Frontier Topics in Empirical Economics: Week 1
Outline of Causal Inference

Zibin Huang [1]

[1] College of Business, Shanghai University of Finance and Economics

September 9, 2024

Plan of This Course

- Basic causal inference and statistical tools (Week 1-4)
  Potential outcome framework, RCT, matching vs regression, non-parametric method, machine learning, DAG framework
- IV (Week 5-7)
  IV, LATE, GMM, MTE, Bartik IV
- Causal inference with panel data (Week 8-9)
  Basic DID and event study, pre-trend testing, synthetic control, staggered DID
- Other topics (Week 10-13)
  RDD, Std err issues, Peer effect and spillover, intro to discrete choice model
- Student Presentation (Week 14-15)

Plan of This Course

- The goal of this course is to get all students stop being reg monkeys
- What is regression monkey? ⇒ Run regs without creativity
  - Running regressions without knowing why
  - Only know very basic statistical off-the-shelf methods
  - Have no economic sense, do not know any economic theory
- This is no economist, this is BAD statistician!
- This course aims to teach you
  - The logic behind regression and causal inference
  - Statistical tools beyond regression in causal inference
  - How to regularize data with your economic theory and intuition

Motivating Example: Female Labor Participation

This is an example from Professor Chao Fu.

- Consider a female labor participation problem
- Utility maximization of female i:

$$\begin{aligned} max \quad & U_i(c_i, 1 - l_i) + \epsilon_{il} \\ s.t. \quad & c_i = w_i l_i \end{aligned} \tag{1}$$

c_i: consumption; l_i: labor supply; ϵ_{il}: unobserved taste shock; w_i: wage

Motivating Example: Female Labor Participation

- Assume that l_i is binary (work, not work)
- $l_i = 1$ if $U(l = 1) \geq U(l = 0)$:

$$U_i(w_i, 0) + \epsilon_{i1} \geq U_i(0, 1) + \epsilon_{i0} \tag{2}$$

- Then given w_i, we have a threshold value of $\epsilon_{i0} - \epsilon_{i1}$ for i to choose to work:

$$\begin{aligned} l_i &= 1 \quad \text{if} \quad \epsilon_{i0} - \epsilon_{i1} < \epsilon^* \\ \epsilon^* &= U_i(w_i, 0) - U_i(0, 1) \end{aligned} \tag{3}$$

Motivating Example: Female Labor Participation

- Assume that shock $\epsilon_{i0} - \epsilon_{i1}$ has a CDF $F_{\epsilon|w}$
- We have the following working probability for i:

$$\begin{aligned} G(w) = Pr(I = 1|w) &= \int_{-\infty}^{\epsilon^*} dF_{\epsilon|w} \\ &= F_{\epsilon|w}(\epsilon^*(w)) \end{aligned} \tag{4}$$

- Two empirical research approaches for this question

Motivating Example: Female Labor Participation

1. We can directly estimate probability function G with linearity assumption
   - Assume that G is a linear function

$$G(w) = \beta_0 + \beta_1 w_i \tag{5}$$

   - Linear Probability Model $\Longrightarrow$ We can use OLS to estimate β
   - This is called "Reduced-form" approach
   - We usually identify it by some research "design" (IV, RDD, DID)
   - Thus, it is also called "Design-based" approach

Motivating Example: Female Labor Participation

2. We can estimate ϵ's CDF F, and utility function U
   - We have the likelihood function as:

$$L(\Theta^U, \Theta^F; data) = \prod_{i=1}^{N} F_\epsilon(\epsilon^*)^{l_i}[1 - F_\epsilon(\epsilon^*)]^{1-l_i} \tag{6}$$

   Θ^U is the parameter set of utility function; Θ^F is the parameter set of shock's CDF

   - We use MLE to estimate Θ^U and Θ^F $\Rightarrow$ Recover choice structure directly
   - This is called "Structural"/"Model-based" approach

Motivating Example: Female Labor Participation

For example,

- Assume a linear utility function $U = \alpha w_i + \phi(1 - l_i)$
- And ϵ follows T1EV distribution
- We have the likelihood function as:

$$
\begin{aligned}
L(\Theta^U, \Theta^F; data) &= \prod_{i=1}^{N} F_\epsilon(\epsilon^*)^{l_i}[1 - F_\epsilon(\epsilon^*)]^{1-l_i} \\
&= \prod_{i=1}^{N} (\frac{exp(\alpha w_i)}{exp(\alpha w_i) + exp(\phi)})^{l_i} \times (\frac{exp(\phi)}{exp(\alpha w_i) + exp(\phi)})^{1-l_i} \qquad (7)
\end{aligned}
$$

