POLS/CS&SS 503:
Advanced Quantitative Political Methodology

CAUSAL INFERENCE

May 26, 2015

Jeffrey B. Arnold

UNIVERSITY of WASHINGTON
DEPARTMENT OF POLITICAL SCIENCE

CENTER for STATISTICS
and the SOCIAL SCIENCES

Overview

Prediction vs. Causal Inference

Consider a relationship between X and Y:

Prediction

- Given values of x predict y.
- Compare values of y for different values of x
- This is a comparison **between** individuals

Causal Inference

- Comparison **within** individuals
- For the **same individual**, what *would happen* as a result of a hypothesized "treatment" value of x

Causal vs. Casual Inference

- Non causal inference
- A typo

What is the effect of hospitals on health?

- Should we compare the health of those in hospitals with those outside hospitals?
- Counterfactual: For individuals, what would have happened if they went to a hospital, or not?

What is the effect of hospitals on health?

The causal effect of going to the hospital is, for two individuals:

- (health of Annie if she goes hospital) - (health if she does not)
- (health of Jeff if he goes to the hospital) - (health if he does not)

Potential Outcomes as imagined by Community

Episode “Remedial Chaos Theory” S3, E4; See |https://www.youtube.com/watch?v=JTsb5hg04Oc|.

Potential Outcome Framework

Consider a **treatment** variable $D_i = \{0, 1\}$?

$$\text{Potential outcome} = \begin{cases} Y_i(1) & \text{if } D_i = 1 \\ Y_i(0) & \text{if } D_i = 0 \end{cases}$$

The treatment variable can be continuous or ordinal, but easier to reason about the binary case for now.

Person	No Hospital $Y_i(0)$	Hospital $Y_i(1)$	Casual Effect $Y_i(0) - Y_i(1)$
Annie	8	10	2
Jeff	5	10	5
Abed	1	7	6
Britta	5	10	5
Chang	3	5	2
Frankie	6	8	2

The fundamental problem of causal inference

We only observe one potential outcome — the observation; we cannot observe both $Y_i(1)$ and $Y_i(0)$.

Person	No Hospital $Y_i(0)$	Hospital $Y_i(1)$	Treatment D_i
Annie	8		0
Jeff	5		0
Abed	1		0
Britta		10	1
Chang		5	1
Frankie		8	1

We can only observe one outcome for any individual.

How does the potential outcome relate to observed outcome?

- Need a way to connect potential outcomes to observed outcomes.
- **SUTVA**: Stable unit treatment value assumption
- Also called a "consistency" assumption
- What is SUTVA?
- The outcome observed for a value of a treatment is equal to the potential outcome for that treatment value.

$$Y_i \text{ if } D_i = d = Y_i(d) \text{ for } d \in \{0, 1\}$$

- No interference between units. Potential outcomes of units is unaffected by the treatments received by other units.
- Not all "treatments" can be used, there may ill-defined counterfactuals.
- If there is interference, you need to think about the problem and redefine treatment or units

Person	No Hospital $Y_i(0)$	Hospital $Y_i(1)$	Treatment D_i	Observed Outcome Y_i
Annie	8		0	8
Jeff	5		0	5
Abed	1		0	1
Britta		10	1	10
Chang		5	1	5
Frankie		8	1	8

- If SUTVA, the observed outcome has to match the potential outcome for that treatment
- Example of failure: health of Chang depends on whether Jeff goes to the hospital.

What causal effects are there?

Suppose population of units $i = 1, \ldots, N$

Individual Causal Effect (ICE)

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

Average Treatment Effect (ATE) The average causal effect

$$\tau = \mathsf{E}(\tau_i) = \frac{1}{N} \sum_{i:X_i=x} Y_i(1) - Y_i(0)$$

Causal effects for subpopulations

Conditional Average Treatment effect (CATE) ATE for a subpopulation

$$\tau(x) = \mathsf{E}(\tau_i | X = x) = \frac{1}{N_x} \sum_{i:X_i = x} (Y_i(1) - Y_i(0))$$

Average treatment effect on the treated (ATT) Causal effect for those that were treated.

$$\tau_{ATT} = \mathsf{E}(\tau_i | D_i = 1) = \frac{1}{\sum D_i} \sum_{i:D_i = 1} (Y_i(1) - Y_i(0))$$

Person	No Hospital $Y_i(0)$	Hospital $Y_i(1)$	ICE $Y_i(0) - Y_i(1)$	Treatment D_i
Annie	8	10	2	0
Jeff	5	10	5	0
Abed	1	7	6	0
Britta	5	10	5	1
Chang	3	5	2	1
Frankie	6	8	2	1

$$\text{ATE} = (2 + 5 + 6 + 5 + 2 + 2)/6 \approx 3.7$$

$$\text{CATE(men)} = (5 + 6 + 2)/3 \approx 4.3$$

$$\text{ATT} = (5 + 2 + 2)/3 = 3$$

Why use different estimands?

