GOV 2001/ 1002/ E-200 Section 9 Causal Inference and Estimation[1]

Anton Strezhnev

Harvard University

March 23, 2016

[1] These section notes are heavily indebted to past Gov 2001 TFs for slides and R code.

Logistics

Reading Assignment- Ho et. al. (2007), King et. al. (2015), and Maertens and Swinnen (2015) **Problem Set 7-** Last Problem Set!

Due by 6pm Wednesday, March 13th on Canvas.

OVERVIEW

- In this section you will...
  - learn how to define a causal effect and what assumptions are needed to estimate it.
  - learn how to estimate a causal effect using data.
  - learn the benefits and drawbacks of different methods of estimating causal effects.

OUTLINE

Causal Inference

Identifying Causal Effects

Causal Effects in Observational Data

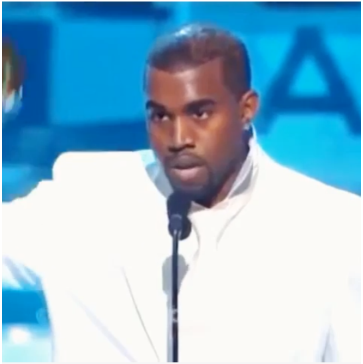
Matching

What is a Causal Effect?

- A causal effect is a comparison of counterfactuals.
- One way of thinking about these counterfactuals is through the concept of "**potential outcomes**"
- For some person i, there exist two unobserved potential outcomes $Y_i(1)$ and $Y_i(0)$.
  - $Y_i(1)$ is the outcome if the person were to receive treatment $T = 1$.
  - $Y_i(0)$ is the outcome if the person were to receive treatment $T = 0$.
- The causal effect τ_i of treatment for that individual is

$$\tau_i = Y_i(1) - Y_i(0)$$

The Fundamental Problem of Causal Inference



"Everybody wanted to know what would happen if I didn't win...I guess we'll never know."
Kanye West, 2005

The Fundamental Problem of Causal Inference

- We cannot estimate τ_i from data! For an individual, Y_i is only *one* of their potential outcomes.
- What do we do? Change the question!
- Most intuitive quantity: the Average Treatment Effect (ATE).

$$\text{ATE} = E[\tau_i] = E[Y_i(1)] - E[Y_i(0)]$$

- **Important:** What causal question does the ATE answer?
  - Does it tell us what the effect will be if *you* get treatment? No!
  - It tells us the average of effects in the population of units.
  - This isn't always what we care about! Sometimes the distribution matters.
  - Example: Two medications have the same ATE, One has a constant small effect. The other has huge benefits but kills 1 in 1000 patients. Which would you prescribe?

Causal vs. Associational Claims

- What's the difference between

$$E[Y_i(1)] - E[Y_i(0)]$$

- and

$$E[Y_i|T_i = 1] - E[Y_i|T_i = 0]$$

?

- First is a causal quantity (counterfactual statement). The second is an association (not counterfactual!)

CAUSAL VS. ASSOCIATIONAL CLAIMS

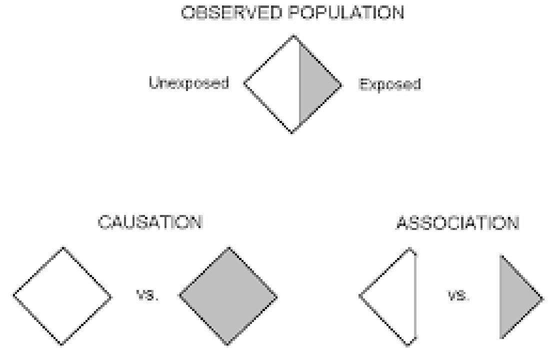


Figure: Difference betweeen causal and association quantities (Hernan and Robins)

Linking Associations to Causal Quantities

- Your difference-in-means is not a causal effect. Your regression coefficient is not a causal effect. Your post-matching linear model is not a causal effect. Your inverse-probability of treatment weighted marginal structural model estimator is not a causal effect.
- **...without additional assumptions.**
- There is *no such thing* as a causal effect derived purely from data alone.

Linking Associations to Causal Quantities

- We need *assumptions* to connect observed associational quantities (like differences in means) to causal effects.
- Basically: When does correlation actually imply causation?
-
- This is often known as **identification** of a causal effect.