Two Approaches: Internal vs External Validity

- Now we compare these two approaches
- First, we need to clarify two important concepts
  - Internal validity
  - External validity
- Internal means the validity within the current specific context or environment
- External means the validity outside the current context or environment
- External refers to our attempt to extrapolate our analysis

Two Approaches: Internal vs External Validity

- There are three layers of policy evaluation (Heckman and Vytlacil, 2007)
- Take One Child Policy (OCP) as an example
  - Evaluating the impact of a historical intervention
    What was the impact of the OCP on fertility rate?
  - Forecasting the impact of an intervention previously happened in environment A to happen in another environment B
    What would be the impact if we restart the OCP in 2023?
  - Forecasting the impact of an intervention never happened in history in any environment
    What would be the impact if we force all women to give birth to at least one child?
- The first one is "internal"
- The second and the third are "external"

Two Approaches: Structural/Model-based Approach

- Target: Primitive parameters ⇒ Choice structure
  Agent's utility function, firm's production function, market structure...
- Advantages
  - Deeper economic thinking: we can understand the original decision-making process
  - Great external validity ⇒ Solid under Lucas' critique
  - More reliable counterfactual analysis
- Disadvantages
  - Need more (untestable) assumptions on functional form, distribution of unobservable...
  - Low internal validity

Two Approaches: Reduced-form/Design-based Approach

- Target: Some marginal effect of conditional expectation function
  What is the impact of A on B?
- Do not care about the mechanism ⇒ A black box of causal effect
- Advantages
  - Very reliable if you have a good exogenous shock
  - Great internal validity, not so many assumptions
- Disadvantages
  - No mechanism is revealed ⇒ More of a statistician than an economist
  - Usually effects are very local ⇒ Low external validity
    The causal effect is estimated for group A. Can it be applied to group B?
  - Hard to have external counterfactual interpretation
    Lucas' critique, General Equilibrium effect...

Two Approaches: Reduced-form/Design-based Approach

- This course will mainly focus on the Reduced-form/Design-based Approach
- Specifically, I will carefully go through traditional regression tools used in Applied Economics
- And introduce tools beyond simple regression
- I will also introduce a little Structural/Model-based Approach (DCM)
- In general, let's try not to be Reg Monkeys!

Potential Outcome Framework and RCT

- Example : Health status and hospitalization

Group	Sample Size	Mean Health Status
Hospital	7,774	3.21
No hospital	90,049	3.93

- Going to hospital makes you more sick?
- No! People go to hospital because they are sick.
- Correlation is NOT causality!!!

Potential Outcome Framework and RCT

Potential Outcome Framework/Rubin Causal Model

- Binary treatment D_i for individual i, some outcome Y_i
- Y_{0i}: The "potential outcome" of i if he/she is not treated, regardless of the treatment status in reality
- Y_{1i}: The "potential outcome" of i if he/she is treated, regardless of the treatment status in reality
- Thus, we have:

$$\begin{aligned} Y_i &= \begin{cases} Y_{1i} & \text{if} \quad D_i = 1 \\ Y_{0i} & \text{if} \quad D_i = 0 \end{cases} \qquad (8) \\ &= Y_{0i} + (Y_{1i} - Y_{0i})D_i \end{aligned}$$

Potential Outcome Framework and RCT

- Individual treatment effect: $Y_{1i} - Y_{0i}$
- Not available: There is only one world! Given i, you see either Y_{0i} or Y_{1i}
- But we can consider averages: By differencing mean outcomes from the two groups

$$\begin{aligned} & E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \\ = & \underbrace{E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]}_{\text{Average Treatment on the Treated (ATT)}} + \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{\text{Selection bias}} \end{aligned} \tag{9}$$

- ATT: Causal effect on the treated group
- Selection bias: Original difference between treated and untreated group
- Give me an example of the selection bias

Potential Outcome Framework and RCT

Randomization can solve the selection problem

- Assume that we randomly assign the treatment to the population:

$$D_i \perp\!\!\!\perp Y_{0i}, Y_{1i} \tag{10}$$

- Then we have selection bias to be zero:

$$E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0] = 0$$

- Thus, simple difference between the mean of treated and untreated group is ATT (and overall ATE)

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] = ATT = ATE$$

Regression, CEF and Causal Inference

- Regression is the most useful tool in applied econometrics
- When can we interpret regression coefficient as causal effect?
- What are the relations among regression, conditional expectation function (CEF) and treatment effect?