Identification

- If you had infinite data (entire population, no sampling variation) could you estimate the parameter uniquely?
- E.g.example of non-identification in regression is collinearity
- **Non parametric identification** Don't require a parametric model of data.
- **Parametric identification** Estimand identified assuming a parametric model of the data, not identified otherwise
- Causal inference concerned with the identification of causal estimands like ATE, ATT.

What's the key to causal inference?

Data + assumptions = causal inference

"What's your identification strategy?" means "What assumptions are required to estimate a causal effect?"

The Selection Problem

We cam rewrite the observed outcome as:

$$
\begin{aligned}
Y_i &= \begin{cases} Y_i(1) & \text{if } D_i = 1 \\ Y_i(0) & \text{if } D_i = 0 \end{cases} \\
&= Y_i(1)D_i + 1 - Y_i(0)(1 - D_i) \\
&= Y_i(0) + (Y_i(1) - Y_i(0))D_i \\
&= \text{(potential outcome for non-treatment)} \\
&\quad + \text{(causal effect of treatment if treated)}
\end{aligned}
$$

The observe value of the individual is the sum of their outcome if not treated and the causal effect of the treatment.

Avg. Causal Effects and the Selection Bias

Observation and causal effects:

$$
\begin{aligned}
&\mathsf{E}(Y_i|D_i=1) - \mathsf{E}(Y_i|D_i=0) && \text{Obs.difference in means} \\
&= \mathsf{E}(Y_i(1)|D_i=1) - \mathsf{E}(Y_i(1)|D_i=1) && \text{ATT} \\
&+ \mathsf{E}(Y_i(0)|D_i=1) - \mathsf{E}(Y_i(0)|D_i=0) && \text{Selection bias}
\end{aligned}
$$

- Selection bias: how diferent treated and untreated groups are under control (when $D_i = 0$)
- Because of the selection bias, ATT is unidentified. E.g.If ATT negative, could chose a large enough selection effect to make the difference in means positive.

How randomization solves the selection Problem

If D_i is randomly assigned then, D_i is independent of the potential outcomes

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

Which means that difference in means between groups simplifies the ATE

$$\begin{aligned}
&\mathsf{E}(Y_i|D_i = 1) - \mathsf{E}(Y_i|D_i = 0) && \text{Difference in means} \\
&= \mathsf{E}(Y_i(1)|D_i = 1) - \mathsf{E}(Y_i(0)|D_i = 0) && \text{SUTVA} \\
&= \mathsf{E}(Y_i(1)|D_i = 1) - \mathsf{E}(Y_i(0)|D_i = 1) && (Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \\
&= \mathsf{E}(Y_i(1) - Y_i(0)|D_i = 1) && \text{Diff means is mean of diffs} \\
&= \mathsf{E}(Y_i(1) - Y_i(0)) && D_i \text{ is indep of potential outcomes}
\end{aligned}$$

How randomization solves the selection problem

- Since D_i is indep of potential outcomes, can measure average causal effect using a simple difference in means
- Randomization makes the treated and untreated groups equal on average, e.g. on average $E(X_i|D_i = 1) = E(X_i|D_i = 0)$ for any known or unknown X_i

Linear Constant Effects, Binary Treatment

Consider an experiment with a single binary treatment

$$\begin{aligned} Y_i &= Y_i(0) + (Y_i(1) - Y_i(0))D_i \\ &= \alpha + \tau D_i + (Y_i(0) - \alpha) \\ &= \alpha + \tau D_i + \eta \end{aligned}$$

Compare

$$\begin{aligned} E(Y_i|D_i = 1) &= \alpha + \tau + E(\eta_i|D_i = 1) \\ E(Y_i|D_i = 0) &= \alpha + E(\eta_i|D_i = 0) \end{aligned}$$

Difference is ATE + selection, but selection effect is 0 due to ignorability

$$\begin{aligned} E(Y_i|D_i = 1) - E(Y_i|D_i = 0) &= \tau + (E(\eta_i|D_i = 1) - E(\eta_i|D_i = 0)) \\ &= \tau \end{aligned}$$

Selection on observables

Key assumption: Conditional on observed covariates X_i, selection bias disappears

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

Means

$$E(Y_i|X_i, D_i = 1) - E(Y_i|X_i, D_i = 0) = E(Y_i(1) - Y_i(0)|X_i)$$

- Given X_i, assignment of D_i is "as if" random
- This assumption allows for a causal interpretation of observational studies
- Also called *conditional independence assumption*, *ignorability*, *no omitted variables*, *no unmeasured confounders*

Estimating Causal Effects Given Observational Data

Using the selection on observables assumption several methods to estimate causal effects:

- Regression
- Matching
- Propensity
- Weighting

Other methods: instrumental variables, panels, difference in difference, regression discontinuity

Regression and Causality

- Regression is causal when the CEF it approximates is causal (Angrist and Pischke 2008)
    - Selection-on-observables assumption is correct
- recovers a causal parameter, not necessarily the one we want.always
    - When individual causal effects are linear and homogeneous: ATE
    - When individual causal effects are non-linear or heterogeneous:
        - weighted average treatement effect
        - weights are the variance of the treatment conditional on the value of X_i, with binary D, highest where Pr$(D_i = 1|X_i) = 0.5$.

