm	Lectures 1–6	Mar 6
L7	Matching estimators	Mar 20
L8	Diff-in-Diff, regression discontinuity	Mar 27
L9	Extending Diff-in-Diff	Apr 3
L10	High-dimensional models	Apr 10

Outline: part IV

	Part IV: special topics	
L11	Practical challenges with inference	Apr 17
L12	Causal inference in networks	Apr 24
L13	Machine learning and causal inference	May 1
Final	Lectures 1–13	May 15

Help me learn about you!

Please take a few minutes to complete the following survey.

http://bit.ly/2BjDIju

Today

Today

Part I. Introduction.

Lecture 1. Causal inference: motivating examples

References:

- MW, Chapter 1
- Holland, P. (1986). "Statistics and Causal Inference." *Journal of the American Statistical Association*.
- Angrist, J. and Pischke, J-S. (2010). "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics," *Journal of Economic Perspectives*.

What is causal inference? Some typical questions

- Does education cause earnings? How large is the effect? If individuals with a college degree had a master's degree, by how much would their earnings increase?
- What is the effect of fertilizer on crop yield?
- How does healthcare affect income?
- How does advertising affect sales/clickthroughs?
- What is the effect of minimum wage on employment?
- What causes individuals to turnout to vote?
- How does race/gender influence hiring decisions?

What is causal inference? Some important concepts

- Association vs causality.
- Spurious relationships; confounding factors; common response/"lurking" variables.
- Experiments vs observational studies
- Potential outcomes, counterfactuals
- Local effects versus population-level effects
- Compliance with treatment

A fun link on spurious correlations.

Warning

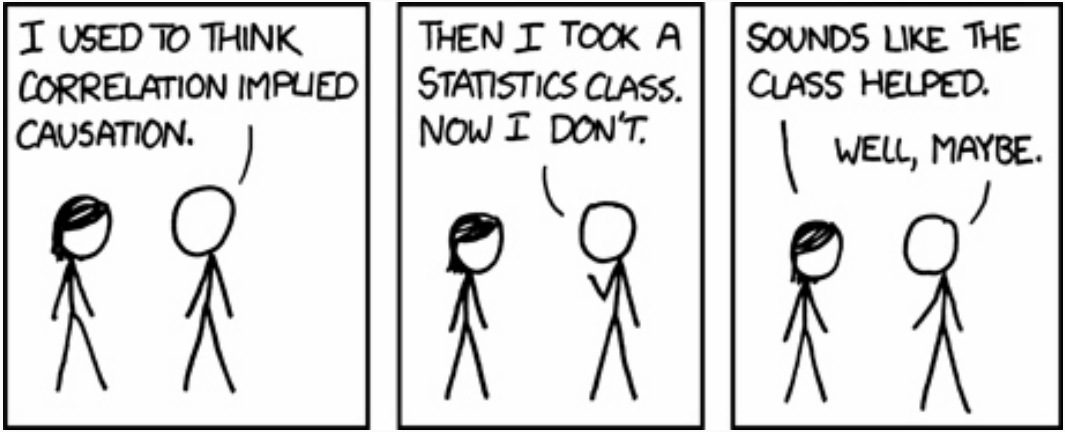


Figure 1: This may become you...

Even strong correlations don't imply causality...

People who hurry are often late $\not\rightarrow$ hurrying makes you late

Hotel occupancy is highest when prices are also high $\not\rightarrow$ increasing prices will increase demand

Places with many police officers have more crime $\not\rightarrow$ reducing police officers will reduce crime

Wine drinkers are wealthier than beer drinkers $\not\rightarrow$ drinking wine increases your income

Many tech CEOs never went to college $\not\rightarrow$ less school helps your career

Controlled experiment vs observational studies

- Ronald Fisher (1890-1962) pioneered **experimental design**, in practice, with work on crop selection, and in theory with his landmark 1935 text, *The design of experiments*. An experiment is controlled, which means that the experimenter has some control on the treatment provided to individuals. In medicine, "interventional" clinical trials.
- By contrast, in **observational studies** the analyst does not have a control on the treatment to provide to individuals. Frequently the case in social sciences. However, sometimes is is enough to understand the determinants of providing the treatment for the purpose of causal inference: "quasi-experiments". In medicine, "observational" clinical trials.

Randomization in business, economics, politics

- Randomized studies in business: AB testing, business experiments.
- Randomized studies in economics and politics: development/campaigning field experiments.

Observational studies in business, economics, politics

- Natural experiments
- Differences-in-differences
- Regression discontinuity
- Instrumental variables

The potential outcome model

The potential outcome model

- The potential outcome model was pioneered by Jerzy Neyman (1923) and extensively developed by Donald Rubin since the mid-1970s.
- Potential outcomes: Y^1 and Y^0 if treated ($1 =$ treatment state) or not treated ($0 =$ control state)
- For each individual i, y_i^0 and y_i^1 denote the potential outcomes for i in the treatment and control states. Individual treatment causal effect is thus

$$\delta_i = y_i^1 - y_i^0$$

Thinking about counterfactuals

Fundamental problem of causal inference: y_i^0 and y_i^1 (and thus δ_i) are *never* simultaneously observed. We have:

Group	Y^1	Y^0
Treatment ($D = 1$)	Observable (Y^*)	(?)
Control ($D = 0$)	(?)	Observable (Y^*)

The (?) are missing data: they indicate *counterfactuals*: "what would have happened if i had not been treated instead of being treated?"

Introduction to causal inference for data scientists

Potential outcomes and treatment effects

Michael Gill

2018-01-30

Center for Data Science | New York University

Today

Today

Part I, lecture 2. The potential outcome model.

References:

- MW, Ch 2
- IR, Chapter 1–3.
- Heckman, J. "The scientific model of causality" Sociological Methodology 2005.
- Dataset from: Lalonde, R. (1986). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review*.

The potential outcome model

- The potential outcome model was pioneered by Jerzy Neyman (1923) and extensively developed by Donald Rubin since the mid-1970s.
- Potential outcomes: Y^1 and Y^0 if treated ($1 =$ treatment state) or not treated ($0 =$ control state)
- For each individual i, y_i^0 and y_i^1 denote the potential outcomes for i in the treatment and control states. Individual treatment causal effect is thus

$$\delta_i = y_i^1 - y_i^0$$

Thinking about counterfactuals

Fundamental problem of causal inference: y_i^0 and y_i^1 (and thus δ_i) are *never* simultaneously observed. We have:

Group	Y^1	Y^0
Treatment ($D = 1$)	Observable (Y^*)	(?)
Control ($D = 0$)	(?)	Observable (Y^*)

The (?) are missing data: they indicate *counterfactuals*: "what would have happened if i had not been treated instead of being treated?"

A note on notation

I will adopt the following notation:

- $Y^* = Y^{Obs} = Y^D$ for the observed outcome, while
- $\mathbf{Y} = (Y^0, Y^1)$ for the vector of potential outcomes.

While this notation removes any possible confusion between the observed outcome and the vector of potential outcomes, it is less standard in the literature.

Observed outcome

- Random treatment $D \in \{0, 1\}$. $d_i = 1$ means i receives the treatment (treatment group); $d_i = 0$ means i does not receive the treatment (control group).
- The obsered outcome is $Y^* = Y^1$ if $D = 1$ and $Y^* = Y^0$ if $D = 0$; therefore

$$Y^* = Y^D = DY^1 + (1 - D)\, Y^0.$$

Average treatment effect

- The *average treatment effect* (ATE) is defined as

$$ATE = E[\delta] = E[Y^1] - E[Y^0].$$

- When the outcome is a binary variable, this is simply $\Pr(Y^1 = 1) - \Pr(Y^0 = 1)$.
- Other measures, such as the *causal risk ratio* $\Pr(Y^1 = 1) / \Pr(Y^0 = 1)$ could also be used as an assessment of the causal effect.

Stable unit treatment value assumption (SUTVA)

SUTVA in a set of restrictions (exclusion restrictions) introduced by Rubin in 1980. Under SUTVA:

- there is *no interference*: a unit's potential outcomes are unaffected by the treatment administered to other individuals;
- there is *no hidden variation of treatment*: treatment is deterministic and exists under only one form.

SUTVA

SUTVA is not always met in practice:

- due to *contagion* (e.g., a vaccination campaign may have effect on other individuals beyond the treated)
- due to *equilibrium effects* (e.g., too many high-skilled job applicants may decrease wages)
- in classrooms, there may be *peer effects*
- in some early drug trials, patients may share a part of the medication
- education may come in *various qualities*

The assignment mechanism

- An assignment mechanism is a process that determines which treatment is administered to units. Typically, it is the distribution of the treatment D_i of unit i conditional on the vector of potential outcomes (Y_i^0, Y_i^1) and the vector of covariates X_i.
- Desirable properties of assignment mechanisms are:
  - *individualistic*: the treatment of a unit i cannot depend on the potential outcomes, or on the covariates associated with another unit i'.
  - *probabilistic*: for every unit i, the probability of each state of the treatment is positive.
  - *unconfounded*: the assignment mechanism is independent of potential outcomes, conditional on the covariates (more on this soon).

The assignment mechanism (continued)

- Some authors distinguish between:
  - classical randomized experiments: the three desirable properties are known, and the form of the assignment mechanism is known and chosen (RCTs).
  - regular assignment mechanisms: same as before, but the form of the assignment mechanism is not known or chosen (quasi-experimental setting in observational studies).
  - irregular assignment mechanisms: observational studies where one of the desirable properties, most typically unconfoundedness, fails.

Formalizing the assignment mechanism

We will start by ignoring covariates. Under perfect randomization: *potential outcomes are independent from treatment*, i.e.,

$$\mathbf{Y} = \left(Y^0, Y^1\right) \perp\!\!\!\perp D$$

This is the case in a perfectly randomized experiment; rarely in observational studies. Of course, this does not imply that the observed outcome Y^{obs} is independent from treatment D.

Under randomized treatment,

$$E\left[Y^* | D = 1\right] = E\left[Y^1 | D = 1\right] = E\left[Y^1\right]$$
$$E\left[Y^* | D = 0\right] = E\left[Y^0 | D = 0\right] = E\left[Y^0\right]$$

Example: should I get a degree?

Take the example where treatment = getting a degree. Assume that $U \sim g(\mu)$ is a random variable capturing unobserved ability, and that earnings without and with treatment are given by

$$Y^0 = f^0(U) \text{ and } Y^1 = f^1(U)$$

Then perfect randomization boils down to

$$U \perp\!\!\!\perp D$$

which means that treatment is perfectly uncorrelated with ability.

Realistic or not?

Example: should I get a degree?

Take the example where treatment = getting a degree. Assume that $U \sim g(\mu)$ is a random variable capturing unobserved ability, and that earnings without and with treatment are given by

$$Y^0 = f^0(U) \text{ and } Y^1 = f^1(U)$$

Then perfect randomization boils down to

$$U \perp\!\!\!\perp D$$

which means that treatment is perfectly uncorrelated with ability.

Realistic or not? We'll get back later on this example, but note that it may not be less reasonable to assume that the more able fraction of the population is selected into getting a degree, so that

$$D = 1\{U \geq \underline{u}\}.$$

A remark

The condition for perfect randomization

$$(Y^0, Y^1) \perp\!\!\!\perp D$$

is stronger than the condition that each of the random variable defining a potential outcome is independent from treatment, i.e.,

$$Y^0 \perp\!\!\!\perp D \text{ and } Y^1 \perp\!\!\!\perp D.$$

- In many cases, we shall need only the weaker form. See Heckman, Ichimura, and Todd, (1998).
- *Why are (or are) the two different?*

DAGs (a minor detour)

DAG: directed acyclic graph

Suppose we have three random variables, X, Y, and Z. These variables could be represented in the following causal graph.

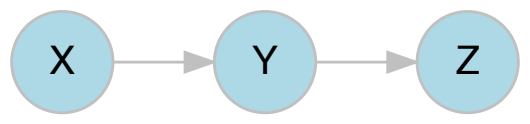
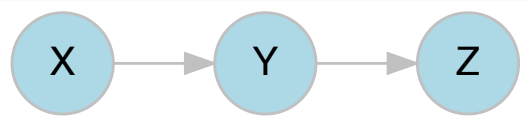


Figure 1: A basic DAG

Here, we think about arrows representing causal dependencies between variables. These needn't be linear, nor have any sign restriction.

d-separation (and conditional independence)

Taking the same example as before, the graph may help to visualize conditional dependence.



Here, the graph above makes the probabilistic claim that

$$Z \perp\!\!\!\perp X | Y$$

In the Pearl parlance, we can directionally-separate Z from X given Y—i.e., if we know Y, X provides no additional information for a prediction about Z. Put another way, X indirectly causes Z, but only through Y.

Confounding

Suppose we have three random variables, X, Y, and U—e.g., schooling, earnings, and (unobserved) student ability. With earnings as our outcome of interest, these variables could be represented in the following causal graph.

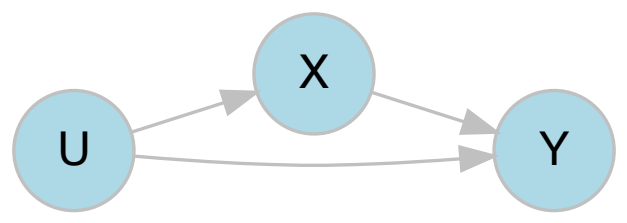


Figure 2: U as a confound

Another confound

But how does this case differ from the prior?

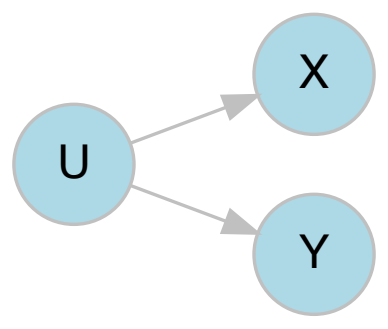
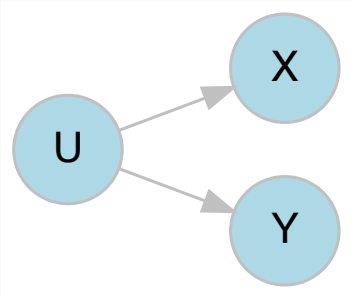


Figure 3: *U* as another confound

What does this graph mean? What are its implied conditional relations?

Implied (causal and probabilistic) conditional relations

- If we read this graph in terms of its *implied probabilistic relations*, there may be an observed relationship between X and Y, but this is *not* causal in nature.
- In other words, it is conceivable that $E[\text{corr}(X, Y)] \neq 0$, or there exhibits some observed association between X and Y, while in truth $X \perp\!\!\!\perp Y | U$.



More on causal vs. probabilistic DAGs

Suppose you a sample dataset with n rows and 2 columns, with a typical row (a_i, b_i) for each unit $i \in 1, \ldots, n$. Your goal is to learn the causal structure between the *unobserved* random variables that generated these data. Call these variables A and B.

Question: how many causal DAGs could we write down for these two random variables (ignoring other possible variables)?

3!

- $A \quad B$
- $A \rightarrow B$
- $A \leftarrow B$

What about with more variables?

Suppose you now had *three* variables, A, B, C. How many DAGs could exist for these data?

What about with more variables?

Suppose you now had *three* variables, A, B, C. How many DAGs could exist for these data?

For those thinking this expands 2^{n-1}, it doesn't...

What about with more variables?

Suppose you now had *three* variables, A, B, C. How many DAGs could exist for these data?

For those thinking this expands 2^{n-1}, it doesn't...

It turns out that given n variables, the total number of possible DAGs to draw is given by the recursion:

$$f(N) = \sum_{k=1}^{n} (-1)^{k+1} \cdot \frac{n!}{(n-k)!k!} \cdot 2^{k(n-1)} \cdot f(n-k)$$

with $f(0) = 1$, and $f(1) = 1$. This was first demonstrated in Robinson (1977).

Possible DAGs for 3 variables: A, B, C

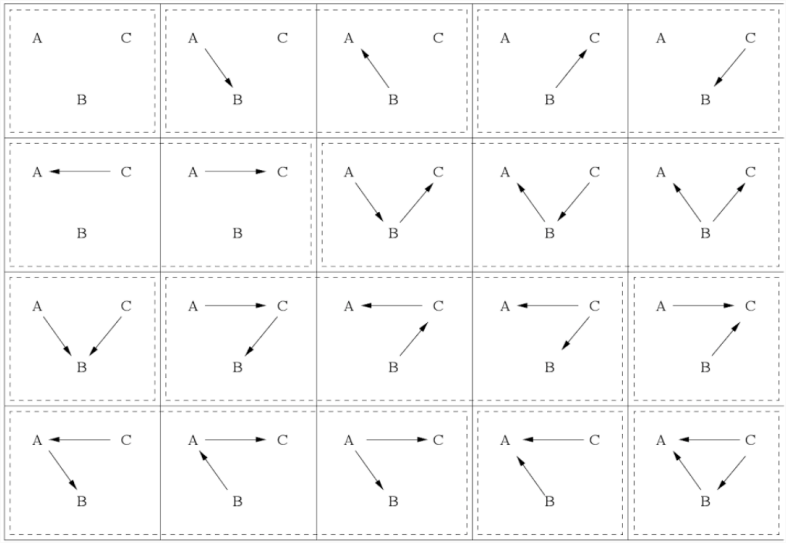


Figure 4: Korb and Nicholson, Bayesian Artificial Intelligence, 2nd Edition, pg. 245

From probabalistic assumptions to DAGs, and visa versa

Suppose all you knew was $X \perp\!\!\!\perp Y | Z$. What DAG(s) satisfy these probabilistic conditions?

From probabalistic assumptions to DAGs, and visa versa

Suppose all you knew was $X \perp\!\!\!\perp Y | Z$. What DAG(s) satisfy these probabilistic conditions?

1. $X \rightarrow Y \rightarrow Z$
2. $X \leftarrow Y \leftarrow Z$
3. $X \leftarrow Z \rightarrow Y$

Note that stochastic constraints do not uniquely identify DAGs, but DAGs will imply specific conditions.

Value of graphical models?

Strengths

- For simple stories, DAGs can be intuitive
- Useful for determining when something can be “non-parametrically identified”
- May help with “smoothing of data” (more on this later, or in HW)

Shortcomings

- Agnostic to functional form
- Hard to go from arbitrary sample to likely DAGs (i.e., “structure learning” in Bayesian networks)
- A given graph is an assumption, and therefore untestable (which is true of observational designs)

Going to R: a simple confounding example

```r
set.seed(808)
N = 5000                          # Number of draws
Z = rnorm(N)                      # N draws from standard normal
X = -1.5 + 0.5*Z + rnorm(N)       # X is function of Z
Y = 1 - 0.8*Z + rnorm(N)          # Y is function of Z

cor(X,Y)                          # Correlation between X and Y

summary(lm(Y ~ X))                # linear regrssion of Y on X
summary(lm(Y ~ Z))                # linear regrssion of Y on Z
summary(lm(Y ~ X + Z))            # linear regrssion of Y on X and Z
```

Residualize the effect of the confounder

```r
X_ = X - predict(lm(X ~ Z))      # X after residualizing Z
Y_ = Y - predict(lm(Y ~ Z))      # Y after residualizing Z

par(mfrow=c(1,2))                # Plotting parameters
plot(X,Y)                        # First plot
abline(lm(Y~X), col="blue")      # X,Y slope w.o controlling for Z
plot(X_,Y_)                      # Second plot
abline(lm(Y_~X_), col="blue")    # X,Y slope after controlling for Z
```

Results

- After controlling for Z, the apparent relationship between X and Y goes away.
- Note: $corr(X, Y) = 0 \not\Rightarrow X \perp\!\!\!\perp Y$, but $X \perp\!\!\!\perp Y \implies corr(X, Y) = 0$

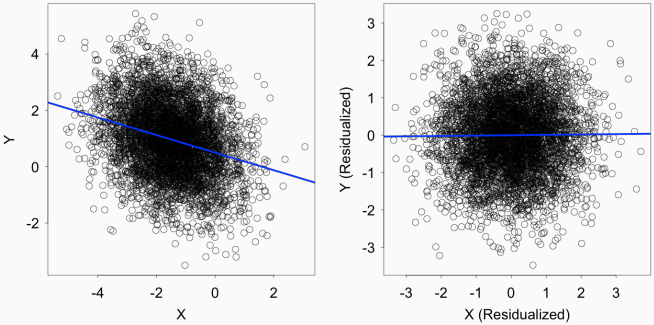


Figure 5: Residualizing X and Y from Z

Returning to Neyman-Rubin

From before, we defined the average treatment effect (ATE) as

Average Treatment Effect (ATE)

$$ATE = E[\delta] = E[Y^1] - E[Y^0]$$

We can also define different conditional average treatment effects. We will begin with two important cases, ATT and ATC.

Average Treatment Effect for the Treated (ATT)

$$ATT = E[\delta|D = 1] = E[Y^1|D = 1] - E[Y^0|D = 1]$$

Average Treatment Effect for the Controls (ATC)

$$ATC = E[\delta|D = 0] = E[Y^1|D = 0] - E[Y^0|D = 0]$$

ATE, ATT, ATC, and the "naive" estimator

ATE, ATT, and ATC are different causal estimands, as they represent different causal quantities of interest. We will distinguish estimands from estimators.

We note that $E[Y^1|D=1]$ and $E[Y^0|D=0]$ are measurable; however, $E[Y^0|D=1]$ and $E[Y^1|D=0]$ are not.

One can form a naive causal estimator by:

Naive causal estimator

$$\delta_{NAIVE} = E[Y^1|D=1] - E[Y^0|D=0] = E[Y^*|D=1] - E[Y^*|D=0].$$

Note that under randomized treatment, a unit's potential outcomes are independent of treatment assignment—i.e., $\mathbf{Y} = (Y^0, Y^1) \perp\!\!\!\perp D$—which implies that $\delta_{NAIVE} = ATE = ATT = ATC$.

Bias of the naive estimator

- The naive estimator, however, does not coincide with ATE. Indeed, letting $\pi = \Pr(D = 1)$,

$$\begin{aligned} ATE = E[\delta] = \; & \pi E[Y^1|D = 1] + (1 - \pi) E[Y^1|D = 0] \\ & - (\pi E[Y^0|D = 1] + (1 - \pi) E[Y^0|D = 0]) \end{aligned}$$

  thus

$$\begin{aligned} ATE = \; & E[Y^1|D = 1] - E[Y^0|D = 1] \\ & - (1 - \pi) \{E[\delta|D = 1] - E[\delta|D = 0]\} \\ = \; & \delta_{NAIVE} - (E[Y^0|D = 1] - E[Y^0|D = 0]) \\ & - (1 - \pi) \{E[\delta|D = 1] - E[\delta|D = 0]\} \end{aligned}$$

- Thus, the difference between ATE and δ_{NAIVE} comes from two terms:
  - $E[Y^0|D = 1] - E[Y^0|D = 0]$, which is a baseline bias.
  - $(1 - \pi) \{E[\delta|D = 1] - E[\delta|D = 0]\}$, a differential treatment effect bias.

Naive estimator vs. ATE

Thus, the difference between ATE and δ_{NAIVE} comes from two terms:

- $E\left[Y^0|D=1\right]-E\left[Y^0|D=0\right]$ which is a baseline bias.
- $(1-\pi)\left\{E\left[\delta|D=1\right]-E\left[\delta|D=0\right]\right\}$ which is a differential treatment effect bias.

Review MW or here if you need more clarity on this derivation, as it will be important to your HW.

The naive estimator in the education decision example

Recall our previous education decision example. The ATE is

$$E\left[f^1\left(U\right)\right] - E\left[f^0\left(U\right)\right]$$

and the naive estimator is

$$\delta_{NAIVE} = E\left[f^1\left(U\right)|D = 1\right] - E\left[f^0\left(U\right)|D = 0\right]$$

- Under perfect randomization ($U \perp\!\!\!\perp D$), these coincide.
- Under the assumption of selection on ability $D = 1\left\{U \geq \underline{u}\right\}$, we get that

$$\delta_{NAIVE} = E\left[f^1\left(U\right)|U \geq \underline{u}\right] - E\left[f^0\left(U\right)|U < \underline{u}\right].$$

The propensity score

- We now introduce a vector of covariates X: (may include age, gender, income, education). The *propensity score* is defined (after Rosenbaum and Rubin, 1983) as the probability of being assigned in the treatment group conditional on the covariates, that is

$$e(x) = \Pr(D_i = 1 | X_i = x) = E[D_i | X_i = X].$$

- For instance, assume that there is an additional variable, "gender" which affects treatment D, i.e.,

$$\Pr(D = 1 | Female) = 0.7$$
$$\Pr(D = 1 | Male) = 0.5,$$

  which means that women are more likely to be treated than men. The propensity score is sometimes known to the researcher, sometimes not.
- It is often assumed that $e(x) \in (0, 1)$ for all x ("overlap" assumption).

Unconfoundedness

In some settings, it is plausible to assume that conditional on the vector of covariates X, the treatment is independent on the potential outcomes, that is

$$(Y^0, Y^1) \perp\!\!\!\perp D \mid X.$$

This assumption is called unconfoundedness after Rubin (1990). A.k.a. ignorability, selection on the observables, data missing at random.

This assumption is often more plausible than perfect randomization; and it has pretty much the same effect: condition on each group x, use the assumption, and re-integrate over x.

An immediate example

Going back to our education decision where treatment=getting a degree. Assume that $X \sim g(\mu)$ is now an observed variable measuring ability, say IQ, and that the potential outcomes are the earnings without and with the degree

$$Y^0 = f^0(X) \text{ and } Y^1 = f^1(X).$$

The treatment (getting a degree) is D, and it is assumed that $e(X) \in (0, 1)$ (overlap). Because (Y^0, Y^1) is a deterministic function of D,

$$(Y^0, Y^1) \perp\!\!\!\perp D \mid X$$

i.e., unconfoundedness holds.

A more realistic example

Let's go back to our education decision where treatment=getting a degree. Maintain that $X \sim g(\mu)$ is an observed variable measuring ability, say IQ, and that the potential outcomes are the earnings without and with the degree are $Y^0 = f^0(X) + \varepsilon^0$ and $Y^1 = f^1(X) + \varepsilon^1$ and that the cost of effort associated with obtaining the degree is $c(X) + \eta$. It is assumed that ε, η, and X are independent.

A more realistic example

Let's go back to our education decision where treatment=getting a degree. Maintain that $X \sim g(\mu)$ is an observed variable measuring ability, say IQ, and that the potential outcomes are the earnings without and with the degree are $Y^0 = f^0(X) + \varepsilon^0$ and $Y^1 = f^1(X) + \varepsilon^1$ and that the cost of effort associated with obtaining the degree is $c(X) + \eta$. It is assumed that ε, η, and X are independent.

The treatment is chosen if the expected benefit associated with the degree exceeds the cost, that is

$$D = 1\{f^1(X) - f^0(X) \geq C(X) + \eta\}$$

and hence, in this model, it is clear that

$$(Y^0, Y^1) \perp\!\!\!\perp D \mid X,$$

i.e., unconfoundedness holds.

Unconfoundedness and the propensity score

- Unconfoundedness implies that

$$\pi_{YD|X}(y, d|x) = \pi_{Y|X}(y|x)\,\pi_{D|X}(d|x),$$

  where $Y = (Y^0, Y^1)$.
- As a result,

$$\pi_{XYD}(x, y, d) = \pi_{XY}(x, y)\,\pi_{D|X}(d|x).$$

- But $\pi_{D|X}(d|x) = e(x)$ if $d = 1$, and $\pi(d|x) = 1 - e(x)$ if $d = 0$, thus

$$\pi_{D|X}(d|x) = 1 - e(x) + d(2e(x) - 1),$$

  so that:
  - $\pi_{D|X}(d|x)$ is a function of $e(x)$ and d only.
  - Under unconfoundedness

$$\pi_{XYD}(x, y, d) = \pi_{XY}(x, y)(1 - e(x) + d(2e(x) - 1)).$$

Unconfoundedness and the propensity score (2)

- Under unconfoundedness, we have

$$(Y^0, Y^1) \perp\!\!\!\perp D \mid e(X),$$

  i.e., instead on conditioning on X (which may be large-dimensional), it is enough to condition on the propensity score. This explain the fundamental importance of the latter.
- Indeed, under unconfoundedness, $\pi_{YD|X}(y, d|x) = \pi_{Y|X}(y|x)\,\pi_{D|X}(d|x)$, but recall that $\pi_{D|X}(d|x) = f(e(x), d)$ is a function of $e(x)$ and of d (in fact, $\pi_{D|X}(d|x) = 1 - e(x) + d(2e(x) - 1)$). Hence

$$\pi_{YD|e}(y, d|e(X) = e) = \frac{\sum_{x':e(x')=e} \pi_{XYD}(x', y, d)}{\#\{x' : e(x') = e\}} = \frac{\sum_{x':e(x')=e} \pi_{XY}(x', y)\,\pi_{D|X}(d|x')}{\#\{x' : e(x') = e\}}$$

$$= f(e, d)\,\frac{\sum_{x':e(x')=e} \pi_{XY}(x', y)}{\#\{x' : e(x') = e\}}$$

$$= f(e, d)\,\pi_{Y|e}(y|e(x) = e), \text{ QED.}$$

ATE under unconfoundedness (1)

- Under unconfoundedness, we have

$$E\left[Y^{*}|D=1,X\right]=E\left[Y^{1}|X\right]$$
$$E\left[Y^{*}|D=0,X\right]=E\left[Y^{0}|X\right]$$

- Indeed, $f_{XYD}\left(x,y,d\right)=f_{X}\left(x\right)f_{Y|X}\left(y|x\right)f_{D|X}\left(d|x\right)$, thus $E\left[Y|D=1,X\right]=E\left[Y^{1}|D=1,X\right]=E\left[Y^{1}|X\right]$, and similarly for the second equality.
- As a result, the conditional ATE is obtained by

$$ATE\left(x\right)=E\left[Y^{1}|X\right]-E\left[Y^{0}|X\right]=E\left[Y|D=1,X\right]-E\left[Y|D=0,X\right].$$

- The ATE is obtained by

$$ATE=E\left[ATE\left[X\right]\right].$$

ATE under unconfoundedness (2)

- Note that under unconfoundedness, we do not have $ATE = E[Y^*|D=1] - E[Y^*|D=0]$!!!
- Indeed, assume that there are only two values of x: 0 and 1. Then

$$ATE(0) = \frac{\mathbb{E}[Y^* 1\{X=0, D=1\}]}{\Pr(X=0, D=1)} - \frac{\mathbb{E}[Y^* 1\{X=0, D=0\}]}{\Pr(X=0, D=0)}$$

$$ATE(1) = \frac{\mathbb{E}[Y^* 1\{X=1, D=1\}]}{\Pr(X=1, D=1)} - \frac{\mathbb{E}[Y^* 1\{X=1, D=0\}]}{\Pr(X=1, D=0)}$$

  and

$$ATE = \Pr(X=0)\, ATE(0) + \Pr(X=1)\, ATE(1)$$

- Both quantities will coincide if $D \perp\!\!\!\perp X$. But one can show that this implies perfect randomization, i.e. $(Y^0, Y^1) \perp\!\!\!\perp D$. (This will be left as an exercise).

Statistical estimation: discrete case

- How to estimate this in practice under the unconfoundedness assumption?
- If X is discrete and can take a finite number of values $x^1, \ldots, x^K$, provided the sample is large enough, a natural estimator of $ATE(x)$ is given by

$$\widehat{ATE}\left(x^k\right) = \frac{1}{\{\#i : x_i = x^k, d = 1\}} \sum_{i:x_i=x^k, d=1} y_i - \frac{1}{\{\#i : x_i = x^k, d = 0\}} \sum_{i:x_i=x^k, d=0} y_i$$

  and the average treatment effect is estimated by

$$\widehat{ATE} = \sum_{k=1}^{K} \frac{\left\{\#i : x_i = x^k\right\}}{\{\#i\}} \widehat{ATE}\left(x^k\right)$$

- If X is continuous, such approach no longer works.

Statistical estimation: continuous case

- When X is continuous, the problem gets more complicated and we will return to it later.
- One possibility is to have a parametric model (e.g. linear) to estimate $E\left[Y^{*}|X, D=0\right]$ and $E\left[Y^{*}|X, D=1\right]$.
- Hahn (1998) has shown that for any estimator $\widehat{ATE}$ of ATE, we have

$$var\left(\widehat{ATE}-ATE\right) \geq \frac{1}{n} E\left[\frac{V^{1}(X)}{e(X)}+\frac{V^{0}(X)}{1-e(X)}+\left(ATE(X)-ATE\right)^{2}\right]$$

where $V^{d}(X)=var\left(Y^{d}|X\right)$, and that this bound can be approximately attained when the sample size n is large.

Statistical estimation: a first linear model

- Assume

$$\begin{cases} Y^0 = \alpha_0 + \beta_0' X + \varepsilon_0 \\ Y^1 = \alpha_1 + \beta_1' X + \varepsilon_1 \end{cases}$$

  where unconfoundness holds: $(\varepsilon_0, \varepsilon_1) \perp\!\!\!\perp D \mid X$.

- Then

$$ATE\,(x) = \alpha_1 - \alpha_0 + (\beta_1 - \beta_0)' x,$$

  and therefore the average treatment effect is obtained by

$$ATE = \alpha_1 - \alpha_0 + (\beta_1 - \beta_0)' E\,[X]\,.$$

- Naive approach; much more on this later.

