Part A: Regression and causality

A2: Potential outcomes and RCTs

Kirill Borusyak

ARE 213 Applied Econometrics

UC Berkeley, Fall 2024

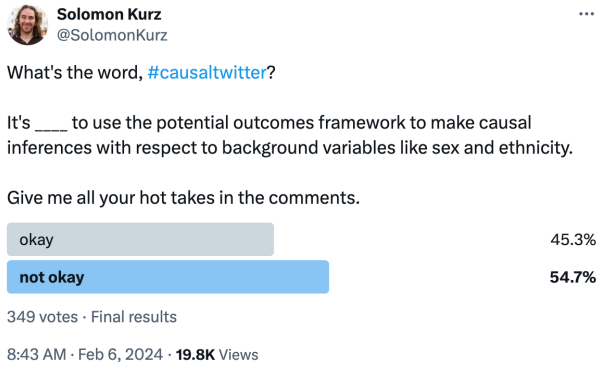
Outline

Rubin causal model

- Consider some population of units i
- Each unit is observed one of several treatment conditions $D_i \in \mathfrak{D}$
  - E.g. $\mathfrak{D} \in \{0, 1\}$: untreated and treated
- Suppose we can imagine each unit under all possible conditions (in the same period)
  - Causality always requires specifying alternatives
  - Corresponding **potential outcomes** are $\{Y_i(d) : d \in \mathfrak{D}\}$
    - e.g. $(Y_i(0), Y_i(1))$ (equivalently written as (Y_{0i}, Y_{1i}))
    - e.g. demand function
  - **Causal effects** $Y_i(d') - Y_i(d)$ are defined by this abstraction
  - Writing $Y_i(d)$ encodes a possibility that D_i impacts Y_i
  - **Realized outcome**: $Y_i = Y_i(D_i)$

What can be a cause/treatment?

Is it meaningful to say *"She did not get this position because she is a woman"* (example from Imbens 2020)*?*



What can be a cause/treatment?

Imagining each unit under all possible conditions is non-trivial:

> *"No causation without* [imagining] *manipulation"* (Holland & Rubin)

1. *"She did not get this position because she is a woman"* ✗
   - Gender is an **attribute**, not a cause; same for race
   - *"She got an orchestra job because of a gender-blind audition"* (cf. Goldin and Rouse 2000) ✓
2. *"She did well on the exam because she was coached by her teacher"* (Holland 1986) ✓
   - *"She did well on the exam because she studied for it"* (Holland 1986) ✗

SUTVA (1)

In writing $Y_i(d_i)$ we implicitly imposed **SUTVA** ("stable unit treatment value assumption")

- Most common meaning: no unmodeled **interference**
  - I.e., treatment statuses of other units, d_{-i} do not affect Y_i
  - Frequently violated: e.g. vaccines and infectious disease; information and technology adoption; equilibrium effects via prices
- Allowing for interference, we'd write $Y_i(d_1, \ldots, d_N)$ for the population of size N
  - We may be interested in own-treatment effects $Y_i(d_i', d_{-i}) - Y_i(d_i, d_{-i})$ and various spillover effects, e.g. $Y_i(d_i, 1, \ldots, 1) - Y_i(d_i, 0, \ldots, 0)$
- No interference is an exclusion restriction: $Y_i(d_i, d_{-i}) = Y_i(d_i, d'_{-i}) \equiv Y_i(d_i)$, $\forall d_i, d_{-i}, d'_{-i}$
- Intermediate case: e.g. $Y_i(\overrightarrow{d}_i)$ for **exposure mapping** $\overrightarrow{d}_i = (d_i, \sum_{k \in \text{Friends}(i)} d_k)$

SUTVA (2)

- Additional meaning of SUTVA: D summarizes everything about the intervention that is relevant for the outcome
- Example 1: *"She got a high wage because she studied for many years"*
  - In writing $Y_i(d)$, we implicitly assume that school quality does not matter
  - To think through violations, we could start from $Y_i(\text{years}, \text{quality})$
- Example 2: $D =$ Herfindahl index of migration origins in a destination region, capturing migrant diversity
  - Assumes that this index summarizes everything about exposure to migration
- Defining treatment variables is imposing a causal model. Don't take it lightly!