Regression, CEF and Causal Inference

Conditional Expectation Function (CEF)

- CEF is the conditional expectation of an outcome Y_i, given some predictor vector X_i

$$E[Y_i|X_i = x] = \int tf_y(t|X_i = x)dt \tag{11}$$

where f_y is pdf

- This is a population concept ($n \to \infty$)
- It describes a prediction of X on Y, but NOT necessarily causal
- We can always decompose Y_i as predicted part (CEF) + error part

$$Y_i = E[Y_i|X_i] + \epsilon_i \tag{12}$$

where $E[\epsilon_i|X_i] = 0$ (conditional mean zero)

Regression, CEF and Causal Inference

- CEF is the best predictor of Y_i given X_i
- It minimizes the mean squared prediction errors

Theorem 3.1.2 in MHE

Let $m(X_i)$ be any function of X_i. The CEF solves

$$E[Y_i|X_i] = argmin_{m(X_i)} E[(Y_i - m(X_i))^2]$$

so it is the MMSE predictor of Y_i given X_i.

Regression, CEF and Causal Inference

Linear Regression

- Regression is a linear prediction that minimizes the mean squared error

$$Y_i = X_i'\beta + \epsilon_i$$

$$\beta = argmin_b E[(Y_i - X_i'b)^2]$$

- We have the first order condition (moment condition) as:

$$E[X_i(Y_i - X_i'\beta)] = 0$$

- The solution can be written as:

$$\beta = E[X_iX_i']^{-1}E[X_iY_i]$$

Regression, CEF and Causal Inference

Tips: Difference between β and $\hat{\beta}_{OLS}$

- Definition

$$\beta = E[X_i X_i']^{-1} E[X_i Y_i]$$
$$\hat{\beta}_{OLS} = (X'X)^{-1} X'Y$$

- $\hat{\beta}_{OLS}$ is an estimator of β (there can be alternative estimators, e.g. MLE)
- Population vs Sample, Identification vs Estimation
- X_i is an $1 \times k$ vector, Y_i is a scalar. They are random variables
- X is an $n \times k$ matrix, Y is an $n \times 1$ vector. They are realizations of random variables (real data)

Regression, CEF and Causal Inference

CEF and linear regression

- $E[\epsilon_i|X_i] = 0$ vs $E[X_i\epsilon_i] = 0$
- Minimizing MMSE: Best predictor (CEF) vs Best linear predictor (linear regression)
- CEF is stronger than linear regression
- If CEF is linear, then linear regression is identical to CEF
- Even if CEF is not linear, regression is the best linear approximation to CEF (Minimize MSE)

Regression, CEF and Causal Inference

- For any data, you can always run a regression (as long as the rank condition is satisfied)
- But the coefficient β is not necessarily a causal effect
- When does a regression coefficient have a causal meaning?

Regression, CEF and Causal Inference

Case 1: We assume randomization (no need for controls) and constant TE

- When we have a random experiment with $D_i \perp\!\!\!\perp Y_{0i}, Y_{1i}$ and regression

$$Y_i = \alpha + \rho D_i + \epsilon_i$$

- Under this randomization, CEF is linear, then:

$$\rho = E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = E[Y_{1i} - Y_{0i}|D_i = 1]$$

- Regression coefficient ρ is the ATT/TE

Regression, CEF and Causal Inference

Case 2: We assume randomization after controls

- Key to go from correlation/prediction to causality: Conditional Independent Assumption (CIA)/Selection on Observables

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

- Treatment is random, after controlling for covariates X_i

Regression, CEF and Causal Inference

Case 2: We assume randomization after controls

- Homogeneous (constant) treatment effect case is simple
- For each $X_i = x$, we have the following regression:

$$Y_i = \alpha + \rho_r D_i + X_i^{\prime} \gamma + \nu_i \tag{13}$$

- With linear CEF, regression coefficient ρ_r is the treatment effect

$$\rho_r = E[Y_i | X_i, D_i = 1] - E[Y_i | X_i, D_i = 0] = E[Y_{1i} - Y_{0i}]$$

Regression, CEF and Causal Inference

Case 2: We assume randomization after controls

- Heterogeneous treatment effect case is more complicated
- Let δ_x be the within group ATE:

$$\delta_x = E[Y_i|X_i, D_i = 1] - E[Y_i|X_i, D_i = 0] = E[Y_{1i}|X_i, D_i = 1] - E[Y_{0i}|X_i, D_i = 1]$$