“Bad controls” lead to bias

Two infamous cases:

1. Conditioning on a variable that is **post-treatment**
2. **Omitted variables** that are correlated with error and treatment

We will formalize these intuitions next week, as we begin thinking about treatment effects.

Introduction to causal inference for data scientists

RCTs, AB testing, and business experiments (1)

Michael Gill

2018-02-06

Center for Data Science | New York University

Today

Field experiments: methods and issues

Core takeaways today:

- Estimating treatment effects via regression
- Overview of seminal field experiments
- Begin to think about inference/uncertainty/hypothesis testing

Regression analysis of treatment effect

- One way to estimate the ATE under perfect randomization is by running a linear regression

$$Y_i^* = \alpha + \beta D_i + \varepsilon_i$$

and note that if the treatment is random, then D_i and ε_i are independent, so the treatment effect can be estimated by linear regression as $\hat{\beta}_{OLS}$.

Unconfoundedness

- Under unconfoundedness, assume

$$\begin{cases} Y^0 = \alpha_0 + \beta_0' X + \varepsilon_0 \\ Y^1 = \alpha_1 + \beta_1' X + \varepsilon_1 \end{cases}$$

  where unconfoundness holds: $(\varepsilon_0, \varepsilon_1) \perp\!\!\!\perp D \mid X$.

- Then, recall that

$$ATE(x) = \alpha_1 - \alpha_0 + (\beta_1 - \beta_0)' x,$$

  and therefore the average treatment effect is obtained by

$$ATE = \alpha_1 - \alpha_0 + (\beta_1 - \beta_0)' E[X].$$

Field experiment

- Field experiments vs lab experiments?

Field experiment

- Field experiments vs lab experiments?
- According to Harrison and List (2006), the following six factor are determinant in a field experiment:
  - The pool of subjects
  - The information brought to the participants
  - The incentive mechanism (what commodity is used to encourage participation). Participants paid to participate? (e.g., Chassang et al.)
  - The task asked from participants
  - The stakes participants have in the outcome
  - The environment

Methods of randomization

- Various types of randomization exist, cf. Duflo, Glennerster and Kremer (2008).
  - *Oversubscription*. Limited budget: size of treatment group is determined by available budget.
  - *Randomized phase-in*. Gradual phase-in of the program across eligible areas. Control group is made of the areas waiting to receive the treatment.
  - *Within-group randomization*. Some subgroups are provided in each targeted area, to minimize spatial inequalities. Greater risk of a spillover problem.
  - *Encouragement design*. Instead of randomizing treatment, announcement of the program, or incentive to participate to the program, is randomly assigned.

Issues with randomization

- Ethical and political issues with randomization

Issues with randomization

- Ethical and political issues with randomization: it may not be ethically or politically acceptable to purposely exclude a group from a treatment. Should treatment be randomly provided or provided to those who need it most?

Issues with randomization

- Ethical and political issues with randomization: it may not be ethically or politically acceptable to purposely exclude a group from a treatment. Should treatment be randomly provided or provided to those who need it most?
- Internal vs. External validity

Issues with randomization

- Ethical and political issues with randomization: it may not be ethically or politically acceptable to purposely exclude a group from a treatment. Should treatment be randomly provided or provided to those who need it most?
- Internal vs. External validity: targeted sample should be representative so that conclusions of a small-scale experiment may not hold on a larger scale. Internal validity: proper randomization; absence of confounding factors.

Issues with randomization

- Ethical and political issues with randomization: it may not be ethically or politically acceptable to purposely exclude a group from a treatment. Should treatment be randomly provided or provided to those who need it most?
- Internal vs. External validity: targeted sample should be representative so that conclusions of a small-scale experiment may not hold on a larger scale. Internal validity: proper randomization; absence of confounding factors.
- Imperfect compliance

Issues with randomization

- Ethical and political issues with randomization: it may not be ethically or politically acceptable to purposely exclude a group from a treatment. Should treatment be randomly provided or provided to those who need it most?
- Internal vs. External validity: targeted sample should be representative so that conclusions of a small-scale experiment may not hold on a larger scale. Internal validity: proper randomization; absence of confounding factors.
- Imperfect compliance: some individuals who are assigned the treatment may not comply. Sometimes, control individuals may get the treatment. A solution is to redefine “treatment” as “probability of being exposed to treatment” – “intention-to-treat”.

Issues with randomization (continued)

- Spillovers.

Issues with randomization (continued)

- Spillovers. Control and treatment groups should be located sufficiently far apart.

Issues with randomization (continued)

- Spillovers. Control and treatment groups should be located sufficiently far apart.
- “Hawthorne effect, ”

Issues with randomization (continued)

- Spillovers. Control and treatment groups should be located sufficiently far apart.
- “Hawthorne effect, ” “John Henry effect”: people modify their behaviour if they know they are part of an experiment, or more generally when they are aware of being observed. This is a threat to external validity.

Some famous field experiments

- We will review three famous field experiments:
  - Effect of class size reduction on academic performance: the STAR experiment in Tenessee. Krueger (1999).
  - Why do people give to charities? DellaVigna, List and Malmendier (2012).
  - What is the effect of incentives on teachers' absenteism? Duflo, Hannaand Ryan (2012).

The STAR experiment: class size reduction

- The question is the causal effect of class size reduction (CSR) on students' academic performances. Two large-scale experiments on the topic: STAR in Tennessee and SAGE in Wisconsin. Focus on the former.
- The STAR experiment: Student/Teacher Assignment Ratio experiment, run in Tennessee in the 1980s.
- 11,600 students and teachers randomly assigned to three groups by class sizes:
  - "small" (13 to 17 students/ teacher),
  - "regular" (22 to 25 students/teacher with no aid)
  - "regular+aid" (22 to 25 students/teacher with full-time teacher aid)
- Randomization was performed within schools, and once initially assigned, students remained assigned to their group for four years.

The regression model

- Krueger (1999) regresses:

$$Y_{ics} = \beta_0 + \beta_1 SMALL_{cs} + \beta_2 REG/A_{cs} + \beta_3 X_{ics} + \alpha_s + \varepsilon_{ics}$$

where Y_{ics} is the average percentile score on the SAT test of student i in class c at school s, $SMALL_{cs}$ is a dummy variable of whether the student was assigned to a small class, REG/A_{cs} is a dummy variable of whether the student was assigned to a "regular+aid" class, and X_{ics} is a vector of student and teacher covariates like gender, etc. Because randomization was done within schools, α_s is a school fixed effect.

Krueger's study: results

- The following figure shows the density of the test score distributions of students (treated/control) in K-3 grades:

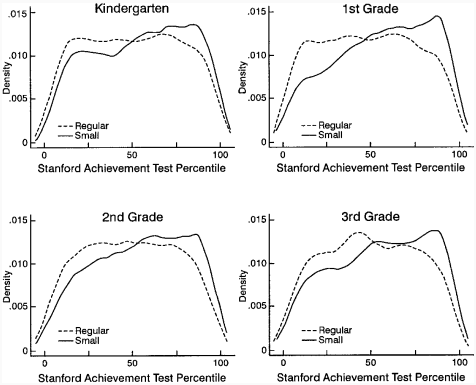


Figure 1:

Krueger's study: results (kindergarten)

Explanatory variable	OLS: actual class size			
	(1)	(2)	(3)	(4)
			A. Kindergarten	
Small class	4.82	5.37	5.36	5.37
	(2.19)	(1.26)	(1.21)	(1.19)
Regular/aide class	.12	.29	.53	.31
	(2.23)	(1.13)	(1.09)	(1.07)
White/Asian (1 = yes	—	—	8.35	8.44
			(1.35)	(1.36)
Girl (1 = yes)	—	—	4.48	4.39
			(.63)	(.63)
Free lunch (1 = yes)	—	—	−13.15	−13.07
			(.77)	(.77)
White teacher	—	—	—	−.57
				(2.10)
Teacher experience	—	—	—	.26
				(.10)
Master's degree	—	—	—	−.51
				(1.06)
School fixed effects	No	Yes	Yes	Yes
R^2	.01	.25	.31	.31

Figure 2:

Krueger's study: results (first grade)

	B. First grade			
Small class	8.57	8.43	7.91	7.40
	(1.97)	(1.21)	(1.17)	(1.18)
Regular/aide class	3.44	2.22	2.23	1.78
	(2.05)	(1.00)	(0.98)	(0.98)
White/Asian (1 = yes)	—	—	6.97	6.97
			(1.18)	(1.19)
Girl (1 = yes)	—	—	3.80	3.85
			(.56)	(.56)
Free lunch (1 = yes)	—	—	−13.49	−13.61
			(.87)	(.87)
White teacher	—	—	—	−4.28
				(1.96)
Male teacher	—	—	—	11.82
				(3.33)
Teacher experience	—	—	—	.05
				(0.06)
Master's degree	—	—	—	.48
				(1.07)
School fixed effects	No	Yes	Yes	Yes
R^2	.02	.24	.30	.30

Figure 3:

Attrition?

- Attrition is the fact that some students may leave the sample (e.g., move to a private school, or to a different school district).

Attrition?

- Attrition is the fact that some students may leave the sample (e.g., move to a private school, or to a different school district).
- Problem: students initially assigned to regular classes and who have higher scores may be more likely to leave the sample, as they have more outside options. Hence attrition may be selective, which will bias the measurement of the treatment effect.

Attrition?

- Attrition is the fact that some students may leave the sample (e.g., move to a private school, or to a different school district).
- Problem: students initially assigned to regular classes and who have higher scores may be more likely to leave the sample, as they have more outside options. Hence attrition may be selective, which will bias the measurement of the treatment effect.
- One way to deal with attrition is to impute the scores of the students who leave the sample. This can be done by prediction based on past score results.

Effects of attrition

TABLE VI
EXPLORATION OF EFFECT OF ATTRITION DEPENDENT VARIABLE: AVERAGE PERCENTILE SCORE ON SAT

	Actual test data		Actual and imputed test data	
Grade	Coefficient on small class dum.	Sample size	Coefficient on small class dum.	Sample size
K	5.32	5900	5.32	5900
	(.76)		(.76)	
1	6.95	6632	6.30	8328
	(.74)		(.68)	
2	5.59	6282	5.64	9773
	(.76)		(.65)	
3	5.58	6339	5.49	10919
	(.79)		(.63)	

Estimates of reduced-form models are presented. Each regression includes the following explanatory variables: a dummy variable indicating initial assignment to a small class; a dummy variable indicating initial assignment to a regular/aide class, unrestricted school effects; a dummy variable for student gender; and a dummy variable for student race. The reported coefficient on small class dummy is relative to regular classes. Standard errors are in parentheses.

Figure 4:

Charitable giving

- 90% of Americans give money to charities every year.
- Why do they give? several theories:
  - because they care about a specific cause: “warm glow” of giving, altruistic motives. This is efficient from a welfare point of view as both utilities of giver and receiver are enhanced (Becker 1974; Andreoni 1989, 1990).
  - because they feel social pressure for it (Akerlof and Kranton 2000). Ambiguous welfare effect: increase receiver’s utility but may decrease giver’s.
- How would you distinguish experimentally between these two motives?

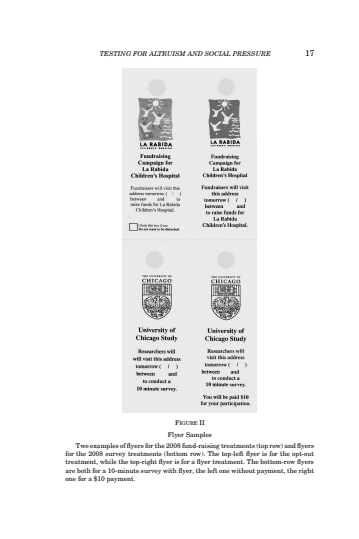
DellaVigna et al.'s field experiment

The field experiment was a door-to-door fundraising drive in the Chicago area for two charities: a local children's hospital, (La Rabida, well-known premier hospital for children in the area), and an out-of-state charity (ECU, unknown to most participants).

The experimental design

- 7,668 households were approached between April and October 2008. The experimental design is carried in order to allow to either seek or avoid the solicitor. Hence, households are randomized into three groups:
  - First treatment: notice. A flyer on the door gives one day prior notice about the one-hour time interval
  - Second treatment: notice with opt-out. The flyer includes a box to be checked if the household does not want to be disturbed.
  - Baseline treatment: solicitors arrive without prior notice.
- Outcome is measured both in terms of % of households who open the door and % of households who give.
- How does this experimental design address the giving motives questions?

The flyer



TESTING FOR ALTRUISM AND SOCIAL PRESSURE 17

LA RABIDA
CHILDREN'S HOSPITAL

Fundraising Campaign for La Rabida Children's Hospital

Fundraisers will visit this address tomorrow (/) between and to raise funds for La Rabida Children's Hospital.

☐ Check this box if you do not want to be detected.

LA RABIDA
CHILDREN'S HOSPITAL

Fundraising Campaign for La Rabida Children's Hospital

Fundraisers will visit this address tomorrow (/) between and to raise funds for La Rabida Children's Hospital.

THE UNIVERSITY OF CHICAGO

University of Chicago Study

Researchers will will visit this address tomorrow (/) between and to conduct a 10 minute survey.

THE UNIVERSITY OF CHICAGO

University of Chicago Study

Researchers will visit this address tomorrow (/) between and to conduct a 10 minute survey.

You will be paid $10 for your participation.

FIGURE II

Flyer Samples

Two examples of flyers for the 2008 fund-raising treatments (top row) and flyers for the 2008 survey treatments (bottom row). The top-left flyer is for the opt-out treatment, while the top-right flyer is for a flyer treatment. The bottom-row flyers are both for a 10-minute survey with flyer, the left one without payment, the right one for a $10 payment.

Figure 5:

DellaVigna et al.'s field experiment

- Determining the motives of charitable giving:
  - If the main motive is altruism, the notice should increase % who are home, % who open the door, and % who give.
  - If the main motive is social pressure, the notice should *decrease* % who open the door, and % who give.
- After the solicitor has collected a gift (or not), s/he asks the individual if s/he wants to complete a survey on charitable giving, and announces a duration randomized into 5 or 10 minutes, and payment for completing the survey randomized into $0, $5, or $10.

Experimental treatments

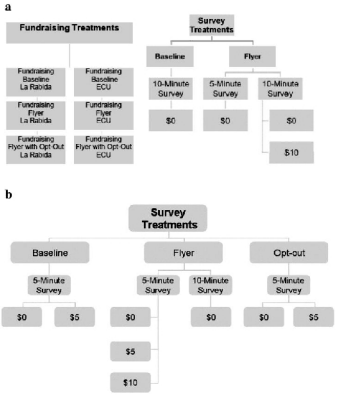


FIGURE III

Experimental Treatments (Top) 2008, (Bottom) 2009

Summary of the treatments run in the door-to-door field experiments in 2008 (charity and survey) and run in 2009. La Rabida and ECU are the names of the two charities for which the funds were raised.

Figure 6:

Regression: design

- Regress the outcome variables (door opening and giving) on the following regressors:
  - Notice
  - Notice+opt out
  - Dummy for charity = ECU (less well-known hospital)
  - Fixed effects for solicitor, date-location, hour, house quality

Regression: results

Specification:	OLS regressions									
Dep. var.:	Indicator for answering the door		Indicator for giving		Indicator for giving: Small amount ($\leq$ \$10)		Indicator for giving: Large amount ($>$ \$10)		Amount given (including \$0)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Flyer treatment	−0.0387		−0.0011		−0.0033		0.0022		−0.1459	
	(0.0137)***		(0.0062)		(0.0052)		(0.0035)		(0.1357)	
Flyer with opt-out treatment	−0.0967		−0.0195		−0.0193		−0.0002		−0.3041	
	(0.0194)***		(0.0084)**		(0.0081)**		(0.0051)		(0.1653)*	
Indicator ECU charity	0.01	0.0041	−0.0249	−0.0263	−0.0127	−0.0107	−0.0123	−0.0155	−0.7611	−0.9767
	(0.0143)	(0.0234)	(0.0049)***	(0.0085)***	(0.0053)**	(0.0085)	(0.0032)***	(0.0052)***	(0.1368)***	(0.2014)***
Flyer treatment * ECU charity		−0.0365		0.0006		−0.0045		0.0051		0.1154
		(0.0313)		(0.0094)		(0.0076)		(0.0045)		(0.1240)
Flyer with opt-out * ECU charity		−0.089		−0.0183		−0.0222		0.0039		−0.0907
		(0.0271)***		(0.0100)*		(0.0098)**		(0.0058)		(0.1268)
Flyer treatment * La Rabida charity		−0.0396		−0.0019		−0.0028		0.0009		−0.2545
		(0.0144)***		(0.0078)		(0.0066)		(0.0046)		(0.1841)
Flyer with opt-out * La Rabida charity		−0.106		−0.0202		−0.0161		−0.0042		−0.4573
		(0.0319)***		(0.0132)		(0.0128)		(0.0087)		(0.2885)
Omitted treatment	No-flyer, La Rabida		No-flyer, La Rabida		No-flyer, La Rabida				No-flyer, La Rabida	
Mean of dep. var. for omitted treatment	0.413		0.0717		0.0414	0.0414	0.0215	0.0215	1.161	1.161
Fixed effects for solicitor, date-location, hour, and area rating	X	X	X	X	X	X	X	X	X	X
N	N = 7668	N = 7668	N = 7668	N = 7668	N = 7668	N = 7668	N = 7668	N = 7668	N = 7668	N = 7668

Notes. Estimates for a linear probability model, with standard errors clustered by solicitor-date, in parentheses. The omitted treatment is the baseline no-flyer fund-raising treatment for the La Rabida charity. The regressions include fixed effects for the solicitor, for the date-town combination, for the hour of day, and for a subjective rating of home values in the block. * significant at 10%; ** significant at 5%; *** significant at 1%.

Figure 7:

Regression: results (continued)

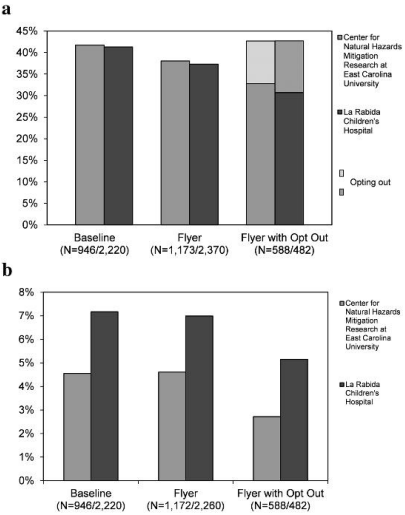


Figure 8:

The effect of incentives on teachers absenteesim

- Teacher absenteeism is a major issue in Indian primary school. Duflo, Hanna and Ryan (2012) run a field experiment to study the effect of incentives to reduce it.
- Study carried in 2003 in the rural villages of Rajasthan, India, where absenteeism rate before the start of the program was 53%.
- In 57 randomly selected, teachers had to be photographed in their classroom with a time stamp. Their salary was a function of attendance:
  - Rs. 500 if attended fewer than 10 days in a month, and
  - Rs. 50 for any additional day attended that month.
- In the 56 comparison schools, teachers were paid a fixed rate for the month (Rs. 1,000).

Impact on teacher's performances

TABLE 1—BASELINE DATA

	Treatment (1)	Control (2)	Difference (3)
Panel A. Teacher attendance			
School open	0.66	0.64	0.02
			(0.11)
	41	39	80
Panel B. Student participation (random check)			
Number of students present	17.71	15.92	1.78
			(2.31)
	27	25	52
Panel C. Teacher qualifications			
Teacher test scores	34.99	33.54	1.44
			(2.02)
	53	54	107
Panel D. Teacher performance measures (random check)			
Percentage of children sitting within classroom	0.83	0.84	0.00
			(0.09)
	27	25	52
Percent of teachers interacting with students	0.78	0.72	0.06
			(0.12)
	27	25	52
Blackboards utilized	0.85	0.89	−0.04
			(0.11)
	20	19	39
F-stat (1,110)			1.21
p-value			(0.27)
Panel E. Baseline test scores			
Took written exam	0.17	0.19	−0.02
			(0.04)
	1,136	1,094	2,230
Total score on oral exam	−0.08	0.00	−0.08
			(0.07)
	940	888	1,828
Total score on written exam	0.16	0.00	0.16
			(0.19)
	196	206	402

Figure 9:

Impact on fractions of schools open

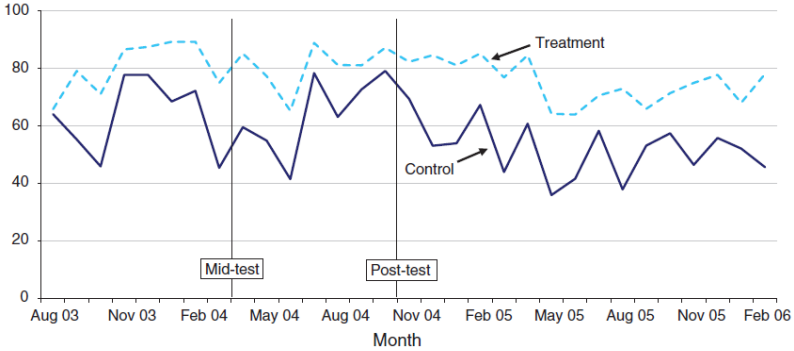


FIGURE 1. PERCENTAGE OF SCHOOLS OPEN DURING RANDOM CHECKS

Notes: The program began in September 2003. August only includes the 80 schools checked before announcement of program. September includes all random checks between August 25 through the end of September. Child learning levels were assessed in a mid-test (April 2004) and a post-test (November 2004). After the post-test, the "official" evaluation period ended. Random checks continued in both the treatment and control schools.

Figure 10:

Impact on teacher attendance

Table 2—Teacher Attendance

September 2003–February 2006			Difference between treatment and control schools		
Treatment (1)	Control (2)	Diff (3)	Until mid-test (4)	Mid- to post-test (5)	After post-test (6)
Panel A. All teachers					
0.79	0.58	0.21	0.20	0.17	0.23
		(0.03)	(0.04)	(0.04)	(0.04)
1,575	1,496	3,071	882	660	1,529
Panel B. Teachers with above median test scores					
0.78	0.63	0.15	0.15	0.15	0.14
		(0.04)	(0.05)	(0.05)	(0.06)
843	702	1,545	423	327	795
Panel C. Teachers with below median test scores					
0.78	0.53	0.24	0.21	0.14	0.32
		(0.04)	(0.05)	(0.06)	(0.06)
625	757	1,382	412	300	670

Notes: Child learning levels were assessed in a mid-test (April 2004) and a post-test (November 2004). After the post-test, the "official" evaluation period was ended. Random checks continued in both the treatment and control schools. Standard errors are clustered by school. Panels B and C only include the 109 schools where teacher tests were available.

Figure 11:

Effect of the nonlinearity of the wage structure

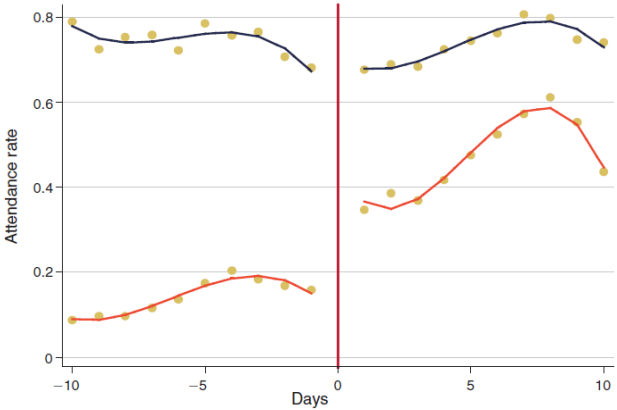


FIGURE 3. RDD REPRESENTATION OF TEACHER ATTENDANCE AT THE START AND END OF THE MONTH

Notes: The top lines represent the months in which the teacher is in the money, while the bottom lines represent the months in which the teacher is not in the money. The estimation includes a third-order polynomial of days on the left and right side of the change of month.

Figure 12:

Let’s begin to think about inference

Parametric vs. non-parametric (Fisherian) tests

- Typically we appeal to CLT in large-sample settings based off of theory about the sampling distribution of sample means (and differences in means). This justifies the t-distribution via asymptotic normality.
- But what if we have small sample sizes? What if we don't like assumptions of normality?

The sharp null

The sharp null implies:

$$\delta_i = \delta \ \forall i$$

The typical frequentist framework testing framework implies that the parameter of interest is equal to some number (often zero, e.g., $\delta = 0$).

In the Fisherian approach, we may want to test something similar:

$$\delta_i = \delta = 0 \ \forall i$$

Steps for testing the sharp null hypothesis

1. Calculate a sample statistic (e.g., $\widehat{\delta_{ATE}}$) using the treatment assignment vector.
2. Consider all possible treatment assignment vectors (using knowledge of the assignment mechanism).
3. For each possible treatment permutation, recalculate the sample statistic as if that had been the true assignment vector. Store each permuted sample statistic.
4. The "exact" p-value is obtained by comparing the sample statistic against the distribution from step 3—i.e., what share of the permuted sample statistics were at least as large as the observed statistic?[1]

[1] Note: if too many permutation vectors exist, you can perform randomization inference by randomly sampling from the set of hypothetical treatment vectors. Then, the p-value is derived by comparing againt the randomization distribution.

Introduction to causal inference for data scientists

RCTs, AB testing, and business experiments (2)

Michael Gill

2018-02-13

Center for Data Science | New York University

Today

Field experiments: methods and issues (2)

Core takeaways today:

- Varieties of experimental designs
- More on permutation/randomization inference
- Neyman confidence intervals
- Bayesian approaches for inference
- Example 1: Resume experiment
- Example 2: GOTV experiment
- Beginning to think about non-compliance

Continuing with randomized experiments

- Business experiment; A/B testing; split testing; bucket testing: same idea with various contexts/names.
- Principle: separate between control/treatment groups and monitor the response.
- Basic recipes:
  - avoid spillovers (geographic/online/across-product substitution)
  - setup a feedback mechanism
  - Often simple is better ($\rightarrow$ avoid comparison across too many groups)
  - beware of unanticipated difficulties (e.g., price discrimination; legal issues; consumer reaction; the potential for unforeseen harms)
  - Anticipate that things will go wrong.

A/B tests and business experiments: methodology

As in IR, there are 4 main types of randomized experiments.

- Bernoulli trials.
- Completely randomized experiments.
- Stratified experiments.
- Paired randomized experiments.

Bernoulli trials

- Recall the conditional probability of being in the treatment group is given by the propensity score $e(x)$, where x is the vector of covariates. In a *Bernoulli trial*, n units are assigned in a i.i.d. manner to the control/treatment groups.
- Letting $\mathbf{D} \in \{0, 1\}^n$, one has in a Bernoulli trial

$$\Pr(\mathbf{D}|X, Y^0, Y^1) = \prod_{i=1}^{n} e(X_i)^{D_i} \left(1 - e(X_i)^{(1-D_i)}\right)$$

- When the treatment probability is uniform $e(x) = q$, one has

$$\Pr(\mathbf{D}|X, Y^0, Y^1) = q^{n_1} (1 - q)^{n_0}$$

where $n_1 = \sum_{i=1}^{n} D_i$ and $n_0 = n - n_0$.

Completely randomized experiments

- In a completely randomized experiment, the size of the treatment group n_1 is fixed, and the n_1 treated individuals are drawn at random without replacement.
- As a result,

$$\Pr\left(\mathbf{D}|X, Y^0, Y^1\right) = \binom{n}{n_1}^{-1} = \left(\frac{n!}{n_1!\,(n-n_1)!}\right)^{-1}$$

if $\sum_{i=1}^{n} D_i = n_1$, $n_1 > 0$, and $\Pr\left(\mathbf{D}|X, Y^0, Y^1\right) = 0$, otherwise.

Stratified experiments

- In stratified experiments, the population is divided into strata or blocks of similar covariates. E.g., male and female; number of children; earnings brackets.
- Within each block $j \in J$, a completely randomized experiment is performed. Let $B_i \in J$ be the index of the block of unit i.
- Each block j has size $n(j) = n_0(j) + n_1(j)$ and

$$\begin{aligned}\Pr(\mathbf{D}|X, Y^0, Y^1) &= \prod_{j \in J} \binom{N(j)}{n_1(j)}^{-1} \text{ if } \sum_{i:B_i=j} D_i = n_1(j) \;\; \forall j \in J \\ &= 0 \text{ else.}\end{aligned}$$

Pairwise experiments

- Pairwise experiments are stratified experiments where the blocks are of size two. In each block one unit (drawn at random) is treated, and the other one is not.
- In this case n is even and $J = \{1, ..., n/2\}$. Each unit has probability $1/2$ of being assigned to the treatment group.
- As a result,

$$\Pr(\mathbf{D}|X, Y^0, Y^1) = 2^{-n/2} \text{ if } \sum_{i:B_i=j} D_i = 1 \; \forall j = 1, ..., n/2$$
$$= 0 \text{ else.}$$

Big picture takeaway

- Details of the assignment mechanism all differently influence $\Pr\left(\mathbf{D}|X, Y^0, Y^1\right)$.
- As a result, different experimental designs have different variance properties and demand different methods for analysis.

Sample uncertainty

Tests for completely randomized experiments

- Assume a fundraising email has been sent to $n = 10,000$ recipients by a political campaign. It comes in two sorts. In the control group (of size $n_0 = 6,000$), no picture is included. In the treatment group (size $n_1 = 4,000$), the picture of the candidate is given. The message asks for a donation of \$500, \$1,000, or \$2,000 dollars.

Tests for completely randomized experiments

- Assume a fundraising email has been sent to $n = 10,000$ recipients by a political campaign. It comes in two sorts. In the control group (of size $n_0 = 6,000$), no picture is included. In the treatment group (size $n_1 = 4,000$), the picture of the candidate is given. The message asks for a donation of \$500, \$1,000, or \$2,000 dollars.
- The question is whether displaying a picture had an effect on donations. The rate of response is as follows:

donation	**control**	**treatment**
\$ 0	2,400	1,200
\$ 500	1800	1,200
\$ 1,000	1200	1,000
\$ 2,000	600	600

Tests for completely randomized experiments

- Fisher's sharp null hypothesis:

$$H_0 : Y_i^0 = Y_i^1 \rightarrow \delta_i = 0 \text{ for all } i = 1, \ldots, N.$$

- In order to test this hypothesis, we need a *test statistic*: a function of D, Y^* and X the distribution of which we can characterize under the null hypothesis. This distribution will tell us "how unlikely" is our observation.

Tests for completely randomized experiments

- Fisher's sharp null hypothesis:

$$H_0 : Y_i^0 = Y_i^1 \rightarrow \delta_i = 0 \text{ for all } i = 1, \ldots, N.$$

- In order to test this hypothesis, we need a *test statistic*: a function of D, Y^* and X the distribution of which we can characterize under the null hypothesis. This distribution will tell us "how unlikely" is our observation.
- A natural test statistic is the absolute value of the estimated ATE, i.e.,

$$T^{diff} = \left| \widehat{\mathbb{E}}\left[Y^* | D = 1\right] - \widehat{\mathbb{E}}\left[Y^* | D = 0\right] \right| = \left| \frac{\sum_{i:D_i=1} Y_i^*}{N_1} - \frac{\sum_{i:D_i=0} Y_i^*}{N_0} \right|.,$$

and here, $\widehat{\mathbb{E}}\left[Y^* | D = 1\right] = 700$, while $\widehat{\mathbb{E}}\left[Y^* | D = 1\right] = 550$.

Tests for completely randomized experiments (continued)

- Another test statistic is the classical t-statistic, given by

$$T^{t-stat} = T^{diff} / \sqrt{s_0^2/n_0 + s_1^2/n_1}$$

where $s_d^2 = \sum_{i:D_i=d} \left(Y_i^* - \widehat{\mathbb{E}}\left[Y^* | D = d\right]\right)^2 / (n_d - 1)$.

Fisher's p-values in our example

- Can we attribute the difference between 550 and 700 to statistical uncertainty alone? in order to do so, one should look at the distribution of T^{diff} under H_0.

Fisher's p-values in our example (continued)

- In order to do this, consider all the draws of $\mathbf{D}$ such that the size of the treatment group is 4,000, and the size of the control group is 6,000. There are $\binom{10,000}{6,000}$ such draws. For each realization $\tilde{\mathbf{D}}$, let us compute the value of T^{diff} that one would get if using $\tilde{\mathbf{D}}$ instead of the actual treatment. Assuming $Y_i^0 = Y_i^1 = Y_i^*$, we would get

$$\tilde{T}^{diff} = \left| \frac{\sum_{i=1}^{10,000} Y_i^* \tilde{D}_i}{\sum_{i=1}^{10,000} \tilde{D}_i} - \frac{\sum_{i=1}^{10,000} Y_i^* \left(1 - \tilde{D}_i\right)}{\sum_{i=1}^{10,000} \left(1 - \tilde{D}_i\right)} \right| .$$

- We could simulate the distribution of $\tilde{T}^{diff}$ and compare with the actual value of T^{diff}. The p-value of the test is given by $\Pr\left(\tilde{T}^{diff} \geq T^{diff}\right)$. If the p-value is higher than the confidence level α, then the observed value of T^{diff} is unusual, and therefore H_0 will be rejected.