Regression and the CEF

- Linear regression approximates the conditional expectation function (CEF), $\mathsf{E}(Y|X)$.

$$Y_i = X_i\beta + \epsilon_i \approx \mathsf{E}(Y_i|X_i) + \epsilon_i \tag{1}$$

- Linear regression estimates the CEF, when the CEF is linear.
- When is the CEF linear?
  - X is multivariate normal
  - X is saturated, e.g. all combinations of binary X
- $X_i\beta$ is minimum MSE predictor of Y_i.
- $X_i\beta$ is minimum MSE predictor of the CEF, $E(Y_i|X_i)$.
- **Agnostic view of regression**. Linear approximation of the CEF, not "true model". Heteroskedasticity will occur use robust standard errors.

Saturated Regression

Suppose X_1, and X_2 are binary variables, the saturated regression is

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \epsilon_i$$

which implies

$$\begin{aligned}
\mathsf{E}(Y_i | X_{1i} = 0, X_{2i} = 0) &= \alpha \\
\mathsf{E}(Y_i | X_{1i} = 0, X_{2i} = 1) &= \alpha + \beta_2 \\
\mathsf{E}(Y_i | X_{1i} = 1, X_{2i} = 0) &= \alpha + \beta_1 \\
\mathsf{E}(Y_i | X_{1i} = 1, X_{2i} = 1) &= \alpha + \beta_1 + \beta_2 + \beta_3
\end{aligned}$$

- There is a parameter for each unique combination of the covariates
- Linear regression **always** fits the CEF because the CEF is a linear combination of the categories. It makes no assumptions about the errors
- Not always feasible to estimate this

Linear Regression, constant effects, binary treatment

- Assume a linear model for the potential outcomes

$$Y_i = \alpha + \tau D_i + \eta_i$$

- The error term η_i is mean 0, and captures the other effects of where $\mathsf{E}(\eta_i) = 0$.
- Suppose $\mathsf{E}(Y_i(1) - Y_i(0))$ is the same for everyone, and linear (true because D_i is binary).

Linear Regression, constant effects, binary treatment

- Assume $\eta_i = X_i\beta + \epsilon_i$ and $E(\eta_i) = X_i\beta$

$$\begin{aligned}\mathsf{E}(Y_i|D_i, X_i) = E(Y_i(d)|X_i) &= \alpha + \tau D_i + \mathsf{E}(\eta_i|X_i) \\ &= \alpha + \tau D_i + X_i\beta + \mathsf{E}(\epsilon_i|X_i) \\ &= \alpha + \tau D_i + X_i\beta\end{aligned}$$

- Suppose selection-on-observables holds

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i|X_i$$

Heterogeneous or nonlinear effects

- Suppose individual causal effects are not equal, or not-linear.
- Regression estimates a single parameter τ_R:

$$Y_i = \tau_R D_i + \beta X_i + \epsilon_i$$

- Is τ_R equal to ATE or ATT?
- No. τ_R is a weighted ATE, weighted by the variance of $D|X$, and with most weight to highest variance. With binary D, this is $\Pr(D_i = 1|X_i) = 0.5$.
- Matching methods can be used to estimate the ATT
- See *Mostly Harmless*, Ch 3.3.

Why Regression for Causal Effects?

- CEF approximation results
- If causal CEF is linear: gives ATE
- If causal CEF is heterogeneous or non-linear: weighted average of ICE
- In all cases can interpret regression coefficient directly
- In other words simple and usually gives you something close to what you want

Important Considerations when Using Regression for Causal Questions

- Define a “treatment variable” and “control variables”
- Consider only one “treatment variable”:
  - It is difficult to reasonably ensure selection on observables for different variables.
  - Study the effects of one cause, not $H_1, \ldots, H_\infty$
  - Defining the causes of an effect is either hard or ill-defined
  - Always ask yourself, what would be the ideal randomized experiment?

References

- Many slides derived from: Matthew Blackwell, GOV 2002 Notes `http://www.mattblackwell.org/files/teaching/gov2002-syllabus.pdf`.
- Many derivations of equations from Angrist and Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Ch 2–3
- Angrist and Pischke, *Mastering Metrics: The Path from Cause to Effect*, Ch 1–2
- Gelman and Hill, Chapter 9 and 10
- Matthew Blackwell, PSC 504 Notes, `http://www.mattblackwell.org/teaching/psc504/`.
- Gelman and Hill, Ch 5. This should have most material you need.
- *Community*, "Remedial Chaos Theory", S3E4.