- It can be shown that ρ_r is the treatment-variance weighted average of δ_x:

$$\rho_r = \frac{E[\sigma_D^2(X_i)\delta_x]}{E[\sigma_D^2(X_i)]} \tag{14}$$

$$\sigma_D^2(X_i) \equiv E[(D_i - E[D_i|X_i])^2|X_i]$$

- Proof see MHE Chapter 3.3.1

Regression, CEF and Causal Inference

- Important! How to understand/interpret equation (14)?
- More weights are assigned to cells with largest treatment variance
- Zero weight if a cell is full of treated/untreated individuals

Regression, CEF and Causal Inference

- Homework: What is the implication of expression (14) when unconditional independence holds (Like in an RCT)? That is, when $D \perp\!\!\!\perp Y_{1i}, Y_{0i}$?

Regression, CEF and Causal Inference

Let's compare assumptions of Regression, CEF and Causal Model

- $y = f(D) + e$
- Linear Regression: $f(D) = \beta D, E(De) = 0$ Uncorrelated
- CEF: $E(e|D) = 0$ Mean Independence
- Causal Model: $e \perp\!\!\!\perp D \quad (D_i \perp\!\!\!\perp y_{0i}, y_{1i})$ Independence
- Tips: When D is dummy, linear regression is CEF

Regression, CEF and Causal Inference

Main takeaways from this part

- Strength of assumptions regarding unobservable e
  Causal model (CIA) > CEF (Mean Independence) > Linear regression (Uncorrelated)
- CEF is the best predictor of Y given X
- Linear regression is the best "linear" predictor of Y given X
- Linear regression is the best linear approximation of CEF
- Under CIA and homogeneous TE, regression coefficient is the TE
- Under CIA and heterogeneous TE, regression coefficient is the treatment-variance weighted average of group ATE

Simpson Paradox, Omitted Variables and Bad Controls

- Consider two treatments A and B for COVID
- We examine the effect of the treatments by patients' conditions (mild/severe)
- We have the death rate by treatments and conditions as:

	Mild	Severe	Total
A	15% (210/1400)	30% (30/100)	16% (240/1500)
B	10% (5/50)	20% (100/500)	19% (105/550)

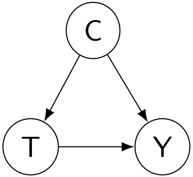
- Total death rate: A < B
- Death rate within condition group: A > B

Simpson Paradox, Omitted Variables and Bad Controls

- Which one is better? A or B? $\Longleftrightarrow$ Should we control for condition (C)?
- It depends on the causal structure!
- Condition is the cause or the consequence of the treatment?

Simpson Paradox, Omitted Variables and Bad Controls

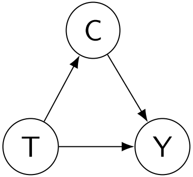
Case 1: When condition C is a cause of treatment T



- C causes T and Y; T causes Y
- C is a pre-determined variable to T ⇒ C is the cause
- We should control for C ⇒ B is better
- If we do not control for C ⇒ Omitted Variable Bias

Simpson Paradox, Omitted Variables and Bad Controls

Case 2: When treatment T is a cause of condition C



- T causes C and Y; C causes Y
- C is a post-determined variable ⇒ C is the consequence
- We should not control for C ⇒ A is better
- If we do control for C ⇒ Bad Control Problem
- Never control a channel!!!