Repeated sampling: Neyman estimator

Recall that the ATE is estimated under perfect randomization by the naive estimator, namely:

$$\begin{aligned}\widehat{ATE} &= \frac{\sum_i Y_i^1 D_i}{n_1} - \frac{\sum_i Y_i^0 (1 - D_i)}{n_0} \\ &= \frac{\sum_i Y_i^1 D_i}{\sum_i D_i} - \frac{\sum_i Y_i^0 (1 - D_i)}{\sum_i (1 - D_i)}\end{aligned}$$

where $n = n_0 + n_1$. This is sometimes called the Neyman estimator.

Repeated sampling: Neyman estimator (2)

- Conditional on potential outcomes as given, when if $\tilde{D}$ drawn uniformly from the set $\{0,1\}^N$ such that $\sum_i \tilde{D}_i = n_1$, the variance of the Neyman estimator can be shown to be

$$V_N = \text{var}\left(\widehat{ATE}\right) = \frac{1}{N-1}\left(\frac{N-n_1}{n_1}\sigma_N^2\left(Y^1\right) + \frac{N-(n-n_1)}{n-n_1}\sigma_N^2(Y^0) + 2\sigma_N\left(Y^1, Y^0\right)\right)$$

  where $N \geq 4$ is the size of the unobserved population, n is the size of the experimental sample, $n_1 \geq 2$ is the size of those randomly assigned to treatment, and $n - n_1 \geq 2$ units are randomly assinged to control. Here, $\sigma^2\left(Y^d\right) = \text{var}\left(Y^d\right)$.
- Neyman showed that as $N \to \infty$, with n and n_1 fixed, $\text{var}\left(\widehat{ATE}\right) \to \frac{1}{n_1}\sigma_N^2(Y^1) + \frac{1}{n-n_1}\sigma_N^2(Y^0)$.

Repeated sampling: Neyman estimator (3)

When $n = N$, the sampling variation of the Neyman estimator reduces to:

$$V_n = \text{var}\left(\widehat{ATE}_n\right) = \frac{1}{n-1}\left(\frac{n-n_1}{n_1}\sigma_n^2(Y^1) + \frac{n_1}{n-n_1}\sigma_n^2(Y^0) + 2\sigma_n(Y^1, Y^0)\right)$$

Neyman proposed a conservative estimator of the variance given

$$\hat{V}_n^a = \frac{n}{n-1}\left(\frac{\hat{\sigma}_n^2(Y^1)}{n_1} + \frac{\hat{\sigma}_n^2(Y^0)}{n-n_1}\right)$$

Repeated sampling: bounds on the variance

- In the previous expression, $\sigma_n^2(Y^d)$ can be easily estimated empirically for $d \in \{0, 1\}$ by

$$s_d^2 = \frac{1}{n_d - 1} \sum_{i:D_i=d} \left(Y_i^* - \bar{Y}_d^*\right)^2$$

- However, $\sigma_n^2(Y^1, Y^0)$ cannot be estimated empirically because one never simultaneously observe Y_i^1 and Y_i^0. Note that when the treatment effect $Y_i^1 - Y_i^0$ is constant, then $\sigma_n^2(Y^1, Y^0)$. This is an upper bound on the variance, and therefore is useful for computing "conservative" confidence intervals. Therefore

$$\text{var}\left(\widehat{ATE}\right) \approx \hat{V}^{Neyman} = \frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}$$

Neyman confidence intervals

Why is this conservative?[1] One can show:

$$\mathbb{E}\left(\hat{V}^{Neyman} - V_n\right) = \frac{1}{n-1}\left[\sigma_n^2(Y^1) + \sigma_n^2(Y^0) - 2\sigma_n(Y^1, Y^0)\right] \geq 0$$

Hence, a conservative 90% confidence interval for the ATE is

$$\left[\widehat{ATE} - 1.645\sqrt{\hat{V}^{Neyman}}, \widehat{ATE} + 1.645\sqrt{\hat{V}^{Neyman}}\right].$$

[1] See Aronow et al. (2014) for more detail on this derivation.

Linear regression and the Neyman estimator (1)

In the setting of completely a randomized experiment, recall the linear regression specification

$$Y_i^* = \alpha + \tau D_i + \beta' X_i + \varepsilon_i$$

and consider the sample OLS estimates $\left(\hat{\alpha}, \hat{\tau}, \hat{\beta}\right)$, given by

$$\min_{\alpha,\tau,\beta} \widehat{\mathbb{E}}\left[\left(Y_i^* - \alpha - \tau D_i - \beta' X_i\right)^2\right],$$

with $(\alpha^*, \tau^*, \beta^*)$ their population analog, where the sample average $\widehat{\mathbb{E}}$ is replaced by the population expectation $\mathbb{E}$ (i.e. $N \to +\infty$).

Linear regression and the Neyman estimator (2)

- Then $\tau^* = ATE = \mathbb{E}\left[Y_i^1 - Y_i^0\right]$, and

$$\sqrt{N}\left(\hat{\tau} - \tau^*\right) \Rightarrow N\left(0, V^{OLS}\right)$$

where

$$V^{OLS} = \frac{\mathbb{E}\left[\left(D_i - \mathbb{E}\left[D_i\right]\right)^2 \left(Y_i^* - \alpha^* - \tau^* D_i - \beta^{*\prime} X_i\right)^2\right]}{\mathbb{E}\left[D_i\right]^2 \mathbb{E}\left[1 - D_i\right]^2}.$$

- Note that we recover $\hat{V}^{Neyman}$ via OLS when there are no covariates.

How “conservative” is it?

- Short answer: can be very
- next HW will have you inspect this fact further

Imputation to deal with inference

- Consider[2] the following 6 observations, taken from the LaLonde dataset:

i	Y_i^0	Y_i^1	D_i	Y_i^*
1	0	?	0	0
2	?	9.9	1	9.9
3	12.4	?	0	12.4
4	?	3.6	1	3.6
5	0	?	0	0
6	?	24.9	1	24.9

- The naive estimator yields $\delta_{naive} = 8.67$.

[2] Cf IR, Ch. 8.4.

Model-based imputation: Bayesian approach

- Assume that we have a model for the potential outcomes. If θ is the parameter, assumed to be drawn from a prior $p(\theta)$, $(\mathbf{Y}^0, \mathbf{Y}^1)$ is the vector of potential outcomes, and if $\mathbf{D}$ is the vector of treatments, and $f(\mathbf{Y}^0, \mathbf{Y}^1|\theta)$ is assumed to be known.
- Further, because we are under complete randomization, one has

$$f(\mathbf{Y}^0, \mathbf{Y}^1, \mathbf{D}|\theta) = f(\mathbf{Y}^0, \mathbf{Y}^1|\theta) f(\mathbf{D}),$$

  which will be our basis for doing Bayesian inference. Note that in observational studies, this independence would not hold.
- We'll denote $\mathbf{Y}^-$ for missing data. Note that $Y_i^- = Y_i^{1-D_i} \; \forall \; i$.

Bayesian inference for ATE

- Step 1: Compute $f(\mathbf{Y}^{-}|\mathbf{Y}^{*}, \mathbf{D}, \theta)$.
- Step 2: Compute $f(\theta|\mathbf{Y}^{*}, \mathbf{D})$.
- Step 3: Compute $f(\mathbf{Y}^{-}|\mathbf{Y}^{*}, \mathbf{D})$.
- Step 4: If the parameter of interest is $\tau = \tau(\mathbf{Y}^{0}, \mathbf{Y}^{1}, \mathbf{D}) = \tau(\mathbf{Y}^{-}, \mathbf{Y}^{*}, \mathbf{D})$, then its distribution can be inferred from the distribution $f(\mathbf{Y}^{-}|\mathbf{Y}^{*}, \mathbf{D})$.

Inference in our example

- Assume that $(Y^0, Y^1) \sim \mathcal{N}\left((\mu^0, \mu^1), \begin{pmatrix} 100 & 0 \\ 0 & 64 \end{pmatrix}\right)$ where $\theta = (\mu^0, \mu^1)$.
- The prior distribution for θ is $\mathcal{N}\left((0, 0), \begin{pmatrix} 10,000 & 0 \\ 0 & 10,000 \end{pmatrix}\right)$.
- The assignment mechanism is

$$\Pr(\mathbf{D} = \mathbf{d} | \mathbf{Y}, \theta) = \binom{N}{N_1}^{-1} 1\left\{\sum_{i=1}^{N} d_i = N_1\right\}$$

An example: step 1

- Compute $f(\mathbf{Y}^{-}|\mathbf{Y}^{*}, \mathbf{D}, \theta)$.
- Letting Y^{-} be the missing data, we have

$$\begin{pmatrix} Y_1^- \\ Y_2^- \\ Y_3^- \\ Y_4^- \\ Y_5^- \\ Y_6^- \end{pmatrix} |\mathbf{Y}^{*}, \mathbf{D}, \mu^0, \mu^1 \sim \mathcal{N}\left(\begin{pmatrix} \mu^1 \\ \mu^0 \\ \mu^1 \\ \mu^0 \\ \mu^1 \\ \mu^0 \end{pmatrix}, \begin{pmatrix} 64 & & & & & \\ & 100 & & & 0 & \\ & & 64 & & & \\ & & & 100 & & \\ & 0 & & & 64 & \\ & & & & & 100 \end{pmatrix} \right)$$

An example: step 2

- Next, we can compute $f(\theta|\mathbf{Y}^*, \mathbf{D})$.
- Here,

$$\begin{pmatrix} \mu^0 \\ \mu^1 \end{pmatrix} |\mathbf{Y}^*, \mathbf{D} \sim \mathcal{N}\left(\mathbb{E}\left[\mu|\mathbf{Y}^*, \mathbf{D}\right], \mathbb{V}\left[\mu|\mathbf{Y}^*, \mathbf{D}\right]\right)$$

where

$$\mathbb{E}\left[\mu|\mathbf{Y}^*, \mathbf{D}\right] = \begin{pmatrix} \widehat{\mathbb{E}}\left[Y^*|D=0\right] \frac{N_0.10{,}000}{N_0.10{,}000+100} \\ \widehat{\mathbb{E}}\left[Y^*|D=1\right] \frac{N_1.10{,}000}{N_1.10{,}000+100} \end{pmatrix} \text{ and}$$

$$\mathbb{V}\left[\mu|\mathbf{Y}^*, \mathbf{D}\right] = \begin{pmatrix} \frac{1}{N_0/100+1/10{,}000} & 0 \\ 0 & \frac{1}{N_1/64+1/10{,}000} \end{pmatrix}$$

- Hence

$$\begin{pmatrix} \mu^0 \\ \mu^1 \end{pmatrix} |\mathbf{Y}^*, \mathbf{D} \sim \mathcal{N}\left(\begin{pmatrix} 4.1 \\ 12.8 \end{pmatrix}, \begin{pmatrix} 5.8^2 & 0 \\ 0 & 4.6^2 \end{pmatrix}\right).$$

An example: step 3

- By combining $f(\theta|\mathbf{Y}^*, \mathbf{D})$ and $f(\mathbf{Y}^-|\mathbf{Y}^*, \mathbf{D}, \theta)$, we find $f(\mathbf{Y}^-|\mathbf{Y}^*, \mathbf{D})$. Here:

$$\begin{pmatrix} Y_1^- \\ Y_2^- \\ Y_3^- \\ Y_4^- \\ Y_5^- \\ Y_6^- \end{pmatrix} |\mathbf{Y}^*, \mathbf{D} \sim \mathcal{N}\left(\begin{pmatrix} 12.8 \\ 4.1 \\ 12.8 \\ 4.1 \\ 12.8 \\ 4.1 \end{pmatrix}, \begin{pmatrix} 85.3 & 0 & 21.3 & 0 & 21.3 & 0 \\ 0 & 133.2 & 0 & 33.2 & 0 & 33.2 \\ 21.3 & 0 & 85.3 & 0 & 21.3 & 0 \\ 0 & 33.2 & 0 & 133.2 & 0 & 33.2 \\ 21.3 & 0 & 21.3 & 0 & 85.3 & 0 \\ 0 & 33.2 & 0 & 33.2 & 0 & 133.2 \end{pmatrix} \right)$$

An example: step 4

- The treatment effect is measured by

$$\frac{1}{N}\sum_{i=1}^{N}\left(Y_i^1 - Y_i^0\right) = \frac{1}{N}\sum_{i=1}^{N}\left(1-2D_i\right)Y_i^- + \frac{1}{N}\sum_{i=1}^{N}\left(2D_i - 1\right)Y_i^*,$$

  whose distribution can be provided using $f\left(\mathbf{Y}^-|\mathbf{Y}^*, \mathbf{D}\right)$.

- One has

$$\frac{1}{N}\sum_{i=1}^{N}\left(Y_i^1 - Y_i^0\right) \sim \mathcal{N}\left(8.7, 5.2^2\right).$$

Examples from real experiments

Case 1: Labor Market Discrimination

- Taken from Bertrand M. and Mullainathan S. (2004). "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination" *American Economic Review*.

Case 1: Labor Market Discrimination

- Taken from Bertrand M. and Mullainathan S. (2004). "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination" *American Economic Review*.
- Abstract: "We study race in the labor market by sending fictitious resumes to help-wanted ads in Boston and Chicago newspapers. To manipulate perceived race, resumes are randomly assigned African-American- or White-sounding names. White names receive 50 percent more callbacks for interviews. Callbacks are also more responsive to resume quality for White names than for African-American ones. The racial gap is uniform across occupation, industry, and employer size. We also find little evidence that employers are inferring social class from the names. Differential treatment by race still appears to still be prominent in the U.S. labor market."

Experimental design: resume experiment

- Field experiment: CV sent in response to help-wanted ads in Boston and Chicago newspapers. Sample size is 1,300.
- Randomize over race perception: White-sounding names (e.g., Emily Walsh or Greg Baker) are assigned randomly to half of the sample, and African-American-sounding names (e.g., Lakisha Washington or Jamal Jones) to the other half.
- Resume quality is also experimentally varied in order to understand how race perception affects the effects of application's other characteristics.

Brief aside

There is a literature about estimating the causal effects of seemingly "immutable characteristics."[3]

Why is is race an interesting causal variable?

[3] See, for example, Greiner and Rubin (2011) and Sen and Wasow (2016) for more on these subjects.

Results of resume experiment

- Applicants with White names: 10 CVs for a callback. With African-American names: 15CVs for a callback.
- The effect of a higher-quality resume on callback rate is smaller for African-American names. According to the authors: “While one may have expected improved credentials to alleviate employers’ fear that African-American applicants are deficient in some unobservable skills, this is not the case in our data.”

Callbacks rates by racial soundingness of names

TABLE 1—MEAN CALLBACK RATES BY RACIAL SOUNDINGNESS OF NAMES

	Percent callback for White names	Percent callback for African-American names	Ratio	Percent difference (p-value)
Sample:				
All sent resumes	9.65	6.45	1.50	3.20
	[2,435]	[2,435]		(0.0000)
Chicago	8.06	5.40	1.49	2.66
	[1,352]	[1,352]		(0.0057)
Boston	11.63	7.76	1.50	4.05
	[1,083]	[1,083]		(0.0023)
Females	9.89	6.63	1.49	3.26
	[1,860]	[1,886]		(0.0003)
Females in administrative jobs	10.46	6.55	1.60	3.91
	[1,358]	[1,359]		(0.0003)
Females in sales jobs	8.37	6.83	1.22	1.54
	[502]	[527]		(0.3523)
Males	8.87	5.83	1.52	3.04
	[575]	[549]		(0.0513)

Notes: The table reports, for the entire sample and different subsamples of sent resumes, the callback rates for applicants with a White-sounding name (column 1) an an African-American-sounding name (column 2), as well as the ratio (column 3) and difference (column 4) of these callback rates. In brackets in each cell is the number of resumes sent in that cell. Column 4 also reports the p-value for a test of proportion testing the null hypothesis that the callback rates are equal across racial groups.

Figure 1:

Callbacks rates by racial soundingness and resume quality

Table 4—Average Callback Rates By Racial Soundingness of Names and Resume Quality

	Panel A: Subjective Measure of Quality (Percent Callback)			
	Low	High	Ratio	Difference (p-value)
White names	8.50	10.79	1.27	2.29
	[1,212]	[1,223]		(0.0557)
African-American names	6.19	6.70	1.08	0.51
	[1,212]	[1,223]		(0.6084)
	Panel B: Predicted Measure of Quality (Percent Callback)			
	Low	High	Ratio	Difference (p- value)
White names	7.18	13.60	1.89	6.42
	[822]	[816]		(0.0000)
African-American names	5.37	8.60	1.60	3.23
	[819]	[814]		(0.0104)

Notes: Panel A reports the mean callback percents for applicant with a White name (row 1) and African-American name (row 2) depending on whether the resume was subjectively qualified as a lower quality or higher quality. In brackets is the number of resumes sent for each race/quality group. The last column reports the p-value of a test of proportion testing the null hypothesis that the callback rates are equal across quality groups within each racial group. For Panel B, we use a third of the sample to estimate a probit regression of the callback dummy on the set of resume characteristics as displayed in Table 3. We further control for a sex dummy, a city dummy, six occupation dummies, and a vector of dummy variables for job requirements as listed in the employment ad (see Section III, subsection D, for details). We then use the estimated coefficients on the set of resume characteristics to estimate a predicted callback for the remaining resumes (two-thirds of the sample). We call "high-quality" resumes the resumes that rank above the median predicted callback and "low-quality" resumes the resumes that rank below the median predicted callback. In brackets is the number of resumes sent for each race/quality group. The last column reports the p-value of a test of proportion testing the null hypothesis that the callback percents are equal across quality groups within each racial group.

Figure 2:

Effect of resume characteristics on likelihood of callback

Dependent Variable: Callback Dummy Sample:	All resumes	White names	African-American names
Years of experience (*10)	0.07	0.13	0.02
	(0.03)	(0.04)	(0.03)
Years of experience^2 (*100)	−0.02	−0.04	−0.00
	(0.01)	(0.01)	(0.01)
Volunteering? (Y = 1)	−0.01	−0.01	0.01
	(0.01)	(0.01)	(0.01)
Military experience? (Y = 1)	−0.00	0.02	−0.01
	(0.01)	(0.03)	(0.02)
E-mail? (Y = 1)	0.02	0.03	−0.00
	(0.01)	(0.01)	(0.01)
Employment holes? (Y = 1)	0.02	0.03	0.01
	(0.01)	(0.02)	(0.01)
Work in school? (Y = 1)	0.01	0.02	−0.00
	(0.01)	(0.01)	(0.01)
Honors? (Y = 1)	0.05	0.06	0.03
	(0.02)	(0.03)	(0.02)
Computer skills? (Y = 1)	−0.02	−0.04	−0.00
	(0.01)	(0.02)	(0.01)
Special skills? (Y = 1)	0.05	0.06	0.04
	(0.01)	(0.02)	(0.01)
Ho: Resume characteristics effects are all zero (p-value)	54.50 (0.0000)	57.59 (0.0000)	23.85 (0.0080)
Standard deviation of predicted callback	0.047	0.062	0.037
Sample size	4,870	2,435	2,435

Figure 3:

Case 2: GOTV experiment

- Taken from Gerber and Green (2000). “The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment”, *American Political Science Review*.

Case 2: GOTV experiment

- Taken from Gerber and Green (2000). “The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment”, *American Political Science Review*.
- Abstract: We report the results of a randomized fieldexperiment involving approximately 30,000 registered voters in New Haven, Connecticut. Nonpartisan get-out-the-vote messages were conveyed through personal canvassing, direct mail, and telephone calls shortly before the November 1998 election. A variety of substantive messages were used. Voter turnout was increased substantially by personal canvassing, slightly by direct mail, and not at all by telephone calls. These findings support our hypothesis that the long-term retrenchment in voter turnout is partly attributable to the decline in face-to-face political mobilization.

Experimental design: GOTV experiment

- $n \approx 30,000$ individuals in New Haven, CT.
- Randomly assign households into different treatments: mail, individual contact, phone call.
- Some individuals were assigned to multiple treatments simultaneously. We will ignore these for now, but this can change the analysis. E.g., Blackwell (2017).
- Within each treatment arm, individuals assigned different messages (e.g., civic duty, neighborhood solidarity, close election), and different dosages.

Considerations

- For phone calls, we know whether a phone was answered.

Considerations

- For phone calls, we know whether a phone was answered.
- For in-person canvassing, we know whether a door was answered.

Considerations

- For phone calls, we know whether a phone was answered.
- For in-person canvassing, we know whether a door was answered.
- Those that were assigned mailers have no measures of compliance (i.e., we don't know if they did/didn't read).

Considerations

- For phone calls, we know whether a phone was answered.
- For in-person canvassing, we know whether a door was answered.
- Those that were assigned mailers have no measures of compliance (i.e., we don't know if they did/didn't read).

Moreover, do we think compliance with respect to treatment is equal across treatment arms?

Intent-to-treat (ITT) effects

- We do, however, observe outcomes for each individual given his/her treatment assignment.
- ITT: estimated differences in outcomes for those assigned to treatment versus assigned to control.
- In the presence of non-compliance, ITT is a biased estimator of the ATE.

GOTV (continued)

We will continue with this example in the next class, as we discuss intrumental variables in the context of a randomized experiment.

Introduction to causal inference for data scientists

Treatment non-compliance and instrumental variables (1)

Michael Gill

2018-02-19

Center for Data Science | New York University

Today

Today's references

References:

- Angrist, J. D., Imbens, G.W., and Rubin, D. B. (1996). "Identification of causal effects using instrumental variables." Journal of the American Statistical Association, 91(434): 444–455.
- IR, Chapters 23, 24, 25
- MW, Chapter 9

Instrumental variables (1)

Core takeaways today:

- Going from ITT to different treatment effects
- How an experiment doesn't always (straightforwardly) give ATE
- Thinking about bounds of treatment effects in these settings
- Missing data in the context of non-compliance

Today's framework

Unit	Z_i	D_i	Y_i^1	Y_i^0
1	0	0	?	✓
2	0	1	✓	?
⋮		⋮	⋮	⋮
n	1	1	✓	?

Today's framework

Unit	Z_i	D_i	Y_i^1	Y_i^0
1	0	0	?	✓
2	0	1	✓	?
⋮		⋮	⋮	⋮
n	1	1	✓	?

- In last class, we motivated a field experiment in which units were randomly assigned treatment in a GOTV experiment.
- Causal effect: compare treatment and control potential outcomes.
- Units in experiments often comply with assignment, but not always
- Human units make compliance with assignment more difficult.

Today's framework

Unit	Z_i	D_i	Y_i^1	Y_i^0
1	0	0	?	✓
2	0	1	✓	?
⋮		⋮	⋮	⋮
n	1	1	✓	?

- In last class, we motivated a field experiment in which units were randomly assigned treatment in a GOTV experiment.
- Causal effect: compare treatment and control potential outcomes.
- Units in experiments often comply with assignment, but not always
- Human units make compliance with assignment more difficult.

Goal: Use Instrumental Variables under potential outcomes framework to identify causal effects.

Notation for compliance

We formalize notation to discuss proper method of analysis.

- Z_i: treatment assignment indicator for unit i.
- $D_i(Z_i)$: treatment received indicator for unit i.
- $\mathbf{Z}, \mathbf{D}$: N-dimensional binary vectors of treatment assignments and treatments received

Notation for compliance

We formalize notation to discuss proper method of analysis.

- Z_i: treatment assignment indicator for unit i.
- $D_i(Z_i)$: treatment received indicator for unit i.
- $\mathbf{Z}, \mathbf{D}$: N-dimensional binary vectors of treatment assignments and treatments received

We can define individuals by their (often unobservable) compliance behavior:

- Complier: $D_i(0) = 0,\ D_i(1) = 1$
- Always-taker: $D_i(0) = D_i(1) = 1$
- Never-taker: $D_i(0) = D_i(1) = 0$
- Defier: $D_i(0) = 1,\ D_i(1) = 0$

Notation (continued)

- Z_i: treatment assignment indicator for unit i.
- $D_i(Z_i)$: treatment received indicator for unit i.
- $D_i(\mathbf{Z}), Y_i(\mathbf{Z}, \mathbf{D})$: potential outcomes for unit i under treatment assignment Z_i and treatment received. Later assumptions will allow us to write this in the form $Y_i(Z_i, D_i(Z_i))$.

What should be our estimand of interest?

Causal effects of Z on D and Z on Y

For unit i, the causal effect of

- Z on D is: $D_i(1) - D_i(0)$

Causal effects of Z on D and Z on Y

For unit i, the causal effect of

- Z on D is: $D_i(1) - D_i(0)$
- Z on Y is: $Y_i(Z_i = 1, D_i(1)) - Y_i(Z_i = 0, D_i(0))$

Unit-Level Causal Effects

$$Y_i(1, D_i(1)) - Y_i(0, D_i(0)) = \begin{cases} Y_i(1,1) - Y_i(0,0) & \text{for compliers} \\ Y_i(1,1) - Y_i(0,1) & \text{for always-takers} \\ Y_i(1,0) - Y_i(0,0) & \text{for never-takers} \\ Y_i(1,0) - Y_i(0,1) & \text{for defiers} \end{cases}$$

Unit-Level Causal Effects

$$Y_i(1, D_i(1)) - Y_i(0, D_i(0)) = \begin{cases} Y_i(1,1) - Y_i(0,0) & \text{for compliers} \\ Y_i(1,1) - Y_i(0,1) & \text{for always-takers} \\ Y_i(1,0) - Y_i(0,0) & \text{for never-takers} \\ Y_i(1,0) - Y_i(0,1) & \text{for defiers} \end{cases}$$

- Always-takers and never-takers have only one treatment received, regardless of assignment.
- Compliers and defiers take different treatments for different assignments.
- Hence, to gauge causal effect of treatment received, we look at compliers and defiers.

An example: Effect of military service on civilian mortality

Taken from Angrist (1990).

- To illustrate assumptions, we use the running example of the effect of military service (Vietnam War) on civilian mortality.

An example: Effect of military service on civilian mortality

Taken from Angrist (1990).

- To illustrate assumptions, we use the running example of the effect of military service (Vietnam War) on civilian mortality.
- Man with low draft lottery number ($Z_i = 1$) will either serve ($D_i = 1$) or not in military.
- In world with perfect compliance, $D_i(\mathbf{Z}) = Z_i$ for all i.

An example: Effect of military service on civilian mortality

Taken from Angrist (1990).

- To illustrate assumptions, we use the running example of the effect of military service (Vietnam War) on civilian mortality.
- Man with low draft lottery number ($Z_i = 1$) will either serve ($D_i = 1$) or not in military.
- In world with perfect compliance, $D_i(\mathbf{Z}) = Z_i$ for all i.
- One example of non-compliance: getting a low draft lottery number ($Z_i = 1$) but not serving in the military ($D_i = 0$).
- Potential outcome $Y_i(z, d)$ equals 1 if person i would have died between 1974-1983 given lottery assignment z and military service indicator d.

Assumptions

Several assumptions are needed for causal inference.

1. SUTVA
2. Random Assignment (of Z)
3. Exclusion Restriction
4. Nonzero Average Causal Effect of Z on D
5. Monotonicity

Assumption 1: SUTVA

No interference among units, well-defined potential outcomes.

1. If $Z_i = Z_i'$, then $D_i(Z) = D_i(Z')$.
2. If $Z_i = Z_i'$ and $D_i = D_i'$, then $Y_i(Z, D) = Y_i(Z', D')$.

Potential outcomes for unit i not related to others' treatments.

SUTVA (continued)

- SUTVA allows us to define the causal effects in standard fashion.
- Causal effect for individual i of Z on D is $D_i(1) - D_i(0)$.
- The causal effect of Z on Y is $Y_i(1, D_i(1)) - Y_i(0, D_i(0))$.

SUTVA (example)

- Veteran status (i.e., treatment received) is not affected by others' draft numbers.
- Mortality rate is not affected by draft status of others.

Can you think of potential violations?

SUTVA (example)

- Veteran status (i.e., treatment received) is not affected by others' draft numbers.
- Mortality rate is not affected by draft status of others.

Can you think of potential violations?

Example: Undrafted people induced to serve in army by drafted friends (or vice versa).

Assumption 2: Random assignment

Assignment is random and does not depend on potential outcomes.

$$\Pr(Z = c) = \Pr(Z = c')$$

for all c and c' such that $\iota^T c = \iota^T c'$

Random assignment: estimate for ITT

Given SUTVA and random assignment, we can estimate the average causal effect of Z on Y.

$$\frac{\sum_i Y_i Z_i}{\sum_i Z_i} - \frac{\sum_i Y_i (1 - Z_i)}{\sum_i (1 - Z_i)}$$

And, the average causal effect of Z on D.

$$\frac{\sum_i D_i Z_i}{\sum_i Z_i} - \frac{\sum_i D_i (1 - Z_i)}{\sum_i (1 - Z_i)}$$

The ratio of these two yields the conventional IV estimator.

$$\widehat{\beta}_{IV} = \left(\frac{\sum_i Y_i Z_i}{\sum_i Z_i} - \frac{\sum_i Y_i (1 - Z_i)}{\sum_i (1 - Z_i)} \right) \Big/ \left(\frac{\sum_i D_i Z_i}{\sum_i Z_i} - \frac{\sum_i D_i (1 - Z_i)}{\sum_i (1 - Z_i)} \right) = \frac{\widehat{\text{cov}}(Y_i, Z_i)}{\widehat{\text{cov}}(D_i, Z_i)}$$

Random assignment: example

Assignment of draft status was based on birth dates (Angrist, Imbens, Rubin, 1996).

- Random assignment probably not violated.
- How could assignment not be random?

Random assignment: example

Assignment of draft status was based on birth dates (Angrist, Imbens, Rubin, 1996).

- Random assignment probably not violated.
- How could assignment not be random?

Example: Draft number depends on (future) health of draftee, as determined by knowledgeable official (e.g., Perfect Doctor).

Assumption 3: Exclusion restriction

Treatment assignment is unrelated to potential outcomes once treatment received is taken into account.

- $Y(Z, D) = Y(Z', D)$ for all Z, Z' and D
- Put another way: instrument only influences outcome through D.

This implies: Potential outcomes for always-takers and never-takers remain the same, regardless of treatment assignment—i.e., $Y_i(1, d) = Y_i(0, d)$ for $d = 0, 1$.

Exclusion restriction: implications

This assumption allows us to define potential outcomes as a function of D only.

$$Y(D) = Y(Z, D) = Y(Z', D) \text{ for all } Z, Z' \text{ and } D$$

We can now define the causal effect of interest—the effect of D on Y for person i is $Y_i(1) - Y_i(0)$.

Exclusion restriction: example

Civilian mortality risk is not affected by draft status once veteran status taken into account, i.e., draft number influences civilian mortality risk only through serving in army/becoming a veteran.

Can you think of violations?

Exclusion restriction: example

Civilian mortality risk is not affected by draft status once veteran status taken into account, i.e., draft number influences civilian mortality risk only through serving in army/becoming a veteran.

Can you think of violations?

Men with low draft numbers might alter their educational plans to get a deferment; in turn, this would likely influence a range of outcomes including civilian mortality risk.

Assumption 4: Nonzero Average Causal Effect of Z on D

At least one person needs to be influenced by instrument.

$$\mathbb{E}[D_i(1) - D_i(0)] \neq 0$$

Requires that we have some compliers or defiers, not just always-takers and never-takers.

Nonzero Average Causal Effect of Z on D: example

Low lottery number increases average probability of service.

Very reasonable assumption: Of men born in 1950, those with low lottery numbers had a 16% higher probability of serving in the military than those with high lottery numbers.

Assumption 5: Monotonicity

No defiers.

$$D_i(1) \geq D_i(0) \text{ for all } i = 1, \ldots, N$$

Assumptions 4 and 5 imply $D_i(1) \geq D_i(0)$ with strict inequality for at least one i.

This condition is referred to as strict (or strong) monotonicity.

Monotonicity: example

There is no one who would have served if given a high lottery number, but not if given a low lottery number.

Potential violations?

Monotonicity: example

There is no one who would have served if given a high lottery number, but not if given a low lottery number.

Potential violations?

Someone would have volunteered for a particular branch of the armed services with a high draft number, but avoids service if compelled to go into the army by a low draft number... AND, the effect of service on health outcome is different from the effect for compliers.

Air Force versus Navy.

Formal definition of an instrument in RCM

A variable Z is an instrumental variable for the causal effect of D on Y if Assumptions 1-5 all hold.

1. SUTVA
2. Random Assignment (of Z)
3. Exclusion Restriction
4. Nonzero Average Causal Effect of Z on D
5. Monotonicity

Interpreting the IV Estimand

Using only SUTVA and exclusion restriction, we have relationship between intention-to-treat effects of Z on Y and D and the causal effect of D on Y at the unit level.

$$
\begin{aligned}
Y_i(1, D_i(1)) - Y_i(0, D(0)) &= Y_i(D_i(1)) - Y_i(D_i(0)) \\
&= [Y_i(1) \cdot D_i(1) + Y_i(0) \cdot (1 - D_i(1))] \\
&\quad - [Y_i(1) \cdot D_i(0) + Y_i(0) \cdot (1 - D_i(0))] \\
&= (Y_i(1) - Y_i(0)) \cdot (D_i(1) - D_i(0))
\end{aligned}
$$

Product of causal effect of D on Y and causal effect of Z on D.