OUTLINE

Causal Inference

Identifying Causal Effects

Causal Effects in Observational Data

Matching

Identifying Causal Effects

- How do we go from something we can estimate

$$E[Y_i|T_i = 1] - E[Y_i|T_i = 0]$$

to something we don't observe

$$E[Y_i(1)] - E[Y_i(0)]$$

Basic Identification Assumptions

- Assumption 1: *Stable Unit Treatment Value Assumption (SUTVA)*
- If a subject i receives treatment $T_i = t$, then the observed outcome Y_i is equal to the potential outcome $Y_i(t)$.
- Seems trivial, but not...
  - Implies that there aren't multiple versions of treatment.
  - Also implies no interference between units - the potential outcome for an individual does not depend on treatments assigned to *other* individuals.
- *Implication:* We can reduce the number of potential outcomes for each unit to two that depend only on the treatment it receives: $Y_i(1)$ and $Y_i(0)$.

Basic Identification Assumptions

- Assumption 2: *Ignorability/Unconfoundedness*
- $\{Y_i(1), Y_i(0)\}$ is independent of T_i
- This is the big assumption...
  - Implies that had the group that received treatment been instead assigned control , the observed outcome in that group would have been the same as the outcome in the group that *did* receive control (and vice-versa).
  - Violated if some third factor affects probability of receiving treatment and $\{Y_i(1), Y_i(0)\}$
- Randomization is valued because it *gives us* ignorability.

Basic Identification Assumptions

- Under these assumptions, the causal effect **is identified** by the observed association between treatment/control.
- By SUTVA

$$E[Y_i|T_i = 1] - E[Y_i|T_i = 0] = E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 0]$$

- By ignorability

$$E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 0] = E[Y_i(1)] - E[Y_i(0)]$$

Outline

Causal Inference

Identifying Causal Effects

Causal Effects in Observational Data

Matching

WHAT IF I DON'T HAVE AN EXPERIMENT!?

- In observational data, unconditional ignorability rarely holds.
  - Sicker people are more likely to take medicine.
  - Wealthier people are more likely to turn out and vote (in the U.S.)
  - Paris of democratic countries are more likely to trade.
- What to do? Make a weaker assumption - conditional ignorability.

Identification in observational data

- Assumption: *Conditional Ignorability*
- $\{Y_i(1), Y_i(0)\}$ are independent of T_i conditional on some set of covariates $\mathbf{X}$
  - Often referred to as "no unobserved confounding"
  - "Theory-driven" assumption - no way of verifying from data alone that it holds!

ESTIMATION IN OBSERVATIONAL DATA

- Under SUTVA and conditional ignorability,

$$\begin{aligned} E[Y_i|T_i = 1, X_i = x] - E[Y_i|T_i = 0, X_i = x] \\ = E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x] \end{aligned}$$

- When X is low-dimensional, can *stratify* by X, take differences in means within strata, and then average those estimates to estimate ATE. No additional modeling assumptions necessary!

Estimation in observational data

- When X is high-dimensional or continuous, what happens to the stratification estimator? Too high variance!
- Some strata might only have a single observation!
- First approach: Make more modeling assumptions. Regression!
- Assume not only conditional ignorability, but also a model for $E[Y|T,X]$.

$$E[Y_i|T_i,X_i] = \beta_0 + \beta_1 T_i + \beta_2 X_{1i} + \beta_3 X_{2i}, \ldots \text{...}$$

- Now we can estimate using OLS...

ESTIMATION IN OBSERVATIONAL DATA

- Under these assumptions, what's the estimate of causal effect?

$$\begin{aligned} E[Y_i|T_i = 1, X_i = x] - E[Y_i|T_i = 0, X_i = x] \\ = \beta_0 + \beta_1 \times 1 + \beta_2 x_1 + \beta_3 x_2 - \beta_0 + \beta_1 \times 0 + \beta_2 x_1 + \beta_3 x_2, \\ = \beta_1 \end{aligned}$$

- Note we've assumed a constant effect! If that doesn't hold, β_1 is still an average of individual treatment effects... but it's a weighted average and may not be representative of the ATE (see Aronow and Samii (2015))
- If we don't like OLS, can use GLMs, GAMs, or whatever to get conditional expectations. How we estimate the CEF is separate from whether it's causally interpretable (identification).

Estimation in observational data

- What's the right regression then?

  ¯_(ツ)_/¯

- If we pick the wrong model, we introduce bias in *estimation*. How do we avoid *model dependence*?