Effects of causes vs. causes of effects

Statistical analysis focuses on effects of causes (treatments) rather than causes of effects (outcomes)

- Causes are not clearly defined

> For example, do bacteria cause disease? Well, yes . . . until we dig deeper and find that it is the toxins the bacteria produce that really cause the disease; and this is really not it either. Certain chemical reactions are the real causes . . . and so on, ad infinitum.
>
> (Holland 1986, p.959)

Outline

Common causal parameters (1)

- We cannot learn the causal effect $Y_i(1) - Y_i(0)$ for any particular unit
  - "Fundamental problem of causal inference": multiple potential outcomes are never observed at once
  - ... but we can sometimes learn some averages
- Average treatment/causal effect: $ATE = \mathbb{E}[Y_i(1) - Y_i(0)]$
  - $ATE = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$
  - $Y_i(1) - Y_i(0)$ is never observed but $Y_i(1)$ and $Y_i(0)$ are: for some but not all units
  - Causal inference can be understood as imputing missing data: e.g. from $\mathbb{E}[Y_i(1) \mid D_i = 1]$ we try to learn $\mathbb{E}[Y_i(1) \mid D_i = 0]$ and thus $\mathbb{E}[Y_i(1)]$
- Conditional average treatment effect $\mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x]$, for predetermined covariates X_i

Common causal parameters (2)

- Average effect on the treated: $ATT = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid D_i = 1\right]$ (a.k.a. TOT, TT)
  - Parameter depends on how selection into treatment happened
  - Yields the aggregate effect of the treatment: Pop Size $\cdot\, P(D_i = 1) \cdot ATT$
- Average effect on the untreated: $ATU = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid D_i = 0\right]$
- All these parameters follow from the distribution of $(Y(1), Y(0), D)$. But are they identified from data on (Y, D)?

Identifying ATT & ATE

$$
\begin{aligned}
ATT &= \mathbb{E}[Y_1 \mid D=1] - \mathbb{E}[Y_0 \mid D=1] \\
&= (\mathbb{E}[Y_1 \mid D=1] - \mathbb{E}[Y_0 \mid D=0]) && \text{(Difference in means)} \\
&\quad - (\mathbb{E}[Y_0 \mid D=1] - \mathbb{E}[Y_0 \mid D=0]) && \text{(Selection bias)}
\end{aligned}
$$

- Thus, $\beta_{\text{OLS}} = \mathbb{E}[Y \mid D=1] - \mathbb{E}[Y \mid D=0] = ATT +$ Selection bias
  - Selection bias $= 0$ iff Y_0 is mean-independent of D
- $ATE = ATT$ iff $(Y_1 - Y_0)$ is mean-independent of D
  - Simple regression identifies ATE and ATT in a randomized control trial (**RCT**) where $(Y_0, Y_1) \perp\!\!\!\perp D$ by **design**
  - Regression with any (fixed set of) predetermined controls X also identifies ATE by FWL or OVB logic

Connecting to linear models

- With a binary treatment, the potential outcomes model implies

$$Y_i = Y_{0i}(1 - D_i) + Y_{1i}D_i = \beta_0 + \beta_{1i}D_i + \varepsilon_i$$

  where $\beta_0 = \mathbb{E}[Y_0]$, $\beta_{1i} = Y_{1i} - Y_{0i}$ and $\varepsilon_i = Y_{0i} - \mathbb{E}[Y_0]$

- With homogeneous effects, $Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$ becomes a causal *model* where $Y_{1i} - Y_{0i} \equiv \beta_1$ (regardless of whether $\varepsilon_i \perp\!\!\!\perp D_i$; think IV)