Interpreting the IV Estimand (continued)

Causal effects for subpopulations with $D_i(0) \neq D_i(1)$ (i.e., compliers and defiers).

$$
\begin{aligned}
&\mathbb{E}[Y_i(1, D_i(1)) - Y_i(0, D(0))] \\
&= \mathbb{E}[(Y_i(1) - Y_i(0))(D_i(1) - D_i(0))] \\
&= \mathbb{E}[Y_i(1) - Y_i(0) | D_i(1) - D_i(0) = 1] \cdot P[D_i(1) - D_i(0) = 1] \\
&\quad - \mathbb{E}[Y_i(1) - Y_i(0) | D_i(1) - D_i(0) = -1] \cdot P[D_i(1) - D_i(0) = -1]
\end{aligned}
$$

Weights sum to $P[D_i(0) \neq D_i(1)]$.

Interpreting the IV Estimand (continued)

Apply the Monotonicity assumption, which requires $D_i(1) \geq D_i(0)$. This removes defiers (i.e., people for whom $D_i(1) - D_i(0) = -1$).

Average causal effect of Z on Y equals product of the average causal effect of D on Y for compliers ($D_i(0) = 0$, $D_i(1) = 1$).

$$
\begin{aligned}
&\mathbb{E}[Y_i(1, D_i(1)) - Y_i(0, D(0))] \\
&= \mathbb{E}[(Y_i(1) - Y_i(0))(D_i(1) - D_i(0))] \\
&= \mathbb{E}[Y_i(1) - Y_i(0) | D_i(1) - D_i(0) = 1] \cdot P[D_i(1) - D_i(0) = 1]
\end{aligned}
$$

Interpreting the IV Estimand (LATE)

Given Assumptions 1-5, the IV estimand is

$$\mathbb{E}[Y_i(1) - Y_i(0)|D_i(1) - D_i(0) = 1] \;=\; \frac{\mathbb{E}[Y_i(1, D_i(1)) - Y_i(0, D_i(0))]}{\mathbb{E}[D_i(1) - D_i(0)]}$$

This follows because monotonicity implies

$$\mathbb{E}[D_i(1) - D_i(0)] \;=\; P[D_i(1) - D_i(0) = 1] \;>\; 0$$

i.e., the proportion of the population who are compliers.

Vitamin A Example: Sommer-Zeger (1991)

RCT of the impact of vitamin A supplements on children's survival.

Villages in Indonesia were assigned to receive vitamin supplements. Deaths were ascertained 12 months following baseline.

Assignment Z_i^{obs}	Supplement D_i^{obs}	Survival Y_i^{obs}	# Units	Type
0	0	0	74	?
0	0	1	11,514	?
1	0	0	34	?
1	0	1	2,385	?
1	1	0	12	?
1	1	1	9,663	?

Vitamin A Example: Sommer-Zeger (1991)

RCT of the impact of vitamin A supplements on children's survival.

Villages in Indonesia were assigned to receive vitamin supplements. Deaths were ascertained 12 months following baseline.

Assignment Z_i^{obs}	Supplement D_i^{obs}	Survival Y_i^{obs}	# Units	Type
0	0	0	74	
0	0	1	$11,514$	
1	0	0	34	N
1	0	1	$2,385$	N
1	1	0	12	C
1	1	1	$9,663$	C

Vitamin A Example: Sommer-Zeger (1991)

RCT of the impact of vitamin A supplements on children's survival.

Villages in Indonesia were assigned to receive vitamin supplements. Deaths were ascertained 12 months following baseline.

Assignment Z_i^{obs}	Supplement D_i^{obs}	Survival Y_i^{obs}	# Units	Type
0	0	0	74	C/N
0	0	1	$11,514$	C/N
1	0	0	34	N
1	0	1	$2,385$	N
1	1	0	12	C
1	1	1	$9,663$	C

Vitamin A Example

Assignment Z_i^{obs}	Supplement D_i^{obs}	Survival Y_i^{obs}	# Units	Type
0	0	0	74	C/N
0	0	1	$11,514$	C/N
1	0	0	34	N
1	0	1	$2,385$	N
1	1	0	12	C
1	1	1	$9,663$	C

Noncompliance: $2,419$ children lived in villages assigned vitamin A supplements, but refused to take them.

How do we proceed with analysis to obtain valid causal inferences?

Intention-to-Treat Analysis

Intention-to-treat analysis: outcomes are compared for units with different assigned treatments.

$$
\begin{aligned}
ITT_Y &= \frac{1}{N}\sum_{i=1}^{N}[Y_i(1, D_i(1)) - Y_i(0, D_i(0))] \\
&= \frac{1}{N}(N_a ITT_Y^a + N_n ITT_Y^n + N_c ITT_Y^c + N_d ITT_Y^d)
\end{aligned}
$$

Without further assumptions, ITT_Y does not summarize the causal effect of treatment received.

The analogue for the observed data summarizes causal effect of treatment assigned, not treatment received.

Intention-to-Treat Analysis (continued)

$$ITT_Y = \frac{1}{N}\left(N_a ITT_Y^a + N_n ITT_Y^n + N_c ITT_Y^c + N_d ITT_Y^d\right)$$

Under exclusion and monotonicity assumptions, we have

$$ITT_Y^c = \frac{ITT_Y}{N_c/N}$$

The conventional instrumental variables estimator is in Imbens and Rubin (1997):

$$\hat{\beta}_{IV} = \frac{\bar{y}_{z=1} - \bar{y}_{z=0}}{\bar{d}_{z=1} - \bar{d}_{z=0}}$$

Under strict monotonicity assumption, $\bar{d}_1 - \bar{d}_0$ is an unbiased estimator of fraction of compliers.

Intention-to-Treat Analysis (continued)

$$ITT_Y = \frac{1}{N}\left(N_a ITT_Y^a + N_n ITT_Y^n + N_c ITT_Y^c + N_d ITT_Y^d\right)$$

Under exclusion and monotonicity assumptions, we have

$$ITT_Y^c = \frac{ITT_Y}{N_c/N}$$

The conventional instrumental variables estimator is in Imbens and Rubin (1997):

$$\hat{\beta}_{IV} = \frac{\bar{y}_{z=1} - \bar{y}_{z=0}}{\bar{d}_{z=1} - \bar{d}_{z=0}}$$

Under strict monotonicity assumption, $\bar{d}_1 - \bar{d}_0$ is an unbiased estimator of fraction of compliers. Furthermore, $\bar{y}_{1\cdot} - \bar{y}_{0\cdot}$ can be interpreted as a causal effect of treatment received only under the exclusion restriction.

From before...

The IV estimand is

$$\mathbb{E}[Y_i(1) - Y_i(0)|D_i(1) - D_i(0) = 1] \;=\; \frac{\mathbb{E}[Y_i(1, D_i(1)) - Y_i(0, D_i(0))]}{\mathbb{E}[D_i(1) - D_i(0)]}$$

Following monotonicity,

$$\mathbb{E}[D_i(1) - D_i(0)] \;=\; P[D_i(1) - D_i(0) = 1] \;>\; 0$$

i.e., the proportion of the population who are compliers.

Hence, how we can estimate the LATE?

Given our assumptions:

Assignment Z_i^{obs}	Supplement D_i^{obs}	Survival Y_i^{obs}	# Units	Type
0	0	0	74	C/N
0	0	1	11,514	C/N
1	0	0	34	N
1	0	1	2,385	N
1	1	0	12	C
1	1	1	9,663	C

If we make the 5 assumptions, our estimate of the LATE is the CACE:

$$\frac{1}{N}\Big(N_c \cdot ITT_c + N_n \cdot ITT_n + N_a \cdot ITT_a + N_d \cdot ITT_d\Big) \rightarrow \frac{\widehat{ITT}_Y}{N_c/N} = \widehat{CACE}$$

Introduction to causal inference for data scientists

Instrumental variables (2) and observational studies

Michael Gill

2018-02-27

Center for Data Science | New York University

Today

Today's references

References:

- Imbens, G. (2010). "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)", Journal of Economic Literature, 399-423.
- Imbens, G. and Rubin, D. (1997.) "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance." The Annals of Statistics 25(1):pp. 305–327.
- MW, Chapter 5
- IR, Chapters 12-13.

Today's goals

- From ITT to LATE... to ATE?
- Extend the ITT framework before to a two-sided non-compliance case
- Motivate observational studies
- Talk a little bit about the midterm

From before...

Unit-Level Intent-to-Treat Effects

$$Y_i(1, D_i(1)) - Y_i(0, D_i(0)) = \begin{cases} Y_i(1,1) - Y_i(0,0) & \text{for compliers} \\ Y_i(1,1) - Y_i(0,1) & \text{for always-takers} \\ Y_i(1,0) - Y_i(0,0) & \text{for never-takers} \\ Y_i(1,0) - Y_i(0,1) & \text{for defiers} \end{cases}$$

- Always-takers and never-takers have only one treatment received, regardless of assignment.
- Compliers and defiers receive different treatments for different assignments.
- Hence, to gauge causal effect of treatment received, we typically look at compliers and defiers.

The LATE estimator

Under basic assumptions (and with binary treatment assigned, Z, and received, D), the IV estimator gives the *LATE* or the *CACE*:

$$\begin{aligned} CACE &= \mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 - D_i^0 > 0] \\ &= \mathbb{E}[Y_i^1 - Y_i^0 | D_i^1 = 1] \\ &= \frac{\sum_{i=1}^n (Y_i^1 - Y_i^0) \cdot d_i^1}{\sum_{i=1}^n d_i^1} \end{aligned}$$

Recall, we can estimate $\widehat{CACE}$ given the ratio of intent-to-treat estimates for our outcome and endogenous treatment of interest:

$$\widehat{CACE} = \frac{\widehat{ITT}_Y}{\widehat{ITT}_D} = \frac{\widehat{\mathbb{E}}(Y^* | Z = 1) - \widehat{\mathbb{E}}(Y^* | Z = 0)}{\widehat{\mathbb{E}}(D^* | Z = 1) - \widehat{\mathbb{E}}(D^* | Z = 0)}.$$

The LATE estimator (continued)

The following provides a consistent estimator of the LATE:

Estimating LATE under assuptions 1-5 (AIR, 1996), no covariates

$$\frac{\widehat{ITT}_Y}{\widehat{ITT}_D} = \frac{\left(\sum_{i=1}^n Z_i Y_i\right) \Big/ \left(\sum_{i=1}^n Z_i\right) - \left(\sum_{i=1}^n (1-Z_i) Y_i\right) \Big/ \left(\sum_{i=1}^n (1-Z_i)\right)}{\left(\sum_{i=1}^n Z_i D_i\right) \Big/ \left(\sum_{i=1}^n Z_i\right) - \left(\sum_{i=1}^n (1-Z_i) D_i\right) \Big/ \left(\sum_{i=1}^n (1-Z_i)\right)}$$

Independence of Z with respect to potential outcomes

- Recall that under non-compliance, units now have a broader set of potential outcomes: $\left(Y_i(1, D_i^1), Y_i(0, D_i^0), D_i^1, D_i^0\right)$
- If we assume the exclusion restriction, this implies $Y_i(z, d) = Y_i(z', d)$.

Challenges presented by non-compliance

- Inability to straightforwardly estimate ATE
- When share of non-compliance is high, precision deteriorates

Hypothetical schedule of potential outcomes

Unit	Y_i^1	Y_i^0	δ_i	$d_i(z=1)$	$d_i(z=0)$	Type
1	6	4	2	1	0	?
2	3	2	1	0	0	?
3	4	5	-1	0	0	?
4	3	0	3	1	0	?
5	2	1	1	0	0	?
6	5	4	1	0	0	?
7	2	1	1	1	0	?
8	6	6	0	1	0	?
9	5	3	2	1	0	?
10	8	7	1	0	0	?

This implies one-sided non-compliance.

Hypothetical schedule of potential outcomes

Unit	Y_i^1	Y_i^0	δ_i	$d_i(z=1)$	$d_i(z=0)$	Type
1	6	4	2	1	0	?
2	3	2	1	0	0	?
3	4	5	-1	0	0	?
4	3	0	3	1	0	?
5	2	1	1	0	0	?
6	5	4	1	0	0	?
7	2	1	1	1	0	?
8	6	6	0	1	0	?
9	5	3	2	1	0	?
10	8	7	1	0	0	?

This implies one-sided non-compliance. Questions: what is CACE? NACE?

Hypothetical schedule of potential outcomes

Unit	Y_i^1	Y_i^0	δ_i	$d_i(z=1)$	$d_i(z=0)$	Type
1	6	4	2	1	0	co
2	3	2	1	0	0	nt
3	4	5	-1	0	0	nt
4	3	0	3	1	0	co
5	2	1	1	0	0	nt
6	5	4	1	0	0	nt
7	2	1	1	1	0	co
8	6	6	0	1	0	co
9	5	3	2	1	0	co
10	8	7	1	0	0	nt

This implies one-sided non-compliance. Questions: what is CACE? NACE?

Hypothetical schedule of potential outcomes

Unit	Y_i^1	Y_i^0	δ_i	$d_i(z=1)$	$d_i(z=0)$	Type
1	6	4	2	1	0	co
2	3	2	1	0	0	nt
3	4	5	-1	0	0	nt
4	3	0	3	1	0	co
5	2	1	1	0	0	nt
6	5	4	1	0	0	nt
7	2	1	1	1	0	co
8	6	6	0	1	0	co
9	5	3	2	1	0	co
10	8	7	1	0	0	nt

This implies one-sided non-compliance. Questions: what is CACE? NACE?

$CACE = (2+3+1+0+2)/5 = 8/5$. $NACE = (1+(-1)+1+1+1)/5 = 4/5$.

CACE, NACE, and ATE

From before, we know the (sample) ATE is the average treatment effect amongst all units. Or, equivalently:

$$
\begin{aligned}
ATE &= \frac{1}{N}\sum_{i=1}^{N}\left(Y_i^1 - Y_i^0\right) \\
&= \frac{N_{co}}{N}\cdot CACE + \frac{N_{nt}}{N}\cdot NACE \\
&= (5/10)\cdot(8/5) + (5/10)\cdot(4/5) = 6/5
\end{aligned}
$$

Group ACE vs. ITT

The prior slides demonstrate that *NACE* is a well-defined causal quantity, but estimation is challenging due to compliance.

With one-sided non-compliance, in general, this implies $NACE \neq ITT_{nt}$, despite the fact that $CACE = ITT_{co}$.

Group ACE vs. ITT

The prior slides demonstrate that *NACE* is a well-defined causal quantity, but estimation is challenging due to compliance.

With one-sided non-compliance, in general, this implies $NACE \neq ITT_{nt}$, despite the fact that $CACE = ITT_{co}$.

We can see this another way by expanding our table from before (assuming the exclusion restriction).

Group ACE vs. ITT (continued)

Unit	Y_i^1	Y_i^0	δ_i	$d_i(z=1)$	$d_i(z=0)$	Type	$Y_i(z=1)$	$Y_i(z=0)$	ITT_i
1	6	4	2	1	0	co	?	?	?
2	3	2	1	0	0	nt	?	?	?
3	4	5	-1	0	0	nt	?	?	?
4	3	0	3	1	0	co	?	?	?
5	2	1	1	0	0	nt	?	?	?
6	5	4	1	0	0	nt	?	?	?
7	2	1	1	1	0	co	?	?	?
8	6	6	0	1	0	co	?	?	?
9	5	3	2	1	0	co	?	?	?
10	8	7	1	0	0	nt	?	?	?

Group ACE vs. ITT (continued)

Unit	Y_i^1	Y_i^0	δ_i	$d_i(z=1)$	$d_i(z=0)$	Type	$Y_i(z=1)$	$Y_i(z=0)$	ITT_i
1	6	4	2	1	0	co	6	4	2
2	3	2	1	0	0	nt	2	2	0
3	4	5	-1	0	0	nt	5	5	0
4	3	0	3	1	0	co	3	0	3
5	2	1	1	0	0	nt	1	1	0
6	5	4	1	0	0	nt	4	4	0
7	2	1	1	1	0	co	2	1	1
8	6	6	0	1	0	co	6	6	0
9	5	3	2	1	0	co	5	3	2
10	8	7	1	0	0	nt	7	7	0

Two-sided non-compliance

Example: CNN townhall on gun policy

CNN's town hall on gun violence was the network at its best — and worst

The teens might not save us, but neither will CNN.

By Todd VanDerWerff | @tvoti | todd@vox.com | Feb 25, 2018, 1:30pm EST

SHARE

Stoneman Douglas High senior Emma Gonzalez asks NRA spokesperson Dana Loesch a question at a CNN town hall. | CNN

Figure 1: A useful title from Vox

Hypothetical experimental design

Suppose you worked for CNN and you wish to know whether viewing the townhall has an effect on viewers' attitudes on gun control.

Hypothetical experimental design

Suppose you worked for CNN and you wish to know whether viewing the townhall has an effect on viewers' attitudes on gun control.

- You obtain a sample of $n = 5,000$ possible TV viewers, and gather their pre-treatment covariates (e.g., age, gender, party affiliation)

Hypothetical experimental design

Suppose you worked for CNN and you wish to know whether viewing the townhall has an effect on viewers' attitudes on gun control.

- You obtain a sample of $n = 5,000$ possible TV viewers, and gather their pre-treatment covariates (e.g., age, gender, party affiliation)
- You randomly assign individuals into either an encouragement condition (e.g., "Please, watch the townhall") or a control condition (e.g., "Please, watch some TV at some point").

Hypothetical experimental design

Suppose you worked for CNN and you wish to know whether viewing the townhall has an effect on viewers' attitudes on gun control.

- You obtain a sample of $n = 5,000$ possible TV viewers, and gather their pre-treatment covariates (e.g., age, gender, party affiliation)
- You randomly assign individuals into either an encouragement condition (e.g., "Please, watch the townhall") or a control condition (e.g., "Please, watch some TV at some point").
- After the townhall, you ask each person whether they watched the townhall (including those in the control arm).

Hypothetical experimental design

Suppose you worked for CNN and you wish to know whether viewing the townhall has an effect on viewers' attitudes on gun control.

- You obtain a sample of $n = 5,000$ possible TV viewers, and gather their pre-treatment covariates (e.g., age, gender, party affiliation)
- You randomly assign individuals into either an encouragement condition (e.g., "Please, watch the townhall") or a control condition (e.g., "Please, watch some TV at some point").
- After the townhall, you ask each person whether they watched the townhall (including those in the control arm).
- For each person in the sample, you solicit their post-treatment attitudes on gun control. E.g., "Would you support a federal ban on assault weapons? Yes, or no?"

A few challenges here

- Those encouraged into treatment may not watch the townhall (i.e., they may be never-takers or defiers)
- Those encouraged into control may watch the townhall (i.e., they may be always takers or defiers), since the show is open for all to view.
- The non-compliance is *two-sided* because it can occur both when assigned to treatment or assigned control.

Hypothetical (observed) data from the CNN experiment

Assuming two-sided non-compliance...

Encourage Z_i^{obs}	Watched D_i^{obs}	Support Y_i^{obs}	# Units	Type
0	0	0	995	?
0	0	1	1,299	?
0	1	0	52	?
0	1	1	154	?
1	0	0	657	?
1	0	1	870	?
1	1	0	242	?
1	1	1	731	?

Hypothetical (observed) data from the CNN experiment

Assuming two-sided non-compliance...

Encourage Z_i^{obs}	Watched D_i^{obs}	Support Y_i^{obs}	# Units	Type
0	0	0	995	co/nt
0	0	1	1,299	co/nt
0	1	0	52	at/de
0	1	1	154	at/de
1	0	0	657	de/nt
1	0	1	870	de/nt
1	1	0	242	at/co
1	1	1	731	at/co

Hypothetical (observed) data from the CNN experiment

Assuming two-sided non-compliance...and monotonicity...

Encourage Z_i^{obs}	Watched D_i^{obs}	Support Y_i^{obs}	# Units	Type
0	0	0	995	co/nt
0	0	1	1,299	co/nt
0	1	0	52	at
0	1	1	154	at
1	0	0	657	nt
1	0	1	870	nt
1	1	0	242	at/co
1	1	1	731	at/co

Hypothetical (observed) data from the CNN experiment

Assuming two-sided non-compliance...and monotonicity...

Encourage Z_i^{obs}	Watched D_i^{obs}	Support Y_i^{obs}	# Units	Type
0	0	0	995	co/nt
0	0	1	1,299	co/nt
0	1	0	52	at
0	1	1	154	at
1	0	0	657	nt
1	0	1	870	nt
1	1	0	242	at/co
1	1	1	731	at/co

From the above, can we estimate CACE?

Computing the CACE under two-sided non-compliance

- As before, we know that $CACE = ITT_Y / \Pr(\text{Complier})$.

Computing the CACE under two-sided non-compliance

- As before, we know that $CACE = ITT_Y / \text{Pr(Complier)}$.
- By LTP (and monotonicity) we know:
  $Pr(\text{Complier}) + Pr(\text{Always-taker}) + Pr(\text{Never-taker}) = 1$.

Computing the CACE under two-sided non-compliance

- As before, we know that $CACE = ITT_Y / \text{Pr(Complier)}$.
- By LTP (and monotonicity) we know:
  $Pr(\text{Complier}) + Pr(\text{Always-taker}) + Pr(\text{Never-taker}) = 1$.
- Hence, $Pr(\text{Complier}) = 1 - Pr(\text{Always-taker}) - Pr(\text{Never-taker})$.

Computing the CACE under two-sided non-compliance

- As before, we know that $CACE = ITT_Y / \text{Pr(Complier)}$.
- By LTP (and monotonicity) we know:
  $Pr(\text{Complier}) + Pr(\text{Always-taker}) + Pr(\text{Never-taker}) = 1$.
- Hence, $Pr(\text{Complier}) = 1 - Pr(\text{Always-taker}) - Pr(\text{Never-taker})$.
- Estimating $Pr(\widehat{\text{Always-taker}})$ and $Pr(\widehat{\text{Never-taker}})$ allows us to identify CACE.

Compliance scores

Define each unit's pre-treatment probability of being a complier, always-taker, never-taker, and defier by:

- $\pi_{i,co}$
- $\pi_{i,at}$
- $\pi_{i,nt}$
- $\pi_{i,de}$

Due to randomization, we know that $Z \perp\!\!\!\perp (\pi_{co}, \pi_{at}, \pi_{nt}, \pi_{de})$

This means the expected share of compliers, defiers, always-takers, and never-takers is equal in both groups. I.e.,

$$\mathbb{E}(\pi_j) = \mathbb{E}(\pi_j|Z) = \mathbb{E}(\pi_j|Z = 1) = \mathbb{E}(\pi_j|Z = 0)$$

, for all $j \in \{co, at, nt, de\}$.

Estimating $Pr(\textbf{Always-taker})$ and $Pr(\textbf{Never-taker})$

Z_i^{obs}	D_i^{obs}	Y_i^{obs}	# Units	Type
0	0	0	995	co/nt
0	0	1	1,299	co/nt
0	1	0	52	at
0	1	1	154	at
1	0	0	657	nt
1	0	1	870	nt
1	1	0	242	at/co
1	1	1	731	at/co

- $\widehat{\mathbb{E}(\pi_{at})} = \mathbb{E}(\widehat{\pi_{at}|Z=0}) = \frac{52+154}{995+1299+52+154} = 0.0824.$
- $\widehat{\mathbb{E}(\pi_{nt})} = \mathbb{E}(\widehat{\pi_{nt}|Z=1}) = \frac{657+870}{657+870+242+731} = 0.6108.$
- $\widehat{\mathbb{E}(\pi_{co})} = 1 - \mathbb{E}(\widehat{\pi_{at}|Z=0}) - \mathbb{E}(\widehat{\pi_{nt}|Z=1}) = 1 - 0.0824 - 0.6108$

At long last, CACE!

We can estimate CACE by:

$$\frac{\widehat{ITT}_Y}{\widehat{\mathbb{E}(\pi_{co})}} = \frac{0.0592}{0.3068} = 0.193$$

At long last, CACE!

We can estimate CACE by:

$$\frac{\widehat{ITT}_Y}{\widehat{\mathbb{E}(\pi_{co})}} = \frac{0.0592}{0.3068} = 0.193$$

If you're interested in designs that accomodate missing data/attrition in the outcome: see, for example, Yau and Little (2001) and related work.

What about uncertainty?

A few choices:

- Closed-form estimation of variance of CACE (see IR)
- Bayesian methods (e.g., Imbens and Rubin, 1997)
- The Jackknife
- Bootstrap methods

The bootstrap?

Main idea: we can approximate the sampling distribution of a given statistic by repeatedly re-drawing points from the observed data.

The bootstrap?

Main idea: we can approximate the sampling distribution of a given statistic by repeatedly re-drawing points from the observed data.

We will focus on the simple nonparametric bootstrap. This will become increasingly important as we move on in this course.

Bootstrap estimation of the variance

Given a sample of size n, and some observed sample statistic T_n, our objective is to estimate the unobserved variance $\mathbb{V}_{\widehat{F}_n}(T_n)$ from our sample.

Bootstrap estimation of the variance

Given a sample of size n, and some observed sample statistic T_n, our objective is to estimate the unobserved variance $\mathbb{V}_{\widehat{F}_n}(T_n)$ from our sample. To do this, we will repeatedly simulate $X'_i, \ldots, X'_n \sim \widehat{F}_n$ and characterize the variance of our bootstrapped collection of sample statistics.

Bootstrap Variance Estimation

1. Draw $X'_i, \ldots, X'_n \sim \widehat{F}_n$, with replacement.
2. Estimate $T'_n = f(X'_i, \ldots, X'_n)$
3. Repeat 1 and 2, M times, and store the collection $T'_{n,1}, \ldots, T'_{n,M}$
4. Your boostrap variance is given by

$$v_{boot} = \frac{1}{M} \sum_{m=1}^{M} \left(T'_{n,m} - \frac{1}{M} \sum_{b=1}^{M} T'_{n,b} \right)^2$$

So, can we learn anything about ATE from IV?

Yes, perhaps...

- Re-weighting our data to estimate ATE (Aronow and Carnegie, 2013)
- Bounding ATE given experiment (Balke and Pearl, 1997)

So, can we learn anything about ATE from IV?

Yes, perhaps...

- Re-weighting our data to estimate ATE (Aronow and Carnegie, 2013)
- Bounding ATE given experiment (Balke and Pearl, 1997)

Two great R packages:

- `icsw`
- `noncompliance`

ICSW (from Aronow and Carnegie, 2013)

Inverse compliance-score weighting

Main idea: if there are pre-treatment covariates that are predictive of compliance status, you can re-weight the sample estimate of the LATE in a manner akin to the Horvitz-Thompson (1952) estimator to obtain:

ICSW estimator for the ATE, $\widehat{ATE_{ICSW}} =$

$$\frac{\left(\sum_{i=1}^{n} \widehat{w_{Ci}} Z_i Y_i\right) \Big/ \left(\sum_{i=1}^{n} \widehat{w_{Ci}} Z_i\right) - \left(\sum_{i=1}^{n} \widehat{w_{Ci}} (1 - Z_i) Y_i\right) \Big/ \left(\sum_{i=1}^{n} \widehat{w_{Ci}} (1 - Z_i)\right)}{\left(\sum_{i=1}^{n} \widehat{w_{Ci}} Z_i D_i\right) \Big/ \left(\sum_{i=1}^{n} \widehat{w_{Ci}} Z_i\right) - \left(\sum_{i=1}^{n} \widehat{w_{Ci}} (1 - Z_i) D_i\right) \Big/ \left(\sum_{i=1}^{n} \widehat{w_{Ci}} (1 - Z_i)\right)}$$

where $\widehat{w_{Ci}} = \frac{1}{\widehat{P_{Ci}}}$, and $\widehat{P_{Ci}} = Pr(D^1 > \widehat{D^0} | X = x_i)$.

Can we infer anything about ATE from our experiment?

We will briefly discuss two approaches next class:

1. Bounding ATE given our experiment (Balke and Pearl, 1997)
2. Re-weighting our data to estimate ATE (Aronow and Carnegie, 2013)

Introduction to causal inference for data scientists

Matched sampling for observational studies

Michael Gill

2018-03-20

Center for Data Science | New York University

Today

Today's references

References:

- Stuart, E. (2010). "Matching Methods for Causal Inference: A Review and a Look Forward". Statistical Science.
- Ho, D., Imai, K., King, G., Stuart, E. (2007). "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." Political Analysis.
- Rubin, D., and Waterman, R. (2006). "Estimating the Causal Effects of Marketing Interventions Using Propensity Score Methodology." Statistical Science.
- IR, Ch. 12, 13, 18

Today's goals

- Introduce section on observational studies
- Motivate matching in light of experimental theory
- Overview several possible approaches
- Discuss their challenges

Observational studies

Observational studies vs. controlled experiments

- In observational studies, the assignment mechanism is typically unknown.
- While unconfoundness was part of the experimental design in a controlled experiment, it is an assumption in observational studies:

$$(Y^0, Y^1) \perp D \mid X$$

- Unconfoundedness is not testable in the following sense: given a random variable (D, X, Y^*), where $D \in \{0, 1\}$, $X \in \mathbb{R}^d$, and $Y^* \in \mathbb{R}$, it is possible to write two potential outcome models (D, X, Y^0, Y^1) compatible with the observation (D, X, Y^*), i.e., such that $Y^* = Y^D$, and such that in the first model, unconfoundedness holds; and in the second one, it does not.

Observational studies vs. controlled experiments

- This means that we'll need to assert judgment whether unconfoundedness is plausible or not. E.g., if X is socio-economic status, if D is smoking, and if Y^* is the health outcome, does it make sense?
- Similarly, the propensity score $e(x)$ (conditional probability of being treated) is unknown. We will have to estimate it.

Estimating treatment effects

- Under unconfoundedness, the ATE is computed by $E\left[ATE\left(X\right)\right]$, where

$$ATE\left(X\right) = E\left[Y|D = 1, X\right] - E\left[Y|D = 0, X\right].$$

- When X is discrete, easy: view study as a stratified experiments with as many blocks as there are values that X can take.
- When X is continuous, this is more difficult, as the expectations are conditional on $X = x$, a zero-probability event. We will explore several attempts to address this difficulty.

Propensity score and unconfoundedness: reminders

- Recall the propensity score is defined as $e(X) = \Pr(D = 1|X)$. Under unconfoundedness, we have

$$(Y^0, Y^1) \perp D \mid e(X).$$

- Overlap condition: $0 < e(x) < 1$ for all x in the support of P_X.
- Under unconfoundedness, the ATE is computed by $E[ATE(X)]$, where

$$ATE(X) = E[Y|D = 1, X] - E[Y|D = 0, X],$$

  and hence $ATE = E[ATE(e)]$

$$ATE(e) = E[Y|D = 1, e(X) = e] - E[Y|D = 0, e(X) = e].$$

- This means that instead of storing the information X, we could keep only $e(X)$, which is less information. However, estimating the propensity score $e(x)$ can be involved. We'll come back to that later.

Efficiency bound

- It is possible to compute what is the best possible accuracy that an estimator can achieve when the sample size gets large.
- Semi-parametric efficiency bound (e.g., Hahn, 1998): for any estimator $\widehat{ATE}$ of ATE, we have as

$$\lim N \, var\left(\widehat{ATE} - ATE\right) \geq E\left[\frac{V^1(X)}{e(X)} + \frac{V^0(X)}{1 - e(X)} + \left(ATE(X) - ATE\right)^2\right]$$

  where $V^d(X) = var\left(Y^d | X\right)$, and that this bound can be approximately attained when the sample size N is large.
- Whether the propensity score $e(x)$ is known or not does not affect this bound.

Parametric regressions

- Regression approach: assume a parametric (typically linear) form of

$$E\left[Y^d|D=d,X=x\right]=\alpha_d+\beta_d'x$$

- Hence, regress Y_i^* on X_i on both the treated and control sets, and the estimator of the ATE will be

$$\hat{\alpha}_1-\hat{\alpha}_0+\left(\hat{\beta}_1-\hat{\beta}_0\right)'\bar{X}.$$

- Probably the most popular approach in applied work.
- However, a problem is that the specification of the regression is unknown: here, for instance, nothing guarantees that it should be linear. We could add higher-order terms but we would then be faced with a model selection issue.

Blocking

- Stratified approach: consider blocks partitioning $\mathcal{X}$, namely $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \ldots \cup \mathcal{X}_K$, and compute

$$\sum_{k=1}^{K} \Pr\left(X \in \mathcal{X}_k\right) ATE_k$$

where

$$ATE_k = E\left[Y^* | D = 1, X \in \mathcal{X}_1\right] - E\left[Y^* | D = 0, X \in \mathcal{X}_k\right]$$

- However, a problem with this methodology is that the size of the blocks that should be set is not obvious: they should not be too large (in the limit where there is only one bin, the estimator would become the naive estimator, which is biased); nor too small (because then some bins would be empty).

Kernel estimation

- Nonparametric approach: $E\left[Y^d | D = d, X = x\right]$ is estimated by kernel estimation using

$$\frac{\sum_{i=1}^{N} K\left(\frac{X_i - x}{h}\right) Y_i 1\left\{D_i = d\right\}}{\sum_{i=1}^{N} K\left(\frac{X_i - x}{h}\right) 1\left\{D_i = d\right\}}$$

  where the parameter h (bandwidth) controls how much information we locally aggregate.
- h should decrease when N gets larger; typically the inverse of a power of N.
- If $K(z) = 1\{|z| \leq 1\}$, the approach is closely related to blocking. Typically, a smoother kernel is used, such as the Gaussian kernel $K(z) = \exp\left(-z^2/2\right)$.