OUTLINE

Causal Inference

Identifying Causal Effects

Causal Effects in Observational Data

Matching

Matching Estimators

- How do we estimate $E[Y_i(1)]$ and $E[Y_i(0)]$ when we have lots of covariates but don't want to assume a model for the outcome?
- For each observation, we observe either $Y_i(1)$ or $Y_i(0)$.
- If we observed everything, then our estimated ATE $\hat{\tau}$ would be

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N} Y_i(1) - \frac{1}{N}\sum_{i=1}^{N} Y_i(0)$$

Matching Estimators

- We could instead imagine coming up with estimates $\hat{Y_i(1)}$ such that

$$\hat{Y_i(1)} = \begin{cases} Y_i & \text{if } T_i = 1 \\ Y_j^* & \text{if } T_i = 0 \end{cases}$$

- What's Y_j^*? It's the value of Y from some other observation j that has treatment $T_i = 0$ and covariates X_j that are "close" to covariates X_i.
- That is, it's the outcome from a "matched" control observation (or observations).
- We could do the same for $\hat{Y_i(0)}$ and use a matching estimator for the ATE

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\hat{Y_i(1)} - \frac{1}{N}\sum_{i=1}^{N}\hat{Y_i(0)}$$

Matching is an Estimation Strategy

- Matching is not an identification strategy.
- Matching is not an identification strategy.
- Matching is not an identification strategy.
- It does **not** eliminate bias from unobserved confounding. If the identification assumptions do not hold, then matching won't magically help.
- However, it *does* ameliorate bias from arbitrary choices of models for $E[Y|T, X]$.
- And gives us a way of assessing how "well" we've adjusted for the observed confounders – imbalance.

HOW DO WE PICK MATCHES?

- Ideal case: **Exact Matching**. For each i find a matched observation j with $X_j = X_i$ but different T_i. Basically impossible in any normal situation.
- Need to allow for some differences between X_j and X_i. One approach is to define a distance metric and match the closest j ("Nearest-neighbor").
  - **Mahalanobis Distance**:

$$M(X_i, X_j) = \left[(X_i - X_j)^T S^{-1}(X_i - X_j)\right]^{-1/2}$$

  - **Propensity Score Difference**:

$$\hat{\pi}_i - \hat{\pi}_j = Pr(T_i = 1|X_i) - Pr(T_i = 1|X_j)$$

  - **Coarsened Exact Matching**:

$$D(X_i, X_j) = \begin{cases} 0 & \text{if } X_i \approx X_j \\ \infty & \text{if } X_i \not\approx X_j \end{cases}$$

How do we evaluate our matching methods?

- Goal of matching is to minimize imbalance (and thereby bias). Often we look at differences in means across controls variables.

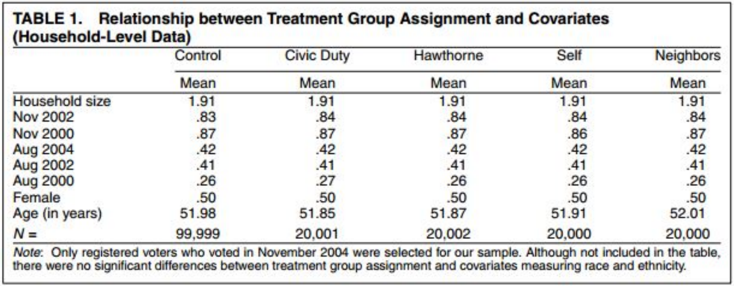


TABLE 1. Relationship between Treatment Group Assignment and Covariates (Household-Level Data)

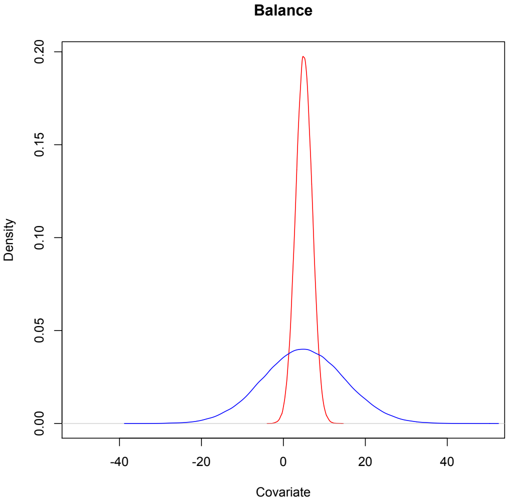
	Control	Civic Duty	Hawthorne	Self	Neighbors
	Mean	Mean	Mean	Mean	Mean
Household size	1.91	1.91	1.91	1.91	1.91
Nov 2002	.83	.84	.84	.84	.84
Nov 2000	.87	.87	.87	.86	.87
Aug 2004	.42	.42	.42	.42	.42
Aug 2002	.41	.41	.41	.41	.41
Aug 2000	.26	.27	.26	.26	.26
Female	.50	.50	.50	.50	.50
Age (in years)	51.98	51.85	51.87	51.91	52.01
N =	99,999	20,001	20,002	20,000	20,000

Note: Only registered voters who voted in November 2004 were selected for our sample. Although not included in the table, there were no significant differences between treatment group assignment and covariates measuring race and ethnicity.

