Causal Inference and Regression

Jeffrey Arnold

May 17, 2016

What is identification?

- We can learn about a quantity of interest (mean, effect, etc.)
- If quantity of interest identified, what would we know if we had access to infinite data?
- Quantity is **identified**, if with infinite data, it takes a single value
- Non-identification example: β with collinear variables in OLS
- E.g. $Y = 1 + 1X_1 + 1X_2$ where X_1 and X_2 are exhaustive dummy variables.
- Even if $N \to \infty$, any β_0, β_1, β_2 where $\beta_1 = \beta_2$, and $\beta_0 + \beta_1 + \beta_2 = 2$ will fit the data equally well.

Causal Identification

- Casual identification is what we can learn about a **causal effect** from the data
- Identification depends on **assumptions** not estimation strategies
- If an effect is not identified, no estimation method will recover it
- "What is your estimation strategy" means what assumptions are you using to claim a causal effect?# Regression and Causal Inference

What do we want to identify?

ATT

Average treatement effect

ATE

Average treatment effect on the treated

$$ATE = \sum_{i \in \text{treated}}$$

Potential Outcomes

- Potential outcome $Y_i(d)$ is

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

- Observed outcome is

$$Y_i = \underbrace{Y_{0i}}_{\text{Potential outcome } D_i = 0} + \underbrace{(Y_{1i} - Y_{0i})}_{\text{casual effect}} \underbrace{D_i}_{\text{treatment}}$$

Observed Average Differences

$$\underbrace{\mathrm{E}(Y_i|D_i=1)-\mathrm{E}(Y_i|D_i=0)}_{\text{Observed avg. diff.}} = \underbrace{\mathrm{E}(Y_{1i}|D_i=1)-\mathrm{E}(Y_{0i}|D_i=1)}_{\text{Avg. treatment effect on treated (ATT)}}$$
$$+\underbrace{E(Y_{0i}|D_i=1)-\mathrm{E}(Y_{0i}|D_i=0)}_{\text{selection bias}}$$

- Need to eliminate selection bias to estimate ATT
- If random assignment $E(Y_{0i}|D_i=1)=Y_(Y_{0i}|D_i=1)$ and everything simplifies

Regression and Potential Outcomes

$$Y_i = \underbrace{\alpha}_{E(Y_{0i})} + \underbrace{\beta}_{Y_{1i}-Y_{0i}} \underbrace{D_i}_{\text{treatment}} + \underbrace{\eta_i}_{Y_{0i}-E(Y_{0i})}$$

What do we need for identication?

- CIA: conditional independence assumption (selection on observables)
- Formally:

$$(Y_{0i}, Y_{1i}) \perp D_i | X_i$$

- Who gets the treatment is independent of their potential outcome values, once we control for X_i.
- D_i assigned "as if" random after we control for X_i
- This is why OVB is so important!
- OVB is what we need for causal identification

How does regression fit into potential outcomes causal inference?

- Regression is widely used
- But what is it doing?

Two Views of Regression and Causal Inference

1. Statistical Model
2. Causal Identification

Regression as Parametric Modeling

- Parameteric modeling
- assume data-generating process that matches theory
- estimate structural parameters of the DGP
- Regression
- CLR assumptions: linearity, iid, zero conditional mean error, homoskedasticity, normal errors
- Regression assumes the following model

$$y_i | \boldsymbol{x}_i \sim N(\boldsymbol{x}_i' \boldsymbol{\beta}, \sigma^2)$$

- Diagnostics of the suitability of those assumptions

Agnostic View

- Focus on causal identification
- Regression approximates the CEF