Estimation by weighting

- Recall, if we know the propensity score, then we can use the Horvitz-Thompson estimator (1952):

$$\tau^{HT} = \frac{1}{N}\sum_{i=1}^{N}\frac{D_i Y_i^*}{e\left(X_i\right)} - \frac{1}{N}\sum_{i=1}^{N}\frac{\left(1-D_i\right)Y_i^*}{1-e\left(X_i\right)}.$$

- Indeed,

$$\mathbb{E}\left[\frac{DY^*}{e\left(X\right)}\right] = \mathbb{E}\left[\frac{\mathbb{E}\left[DY^1|X\right]}{e\left(X\right)}\right] = \mathbb{E}\left[\frac{\mathbb{E}\left[D|X\right]\mathbb{E}\left[Y^1|X\right]}{e\left(X\right)}\right] = \mathbb{E}\left[Y^1\right]$$

  and similarly,

$$\mathbb{E}\left[\frac{(1-D)Y^*}{1-e\left(X\right)}\right] = \mathbb{E}\left[Y^0\right].$$

- In a perfectly randomized experiment, $e\left(x\right) = N_1/N$, and we recover the naive estimator. However, in observational studies, often $e\left(x\right)$ is unknown.

Matching estimators

- Matching idea: for each treated unit, look for a nontreated unit which offers closest possible resemblance in terms of covariates x, and record the difference in outcomes.
- Some matching estimators match based on the similarity in the covariates x; some match based on the similarity in the propensity score $e(x)$ (propensity score matching).
- Several variants:
  - Exact matching
  - Nearest-neighbor, caliper, and radius matching
  - Interval matching
  - Kernel matching

Matching estimators

Matching

- See Heckman, Ichimura, and Todd (1997).
- Typical estimator is

$$\frac{1}{N_1}\left[\sum_{i\in I_1}\left(Y_i^* - \sum_{j\in I_0}\omega_{ij}Y_j^*\right)\right]$$

  where ω_{ij} captures the similarity between the propensity score of units i and j. We'll see examples of ω_{ij} later.
- Advantage of matching methods: no need to estimate the propensity score (unless we match on propensity score).

Nearest neighbor matching

- Nearest-neighbor matching: for a given unit i, rank the units j such that $D_j \neq D_i$ by increasing $\left|X_i - X_j\right|$, (i.e., the first ones are the most similar to i in terms of covariates to have received a different treatment).
- Let $J_m(i)$ be the set of the first m such units.
- Then set $\omega_{ij} = 1\{j \in J_m(i)\}$. A unit will be matched with an average over the m closest neighbors that have received different treatment.
- Important bias as soon as dimension of covariates is above 2.

Kernel matching

- In kernel methods, can take

$$\omega(i,j) = \frac{K\left(\frac{x_j - x_i}{h_n}\right) 1\left\{D_i \neq D_j\right\}}{\sum_{k \in I} K\left(\frac{x_k - x_i}{h_n}\right) 1\left\{D_i \neq D_k\right\}}$$

- Same spirit as nearest neighbors but weights all observations.

Propensity score matching

- Idea: replace X_i by propensity score $e(X_i)$.
- Advantage: if x is high dimensional, handles the "curse of dimensionality".
- Drawback: $e(x)$ is typically not known and must be estimated.
- 3 steps:
  1. Estimate $\hat{e}(x)$
  2. Define the region of common support, i.e., the set of x such that $0 < e(x) < 1$.
  3. Match participants to nonparticipants using one of the methods above

Propensity score estimation: logistic specification

- In some settings, the propensity score $e(x) = \Pr(D = 1|X = x)$ is known and explicit.
- In others, it is not, in which case it should be estimated. As a result,

$$e_\phi(x) = \frac{\exp(X'\phi)}{1 + \exp(X'\phi)}.$$

- Note that here, $X'\phi = \sum_k \phi_k X_k$ is linear with respect to the entries of X_k, but this is without loss of generality, as we could replace X by $h(X)$. For instance one could have

$$h(X) = (X, X_1^2, X_1X_2, ...X_1X_K, X_1X_2, X_2^2, ...)$$

however the dimension of the regressors should increase moderately with the sample size: beware of overfitting.

Propensity score estimation: logistic specification (2)

- The likelihood of $(D_1, ..., D_N)$ is

$$\prod_{i=1}^{N} \frac{\exp\left(D_i X' \phi\right)}{1 + \exp\left(X_i' \phi\right)}$$

- Hence $\hat{\phi}$ can be estimated by maximum likelihood by

$$\max_{\phi} \sum_{i=1}^{N} D_i X_i' \phi - \log\left(1 + \exp\left(X_i' \phi\right)\right),$$

  which is the maximization of a concave function.
- By first order conditions, at the optimal $\hat{\phi}$,

$$\sum_{i=1}^{N} \left(D_i - e_{\hat{\phi}}\left(X_i\right)\right) X_i = 0$$

Propensity score estimation: logistic specification (3)

- $\hat{\phi}$ can be computed by gradient descent using

$$\phi^{t+1} = \phi^t - \epsilon^t \left(\sum_{i=1}^{N} \left(D_i - e_{\hat{\phi}}(X_i) \right) X_i \right)$$

- Consider a random variable U with the logistic distribution, whose c.d.f. is

$$F_U(z) = \frac{\exp(z)}{1 + \exp(z)} = \frac{1}{1 + e^{-z}},$$

  and assuming U is independent from X; we can interpret the treatment assignment mechanism by

$$D = 1\{X'\phi \geq U\}.$$

Example: Card-Krueger study on minimum wage

- Card, D. and Krueger, A. (1994). "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *American Economic Review* 84 (4): 772–793.
- Abstract: "On April 1, 1992, New Jersey's minimum wage rose from $4.25 to $5.05 per hour. To evaluate the impact of the law we surveyed 410 fast-food restaurants in New Jersey and eastern Pennsylvania before and after the rise. Comparisons of employment growth at stores in New Jersey and Pennsylvania (where the minimum wage was constant) provide simple estimates of the effect of the higher minimum wage. We also compare employment changes at stores in New Jersey that were initially paying high wages (above $5) to the changes at lower-wage stores. We find no indication that the rise in the minimum wage reduced employment."

A natural experiment

- On April 1, 1992, New Jersey raised the minimum wage from \$4.25 (federal minimum) to \$5.05 per hour. The adjacent state of Pennsylvania did not. Remained at the federal minimum.
- 410 fast-food restaurants in New Jersey and eastern Pennsylvania surveyed.
- Treatment: $D_i = 1$ if store i is in NJ, $D_i = 0$ if in PA.
- Outcome Y_i is employment in store i.
- Covariates X_i: chain (Burger King, KFC, Roy Rogers, Wendy's, Company-owned); region.

Results: design and response rates

TABLE 1—SAMPLE DESIGN AND RESPONSE RATES

	All	Stores in: NJ	Stores in: PA
Wave 1, February 15 – March 4, 1992:			
Number of stores in sample frame:[a]	473	364	109
Number of refusals:	63	33	30
Number interviewed:	410	331	79
Response rate (percentage):	86.7	90.9	72.5
Wave 2, November 5 – December 31, 1992:			
Number of stores in sample frame:	410	331	79
Number closed:	6	5	1
Number under rennovation:	2	2	0
Number temporarily closed:[b]	2	2	0
Number of refusals:	1	1	0
Number interviewed:[c]	399	321	78

[a]Stores with working phone numbers only; 29 stores in original sample frame had disconnected phone numbers.

[b]Includes one store closed because of highway construction and one store closed because of a fire.

[c]Includes 371 phone interviews and 28 personal interviews of stores that refused an initial request for a phone interview.

Figure 1:

Results: distributions of starting wages

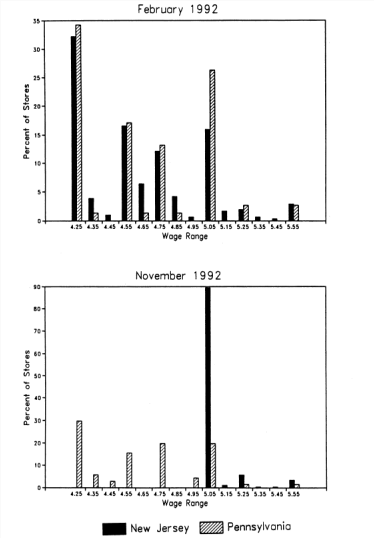


Figure 2:

Results: change in employment

TABLE 3—AVERAGE EMPLOYMENT PER STORE BEFORE AND AFTER THE RISE IN NEW JERSEY MINIMUM WAGE

	Stores by state			Stores in New Jersey[a]			Differences within NJ[b]	
Variable	PA (i)	NJ (ii)	Difference, NJ − PA (iii)	Wage = $4.25 (iv)	Wage = $4.26–$4.99 (v)	Wage ≥ $5.00 (vi)	Low–high (vii)	Midrange–high (viii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	−2.89 (1.44)	19.56 (0.77)	20.08 (0.84)	22.25 (1.14)	−2.69 (1.37)	−2.17 (1.41)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	−0.14 (1.07)	20.88 (1.01)	20.96 (0.76)	20.21 (1.03)	0.67 (1.44)	0.75 (1.27)
3. Change in mean FTE employment	−2.16 (1.25)	0.59 (0.54)	2.76 (1.36)	1.32 (0.95)	0.87 (0.84)	−2.04 (1.14)	3.36 (1.48)	2.91 (1.41)
4. Change in mean FTE employment, balanced sample of stores[c]	−2.28 (1.25)	0.47 (0.48)	2.75 (1.34)	1.21 (0.82)	0.71 (0.69)	−2.16 (1.01)	3.36 (1.30)	2.87 (1.22)
5. Change in mean FTE employment, setting FTE at temporarily closed stores to 0[d]	−2.28 (1.25)	0.23 (0.49)	2.51 (1.35)	0.90 (0.87)	0.49 (0.69)	−2.39 (1.02)	3.29 (1.34)	2.88 (1.23)

Notes: Standard errors are shown in parentheses. The sample consists of all stores with available data on employment. FTE (full-time-equivalent) employment counts each part-time worker as half a full-time worker. Employment at six closed stores is set to zero. Employment at four temporarily closed stores is treated as missing.

[a]Stores in New Jersey were classified by whether starting wage in wave 1 equals $4.25 per hour ($N = 101$), is between $4.26 and $4.99 per hour ($N = 140$), or is $5.00 per hour or higher ($N = 73$).

[b]Difference in employment between low-wage ($4.25 per hour) and high-wage ($\geq$ $5.00 per hour) stores; and difference in employment between midrange ($4.26–$4.99 per hour) and high-wage stores.

[c]Subset of stores with available employment data in wave 1 and wave 2.

[d]In this row only, wave-2 employment at four temporarily closed stores is set to 0. Employment changes are based on the subset of stores with available employment data in wave 1 and wave 2.

Figure 3:

Causal estimands and matching

Oftentimes researchers choose to estimate the ATT rather than the ATE using matching methods.

Causal estimands and matching

Oftentimes researchers choose to estimate the ATT rather than the ATE using matching methods. When/why is this?

More on this: Imbens (2004) and Stuart (2010)

Desirable properties of a matching estimator

From Rubin and Thomas (1992) and Rubin and Thomas (1996):

- Affinely invariant (same matches given linear transormations of the covariate space)
- Equal percent bias reducing: EPBR (Rubin, 1974; Rubin and Stuart, 2006)

General framework for matching methods

1. Define “closeness”
2. Implement matches, given closeness definition
3. Assess quality of matches, and estimate treatment effects

General considerations

- All approaches rely on a strong selection on observables assumption (i.e., strong ingnorability, or no unobserved differences between groups conditional on covariates)
- Important to include all (pre-treatment) variables related both to the treatment assignment and outcome
- Multivariate methods prone to variance increases given irrelevant controls; PSM less prone to irrelevant variables
- Naive computation of standard errors generally too low after matching (Imbens and Wooldridge, 2009)
- Small sample settings prevent conditioning on large number of controls, and may overfit via PSM.

Other common distance measures

- Exact: $D_{ij} = 0$ if $X_i = X_j$; else $D_{ij} = \infty$.
- Mahalanobis: $D_{ij} = (X_i - X_j)'\Sigma^{-1}(X_i - X_j)$, where Σ is the variance covariance matrix.
- Propensity score: $D_{ij} = |e_i - e_j|$
- Linear propensity score: $D_{ij} = |\text{logit}(e_i) - \text{logit}(e_j)|$

Caliper matching

Oftentimes, you will see combinations of heuristics:

$$D_{ij} = \begin{cases} (X_i - X_j)'\Sigma^{-1}(X_i - X_j), & \text{if } |\text{logit}(e_i) - \text{logit}(e_j)| \leq c \\ \infty, & \text{if } |\text{logit}(e_i) - \text{logit}(e_j)| > c \end{cases}$$

Selection of caliper may vary depending on methods used, but Rosenbaum and Rubin (1985) suggest 0.25 standard deviations of linear propensity score.

Good settings for matching methods

- Lots of data, with rich pre-treatment covariates
- When there is considerable imbalance between number of treated and control units
- When outcome values are not yet available (e.g., in an experiment), and matching used to guide follow up
- In settings with cost constraints
- Randomized trials in smaller sample settings (match units a priori, then randomize)

Limitations/challenges of methods

- Variance estimation: this is still a hot topic (e.g., Imbens, 2014)
- Selection among potential methods
- Strong ignorability is untestable
- What to do in presence of missing data (e.g., multiple imputaion)

Practical guidance

- If wanting to estimate ATE, estimate via IPTW or full matching (Hansen, 2004)
- If goal is ATT: nearest neighbor matching, subclassification
- Examine balance of pre-treatment covariates; it is possible for matching methods to fail, since on units weights are determined by convex hull.

Introduction to causal inference for data scientists

Differences-in-differences, regression discontinuity

Michael Gill

2018-03-27

Center for Data Science | New York University

Today

Today's references

References:

- MW, Chapters 6-7
- Lee, D., and Lemieux, T. (2010). "Regression discontinuity designs in economics" Journal of Economic Literature.
- Calonico, S., Cattaneo, M., and Titiunik, R. (2014). "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." Econometrica.

Today's goals

- Motivate time series analysis via Diff-in-Diff
- Introduce regression discontinuity (and related designs)
- Think about earlier experimental assumptions in this light
- Briefly discuss problems that commonly arise (and ways to address)

Differences-in-differences

Basic idea

- The idea in difference-in-difference (DID) is that heterogeneity is captured by a time-invariant additive factor.
- By looking at the double difference in outcomes treated minus control, after minus before, we can get rid of this factor, and hence, evaluate the treatment effect.
- The DID estimator is therefore

$$DID = \mathbb{E}\left[Y^1(t_1) - Y^1(t_0) \,|D = 1\right] - \mathbb{E}\left[Y^0(t_1) - Y^0(t_0) \,|D = 0\right]$$

Regression framework

- Specification:

$$Y_i^d(t) = \alpha(t) + \rho(t)\, d + \varepsilon_i(t)$$

and one can assume $\alpha(t) = \alpha + \gamma t$ and $\rho(t) = \rho + \beta(t)\, t$, that is

$$Y_i^d(t) = (\alpha + \gamma t) + (\rho + \beta(t)\, t)\, d + \varepsilon_i(t),$$

where $\mathbb{E}[\varepsilon_i(t)] = 0$ and $cov(\varepsilon_i(t), D_i) = 0$.

- Hence, if $t_0 = 0$ and $t_1 = 1$,

$$Y_i^d(t_1 = 1) - Y_i^d(t_0 = 0) = \gamma + \beta d + \varepsilon_i(t_1 = 1) - \varepsilon_i(t_0 = 0),$$

and hence the treatment effect is estimated by β.

Parallel trend assumption

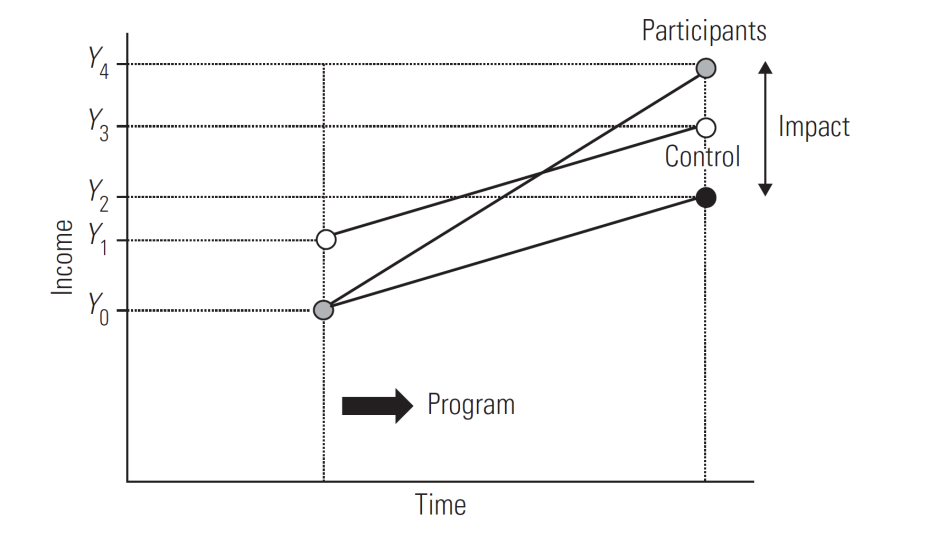
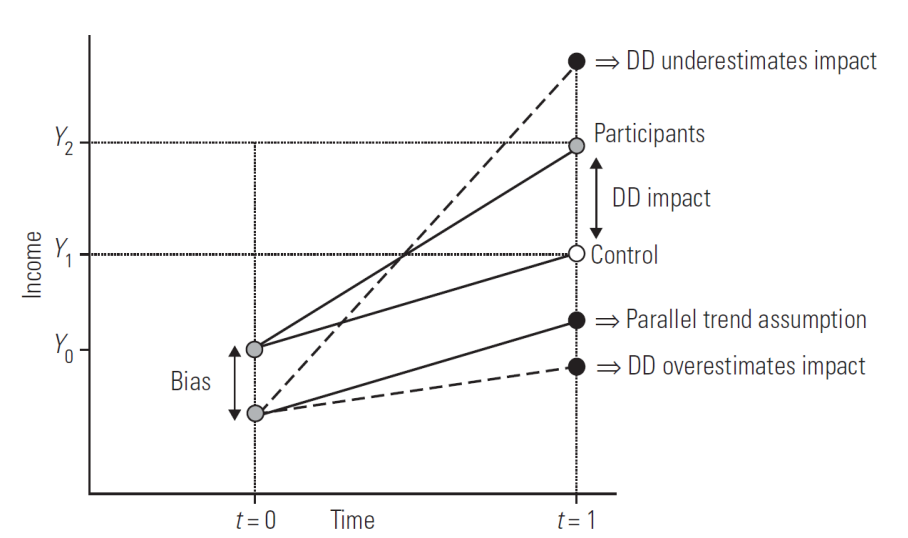


Figure 1: Source: Khandker et al. (2010)

DID: possible biases

- Time-varying unobserved heterogeneity:



Panel fixed-effects

- A generalization of diff-in-diffs is called *panel fixed-effects models*. In these models,

$$Y_i^*(t) = \rho D_{it} + \gamma X_{it} + \eta_i + \varepsilon_i(t)$$

  where there is time-invariant unobservable heterogeneity η_i at the unit level.
- By differentiation, if $\Delta W_{it} = W_{it} - W_{i(t-1)}$, one has

$$\Delta Y_i^*(t) = \rho \Delta D_{it} + \gamma \Delta X_{it} + \Delta \varepsilon_i(t)$$

  and the treatment effect, measured by ρ, can be obtained by OLS. Standard errors should be corrected for autocorrelation.

Back to Card-Krueger (1)

- The Card-Krueger study is emblematic of the diff-in-diffs methodology.
- Two groups: control (PA) and treatment (NJ); before / after raise of NJ minimum wage.
- Focus on Full Time Equivalent employment.

Back to Card-Krueger (2)

TABLE 3—AVERAGE EMPLOYMENT PER STORE BEFORE AND AFTER THE RISE IN NEW JERSEY MINIMUM WAGE

	Stores by state			Stores in New Jersey[a]			Differences within NJ[b]	
Variable	PA (i)	NJ (ii)	Difference, NJ − PA (iii)	Wage = $4.25 (iv)	Wage = $4.26–$4.99 (v)	Wage ≥ $5.00 (vi)	Low–high (vii)	Midrange–high (viii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	−2.89 (1.44)	19.56 (0.77)	20.08 (0.84)	22.25 (1.14)	−2.69 (1.37)	−2.17 (1.41)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	−0.14 (1.07)	20.88 (1.01)	20.96 (0.76)	20.21 (1.03)	0.67 (1.44)	0.75 (1.27)
3. Change in mean FTE employment	−2.16 (1.25)	0.59 (0.54)	2.76 (1.36)	1.32 (0.95)	0.87 (0.84)	−2.04 (1.14)	3.36 (1.48)	2.91 (1.41)
4. Change in mean FTE employment, balanced sample of stores[c]	−2.28 (1.25)	0.47 (0.48)	2.75 (1.34)	1.21 (0.82)	0.71 (0.69)	−2.16 (1.01)	3.36 (1.30)	2.87 (1.22)
5. Change in mean FTE employment, setting FTE at temporarily closed stores to 0[d]	−2.28 (1.25)	0.23 (0.49)	2.51 (1.35)	0.90 (0.87)	0.49 (0.69)	−2.39 (1.02)	3.29 (1.34)	2.88 (1.23)

Notes: Standard errors are shown in parentheses. The sample consists of all stores with available data on employment. FTE (full-time-equivalent) employment counts each part-time worker as half a full-time worker. Employment at six closed stores is set to zero. Employment at four temporarily closed stores is treated as missing.

[a]Stores in New Jersey were classified by whether starting wage in wave 1 equals $4.25 per hour ($N = 101$), is between $4.26 and $4.99 per hour ($N = 140$), or is $5.00 per hour or higher ($N = 73$).

[b]Difference in employment between low-wage ($4.25 per hour) and high-wage ($\geq$ $5.00 per hour) stores; and difference in employment between midrange ($4.26–$4.99 per hour) and high-wage stores.

[c]Subset of stores with available employment data in wave 1 and wave 2.

[d]In this row only, wave-2 employment at four temporarily closed stores is set to 0. Employment changes are based on the subset of stores with available employment data in wave 1 and wave 2.

Figure 3:

Key assumptions in difference in differences

- Unconfoundedness
- Parallel trends
- SUTVA

If assignment to treatment correlated with unobserved (non-additive) heterogeneity, diff-in-diff estimator will be biased.

Parallel trends: necessary for unbiased estimation, since treatment effects are estimated given shifts in gaps between groups.

SUTVA needed for standard reasons

Extensions

One can generalize to settings with more than two time periods, more than two treatment groups (triple-difference, panel fixed effects).

Flexible framework for large-scale interventions (e.g., tax policies, medical insurance,)

Challenges: standard errors need to be adjusted (e.g., Moulton, 1986). Can use permutation tests, cluster/block bootstraps to address concerns. More discussion in Bertrand et al. (2004).

"Ashenfelter's dip"

More on parallel trends

Is this a testable assumption? Perhaps, somewhat:

What if parallel trends not satisfied?

Semi-parametric weighting schemes introduced by Abadie (2005).

More on this next week, as we get into synthetic controls.

Regression discontinuity

- Idea: sometimes, a threshold determines eligibility to program participation. Regression discontinuity (RD) compares participants and nonparticipants locally in a neighborhood of the threshold.
- Allows for observed and nonobserved heterogeneity.
- Mostly applicable when the criterion is clear and explicit.
- Discontinuities examples:
  - households with landholdings less than a certain size are eligible to microcredits
  - students above a certain standardized score are eligible to scholarships
  - individuals above a certain age are eligible to certain pension programs

Regression discontinuity

- Take an example where treatment=college scholarship, X is standardized score, Y^1 is earnings if was eligible to scholarship, Y^0 is earnings if not.
- Recall $Y_i^* = D_i Y_i^1 + (1 - D_i) Y_i^0$.
- Two types of regression discontinuity:
  - Sharp regression discontinuity: $D_i = 1\{X_i \geq \bar{x}\}$, in which case $e_i(x) = 1\{x \geq \bar{x}\}$
  - Fuzzy regression discontinuity design: $\lim_{x \to \bar{x}^+} e_i(x) \neq \lim_{x \to \bar{x}^-} e_i(x)$

Sharp regression discontinuity: example (1)

- What is the effect of incumbency on electoral outcomes? i.e., do electors want change or continuity? In other words, what is the probability of a Democrat winning a House election given that a Democrat won the previous election?
- Lee (2008) studies a RDD using the margin of victory at the previous election. Here, sharp **discontinuity**: if margin>0, then a Dem was elected, if margin<0, then a Dem was not elected.
- If there is no incumbency effect, then there should be no jump. The jump is informative about the incumbency effect.

Sharp regression discontinuity: example (2)

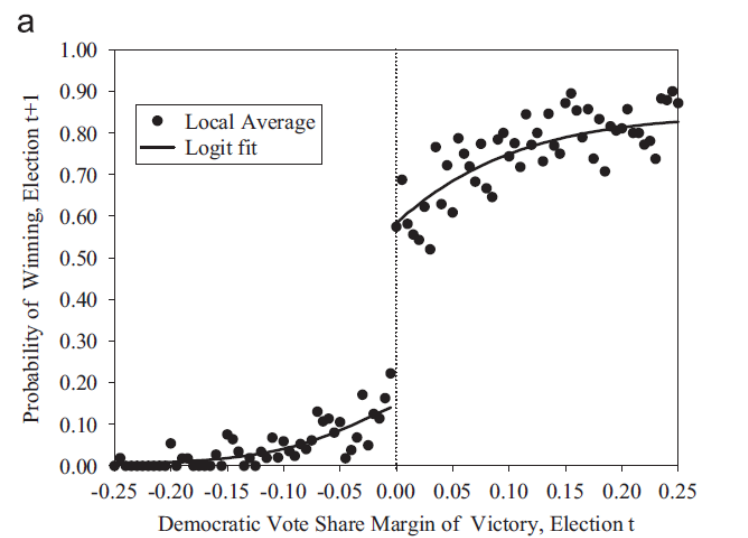


Figure 4: Source: Lee (2007).

Sharp regression discontinuity: treatment effect

- Key assumption: $\mathbb{E}\left[Y^0|X=x\right]$ and $\mathbb{E}\left[Y^1|X=x\right]$ are continuous at $\bar{x}$.
- In this case, the treatment effect is estimated by

$$\delta_{SRD} = \lim_{x\to\bar{x}^+} \mathbb{E}\left[Y^*|X=x\right] - \lim_{x\to\bar{x}^-} \mathbb{E}\left[Y^*|X=x\right].$$

- Note that this only measures the treatment effect *locally* at $\bar{x}$.

Sharp regression discontinuity: locally linear regression

- Rewrite the previous situation as $D_i = 1\{X_i \geq \bar{x}\}$, and

$$Y_i^1 = \alpha^1 + \beta^1 X_i + \varepsilon_i$$
$$Y_i^0 = \alpha^0 + \beta^0 X_i + \varepsilon_i$$

- Then one can estimate α^1 and β^1 by local linear regression: take a kernel $K(u)$ such that $K(u) = 0$ if $u < 0$, for instance $K(u) = \exp(-u^2/2)\,1\{u \geq 0\}$ and compute $(\hat{\alpha}^1, \hat{\beta}^1)$ that solve

$$\min_{\alpha,\beta} \sum_i K\left(\frac{X_i - \bar{x}}{h}\right)(Y_i - \alpha - \beta X_i)^2,$$

  where h controls the bandwidth size.
- For $\hat{\alpha}^0$ and $\hat{\beta}^0$, do the same with another kernel $K(u)$ such that $K(u) = 0$ for $u > 0$.
- The estimator δ_{SRD} is then obtained as $\hat{\delta}_{SRD} = \hat{\alpha}^1 - \hat{\alpha}^0$.

Fuzzy regression discontinuity: example (1)

- Van der Klauuw (2002) uses financial aid data from an East Coast college. Financial aid officers rank students according to a score given by

$$S = \phi_0 \times (\text{first three digit of SAT score}) + \phi_1 \times GPA$$

  and rank them into four groups divided by cutoffs $S_1 < S_2 < S_3$.
- Groups determine which type of financial aid students are eligible to; however adjustments (based on merit, affirmative action, or need) are possible.
- The FRD approach will consist in comparing individuals with close scores on each side of the cutoff.

Fuzzy regression discontinuity: example (2)

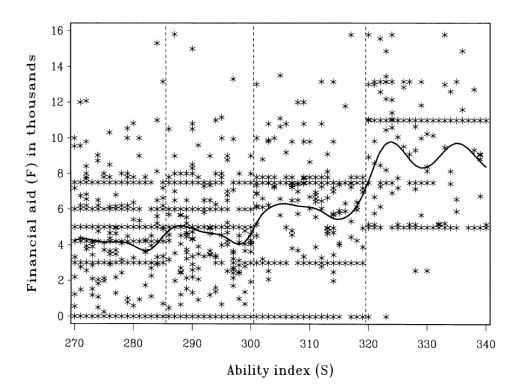


Figure 5: Source: van der Klauuw (2002)

Fuzzy regression discontinuity: treatment effect

- Assume $\mathbb{E}\left[Y^0|X=x\right]$ and $\mathbb{E}\left[Y^1|X=x\right]$ are continuous at $\bar{x}$, and the propensity score $e\left(x\right)$ is discontinuous at $\bar{x}$.
- The treatment effect is estimated by

$$\delta_{FRD} = \frac{\lim_{x\to\bar{x}^+}\mathbb{E}\left[Y^*|X=x\right] - \lim_{x\to\bar{x}^-}\mathbb{E}\left[Y^*|X=x\right]}{\lim_{x\to\bar{x}^+}\mathbb{E}\left[D|X=x\right] - \lim_{x\to\bar{x}^-}\mathbb{E}\left[D|X=x\right]}.$$

- Note that in the case of sharp regression discontinuity, the denominator is 1, and one recovers the formula for δ_{SRD}.

Fuzzy regression discontinuity and unconfoundedness

- Recall the definition of unconfoundedness: $(Y^0, Y^1) \perp D|X$, hence

$$\mathbb{E}\left[Y^d|D=1, X=x\right] = \mathbb{E}\left[Y^d|D=0, X=x\right] = \mathbb{E}\left[Y^d|X=x\right],$$

  and hence

$$\begin{aligned}&\mathbb{E}\left[Y^1|X=x\right] - \mathbb{E}\left[Y^0|X=x\right]\\ &= \mathbb{E}\left[Y^*|D=1, X=x\right] - \mathbb{E}\left[Y^*|D=0, X=x\right].\end{aligned}$$

- Unconfoundedness does not require a discontinuity in the treatment; instead it assumes that similar units will receive similar treatment.

Locally linear regression

- As for SRD, take a kernel $K(u)$ such that $K(u) = 0$ if $u < 0$, and compute $(\hat{\alpha}^1, \hat{\beta}^1)$ that solve

$$\min_{\alpha,\beta} \sum_i K\left(\frac{X_i - \bar{x}}{h}\right)(Y_i - \alpha - \beta X_i)^2 ,$$

  and $\left(\hat{a}^1, \hat{b}^1\right)$ that solve

$$\min_{\alpha,\beta} \sum_i K\left(\frac{X_i - \bar{x}}{h}\right)(D_i - a - bX_i)^2 ,$$

  and similarly for $(\hat{\alpha}^0, \hat{\beta}^0)$ and $\left(\hat{a}^0, \hat{b}^0\right)$.
- δ_{FRD} is then estimated by

$$\delta_{FRD} = \frac{\hat{\alpha}^1 - \hat{\alpha}^0}{\hat{a}^1 - \hat{a}^0}.$$

Example: Duflo-Hanna-Ryan (2012)

- Study on the effect of incentives in Indian primary school.
- Teacher's salary was a function of attendance:
  - Rs. 500 if attended fewer than 10 days in a month, and
  - Rs. 50 for any additional day attended that month.
- Expect discontinuity around the 10 days cutoff.

RDD around the 10 day cutoff

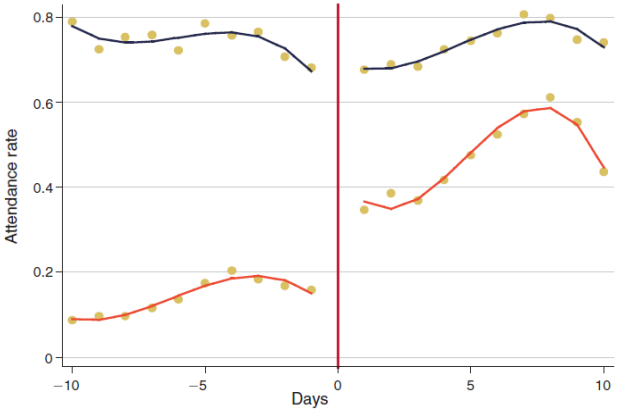


FIGURE 3. RDD REPRESENTATION OF TEACHER ATTENDANCE AT THE START AND END OF THE MONTH

Notes: The top lines represent the months in which the teacher is in the money, while the bottom lines represent the months in which the teacher is not in the money. The estimation includes a third-order polynomial of days on the left and right side of the change of month.