- With heterogeneous effects, can rederive our result about RCTs: if $(\varepsilon_i, \beta_{1i}) \perp\!\!\!\perp D_i$ and denoting $\mu = \mathbb{E}[D_i]$,

$$\beta_{OLS} = \frac{\mathbb{E}[(D_i - \mu)\, Y_i]}{\mathrm{Var}[D_i]} = \frac{\mathrm{Cov}[D_i, \varepsilon_i]}{\mathrm{Var}[D_i]} + \frac{\mathbb{E}[D_i (D_i - \mu_i)\, \beta_{1i}]}{\mathrm{Var}[D_i]} = \mathbb{E}[\beta_i] \equiv ATE$$

RCT with ordered or continuous treatments

Consider a RCT where D takes more than two values (e.g. different dosages)

- $D \perp\!\!\!\perp \{Y(d)\}_{d \in \mathfrak{D}} \Longrightarrow \mathbb{E}[Y \mid D = d] = \mathbb{E}[Y(d) \mid D = d] = \mathbb{E}[Y(d)]$
- A saturated regression of Y on dummies for all values of D (or a nonparametric regression with continuous D) traces the average structural function $\mathbb{E}[Y(d)]$
- A simple regression of Y on D identifies a convexly-weighted average of $\partial \mathbb{E}[Y(d)] / \partial d$ (or its discrete version):

$$\beta_{OLS} = \int_{-\infty}^{\infty} \omega(\tilde{d}) \frac{\partial \mathbb{E}\left[Y(\tilde{d})\right]}{\partial \tilde{d}} \mathrm{d}\tilde{d}, \qquad \omega(\tilde{d}) = \frac{\mathrm{Cov}\left[\mathbf{1}\left[D \geq \tilde{d}\right], D\right]}{\mathrm{Var}[D]}$$

$$\text{or } \beta_{OLS} = \sum_{k=1}^{K} \omega_k \frac{\mathbb{E}[Y(d_k) - Y(d_{k-1})]}{d_k - d_{k-1}}, \quad \omega_k = \frac{(d_k - d_{k-1}) \mathrm{Cov}[\mathbf{1}[D \geq d_k], D]}{\mathrm{Var}[D]}$$

The importance of convex weighting

- Imagine $\mathbb{E}[Y(0)] = 0$, $\mathbb{E}[Y(1)] = 3$, $\mathbb{E}[Y(2)] = 4$
  - Higher dosage is always good (on average)
- OLS of Y on D in an RCT will produce a coefficient between $\mathbb{E}\left[\frac{Y(2)-Y(1)}{2-1}\right] = 1$ and $\mathbb{E}\left[\frac{Y(1)-Y(0)}{1-0}\right] = 3$
- An estimator without convex weighting may not: e.g.

$$2 \cdot \mathbb{E}[Y(2) - Y(1)] - 1 \cdot \mathbb{E}[Y(1) - Y(0)] = -1,$$

  as if higher treatment is bad
  - Convex weighting avoids **sign reversals**. Defines a **weakly causal** estimand.

Distribution of gains

Heckman, Lalonde, Smith (1999) list some other interesting parameters:

1. How widely are the gains distributed?
    a. The proportion of people taking the program who benefit from it: $P(Y_1 > Y_0 \mid D = 1)$
    b. Median gains among participants (and other quantiles)
2. Does the program help the lower tail?
    a. Distribution of gains by untreated value: e.g. $\mathbb{E}\left[Y_1 - Y_0 \mid Y_0 = \bar{y}, D = 1\right]$
    b. Increase in % above a threshold due to a policy: $P(Y_1 > \bar{y} \mid D = 1) - P(Y_0 > \bar{y} \mid D = 1)$

Distribution of gains: Identification

Does an RCT identify these other parameters, e.g. median gains?