Simpson Paradox, Omitted Variables and Bad Controls

- Rule of thumb: Control pre-determined variables, not post-determined ones
- But sometimes controlling for pre-determined variables can also be wrong
- Let's discuss this "bad control" issue in more details in Week 4
- DAG will offer you a clear and powerful tool to determine which variables to control, given your proposed causal structure

Simpson Paradox, Omitted Variables and Bad Controls

- Quiz: Should we control for X?
    - Y=wage, D=education, X=natural ability
    - Y=wage, D=education, X=labor participation decision
    - Y=GDP at t+1, D=R&D expenditure at t, X=trade volume at t+1

Matching

- Regression is only one of the tools we use to tackle causal effect
- Matching is another common tool
- It is simple and non-parametric
- Basic idea
  - (1) Compare treated and control units with same covariates;
  - (2) Put together to produce a single overall weighted average treatment effect
- Regression is a particular sort of weighted matching estimator

Matching

- Assume that for treatment D_i, we have CIA: $Y_{0i}, Y_{1i} \perp\!\!\!\perp D_i | X_i$
- We can express treatment on the treated (TOT) as:

$$\begin{aligned}\delta_{TOT} &= E[Y_{1i} - Y_{0i} | D_i = 1] = E[E[Y_{1i} - Y_{0i} | X_i, D_i = 1] | D_i = 1] \\ &= E[E[Y_{1i} | X_i, D_i = 1] - E[Y_{0i} | X_i, D_i = 1] | D_i = 1] \\ &= E[E[Y_i | X_i, D_i = 1] - E[Y_i | X_i, D_i = 0] | D_i = 1] \\ &= E[\delta_x | D_i = 1]\end{aligned}$$

- The corresponding matching estimator (sample analog) is:

$$\hat{\delta}_{TOT} = \sum_x \hat{\delta}_x \hat{P}(X_i = x | D_i = 1)$$

- Similarly, we can derive a matching estimator for ATE:

$$\hat{\delta}_{ATE} = \sum_x \hat{\delta}_x \hat{P}(X_i = x)$$

Matching vs Regression

Regression is one of the matching estimators!

- Matching estimator of TOT: $\hat{\delta}_{TOT} = \sum_x \hat{\delta}_x P(X_i = x | D_i = 1)$
  Weighted by probability mass for treated group
- Matching estimator of ATE: $\hat{\delta}_{ATE} = \sum_x \hat{\delta}_x P(X_i = x)$
  Weighted by probability mass for all units
- Regression estimator: $\frac{\sum_x \hat{\sigma}^2_D(X_i)\hat{\delta}_x}{\sum_x \hat{\sigma}^2_D(X_i)}$
  Weighted by treatment variances

Matching vs Regression

- Homework: Explain the meaning of the weights in these three estimators. To which observation/cell are they going to give the largest weights?

Propensity Score Matching

- Assume that we want to estimate college premium on wages
- To have CIA, we need a lot of controls:
  Gender, race, nationality, birth weight, IQ, parents' education, parents' income...
- Curse of dimensionality: There are too many dimensions in X_i
- We will not have enough observations for each value of X_i to estimate $\hat{\delta}_x$
- Maybe you have 10,000 observations
- But only 2 of them are Han boys with IQ 150, family income 100,000 RMB/year, parents are high-school educated
- Very hard to implement the matching estimator (but regression is still feasible)

Propensity Score Matching

- Propensity Score Matching (PSM) is a simple method to reduce the dimensionality
- Assumption 1 (CIA): $Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i | X_i$
- Assumption 2 (Overlap): $0 < P(D_i = 1 | X_i) < 1$
- PSM Theorem: If Assumptions 1 and 2 hold, we have $Y_{1i}, Y_{0i} \perp\!\!\!\perp D_i | P(X_i)$, where $P(X_i) = P(D_i = 1 | X_i)$
- We are fine, as long as we control for the propensity score $P(X)$

Propensity Score Matching

- Go back to the college premium example
- Instead of matching across all controls (gender, family income...)
- We can match for the predicted probability $P(X)$ for each person to go to college
- We just replace all X_i with $P(X_i)$ in the matching estimator, and get the PSM estimator.

Regression vs PSM

- Regression usually does not suffer from the curse of dimensionality
- Since we are regularizing controls by linear function (next class)
- We can also combine regression and PSM by running a regression, controlling for propensity score (but not each variable)

Regression vs PSM

- In general, Angrist prefers regression
- Because some parts of the process to implement PSM are not standardized
- e.g. how to estimate the propensity score $P(X)$? (Logit? LPM? Probit?)
- PSM CANNOT solve the endogeneity issue!!!!!!
- PSM CANNOT solve the endogeneity issue!!!!!!
- PSM CANNOT solve the endogeneity issue!!!!!!

References

Heckman, James J and Edward J Vytlacil. 2007. “Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation.” *Handbook of Econometrics* 6:4779–4874.