Figure 6:

Other fun RDD applications in data science

- Yelp reviews: effect of stars on consumer demand
- Effects of page location on product demand

What about sorting?

McCrary (2008) for bunching heuristics.

RDD design is equivalent to local level-randomization, so sorting would be equivalent to local selection into treatment.

Evaluate distribution of pre-treatment covariates for those on units on either side of discontinuity.

Introduction to causal inference for data scientists

Extending differences-in-differences

Michael Gill

2018-04-03

Center for Data Science | New York University

Today

Today's references

References:

- Athey, S. and Imbens, G. (2006) "Identification and Inference in Non-Linear Difference-in-Differences Models," Econometrica.
- Abadie, A., Diamond, A., and Hainmueller, J. (2010) "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." JASA.

Today's goals

- Extensions of diff-in-diff
- What if we care about non-linear differences (e.g., "changes-in-changes", quantile DID)?
- Bridge discussions of matching with DID
- What happens if only one unit is treated?
- How to approach inference?

Extending differences-in-differences

Recall from before

- Last week we discussed how a difference-in-difference approach can account for unobserved confounding, by restricting the way in which confounding can influence outcomes.
- Consider a model of the form:

$$Y_{it} = \delta_{it} D_{it} + \mu_i + \gamma_t + \varepsilon_{it}$$

where potential outcomes for unit i in t are given by

$$Y_{it}^{0} = \mu_i + \gamma_t + \varepsilon_{it} \quad (1)$$

$$Y_{it}^{1} = \delta_{it} + \mu_i + \gamma_t + \varepsilon_{it} \quad (2)$$

- Here, $\delta_{it} = Y_{it}^{1} - Y_{it}^{0}$.

More on diff-in-diff

$$Y_{it} = \delta_{it} D_{it} + \mu_i + \gamma_t + \varepsilon_{it}$$

- Because μ_i is fixed, it is not independent of D_{it}.
- We will need to assume that $E[\varepsilon_{it}|D_{it}] = E[\varepsilon_{it}]$.
- **Note:** identification also possible assuming $E[\Delta\varepsilon_{it}|D_{it}] = E[\Delta\varepsilon_{it}]$

Diff-in-diff in panel data model

- Given structure from before, we can write outcomes in the following manner:

$$
\begin{aligned}
E[Y_{i1}|D_{i1}=1] &= E[\delta_{i1}|D_{i1}=1] + E[\mu_i|D_{i1}=1] + \gamma_1 + E[\varepsilon_{i1}] \\
E[Y_{io}|D_{i1}=1] &= E[\mu_i|D_{i1}=1] + \gamma_o + E[\varepsilon_{io}] \\
E[Y_{i1}|D_{i1}=0] &= E[\mu_i|D_{i1}=0] + \gamma_1 + E[\varepsilon_{i1}] \\
E[Y_{io}|D_{i1}=0] &= E[\mu_i|D_{i1}=0] + \gamma_o + E[\varepsilon_{io}]
\end{aligned}
$$

Identifying ATT in panel framework

$$
\begin{aligned}
\delta_{\text{ATT}} &= E[\delta_{i1}|D_{i1}=1] \\
&= \Big[E[Y_{i1}|D_{i1}=1] - E[Y_{i1}|D_{i1}=0]\Big] - \Big[E[Y_{i0}|D_{i1}=1] - E[Y_{i0}|D_{i1}=0]\Big] \\
&= \Big[E[\Delta Y_{i1}|D_{i1}=1]\Big] - \Big[E[\Delta Y_{i1}|D_{i1}=0]\Big]
\end{aligned}
$$

Which highlights the difference-in-differences component.

- Given: $E[\Delta\varepsilon_{i1}|D_{it}] = E[\Delta\varepsilon_{i1}] \implies E[\Delta Y^0_{i1}|D_{it}=1] = E[\Delta Y^0_{i1}|D_{it}=0]$
- Meaning: in absence of treatment, expected outcomes (i.e., differences) for treated and control would be the same. This is equivalent to invoking common trends assumption.

Common extensions

- More than two time periods
- Unit-specific time trends—e.g., Bertrand et al. (2004). Note: need at least 3 time periods for this.
    - Good example: Angrist and Pischke (2009), re-analyzing study from Besley and Burgess (2004). Presence of unit-time trends removes DID estimate.
- Relaxations of common-trends: Abadie (2005)
- If multiple pre-treatment time periods exist, test common trends using placebo diff-in-diff.

Athey and Imbens (2006) (1)

(Very briefly. For much more, see paper.)

- Diff-in-diff estimates are not invariant to non-linear transformations of outcome variable (e.g., $\log(\text{Clicks}_{it})$ vs. Clicks_{it})
- Generalizes the changes-in-changes model, of which diff-in-diff is a special case.
- "[A]llow the effects of both time and the treatment to differ systematically across individuals..."
- Nice implementation in `qte` package in R.

Athey and Imbens (2006) (2)

$$\delta_{\text{CIC}} = E[Y_{11}^{I} - Y_{11}^{N}] = E[Y_{11}^{I}] - E[k^{\text{CIC}}(Y_{10})] \quad (3)$$

$$= E[Y_{11}^{I}] - E[F_{Y,01}^{-1}(F_{Y,00}(Y_1 0)) \quad (4)$$

Basic idea:

- Begin with a value of y, and find its quantile q in Y_{10}
- Find the corresponding quantile for y in the Y_{00} distribution, $q' = F_{Y,00}(y)$
- Find change in y from k^{CIC}, where $k^{CIC}(y) = F_{Y,01}^{-1}(F_{Y,00}(y))$, by finding value for y at quantile q' in Y_{01} to yield:

$$\Delta^{\text{CIC}} = F_{Y,01}^{-1}(q\prime) - F_{Y,00}^{-1}(q\prime) = F_{Y_{Y,01}}^{-1}(F_{Y,00}(y)) - y$$

- Last: compute counterfactual $Y_{11}^{N} = y + \Delta^{\text{CIC}}$, s.t. $k^{\text{CIC}}(y) = y + \delta^{\text{CIC}} = F_{Y,01}^{-1}(F_{Y,00}(y))$

Athey and Imbens (2006) (3)

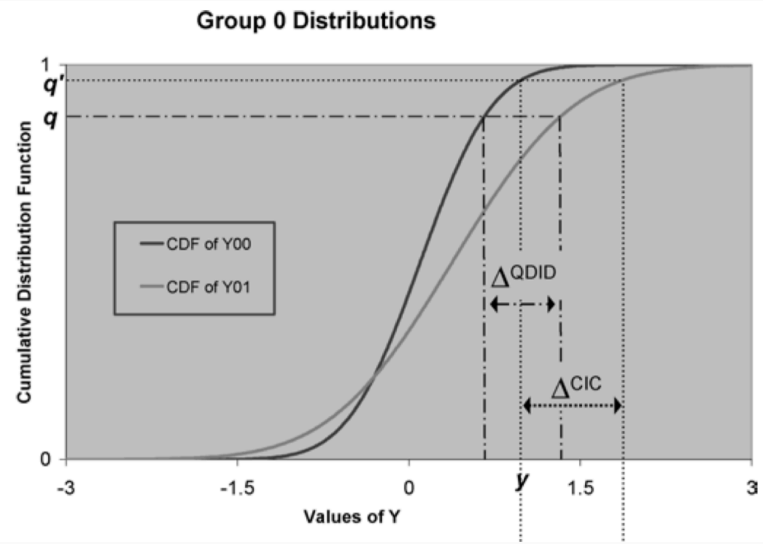


Figure 1: Source: Athey and Imbens (2006).

Athey and Imbens (2006) (4)

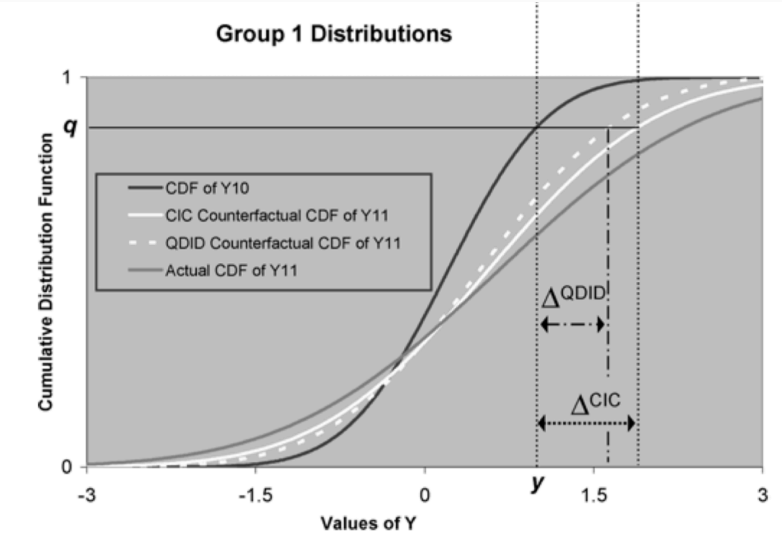


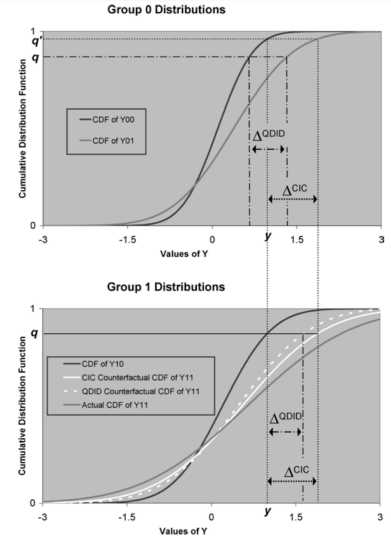
Figure 2: Source: Athey and Imbens (2006).

Athey and Imbens (2006) (5)

1. Given y, find its quantile q in Y_{10}
2. Find corresponding quantile for y in Y_{00}, $q' = F_{Y,00}(y)$
3. Find change in y from k^{CIC}, where $k^{CIC}(y) = F^{-1}_{Y,01}(F_{Y,00}(y))$, by finding value for y at quantile q' in Y_{01} to yield:

$$\begin{aligned}\Delta^{\text{CIC}} &= F^{-1}_{Y,01}(q\prime) - F^{-1}_{Y,00}(q\prime) \\ &= F^{-1}_{Y_{Y,01}}(F_{Y,00}(y)) - y\end{aligned}$$

4. Compute counterfactual $Y^N_{11} = y + \Delta^{\text{CIC}}$, s.t. $k^{\text{CIC}}(y) = y + \delta^{\text{CIC}} = F^{-1}_{Y,01}(F_{Y,00}(y))$



Synthetic controls

The general setting (1)

- As in diff-in-diff, oftentimes we are interested in effects of interventions on aggregate units, where a set of control units are used to estimate counterfactual trajectory.
    - "Comparative case studies"
- But what if average of control units doesn't well-approximate the treated unit (before the intervention)?

The general setting (2)

- We are interested in effect of an intervention
- Set of treated units (possibly only one)
- Set of control units (numerous)
- Observe pre-treatment and post-treatment measures of outcomes for all units
- Preferably, numerous pre- and post-treatment periods

Case study: effect of tobacco tax on cigarette sales in CA

- Taken from Abadie et al. (2010).
- Setup to problem originally formulated in Abadie et al. (2003), which looked at effects of terrorism on economic growth in Basque country.

Problem

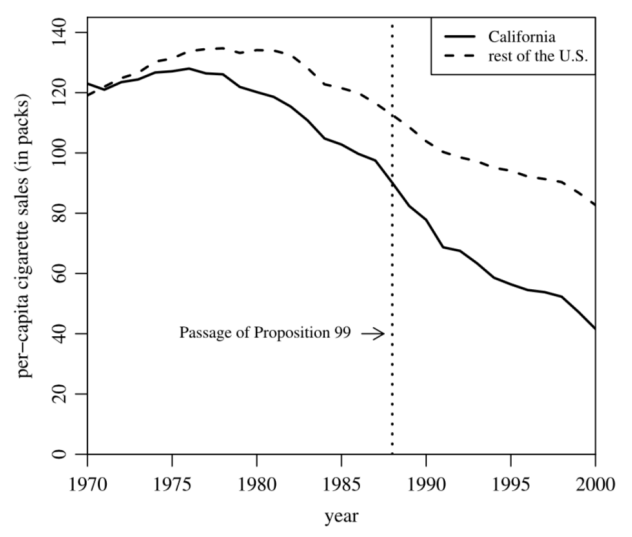


Figure 3: Source: Abadie et al. (2010).

Setup and notation

- WLOG, assume there is only one treated unit
- There are $J + 1$ total units, and time periods $1, 2, \ldots, T$
- The treated unit receives treatment in periods $T_0 + 1, \ldots, T$
- J other units are potential controls ("donor pool")
- The synthetic control estimator is

$$\widehat{\delta}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}$$

Recall: this should look fairly familiar to the matching estimator motivated several lectures ago.

How to construct weights?

How to construct weights?

There are many choices!

- Let $w = (w_2, \ldots, w_{j+1})'$
- Weights are $w_j \geq 0$ for $j = 2, \ldots, J+1$, and $\sum_{j=2}^{J+1} w_j = 1$.
- X_1 is a $k \times 1$ vector of pre-treatement variables for treated unit
- X_0 is a $k \times J$ matrix of same variables for donor pool
- The vector $w^* = (w_2^*, \ldots, w_{J+1}^*)'$ is chosen to minimize $||X_1 - X_0 w||$ (subject to specified constraints)

Weighted Euclidean norm

- Weighted Euclidean norm:

$$||X_1 - X_0 w|| = \sqrt{(X_1 - X_0 w)' V (X_1 - X_0 w)}$$

- V is diagonal matrix with non-negative values, which control for relative importance of predictors
- Other choices are possible (e.g., Mahalanobis distance), so long as V is appropriately defined (symmetric, positive semi-definite)
- In many cases, solution is unique (e.g., X_1 not in convex hull of X_0).

Back to example: tobacco tax

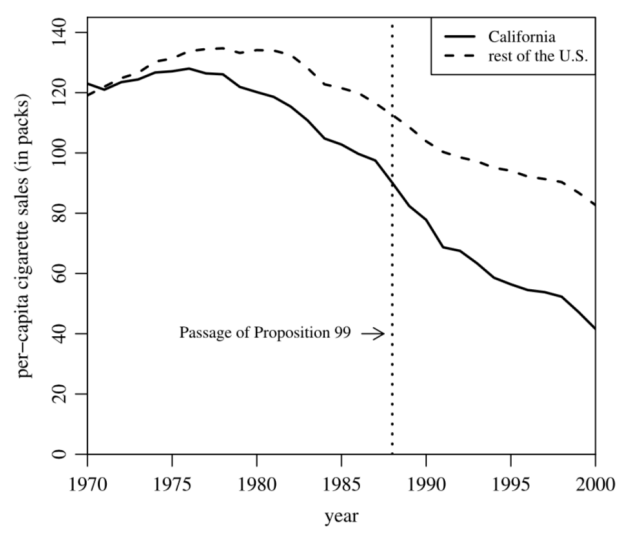


Figure 4: Source: Abadie et al. (2010).

Estimated weights in tobacco example

Table 2. State weights in the synthetic California

State	Weight	State	Weight
Alabama	0	Montana	0.199
Alaska	–	Nebraska	0
Arizona	–	Nevada	0.234
Arkansas	0	New Hampshire	0
Colorado	0.164	New Jersey	–
Connecticut	0.069	New Mexico	0
Delaware	0	New York	–
District of Columbia	–	North Carolina	0
Florida	–	North Dakota	0
Georgia	0	Ohio	0
Hawaii	–	Oklahoma	0
Idaho	0	Oregon	–
Illinois	0	Pennsylvania	0
Indiana	0	Rhode Island	0
Iowa	0	South Carolina	0
Kansas	0	South Dakota	0
Kentucky	0	Tennessee	0
Louisiana	0	Texas	0
Maine	0	Utah	0.334
Maryland	–	Vermont	0
Massachusetts	–	Virginia	0
Michigan	–	Washington	–
Minnesota	0	West Virginia	0
Mississippi	0	Wisconsin	0
Missouri	0	Wyoming	0

Figure 5: Source: Abadie et al. (2010).

Pre-treatment covariate balance from synthetic matches

Table 1. Cigarette sales predictor means

Variables	California Real	California Synthetic	Average of 38 control states
Ln(GDP per capita)	10.08	9.86	9.86
Percent aged 15–24	17.40	17.40	17.29
Retail price	89.42	89.41	87.27
Beer consumption per capita	24.28	24.20	23.75
Cigarette sales per capita 1988	90.10	91.62	114.20
Cigarette sales per capita 1980	120.20	120.43	136.58
Cigarette sales per capita 1975	127.10	126.99	132.81

NOTE: All variables except lagged cigarette sales are averaged for the 1980–1988 period (beer consumption is averaged 1984–1988). GDP per capita is measured in 1997 dollars, retail prices are measured in cents, beer consumption is measured in gallons, and cigarette sales are measured in packs.

Figure 6: Source: Abadie et al. (2010).

Synthetic control vs. observed series

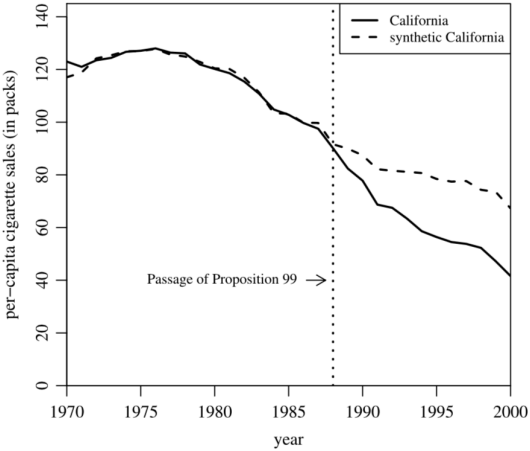


Figure 2. Trends in per-capita cigarette sales: California vs. synthetic California.

Figure 7: Source: Abadie et al. (2010).

The period-level differences

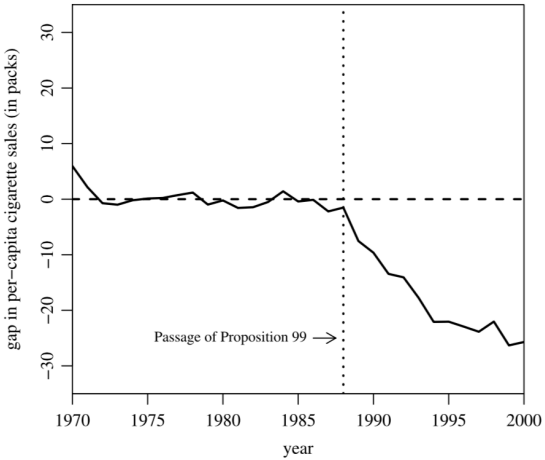


Figure 3. Per-capita cigarette sales gap between California and synthetic California.

Figure 8: Source: Abadie et al. (2010).

Inference? Placebo tests

- For each of the J control units, repeat synthetic control estimation (but pretend as if control units are treated), and estimate treatment effect
- Compare collection of J placebo estimates against original estimate(s)
- Compare rank of $\widehat{\delta}_{it}$ against placebo estimates.
- Akin to a randomization inference approach.

Results of placebo tests (1)

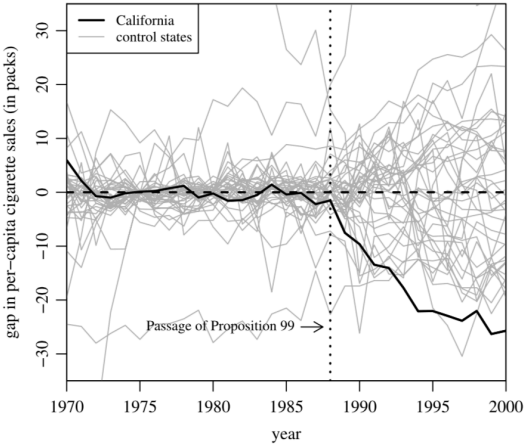


Figure 4. Per-capita cigarette sales gaps in California and placebo gaps in all 38 control states.

Figure 9: Source: Abadie et al. (2010).

Results of placebo tests (2)

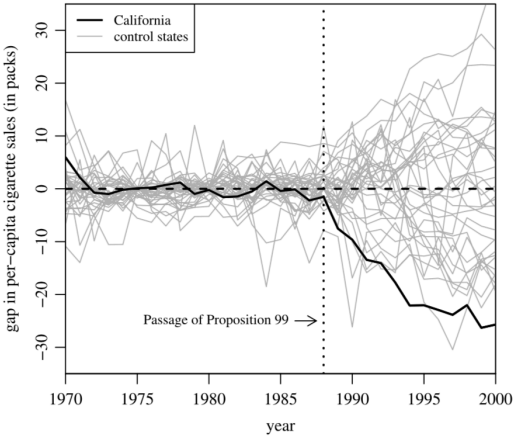


Figure 5. Per-capita cigarette sales gaps in California and placebo gaps in 34 control states (discards states with pre-Proposition 99 MSPE twenty times higher than California's).

Figure 10: Source: Abadie et al. (2010).

Results of placebo tests (3)

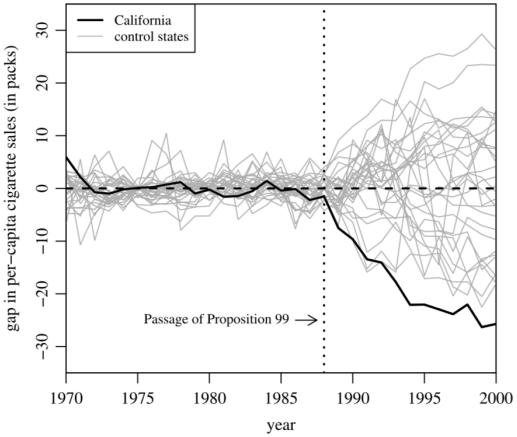


Figure 6. Per-capita cigarette sales gaps in California and placebo gaps in 29 control states (discards states with pre-Proposition 99 MSPE five times higher than California's).

Figure 11: Source: Abadie et al. (2010).

Results of placebo tests (4)

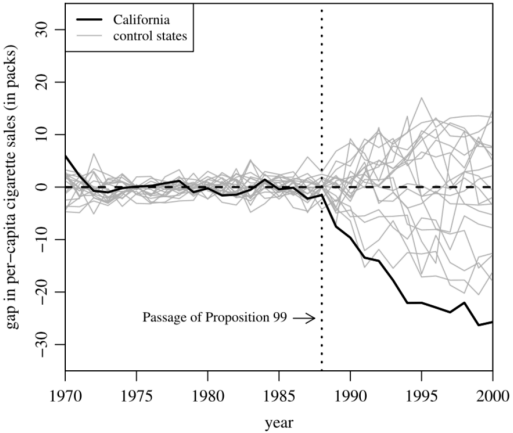


Figure 7. Per-capita cigarette sales gaps in California and placebo gaps in 19 control states (discards states with pre-Proposition 99 MSPE two times higher than California's).

Figure 12: Source: Abadie et al. (2010).

Extensions of synthetic control

- Abadie and L'Hour (2017): estimator minimizes:

$$||X_1 - X_0 w||^2 + \lambda \sum_{j=2}^{J+1} w_j ||X_j - X_1||^2$$

  , with $\lambda > 0$. Helps with uniqueness via sparse penalty.
- Athey et al. (2017): matrix completion methods, with missing data
- Hahn and Shi (2017): superpopulation frequentist methods for inference
- Doudchenko and Imbens (2017): allows weights to vary overtime

Next lecture

- We will think about settings when there are possibly many controls relative to time periods (i.e., $N >> T_0$)
- Or relatively few controls relative to time periods (i.e., $T_0 >> N$)

Introduction to causal inference for data scientists

High-dimensional models

Michael Gill

2018-04-10

Center for Data Science | New York University

Today

Today's references

References:

- Belloni, A., Chernozhukov, V., Fernandez-Val, I., and Hansen, C. (2017). "Program Evaluation and Causal Inference with High-Dimensional Data." Econometrica.
- Brodersen, K., Galluser, F., Koehler, J., Remy, N., and Scott, S. (2015). "Inferring causal impact using Bayesian structural time series models." Annals of Applied Statistics.
- Varian, H. (2016). "Causal inference in economics and marketing," PNAS.

Today's goals

- Consider cases in which there may be many possible controls; in IV, there may be many instruments.
- Motivate model selection in these contexts (cross validation vs. rate-optimal penalties)
- Focus will be on lasso-type regularization methods (and their close cousins)
- Consider synthetic controls and IV in high-dimensional settings

High-dimensional models

Recall from before

- Last class we discussed that challenges may arise when $N >> p$ (e.g., more units than periods, which make evaluation of parallel trends difficult)
- Also difficult when $p >> N$ (e.g., Census data, text/language data, social media data)
- Richer data sources call for new techniques to apply to classic settings

Framework

$$Y_i^D = \delta D_i + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$$

- Assume, as in prior weeks, we have data generated from above.
- $E[Y_i^1 - Y_i^0] = \delta$
- Here, the collection x are our pre-treatment "control" variables.
- In the IV framework, we might also invoke: $E[\varepsilon_i | x_i, Z_i] = 0$

Which “controls” should be in our equation?

- Concerns of overfitting and underfitting
- Excluding relevant controls can lead to bias
- Including irrelevant controls can increase variance
- If there are many “non-zero” controls, model fit may not be feasible (degrees of freedom)
- But if number of key controls is sparse, we have some options

Ad hoc approaches to selecting relevant regressors

- t-test for covariates significantly correlated with outcome, then drop insignificant regressors
- F-tests
- Cross-validation techniques (but over which prediction function?)
- “Wave your hands” and ignore the problem (a traditional choice)

Each of the above can go wrong.

A sparse data-generation process

- $y_i = x_i' \beta_0 + \varepsilon_i$
- $E[\varepsilon_i x_i] = 0$
- $x_i = (x_{ij}, j = 1, \ldots, p)'$, $\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$ (standardization)
- Note: this DGP can be written down with $p >> N$
- Sparsity: $s := ||\beta_0||_0 = \sum_{j=1}^{p} 1\{\beta_{0j} \neq 0\} << n$
- **Key idea:** number of relevant covariates smaller than sample size.

Even with seemingly low p

- Dimensionality can easily change (transformations of variables, interactions)
- Oftentimes we knowingly simplify model because of d.o.f. concerns.

Approximate sparsity

$$y_i = x_i' \beta_0 + r_i + \varepsilon_i$$

- Here, r_i is a regularization term ("bias"), which is non-zero.
- Can be shown that the bias is smaller than the estimation error, which in the limit is:
  - $$\frac{s \log p}{n} \to 0$$
  - $$\sqrt{\frac{1}{n} \sum_{i=1}^{n} r_i^2} \leq \sigma \sqrt{\frac{s}{n}} \to 0$$

Lasso (1)

- Before, motivated objective function with form:

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i'\beta\right)^2 + \lambda||\beta||_0, \text{ where } ||\beta||_0 = \sum_{j=1}^{p}\mathbf{1}\{\beta_{0j} \neq 0\}$$

  - The penalty parameter λ is a scalar that controls the sparsity

- Solution is not feasible with large p (NP hard)
- Instead, we estimate

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i'\beta\right)^2 + \lambda||\beta||_1, \text{ where } ||\beta||_1 = \sum_{j=1}^{p}|\beta_j|$$

Lasso (2)

Least absolute shrinkage and selection operator: Tibshirani (1996).

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i'\beta\right)^2 + \lambda||\beta||_1$$

$$\min_{\beta}\left\{||\mathbf{y} - \mathbf{x}\beta||_2^2 + \lambda||\beta||_1\right\}$$

Reasonable tradeoff, relative to l_0 norm. Computationally feasible.

Lasso variants

- Elastic net: Zou and Hastie (2005)

$$\min_{\beta}\left\{||\mathbf{y}-\mathbf{x}\beta||_2^2+\alpha||\beta||^2+(1-\alpha)||\beta||_1\right\},\ \text{where } \alpha=\frac{\lambda_2}{\lambda_1+\lambda_2}$$

- Adaptive lasso: Zou (2006)

$$\min_{\beta}\left\{||\mathbf{y}-\mathbf{x}\beta||_2^2+\lambda\sum_{j=1}^{p}\hat{w}_j|\beta_j|\right\}$$

- Group lasso: Yuan and Lin (2006), Friedman et al. (2010)

$$\min_{\beta}\left\{||\mathbf{y}-1-\sum_{l=1}^{L}\mathbf{x}_l\beta_l||_2^2+\lambda\sum_{l=1}^{L}\sqrt{p_l}||\beta_l||_2\right\}$$

- Square-root lasso: Belloni et al. (2011)

$$\min_{\beta}\left\{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-x_i'\beta)^2}+\lambda||\beta||_1\right\}$$

How to tune the penalty parameter?

- Two primary choices:
  1. Cross-validation techniques: Chetverikov et al. (2016)
  2. Rate-optimal penalty choices: Bickel et al. (2009)

(K-fold) cross-validation procedure

- Randomly divide data into K subsamples
- Holding out part k, fit model on remaining data, then predict on part k
- Repeat for all $k = 1, \ldots, K$
- Combine results:

$$CV_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} \mathscr{L}(k)$$

- Common loss is the mean-squared error: $\mathscr{L}(k) = \frac{1}{n_k} \sum_{i \in C_k} (y_i - \hat{y}_i)^2$.

Challenges with cross-validation

- In relatively small samples, behavior of CV is still sensitive to initial sample selection (i.e., sampling noise)
- If $p >> N$, validity of cross-validation for still an open area of reseach for lasso-type models
- Two good papers: Yu and Feng (2013); Chetverikov et al. (2016)

Chetverikov et al. (2016)

"In this paper, we derive a rate of convergence of the Lasso estimator when the penalty parameter λ for the estimator is chosen using K-fold cross-validation; in particular, we show that in the model with the Gaussian noise and under fairly general assumptions on the candidate set of values of λ, the prediction norm of the estimation error of the cross- validated Lasso estimator is with high probability bounded from above up to a constant by $(s \log p/n)^{1/2} \cdot \log^{7/8}(pn)$, where n is the sample size of available data, p is the number of covariates, and s is the number of non-zero coefficients in the model. Thus, the cross-validated Lasso estimator achieves the fastest possible rate of convergence up to a small logarithmic factor $\log^{7/8}(pn)$."

- Gives sparsity bound under general conditions for K-fold CV-lasso estimator

Rate-optimal selection of λ for lasso

- From Bickel et al. (2009):

$$\lambda = \sigma \cdot 2\sqrt{2\log(pn)/n}$$

- However, needs value for σ
- One approach: estimate σ via intitialization around standard deviation of sample mean (of outcome of interest). See: Belloni and Chernozhukov (2009).

Rate-optimal selection of λ for $\sqrt{lasso}$

- Rate-optimal penalty: $\sqrt{2\log(pn)/n}$
- Different objective function yields: globally convex, polynomial time, tuning free solution.
- Ideal for settings in which $n << p$, or when n is fairly small in general.
- Original article: Belloni, Chernozhukov, and Wang (2010)

Post-double selection procedure

Taken from: Belloni, Chernozhukov, and Hansen (2013)

Works in low-dimensional settings, and in high-dimensions with approximate sparsity in controls.

- Lasso y and x. Keep x if passes threshold.
- Lasso d and x. Keep x if passes threshold. (If an IV model, lasso z and x as well.)
- Taking union of selected covariates, refit desired model. Utilize standard CIs.

Simulation study from sparse data generation process

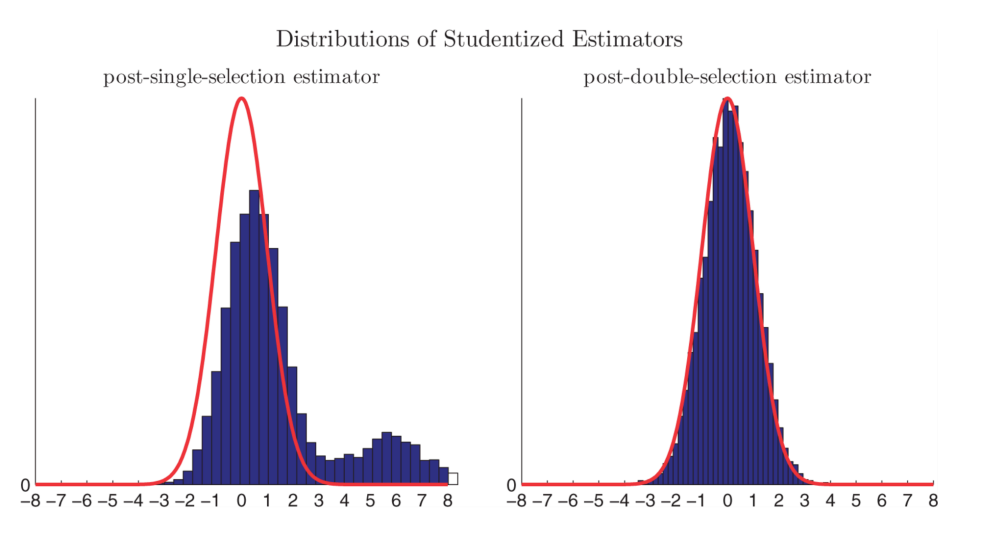


Figure 1: Source: Belloni et al. (2013)

Confidence intervals?