- Not without extra restrictions!
- E.g. imagine an RCT where Y takes values 0, 1, 2 with equal prob. in both treated and control groups
- This is consistent with (Y_0, Y_1) taking values $(0,0), (1,1), (2,2)$ with equal prob.
  - No casual effect for anyone. Median gain $= 0$
- Or with (Y_0, Y_1) taking values $(0,1), (1,2), (2,0)$ with equal prob.
  - Median gain $= 1$

Exception: $P(Y_1 > \bar{y} \mid D = 1) - P(Y_0 > \bar{y} \mid D = 1)$ is identified — *how?*

Outline

Criticisms by Heckman and Vytlacil 2007

1. Estimated effects cannot be transferred to new environments (limited external validity) and to new programs never previously implemented
   - Interventions are black boxes, with little attempt to unbundle their components
   - Mechanisms are not possible to pin down
   - Knowledge does not cumulate across studies (contrast with estimates of a labor supply elasticity — a structural parameter)
     - ⋆ Counterpoint from Angrist and Pischke (2010): *"Empirical evidence on any given causal effect is always local, derived from a particular time, place, and research design. Invocation of a superficially general structural framework does not make the underlying variation more representative. Economic theory often suggests general principles, but extrapolation to new settings is always speculative. A constructive response to the specificity of a given research design is to look for more evidence, so that a more general picture begins to emerge."*

Criticisms by Heckman and Vytlacil 2007 (cont.)

2. Estimands need not be relevant even to analyze the observed policy
   - Informative on whether to throw out the program entirely (ATT) and whether to extend it forcing it on everyone not covered yet (ATU)
   - But not whether to extend/shrink it on the margin (Heckman et al. 1999, Sec.3.4)
   - Or a policy change that affects the assignment mechanism, e.g. available options
   - No analysis from the social planner's point of view, e.g. accounting for externalities
   - No analysis of causal parameters other than means, e.g. median gains

Optional exercise: read Heckman-Vytlacil's Sec. 4.4. Do you agree with everything?

Roy model

Alternative "structural" approach: to model self-selection explicitly

- Original Roy (1951) model: self-selection based on outcome comparison
  - D = choice of occupation (e.g. agriculture vs not) or education level
  - $Y(d)$ = earnings for a given occupation/education
  - People vary by occupational productivities/returns to education, known to them
  - They choose based on them: $D = \arg\max_{d \in \mathfrak{D}} Y_i(d)$, perhaps with homogeneous costs
- Extended Roy model: costs are heterogeneous but fully determined by observables
  - which may or may not affect the outcome at the same time
- Generalized Roy model: self-selection based on unobserved preferences
  - $D = \arg\max_{d \in \mathfrak{D}} R_i(d)$ where e.g. $R_i(d) = Y_i(d) - C_i(d)$ for costs $C_i(d)$

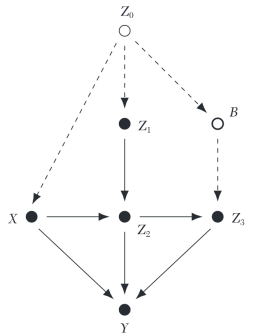
Roy model: Identification

What does this structure buy us?

- No free lunch: *"for general skill distributions* [i.e., without parametric restrictions], *the* [original Roy] *model is not identified* [from a single cross-section] *and has no empirical content"* (Heckman and Honore 1990)
- But with more data and restrictions can identify the ATE and even the distribution of $(Y_0, Y_1, R_1 - R_0) \Longrightarrow$ distribution of gains
- Assumptions are often parametric: e.g. Heckman correction via normality of potential outcomes
  - Not living up to the goal of using economic theory for identification?
- Can do better with cost shifters that shift selection but not outcomes
  - Value over traditional IV methods is not so clear?

Another alternative: Directed acyclic graphs (DAGs)

Directed acyclic graphs of Judea Pearl represent causal relationships graphically: e.g.