Figure: Balance table from actual experiment by Gerber, Green and Larimer (2008)

How do we evaluate our matching methods?

- However, means can be misleading:



How do we evaluate our matching methods?

- Also, don't ignore variance! We can get perfect balance...by dropping all observations except for one identically matched pair. Who wants to publish with $n = 2$?
- One advantage of propensity score matching is that it has an implicit variable selection method so we're not matching on irrelevant covariates (that are not predictive of treatment).
- However, it still relies on untestable modelling assumptions about the propensity score.
- And only approximates complete randomization rather than blocking (and so suffers in efficiency/bias reduction).
- Open research question to find best methods of variable selection in matching without relying on models like propensity score.

Other matching dilemmas

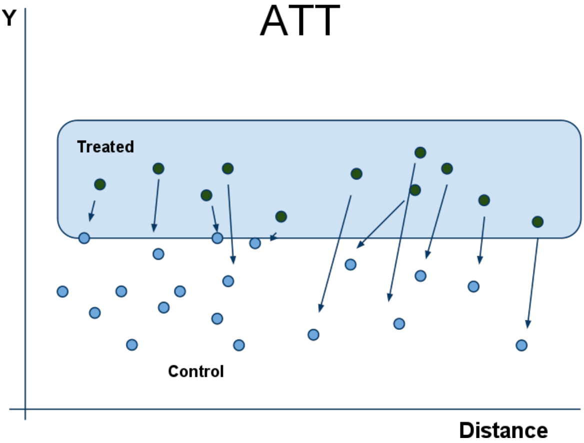
- **When do we match versus throw away observations?** How bad of a match are we willing to tolerate before we change the quantity of interest?
- **How many matches do we match to each observation?** Matching more increases bias (because we're willing to tolerate greater distance) but reduces variance (more data).
- **Do we match with or without replacement?** Matching with replacement reduces bias (ensures that closest observations always get matched). However, it increases variance (we're reusing observations in the matched data, reducing effective sample size).

MORE ON DROPPING OBS AND CHANGING QOI

- Which observations are you going to find matches for and which ones will you drop?
- Dropping observations changes the quantity of interest.
  - Often, we have few treateds and lots of controls.
  - Matching every control would lead to a lot of bias, so we narrow QOI to "ATT" - Average Treatment Effect on the Treated.

$$\text{ATT} = E[Y_i(1)|A_i = 1] - E[Y_i(0)|A_i = 1]$$

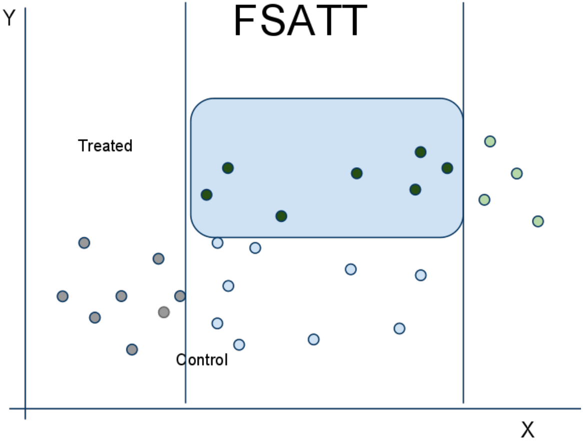
More on dropping obs and changing QOI



More on dropping obs and changing QOI

- But what if finding matches for some treateds is too unreasonable.
- Can narrow estimand even further by dropping treateds... to the Feasible Sample Average Treatment Effect on the Treated (FSATT)!
- Careful - this improves internal validity (reduces bias) but might limit external validity (by changing the population of interest).
- How do we drop? Common support threshold. Caliper/maximum match distance.

MORE ON DROPPING OBS AND CHANGING QOI



Summary

- Causal effects are counterfactual comparisons - need assumptions to connect them to quantities we can estimate.
- Causal inference is composed of two sets of assumptions:
  - Identification: What needs to be true in order for differences in conditional expectations to be interpretable as causal quantities?
  - Estimation: How do we estimate the conditional expectation?
- Matching is a tool for **estimation** not **identification**.Relies on the same “no unobserved confounders” assumption as a regression estimator.
- However, it allows us to directly assess the amount of estimation bias we introduce when we adjust for observed confounders (in contrast to regression).
- But if you have strong unobserved confounding, no amount of matching will save you. Change your research design!

QUESTIONS

Questions?