- Authors demonstrate that it is appropriate to use traditional standard errors after double-selection (e.g., conventional, cluster-robust, bootstrap)
- Related to: " the number of nonzero coefficients is an unbiased estimate for the degrees of freedom of the lasso" Zou, Hastie, and Tibshirani (2007)

Bayesian structural time series

BSTS

- Introduced in Brodersen et al. (2015)
- Parametric extension of synthetic control method
- Builds Bayesian state-space model, under assumptions of common panel-data framework
- Variable selection occurs via Spike-and-Slab penalization
- Effective model size has analogues to sparsity parameter in lasso
- Weights on variables (e.g., control time series) determined by proportion of time variable is included in model

Differences with traditional synthetic control

- (Model-based) extrapolation outside of convex hull of outcome
- Simulation-based method: more simulations, better inferences
- Credible posterior-inference on estimated counterfactual given variable selection
- Can generate predictions even in absence of control series.

Graphical model of BSTS

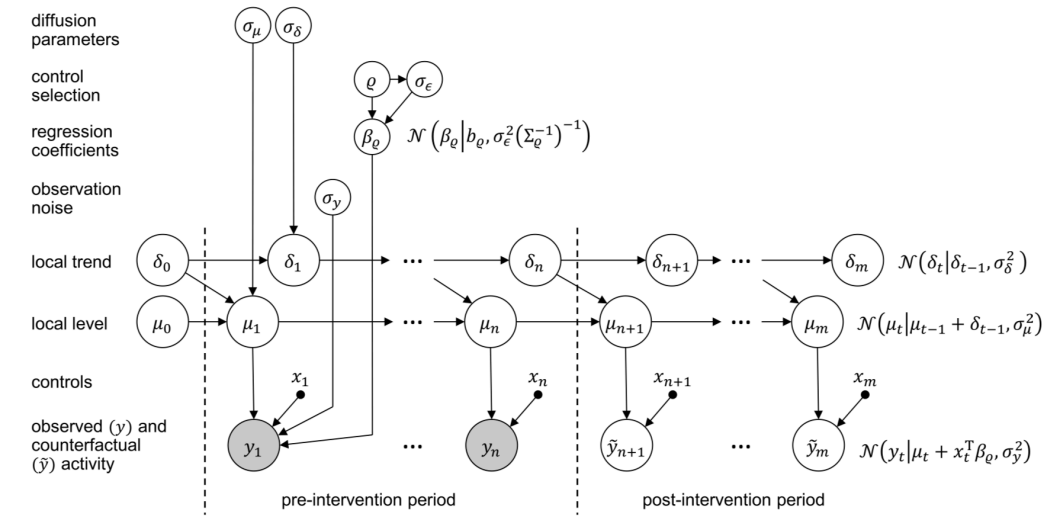


Figure 2:

Example (1): simluation

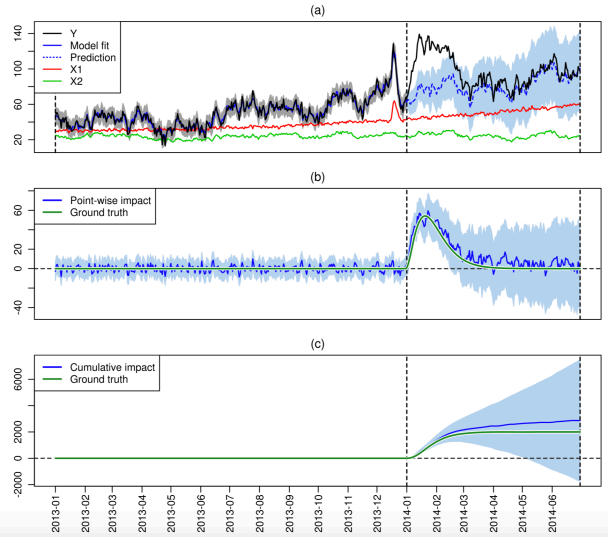


Figure 3:

Example (2): Google Ads, click data

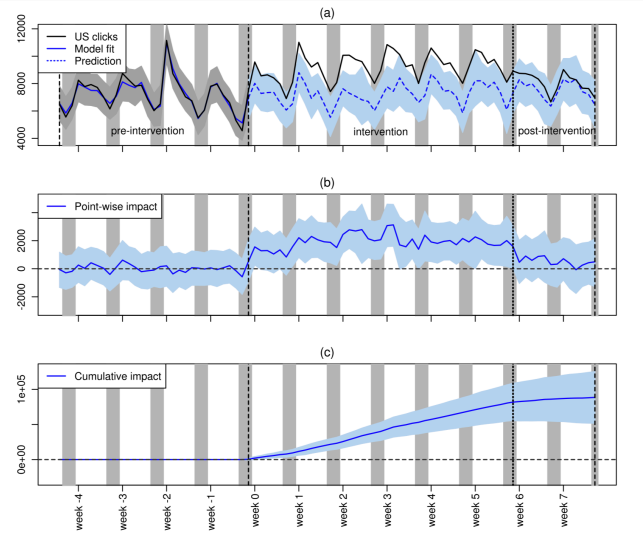


Figure 4:

Challenges with BSTS

- Tuning effective model size
- Elicitation of priors? Independent Bernoullis with uniform weights. (Empirical Bayes?)
- Not a “silver bullet”: still sensitive to initial control variables
- Many draws needed to generate meaningful posterior inference.

Two great packages for today's methods

- `hdm`: Chernozhukov et al. (2016)
- `CausalImpact`: https://google.github.io/CausalImpact/CausalImpact.html

Cases like:

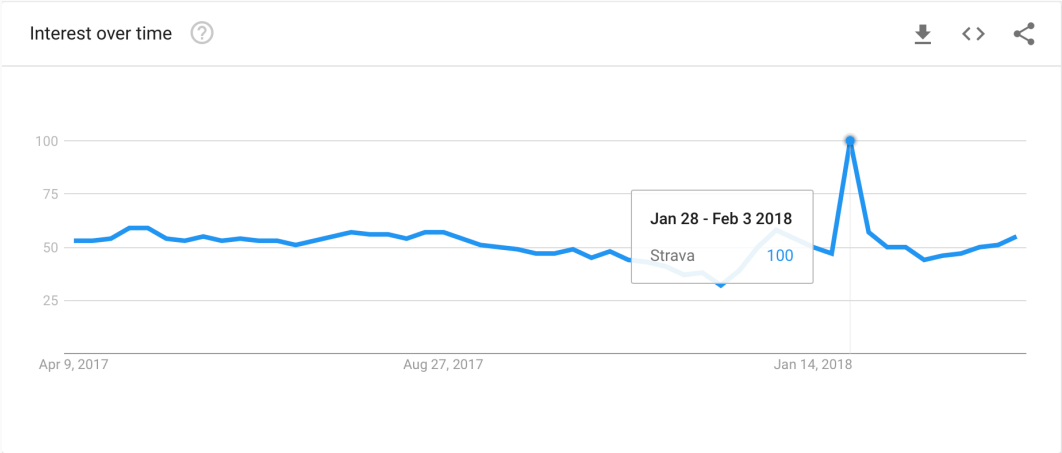


Figure 13: Strava Google trends following news

Introduction to causal inference for data scientists

Practical challenges with inference

Michael Gill

2018-04-17

Center for Data Science | New York University

Today

(Some of) today's references

References:

- Ding, P., and VanderWeele, T. (2016). "Sensitivity Analysis Without Assumptions." Epedemiology.
- Young, A. (2017). "Consistency without Inference: Instrumental Variables in Practical Application." Working Paper.

- Others on syllabus

Today's goals

- Theme today: common ways things go wrong (and proposed approaches for dealing with them)
- **Many** reasons for spurious effects
- Sensitivity analysis (for unmeasured confounding)
- The problem of multiple comparisons—and "p-hacking"
- The problem of "weak instruments"
- Adjusting for estimation error in matching and IV

Unobserved confounding and sensitivity analysis

A DAG from before

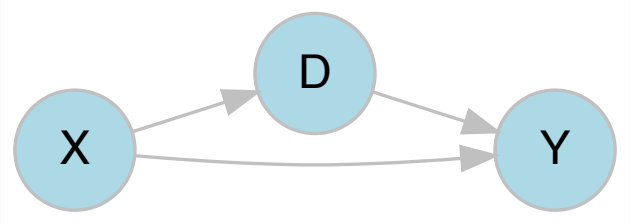


Figure 1: X as a (measurable) confound

- Assuming the data generation process above, conditioning on X allows the direct effect of $D \to Y$ to be nonparametrically identified.
- Having a measure of X, in other words, is necessary to satify unconfoundedness.

A DAG with unobserved confounding

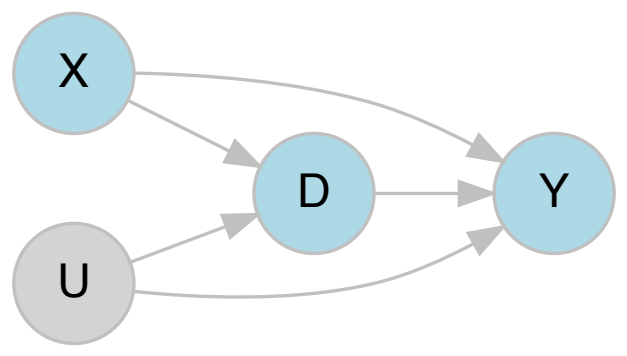


Figure 2: *U* as an unobserved confound

When unobserved missing) data is not a problem

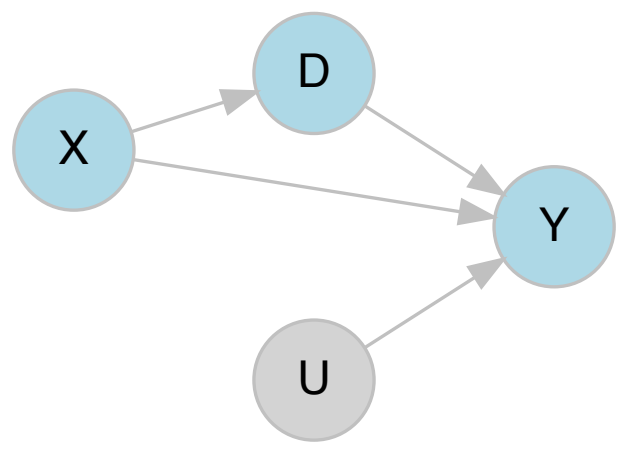


Figure 3: *U* as an unobserved variable, but doesn't confound *D*

Sensitivity analysis

- **Ideally:** by how much would estimated results change if one were to adjust for unobserved confounds? Sometimes called “bias analysis” (e.g., epidemiology, biostatistics).

Sensitivity analysis

- **Ideally:** by how much would estimated results change if one were to adjust for unobserved confounds? Sometimes called “bias analysis” (e.g., epidemiology, biostatistics).
- **Criticisms:** the presence of unobserved confounding is untestable, so people may may ad hoc arguments about the degree of the problem
- **Practically:** in observational studies, a good idea to think about the sensitivity of your result to a range of scenarios

An example for conservative bounds

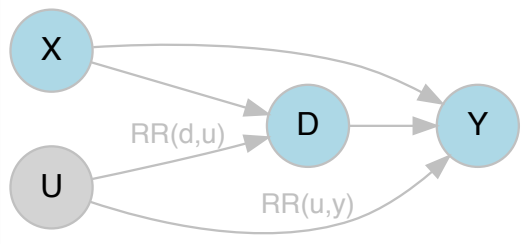
- Taken from Ding and VanderWeele (2016, 2017)—but changing notation slightly, to be consistent with rest of the course.
- Y is the outcome, D is the treatment of interest, X is the set of observed confounds, and U are the unobserved confounds.
- For this example, we will think about the relative risk:

$$RR^{obs}_{d,y|x} = \frac{E(Y^*|D=1, X=x)}{E(Y^*|D=0, X=x)} = \frac{Pr(Y=1|D=1, X=x)}{Pr(Y=1|D=0, X=x)}$$

- Recall (from lecture 2) we can define treatment effects in multiple ways.

Further notation

- $RR(u, y)$: maximum risk ratio for the outcome comparing any 2 categories of the unmeasured confounders, within either treatment group, conditional on the observed covariates
- $RR(d, u)$: how imbalanced the treatment groups are in the unmeasured confounder, U



- $B = \frac{RR(u,y) \times RR(d,u)}{RR(u,y) + RR(d,u) - 1}$
- RR/B: maximum amount the set of unmeasured confounders could alter an observed risk ratio, RR. Holds for $RR > 1$.
- If $RR < 1$, the value is $RR \times B$.

More formally

Define:

$$
\begin{aligned}
RR(u, y|D = 1) &= \frac{\max_k Pr(Y = 1|D = 1, X = x, U = k)}{\min_k Pr(Y = 1|D = 1, X = x, U = k)} \\
RR(u, y|D = 0) &= \frac{\max_k Pr(Y = 1|D = 0, X = x, U = k)}{\min_k Pr(Y = 1|D = 0, X = x, U = k)} \\
RR(d, u, k) &= \frac{Pr(U = k|D = 1, X = x)}{Pr(U = k|D = 0, X = x)}
\end{aligned}
$$

$$
\begin{aligned}
RR(u, y) &= \max\Big(RR(u, y|D = 1), RR(u, y|D = 0)\Big) \\
RR(d, u) &= \max_k RR(d, u, k)
\end{aligned}
$$

Getting to true relative risk

- If knowing X, U can control for confounding for the effect of D on Y:

$$RR^{\text{true}}_{d,y|x} = \frac{\sum_{k=0}^{K-1} Pr(Y=1|D=1, X=x, U=k) \cdot Pr(U=k|X=x)}{\sum_{k=0}^{K-1} Pr(Y=1|D=0, X=x, U=k) \cdot Pr(U=k|X=x)}$$

- The authors demonstrate:

$$RR^{true}_{d,y} \geq RR^{obs}_{d,y} \Big/ \frac{RR_{d,u} \times RR_{u,y}}{RR_{d,u} + RR_{u,y} - 1}$$

- In words: even with unmeasured confounding, the true relative risk must be at least as large as this quantity.

Example: effects of breastfeeding on infant health outcomes (1)

- D: breastfeeding
- X: age, birth-weight, social status, maternal education, and family income
- Y: infant death from respiratory infection
- U: family smoking

Example: effects of breastfeeding on infant health outcomes (1)

- D: breastfeeding
- X: age, birth-weight, social status, maternal education, and family income
- Y: infant death from respiratory infection
- U: family smoking

Without controlling for family smoking, original article found that breast-feeding contributed to a $RR = 3.9(\text{CI}, 1.8 \text{ to } 8.7)$.

We are interest in how strongly an unmeasured confounder must be (related to the treatment and outcome) to explain away an effect estimate.

Example: effects of breastfeeding on infant health outcomes (2)

Suppose:

- $RR_{u,y} = 4$: maximum ratio by which smoking could increase respiratory death
- $RR_{d,u} = 2$: max by which smoking differed by breastfeeding status

This leads to a bias factor:

$$\begin{aligned} B &= \frac{RR_{u,y} \times RR_{d,u}}{RR_{u,y} + RR_{d,u} - 1} \\ &= 4 \times 2/(4 + 2 - 1) \\ &= 1.6 \end{aligned}$$

$\therefore$, dividing $\widehat{RR}$ and its CI by 1.6 is insufficient to overturn the result: $\widehat{RR}_B = 2.43$, and $CI_B = (1.1, 5.4)$.

The E-value

- "The E-value is the minimum strength of association, on the risk ratio scale, that an unmeasured confounder would need to have with both the treatment and outcome, conditional on the measured covariates, to fully explain away a specific treatment–outcome association."
- Taken from Ding and VanderWeele (2017)

$$E = RR\sqrt{RR \times (RR - 1)}$$

Computation of the E-value

- **Very simple:** $E = 3.9\sqrt{3.9 \times (3.9 - 1)}$
- **Interpretation:** "The observed risk ratio of 3.9 could be explained away by an unmeasured confounder that was associated with both the treatment and the outcome by a risk ratio of 7.2-fold each, above and beyond the measured confounders, but weaker confounding could not do so."
- **Suggestion:** in many cases, simple to report this value as a general sensitivity parameter in observational settings

More on calculation

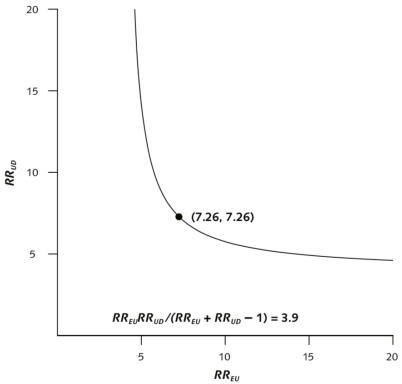
Table 1. Calculating the E-Value for Risk Ratios

Estimate or CI, by Direction of Risk Ratio	Computation of the E-Value
***RR* >1**	
Estimate	E-value = $RR + \text{sqrt}\{RR \times (RR - 1)\}$
CI	If $LL \leq 1$, then E-value = 1 If $LL > 1$, then E-value = $LL + \text{sqrt}\{LL \times (LL - 1)\}$
***RR* <1**	
Estimate	Let $RR^* = 1/RR$ E-value = $RR^* + \text{sqrt}\{RR^* \times (RR^* - 1)\}$
CI	If $UL \geq 1$, then E-value = 1 If $UL < 1$, then let $UL^* = 1/UL$ and E-value = $UL^* + \text{sqrt}\{UL^* \times (UL^* - 1)\}$

LL = lower limit of the CI; *RR* = risk ratio; *RR** = inverse of *RR*; *UL* = upper limit of the CI; *UL** = inverse of *UL*.

Figure 4: Extensions for different outcomes

More on calculation (2)



The E-value essentially sets the 2 parameters, RR_{UD} and RR_{EU}, equal to each other to determine the required minimum for both. The E-value for Victora and colleagues' (23) estimate corresponds to the point (7.26, 7.26). RR_{EU} = maximum risk ratio for any specific level of the unmeasured confounders comparing those with and without treatment, with adjustment already made for the measured covariates; RR_{UD} = maximum risk ratio for the outcome comparing any 2 categories of the unmeasured confounders, with adjustment already made for the measured covariates.

Figure 5: Sensitivity given different values of unobserved *RR*

Extensions from paper (1)

Table 2. E-Values for Other Effect Measures

Effect Measure	**Computation of Approximate E-Value**
OR or HR for rare outcomes	When the outcome is relatively rare (e.g., <15%) by the end of follow-up, the E-value formula in **Table 1** may be used (37). In a case–control study, the outcome only needs to be rare in the underlying population, not in the case-control study.
Rate ratio for count and continuous outcomes	For ratio measures for count outcomes (or nonnegative continuous outcomes), the E-value may be found by replacing the risk ratio with the rate ratio (or the ratio of expected values) in the E-value formula in **Table 1** (37).
OR for common outcomes	When the outcome is common (>15% at the end of follow-up), an approximate E-value may be obtained by replacing the risk ratio with the square root of the OR (45), i.e., $RR \approx \text{sqrt(OR)}$, in the E-value formula in **Table 1**.
HR for common outcomes	When the outcome is common (>15% at the end of follow-up), an approximate E-value may be obtained (45) by applying the approximation $RR \approx (1 - 0.5^{\text{sqrt(HR)}})/(1 - 0.5^{\text{sqrt}(1/\text{HR})})$ in the E-value formula in **Table 1**.
Difference in continuous outcomes	With standardized effect sizes d (mean of the outcome variable divided by the SD of the outcome) and an SE for this standardized effect size s_d, an approximate E-value may be obtained (45–47) by applying the approximation $RR \approx \exp(0.91 \times d)$ in the E-value formula. An approximate CI for the risk ratio may be found by using the approximation $(\exp\{0.91 \times d - 1.78 \times s_d\}, \exp\{0.91 \times d + 1.78 \times s_d\})$. This approach relies on additional assumptions and approximations. Other sensitivity analysis techniques have been developed for this setting (27–29), but they generally require additional assumptions, and the variables do not necessarily have a corresponding E-value.
Risk difference	If the adjusted risks for the treated and untreated are p_1 and p_0, then the E-value may be obtained by replacing the risk ratio with p_1/p_0 in the E-value formula. The E-value for the CI on a risk difference scale is more complex, and software to obtain this is described in the **Supplement** (available at Annals.org). Alternatively, if the outcome probabilities p_1 and p_0 are not very small or very large (e.g., if they are between 0.2 and 0.8), then the approximate approach for differences in continuous outcomes given previously may be used. Other sensitivity analysis techniques have been developed for this setting (27–29) but generally require additional assumptions and do not provide a corresponding E-value.

HR = hazard ratio; OR = odds ratio; *RR* = risk ratio.

Figure 6: Extensions for different outcomes

Extensions from paper (2)

Table 3. Issues of Interpretation of the E-Value

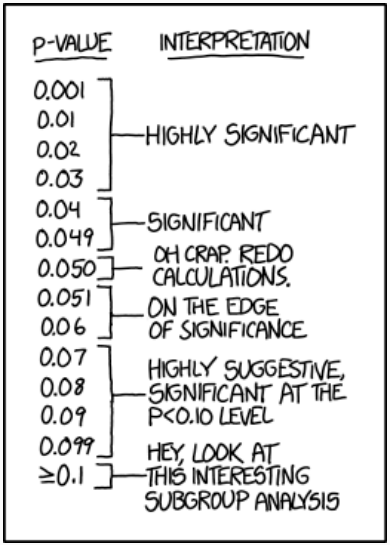
Issue	Interpretation
Likely effect sizes	The E-value should be interpreted in the context of the effect sizes that an unmeasured confounder is likely to have with respect to the outcome and treatment. In the context of biomedical and social sciences research, effect sizes ≥2- or 3-fold occasionally occur but are not particularly common; a variable that affects both treatment and outcome each by 2- or 3-fold would likely be even less common. For purposes of comparison, calculating the analogous E-value for each of the measured covariates if they had been omitted may be helpful.
E-values and sensitivity analysis	The E-value for the respiratory death example was 7.2. In the formula for the bias factor *B*, a confounder that was associated with the respiratory death by less than 7.2-fold might explain away the effect estimate but would have to be associated with the treatment by a risk ratio more than 7.2-fold. Values of the sensitivity analysis variables with a less extreme confounder–outcome association will require a more extreme treatment–confounder association, and vice versa.
Sample size, E-values, and *P* values	A large study with a precisely estimated association often has a very small *P* value; the *P* value may be made arbitrarily small by increasing the sample size. However, if the effect size is small, then the E-value will be small. The E-value depends on the magnitude of the association; it cannot be made arbitrarily large simply by increasing the sample size. The E-value for the CI does depend on the sample size. However, as the sample size increases, the E-value for the CI does not get arbitrarily large; it is bounded by the strength of the association (the limit sometimes is referred to in other contexts as the "design sensitivity" [17, 18]). A large sample size may give a small *P* value; a large effect size will give a large E-value.

Figure 7: Extensions for different outcomes

Multiple comparisons

If you test many questions, you'll find false positives

- Big problem in "big data"/"data mining"



Jelly beans (and more memes)

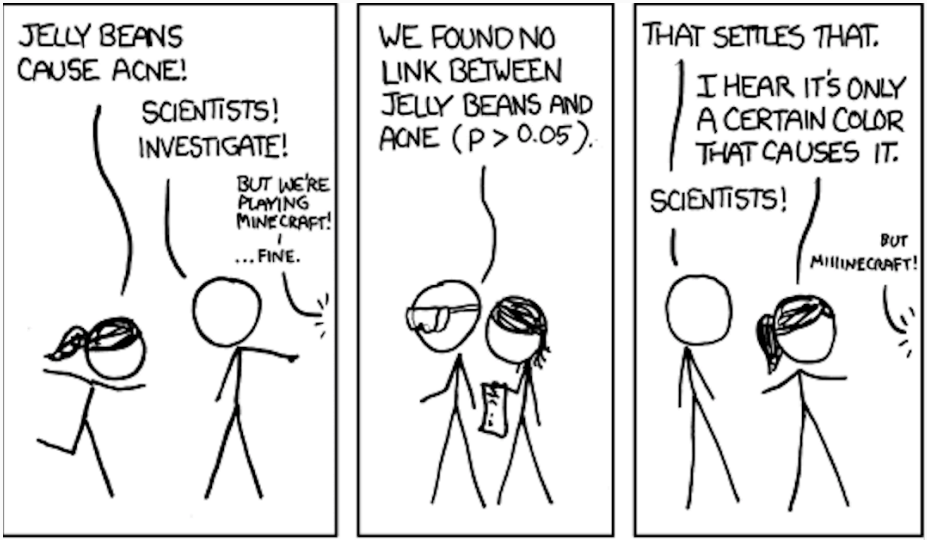


Figure 9:

Jelly beans (2)

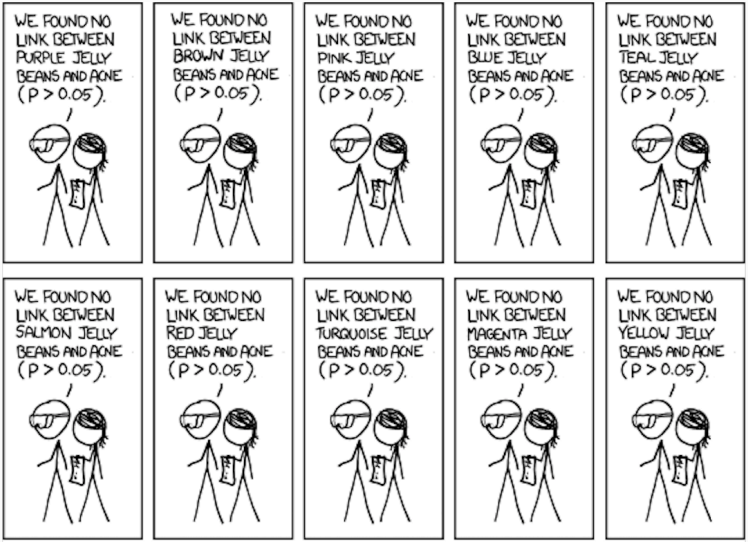


Figure 10:

Jelly beans (3)

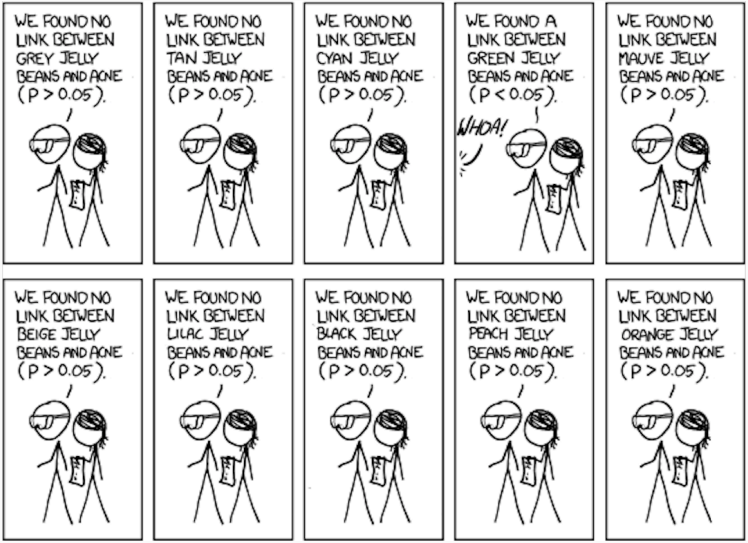


Figure 11:

Jelly beans (4)

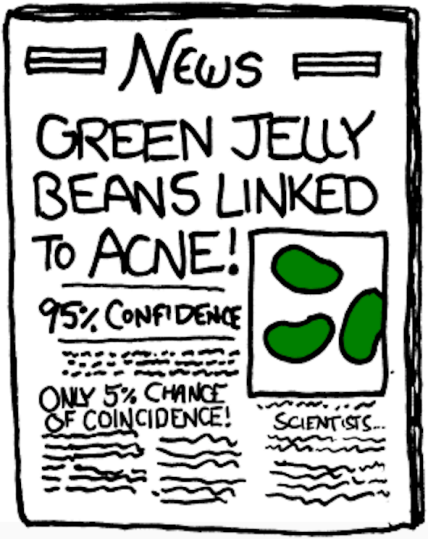


Figure 12:

Type I/II errors

- Type I: probability of falsely assuming an “effect exists” when it doesn’t
- Type II: probability of falsely assuming an “effect doesn’t exist” when it does
- Related, but not identical, to discussions of precision/recall in CS literatures

Demonstration in R

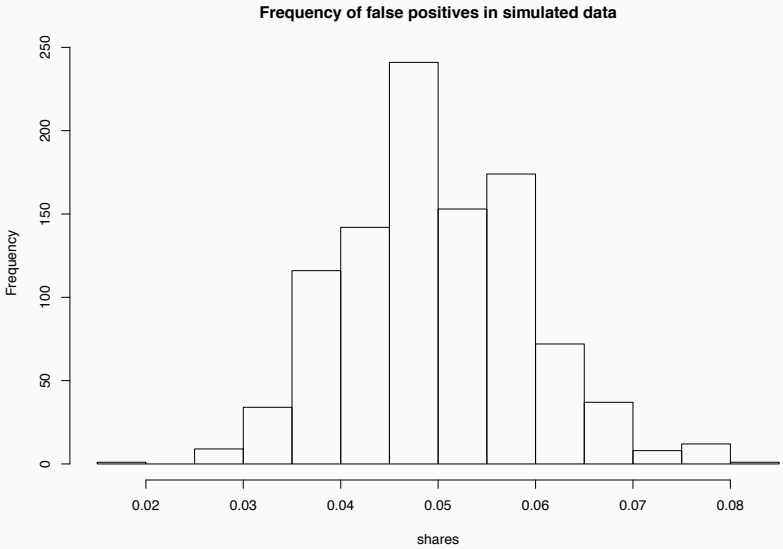
```
n_hyp = 500
n = 100

share_spurious = function(n,n_hyp){
out = replicate(n_hyp,t.test(rnorm(n),rnorm(n)))
p.vals = unlist(out[rownames(out)==”p.value”,])
sum(sort(p.vals)<0.05)/n_hyp
}

shares = replicate(1000,share_spurious(n,n_hyp))
```

Demonstration in R: results

```r
hist(shares, main="Frequency of false positives in simulated data")
```



Bonferroni test

- **Very simple**, conservative adjustment (but arguably *too conservative* in some contexts, e.g., Holm-Bonferroni)
- Given m independent hypotheses:
  - Re-scale critical value of $p-$value to $1 - \alpha/m$
  - Also valid for confidence interval adjustment
- As a general rule: "weaker" effects (in terms of ratio of effect to estimated error), not magnitiude of point estimate, are more likely to be overturned by Bonferroni corrections.

Common adjustment for controlling the family-wise error rate (FWER).

Consider last HW: effect of Trump's election on individual stocks

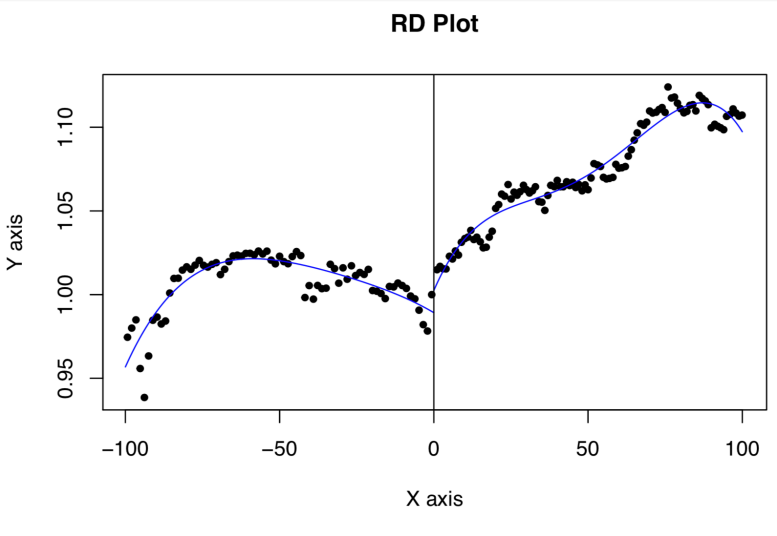


Figure 13: Aggregate RDD response

Effect of Trump's election on individual stocks (2)

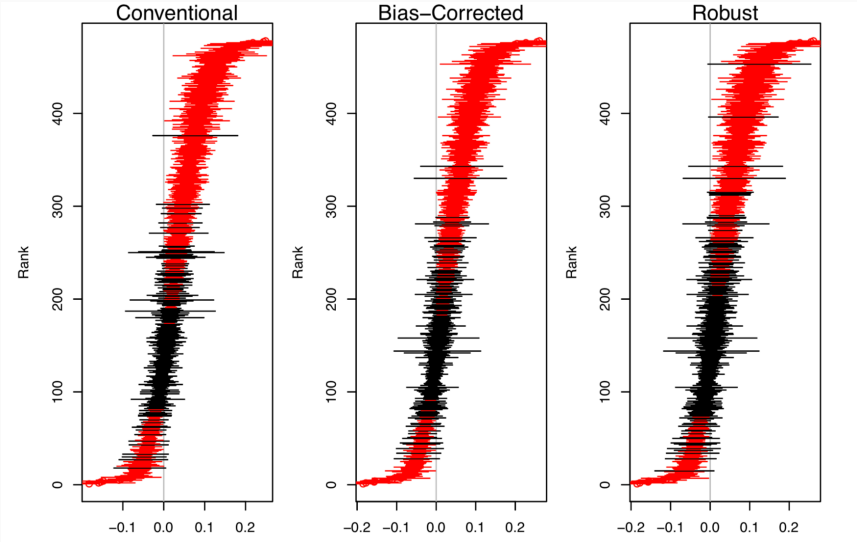


Figure 14: Statistical significance by RDD model

Effect of Trump's election on individual stocks (3)

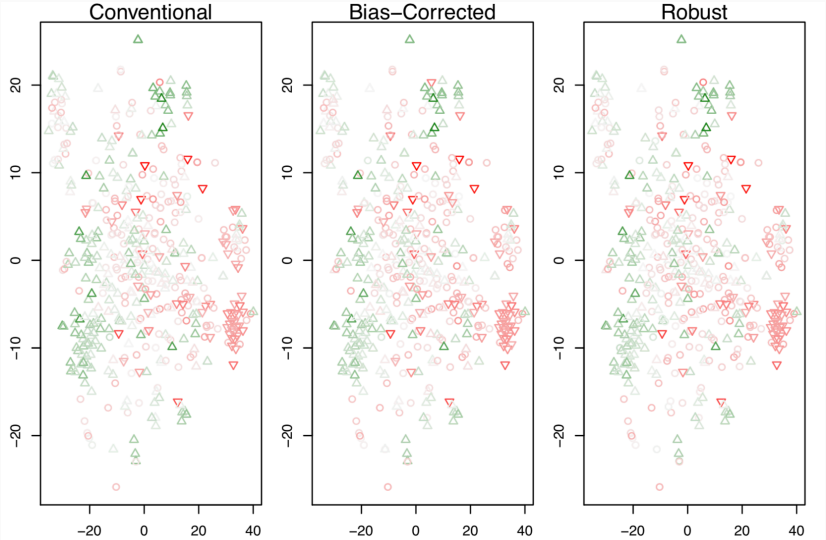


Figure 15: Effects by TSNE projection

Other errors/corrections:

- Type S and M errors. See, e.g., Gelman and Carlin (2014)
  - Type S: probability of estimating the wrong sign of the effect
  - Type M: "exaggeration ratio"—expected degree of effect exaggeration given sample size
- Multiple pairwise comparisons: Tukey multiple-testing adjustment. Important in many cases in which one may be interested in comparing all units'/groups' estimated means (or treatment effects).