X = soil treatment (fumigation)
Y = crop yield
Z_1 = eelworm population before the treatment
Z_2 = eelworm population after the treatment
Z_3 = eelworm population at the end of season

Z_0 = eelworm population last season (unobserved)
B = bird population (unobserved)

- "Do-calculus" allows to verify whether the average total effect of X on Y is identified from observing (X, Y, Z_1, Z_2, Z_3)
- Popular in epidemiology but not in economics. Why?

Some limitations of DAGs

Imbens (JEL 2020) lists some pitfalls of DAGs relative to potential outcomes:

1. Economists avoid complex models with many variables
2. Randomization and manipulability have no special value in DAGs
3. Too nonparametric:
    a. Not possible to incorporate additional assumptions, such as continuity (important in RDD) ~~and monotonicity (important in IV)~~ *(see Maiti, Plecko, Bareinboim 2024)*
    b. Too much focus on identification, relative to estimation and inference
4. Difficult to model interference
5. Clunky to model simultaneity, e.g. demand and supply

Outline

Causality vs. prediction

- Economists obsess with causality but sometimes prediction is the relevant goal
- The choice should be guided by the ultimate goal: decision making
- Two scenarios (see Kleinberg et al., 2015):

1. The action/policy $D \in \{0, 1\}$ affects the outcome Y, and the payoff (i.e., utility) π depends on Y
   - E.g. $D =$ rain dance in a drought, $Y =$ it rains

$$\pi(d) = aY(d) - bd \quad \Longrightarrow \quad \mathbb{E}\left[\pi(1) - \pi(0)\right] = a\mathbb{E}\left[Y(1) - Y(0)\right] - b$$

   - Optimal decision: $D = \mathbf{1}\left[\mathbb{E}\left[Y(1) - Y(0)\right] \geq b/a\right]$
   - This is a causal problem. Running an RCT is very helpful
   - Better knowledge of heterogeneous causal effects $\mathbb{E}\left[Y(1) - Y(0) \mid X\right]$ based on observed covariates X also yields better decisions $D(X)$

Causality vs. prediction (2)

2. Y is unaffected by D but the marginal payoff of actions, $\partial\pi/\partial D$, depends on Y
   - E.g. $D =$ take an umbrella, $Y =$ it rains

$$\pi(d) = aY \cdot d - bd \quad \Longrightarrow \quad \mathbb{E}\left[\pi(1) - \pi(0)\right] = a\mathbb{E}\left[Y\right] - b$$

   - Optimal decision: $D = \mathbf{1}\left[\mathbb{E}\left[Y\right] \geq b/a\right]$
   - This is a prediction problem. Running an RCT is not helpful
   - Better prediction $\mathbb{E}\left[Y \mid X\right]$ yields better decisions $D(X)$

- *Note:* This scenario can also be recast as a causal problem:
  - D affects $\tilde{Y}(D) =$ you get wet $= Y \cdot (1 - D)$
  - But we know potential outcome $\tilde{Y}(1) = 0$
  - And we have data on $\tilde{Y}(0) = Y$ to make a *prediction* of $\tilde{Y}(1) - \tilde{Y}(0)$

Policy-relevant prediction problems: Examples

1. Eliminating futile hip and knee replacement surgeries
   - Surgery has costs: monetary + painful recovery
   - Benefits depend on life expectancy
   - Kleinberg et al. (2015) show 10% (1%) of patients have *predictable* probability of dying within a year of 24% (44%) for reasons unrelated to this surgery
2. Improving admissions by predicting college success
   - Geiser and Santelices (2007) show that high-school GPA is a better predictor of performance at UC colleges than SAT
   - If UC had to reduce admissions, rejecting applicants with marginal GPAs would result in losing fewer good students than rejecting marginal SAT applicants
3. See Kleinberg et al. "Human Decisions and Machine Predictions" (2018) for a more subtle example on bail decisions by judges