Common problems in IV and matching

“Weak instruments” in IV

- Recent paper by Young (2017), on syllabus
- Re-replication of dozens of 2SLS/IV models in top economics journals (in primarily observational settings)
- Author shows sensitivity first-stage regression models. In many cases, causal estimates produced by 2SLS not meaningfully different from what obtained in OLS.
- Builds on big literature on the problem of “weak instruments”: e.g., Chernozhukov and Hansen (2008)

Uncertainty in matching methods

- Abadie and Imbens (2006): in general, matching estimators not $\sqrt{N}$ consistent; propose asymptotic standard error adjustment (implementation in `Matching` package)
- Abadie and Imbens (2008): naive bootsrap (repeated-resampling units with equal weights) fails to give valid confidence intervals
- Abadie and Imbens (2011): bias correction method based on nonparametric regression techniques
- Otsu and Rai (2018): valid weights for bootstrapped confidence intervals

Introduction to causal inference for data scientists

Inference in causal graphs

Michael Gill

2018-04-24

Center for Data Science | New York University

Today

(Some of) today's references

References:

MW Morgan, S., and Winship, C. (2014). Counterfactuals and Causal Inference: Methods and Principles for Social Research, 2nd Edition. Cambridge.

PGJ Pearl, Glymour, and Jewell (2016). Causal Inference in Statistics: A Primer.

PM Pearl, Mohan (2012). Graphical models for causal inference.

Today's goals

- Motivate graphical models for non-parametric causal identification
- Hands-on practice for interpreting such models
- Backdoor, and frontdoor criteria
- Minimal adjustment sets

Causal graphs

Graphs?

- A joint probability distribution $= f(x_1, \ldots, x_n)$ always has the factorization
$$f(x_1, \ldots, x_n) = \prod_j f\left(x_j | x_1, \ldots, x_{j-1}\right)$$
- For any pair of variables x_i and x_j ($i < j$), we draw an arc from x_i to x_j if $f\left(x_j | x_1, \ldots, x_{j-1}\right)$ actually depends on x_i.
- Without any restriction, we get a graph with the "triangular" structure where $x_i \to x_j$ whenever $i < j$.
- But in many other cases (like our previous examples), can do much better.
- **Note:** the construction is not unique, and depends on the ordering of the indices. Clearly, some are "better" (i.e., induce more sparsity) than others.

Directed graphs: basic definitions

- The causal relationships are represented by a directed graph.
- Nodes are random variables
- Directed edges: single-headed arrows. $X \to Y$ means loosely "r.v. X has a causal effect on r.v. Y". Precise definition later.
- Vocabulary: path; directed path; descendant; parent; child.

Directed acyclic graphs

- Cycles=a directed path from a node to itself
- Directed acyclic graphs=directed graph without a cycle.
- Thus, we rule out simultaneous causation: “I like you because you like me”.
- **Property:** there is always at least one node without a parent.

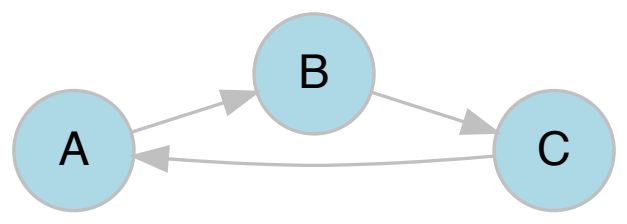


Figure 1: A directed graph with a cycle

Latent variables

- Two types of nodes: solid circle • indicates observed random variable, while hollow circle ○ indicates latent (unobserved) random variable.
- **Shorthand:** a curved dashed bidirected edges between two variables indicates that those variables have a common latent variable among their ancestors. As in the below example from [MW].

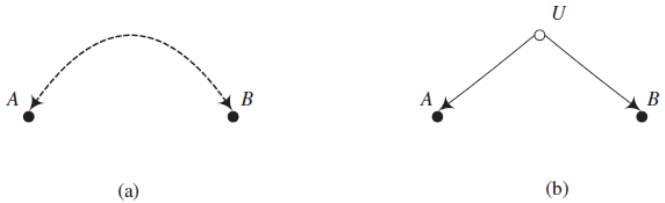


Figure 3.2 Two representations of the joint dependence of A and B on unobserved common causes.

DAGs with three variables

- With three variables, three possibilities:
  1. Mediation / chain: $a \to b \to c$:

$$\text{e.g.,} \qquad f(a,b,c) = f(a)f(b|a)f(c|b)$$

  2. Mutual dependence / fork: $a \leftarrow c \to b$:

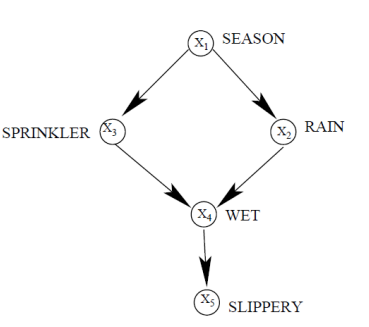
$$\text{e.g.,} \qquad f(a,b,c) = f(c)f(a|c)f(b|c)$$

  3. Mutual causation / collider: $a \to c \leftarrow b$:

$$\text{e.g.,} \qquad f(a,b,c) = f(a)f(b)f(c|a,b)$$

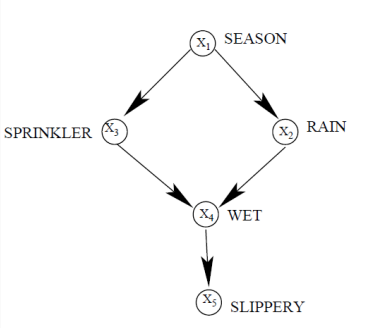
DAGs: example

- Consider the following example, found in [PM]: Season is dry or wet. Sprinker is on or off, etc. What is the probability distribution associated with the graph above?



DAGs: example

- Consider the following example, found in [PM]: Season is dry or wet. Sprinker is on or off, etc. What is the probability distribution associated with the graph above?
- $P(x_1, \ldots, x_5) =$
  $P(x_1)\, P(x_2|x_1)\, P(x_3|x_1)\, P(x_4|x_2, x_3)\, P(x_5|x_4)$.



Blocking

- Two sets X and Y are blocked (or d-separated) by Z if and only if Z blocks every path between X and Y. Denoted $X \perp Y|Z$.
- Z blocks a path from a to b if:
  - there is $c \in Z$ such that path contains chain $a \rightarrow c \rightarrow b$ or fork: $a \leftarrow c \rightarrow b$; or
  - path contains a collider: $a \rightarrow c \leftarrow b$ such that c is not in Z and neither are the descendants of c

Blocking: example 1

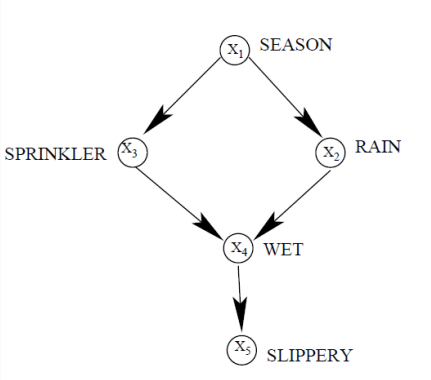


Figure 2: Source: [PM].

- $X_1 \perp X_4 | X_2, X_3$; $X_3 \perp X_5 | X_4$; $X_1 \perp X_5 | X_4$

Blocking: example 1

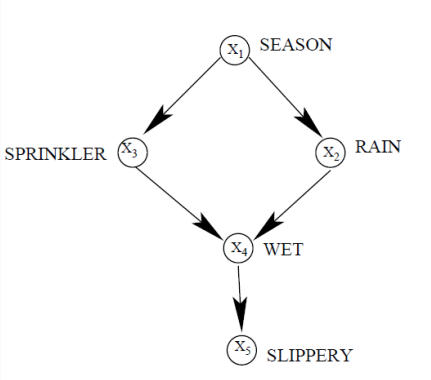


Figure 2: Source: [PM].

- $X_1 \perp X_4 | X_2, X_3$; $X_3 \perp X_5 | X_4$; $X_1 \perp X_5 | X_4$
- $X_3 \perp X_2 | X_1$

Blocking: example 2

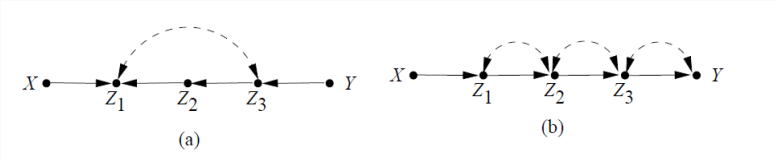


Figure 3: Source: [PM].

- In (a), $X \perp Y$

Blocking: example 2

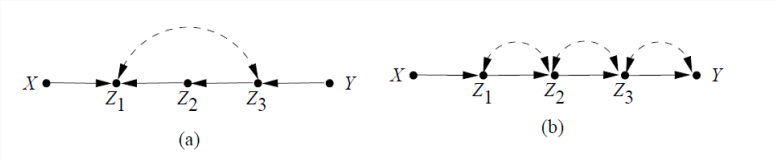
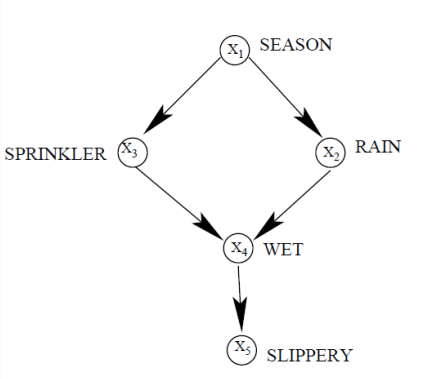


Figure 3: Source: [PM].

- In (a), $X \perp Y$
- In (b) $X \perp Y$ does not hold

Pearl's do operator (1)



- What is the probability that road is slippery if we observe that sprinkler is on?
- What is the probability that road is slippery if we ensure that sprinkler is on?

Pearl's do operator (2)

- Recall $P(x_1, \ldots, x_5) = P(x_1)\, P(x_2|x_1)\, P(x_3|x_1)\, P(x_4|x_2, x_3)\, P(x_5|x_4)$.

Pearl's do operator (2)

- Recall $P(x_1, \ldots, x_5) = P(x_1)\, P(x_2|x_1)\, P(x_3|x_1)\, P(x_4|x_2, x_3)\, P(x_5|x_4)$.
- Probability that road is slippery if we observe that sprinkler is on:

$$\Pr(x_5|x_3) = \frac{\sum_{x_1,x_2,x_4} P(x_1)\, P(x_2|x_1)\, P(x_3|x_1)\, P(x_4|x_2, x_3)\, P(x_5|x_4)}{\sum_{x_1,x_2,x_4,x_5} P(x_1)\, P(x_2|x_1)\, P(x_3|x_1)\, P(x_4|x_2, x_3)\, P(x_5|x_4)}$$

Pearl's do operator (2)

- Recall $P(x_1, \ldots, x_5) = P(x_1)\, P(x_2|x_1)\, P(x_3|x_1)\, P(x_4|x_2, x_3)\, P(x_5|x_4)$.
- Probability that road is slippery if we observe that sprinkler is on:

$$\Pr(x_5|x_3) = \frac{\sum_{x_1,x_2,x_4} P(x_1)\, P(x_2|x_1)\, P(x_3|x_1)\, P(x_4|x_2, x_3)\, P(x_5|x_4)}{\sum_{x_1,x_2,x_4,x_5} P(x_1)\, P(x_2|x_1)\, P(x_3|x_1)\, P(x_4|x_2, x_3)\, P(x_5|x_4)}$$

- Probability that road is slippery if we ensure that sprinkler is on: replace $P(x_3|x_1)$ by one in the expression of $P(x_1, \ldots, x_5)$

$$\Pr(x_5|do(x_3)) = \sum_{x_1,x_2,x_4} P(x_1)\, P(x_2|x_1)\, P(x_4|x_2, x_3)\, P(x_5|x_4)$$

Pearl's do operator (3)

- Graphical interpretation: remove edge $x_1 \to x_3$

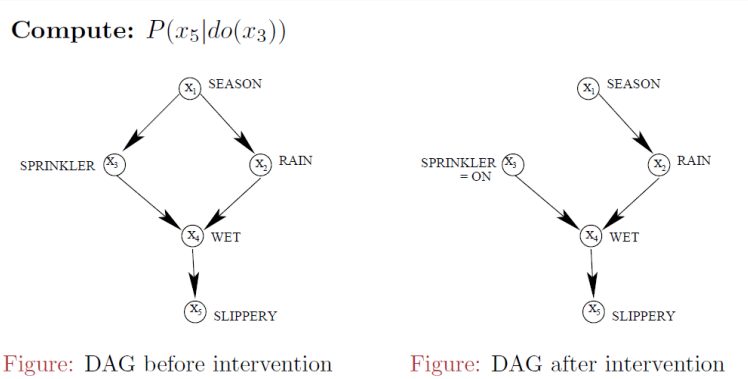
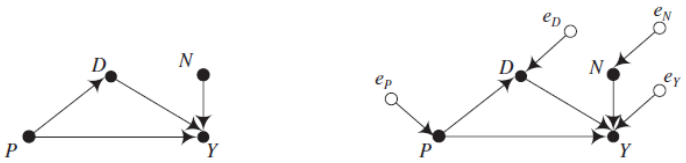


Figure 4: Source: [PM].

Magnification

- To each variable X, we associate an "error term" e_X. The e_X are unobserved and independent.
- The error terms are not represented under the "standard representation", and represented under the "explicit representation".



(a) Standard representation

(b) Under magnification

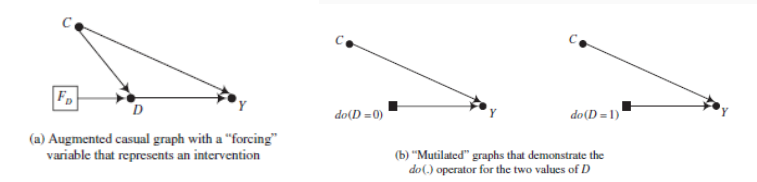
Figure 3.7 Equivalent directed graph representations of the effects of parental background (P), charter schools (D), and neighborhoods (N) on test scores (Y).

Causal graphs

- In Pearl's formal definition of a causal graph:
  1. all variables (other than error terms) are observed
  2. each variable X is associated with an unique error term e_X, and the e_X are independent
  3. for each variable X, there exists a $do\,(X = x)$ operation, which replaces X by the constant x and removes all the arcs pointing at X, leaving everything else unchanged.

Causal graphs (2)

- In the figure below, the causal effect of D on Y is given by $E\left[Y|do\left(D=1\right)\right]-E\left[Y|do\left(D=0\right)\right]$.



(a) Augmented casual graph with a "forcing" variable that represents an intervention

(b) "Mutilated" graphs that demonstrate the $do(.)$ operator for the two values of D

Backdoor paths

- A *back-door path* is a path between any causally ordered sequences of variables that begins with a directed edge that points to the first variable.
- In the graph below, the causal effect of D on Y is confounded by the backdoor path $D \leftarrow C \rightarrow O \rightarrow Y$. In order to evaluate this causal effect, we should condition on $\{C\}$, on $\{O\}$, or on $\{C, O\}$.

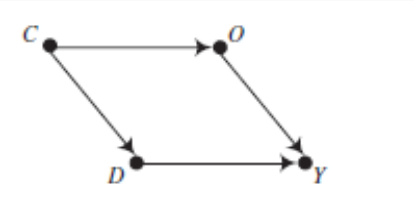


Figure 5: Source: MW

“Collider”” variable and their pitfalls

- As we saw before, in case of a fork, conditioning on the parent variable allows us to identify the causal effect. What about the case of a collider?
- Consider a collider $A \to C \leftarrow B$, where A=SAT score $\in \{0, 1\}$ (low/high); B=interview outcome $\in \{0, 1\}$ (unfavorable/favorable); and C=college admission decision $\in \{0, 1\}$ (rejected/admitted); e.g.

$$C = 1\{A + B \geq 1\}.$$

  Assume A and B are unconditionally independent. Conditional on $C = 1$, $\Pr(A = B | C = 1) = 2/3$ there will be negative correlation between A and B.
- Hence, conditioning on a collider creates dependence.

Be careful when conditioning on a collider.

The back-door criterion (and adjustment sets)

- **Goal:** block paths that generate noncausal associations without blocking paths that generate causal effect.
- Back-door criterion: the causal effect is identified by conditioning on a set of *observed* variables Z if two conditions are met:
  1. All back-door paths between the causal variable and the outcome variable are blocked after conditioning on Z, which will will be the case if each back-door path: (a) contains $A \rightarrow C \rightarrow B$, where C is in Z; or (b) contains a fork of mutual independence $A \leftarrow C \rightarrow B$, where C is in Z, or (c) contains an inverted fork of mutual causation $A \rightarrow C \leftarrow B$, where C and all its descendants are not in Z.
  2. No variables in Z are descendants of the causal variable that lie on (or descend from other variables that lie on) any of the directed path that begin at the causal variable and reach the outcome variable.
- Conditioning on a set that satisfies the back-door criterion identifies the causal effect.

The back-door criterion: example 1

- Example: in the figure below:
  - the only back-door path from D to Y is $D \leftarrow C \rightarrow O \rightarrow Y$.
  - there is only one directed path from D to Y, which is $D \rightarrow Y$.
- Therefore $\{C\}$, $\{O\}$, or $\{C, O\}$ satisfy the back-door criterion.

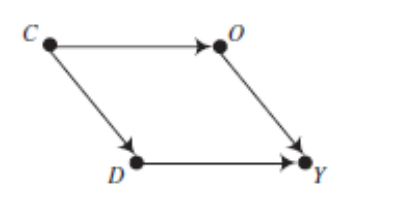


Figure 6: Source: MW

The back-door criterion: example 2

- Consider the figure below, where the dependence in a lag variable Y_{t-1} is considered.
- Is the back-door criterion satisfied?

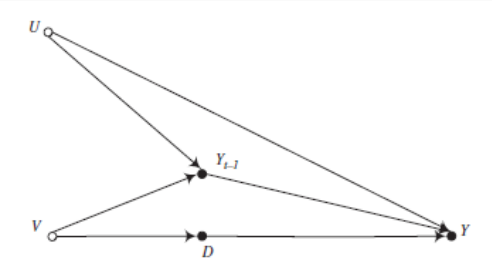


Figure 7: Source: MW

The back-door criterion: example 2 (continued)

- There are two back-door paths from D to Y: $D \leftarrow V \rightarrow Y_{t-1} \rightarrow Y$, and $D \leftarrow V \rightarrow Y_{t-1} \leftarrow U \rightarrow Y$
- Y_{t-1} blocks the first back-door path, but not the second one because it is a collider in the latter. Y_{t-1} being the only observable variable other than D and Y, the back-door criterion is not satisfied.

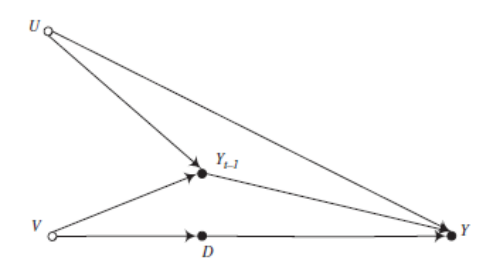


Figure 8: Source: MW

The back-door criterion: example 3

- Consider the figure below. Is the back-door criterion satisfied?

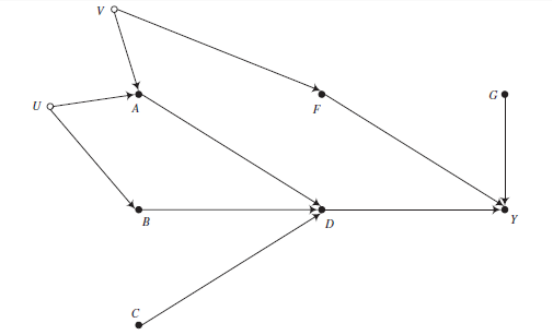


Figure 9: Source: MW

The back-door criterion: example 3 (continued)

- There are two back-door paths between D and Y: $D \leftarrow A \leftarrow V \rightarrow F \rightarrow Y$, and $D \leftarrow B \leftarrow U \rightarrow A \leftarrow V \rightarrow F \rightarrow Y$
- A is a collider in the second path but not in the first one. The problem here is to block the first path without unblocking the second one.
- It turns out that:
  - $\{F\}$ satisfies the back-door criterion as it is part of the chain $V \rightarrow F \rightarrow Y$ in every back-door paths.
  - $\{A, B\}$ satisfies the back-door criterion as A appears in the chain $D \leftarrow A \leftarrow V$ in the first path, and B appear in the chain $D \leftarrow B \leftarrow U$ in the second path.
- Thus, conditioning on either F or on (A, B) will identify the causal effect.

The back-door criterion: example 4

- Consider the figure below. Is the back-door criterion satisfied?

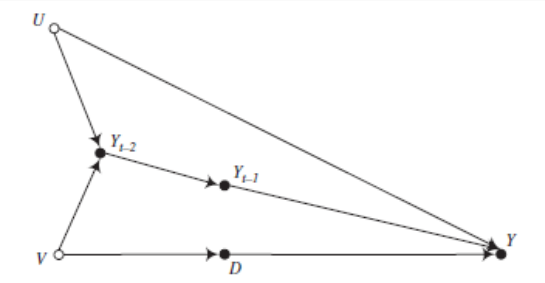


Figure 10: Source: MW

The back-door criterion: example 4 (continued)

- There are two back-door paths between D and Y: $D \leftarrow V \rightarrow Y_{t-2} \rightarrow Y_{t-1} \rightarrow Y$, and $D \leftarrow V \rightarrow Y_{t-2} \leftarrow U \rightarrow Y$,
- There is no collider in the first path, and Y_{t-2} is a collider in the second path. In order to block the first path, one need to condition either on Y_{t-1} or on Y_{t-2}. One cannot condition on Y_{t-2} as it is a collider in the second path. However, one cannot condition on Y_{t-1} either as it is a descendant of Y_{t-2}.
- Hence, the back-door criterion is not satisfied.

The back-door criterion: example 5

- Consider the figure below. Is the back-door criterion satisfied?

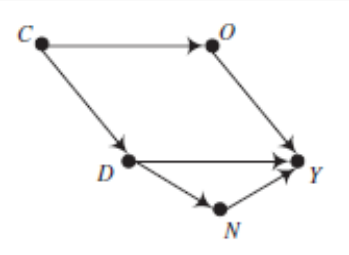


Figure 11: Source: MW

The back-door criterion: example 5 (continued)

- There is only one back-door paths between D and Y: $D \leftarrow C \rightarrow O \rightarrow Y$
- Here, there are two directed paths from D to Y: $D \rightarrow Y$ and $D \rightarrow N \rightarrow Y$. Hence, $\{C\}$, $\{O\}$, or $\{C, O\}$ satisfy the back-door criterion; any set that would contain N would not.

The back-door criterion: example 6

- Consider the figure below. Is the back-door criterion satisfied?

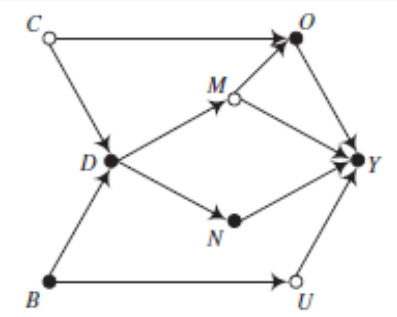


Figure 12: Source: MW

The back-door criterion: example 6 (continued)

- There are three back-door paths between D and Y: $D \leftarrow C \rightarrow O \rightarrow Y$, and $D \leftarrow C \rightarrow O \leftarrow M \rightarrow Y$, and $D \leftarrow B \rightarrow U \rightarrow Y$.
- The candidate conditioning variables are O, B and N. B is needed to block the third path. O is needed to block the first path, but it will unblock the second path (collider).
- Actually, the second part of the back-door criterion assumption is not met either, as O is on a directed path from D to Y.
- Hence, the back-door criterion is not met here.

Self-selection bias

- Consider potential outcomes: $Y = DY^1 + (1 - D)\,Y^0$ which rewrites into

$$Y = Y_0 + \delta D$$

with $\delta = Y^1 - Y^0$. Heckman and Robb (1989) assume $D = 1\{Z\phi + U \geq 0\}$, where Z is a random vector of observable variables, ϕ is a vector of coefficient, and U is some unobserved factor. If $e_U \perp e_Y$, then there is unconfoundedness; otherwise there is selection on the unobservables.

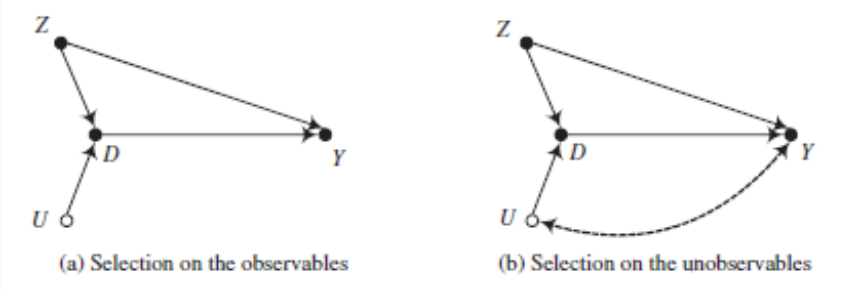


Figure 13: Source: MW

Self-selection bias (continued)

- Consider the IV regression

$$Y = \alpha + \delta D + \varepsilon$$

where $\varepsilon \perp Z$. This can be expressed as a causal graph as follows:

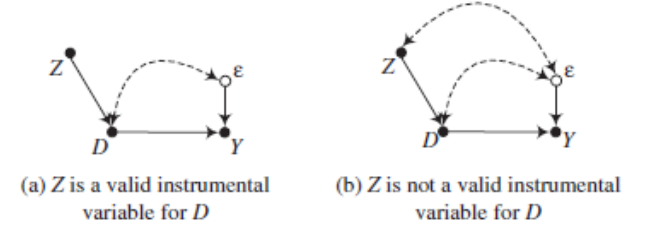
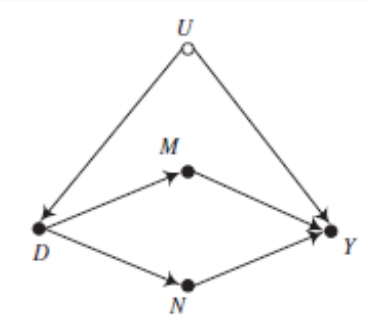


Figure 14: Source: MW

The front-door criterion: an example

- Consider the directed graph below. The back-door criterion does not apply because U is unobservable, so the path $D \leftarrow U \rightarrow Y$ cannot be blocked.
- However, the full effect of D on Y goes through M and N, which are both observed. Are the four causal effects $D \rightarrow M$, $D \rightarrow N$, $M \rightarrow Y$ and $N \rightarrow Y$ identified?



The front-door criterion: an example (continued)

- By the back-door criterion, the causal effects $D \to M$, $D \to N$ are identified (unconditionally).
- By the back-door criterion again, the causal effects $M \to Y$ and $N \to Y$ are identified (by conditioning on D).

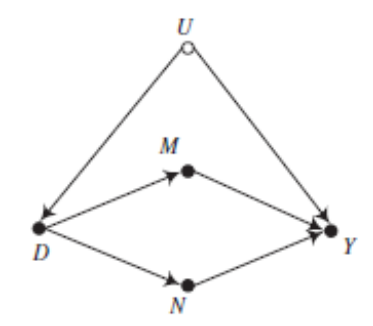
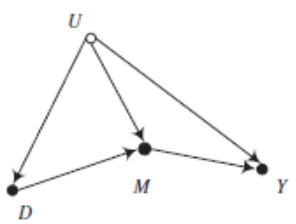


Figure 16: Source: MW

Front-door criterion

- If there is at least one unblocked back-door path connecting a causal variable D to an outcome variable Y, the causal effect is identified by a set $\{X_i\}$ of observed variables if the following two conditions are met:
  - **Exhaustiveness:** any directed path from D to Y contains at least one of the X_i's, and
  - **Isolation:** there is no unblocked back connecting D to any of the X_i, and all back-door paths from any X_i to Y can be blocked by D.
- For example, the following graph does not satisfy isolation:



Wrapping up

- Causal graphs are useful ways to express conditional relations
- Can be clarifying as to various strategies for identifying links between nodes
- **Challenge:** graphs are assumptions, and generally not known. Careful meta-analysis or theory must be taken to an individual case. Even if links are identifiable given adjustment sets, functional forms may still be ambiguous. "Structure learning" is difficult and also subject to power/model specification.
- **Practical advice**: think through possible DAG structures and functional relationships. If relevant variables are not present in your sample, consider sensitivity analysis (e.g., from last week)

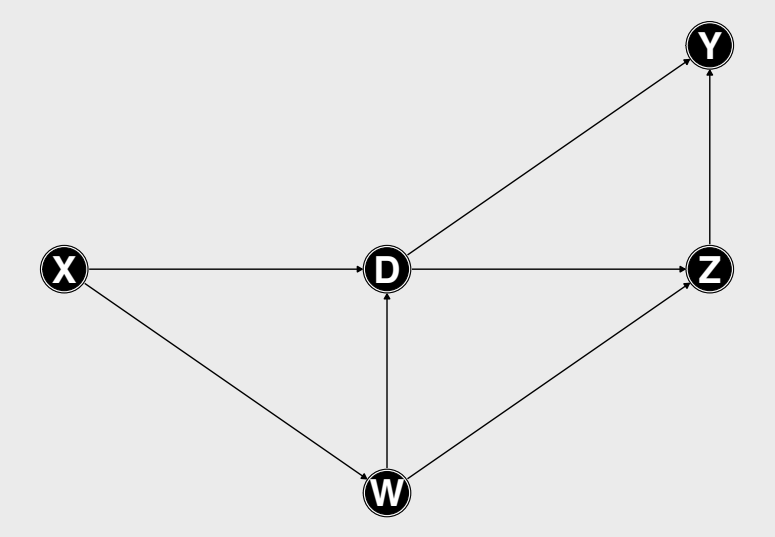
Bonus/encore: Implementations in R!

```r
library(dagitty)
library(lavaan)
library(ggdag)

g <- dagitty('dag {
    X [pos="0,1"]
    D [pos="1,1"]
    Z [pos="2,1"]
    W [pos="1,0"]
    Y [pos="2,2"]
    X -> D -> Z -> Y
    X -> W -> D -> Y
    W -> Z
}')
```

Implementations in R! View results

```
ggdag(g,node_size = 20, text_size=12, label_col = "darkgray")
```



Parents/ancestors

```r
parents(g, "D")
```

```
## [1] "W" "X"
```

```r
ancestors(g, "D")
```

```
## [1] "D" "X" "W"
```

Implied conditional independences

```r
impliedConditionalIndependencies(g)
```

```
## W _||_ Y | D, Z
## X _||_ Y | D, Z
## X _||_ Y | D, W
## X _||_ Z | D, W
```

Finding minimum adjustment sets

```r
adjustmentSets(g, "D", "Y", type="all")
```

```
##  { W }
##  { W, X }
```

How cool is that?

Figure 18: Source: whoever makes emojis