Program Evaluation (Causal Inference) 1: Introduction

Yuta Toyama

Last update: July 2, 2019

Introduction

- Program Evaluation, or Causal Inference
    - Estimation of "treatment effect" of some intervention (typically binary)
    - Example:
        - effects of job training on wage
        - effects of advertisement on purchase behavior
        - effects of distributing mosquito net on children's school attendance
- Difficulty: treatment is **endogenous decision**
    - selection bias, omitted variable bias.
    - especially in observational data (in comparison with experimental data)

Overview

- Introduce Rubin's causal model (potential outcome framework)
  - Generalization of the linear regression model: Nonparametric

- Solutions to the selection bias
  1. Randomized control trial (today)
  2. Matching (today)
  3. Instrumental Variable Estimation (today)
  4. Difference-in-differences (next week)
  5. Regression Discontinuity Design (week after next)

Reference

- Angrist and Pischke:
  - Mostly harmless econometrics : advanced undergraduate to graduate students
  - Mastering Metrics: good for undergraduate students after taking econometrics course.
- Ito: Data Bunseki no Chikara (in Japanese)

Framework

- Y_i: observed outcome for person i

- D_i: treatment status

$$D_i = \begin{cases} 1 & \textit{treated (treatment group)} \\ 0 & \textit{not treated (control group)} \end{cases}$$

- Define *potential outcomes*
  - Y_{1i}: outcome for i when she is treated (treatment group)
  - Y_{0i}: outcome for i when she is not treated (control group)

- With this, we can write

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

Two Key points

- Point 1: Fundamental problem of program evaluation
  - We can observe (Y_i, D_i), but never observe Y_{0i} and Y_{1i} **simultaneously**.
  - **Counterfactual** outcome.
- Point 2: Stable Unit Treatment Value Assumption (SUTVA)
  - Treatment effect for a person does **not depend on the treatment status of other people.**
  - Rules out externality / general equilibrium effects.
    - Ex: If everyone takes the job training, the equilibrium wage would change, which affects the individual outcome.

Parameters of Interest

- Define the individual treatment effect $Y_{1i} - Y_{0i}$
    - Key: allowing for heterogenous effects across people
- Individual treatment effect cannot be identified due to the fundamental problem.
- Instead, we focus on the average effects
    - Average treatment effect: $ATE = E[Y_{1i} - Y_{0i}]$
    - Average treatment effect on treated: $ATT = E[Y_{1i} - Y_{0i}|D_i = 1]$
    - Average treatment effect on untreated: $ATT = E[Y_{1i} - Y_{0i}|D_i = 0]$
    - Average treatment effect conditional on covariates X_i: $ATE(x) = E[Y_{1i} - Y_{0i}|D_i = 1, X_i = x]$

Relation to Regression Analysis

- Assume that
  1. linear (parametric) structure in Y_{0i}, and
  2. constant (homogenous) treatment effect,

$$\begin{aligned} Y_{0i} &= \beta_0 + \epsilon_i \\ Y_{1i} - Y_{0i} &= \beta_1 \end{aligned}$$

- You will have

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i$$

- Program evaluation framework is nonparametric in nature.
  - Though, in practice, estimation of treatment effect relies on a parametric specification.

Selection Bias

- Consider the comparison of average outcomes between treatment and control group

- Does this tell you average treatment effect? No in general!

$$\underbrace{E[Y_i|D_i=1]-E[Y_i|D_i=0]}_{simple\ \ comparison} = E[Y_{1i}|D_i=1]-E[Y_{0i}|D_i=0]$$

$$= \underbrace{E[Y_{1i}-Y_{0i}|D_i=1]}_{ATT}$$

$$+ \underbrace{E[Y_{0i}|D_i=1]-E[Y_{0i}|D_i=0]}_{selection\ \ bias}$$

- The bias term $E[Y_{0i}|D_i=1]-E[Y_{0i}|D_i=0]$
  - not zero in general: Those who are taking the job training **would do a good job even without job training**
  - Cannot observe $E[Y_{0i}|D_i=1]$: the outcome of people in treatment group when they are NOT treated (counterfactual).

Solutions

- The core of program evaluation is how to identify (estimate) the treatment effect parameters.

- Randomized Control Trial (A/B test):
  - Assign treatment D_i randomly
- Matching (regression):
  - Using observed characteristics of individuals to control for selection bias
- Instrumental variable
  - Use the variable that affects treatment status but is not correlated to the outcome
- Difference-in-differences
  - Use the panel data to control for individual heterogeneity by fixed effects.
- Regression Discontinuity Design
  - Exploit the randomness around the thresholds.

- Others: Bound approach, synthetic control method, regression kink design, etc..

What is RCT ?

- RCT: Randomized Controlled Trial
- Measure the effect of “treatment” by
  1. randomly assigning treatment to a particular group (treatment group)
  2. measure outcomes of subjects in both treatment and “control” group.
  3. the difference of outcomes between these two groups is “treatment” effect.
- Starts with clinical trial: measure the effects of medicine.

Example from Development Economics

- Esther Duflo “Social experiments to fight poverty”
  - `https://www.ted.com/talks/esther_duflo_social_experiments_to_fight_poverty?language=en`

Framework

- Key assumption: Treatment D_i is independent with potential outcomes (Y_{0i}, Y_{1i})

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

- Under this assumption,

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

- The sample selection does not exist! Thus,

$$\underbrace{E[Y_i|D_i = 1] - E[Y_i|D_i = 0]}_{simple\ \ comparison} = \underbrace{E[Y_{1i} - Y_{0i}|D_i = 1]}_{ATT}$$

- Difference of the sample average is consistent estimator for the ATT

$$\frac{\frac{1}{N}\sum_{i=1}^{N} Y_i \cdot \mathbf{1}\{D_i = 1\}}{\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{D_i = 1\}} - \frac{\frac{1}{N}\sum_{i=1}^{N} Y_i \cdot \mathbf{1}\{D_i = 0\}}{\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{D_i = 0\}}$$

Example: RAND Health Insurance Experiment (HIE)

- Taken from Angrist and Pischke (2014, Sec 1.1)
- 1974-1982, 3958 people, age 14-61
- Randomly assigned to one of 14 insurance plans.
  - No insurance premium
  - Different provisions related to cost sharing

- 4 categories
  - Free
  - Co-insurance: Pay 25-50% of costs
  - Deductible: Pay 95% of costs, up to $150 per person ($450 per family)
  - Catastrophic coverage: 95% of health costs. No upper limit. Approximate "no insurance"

First step: Balance Check

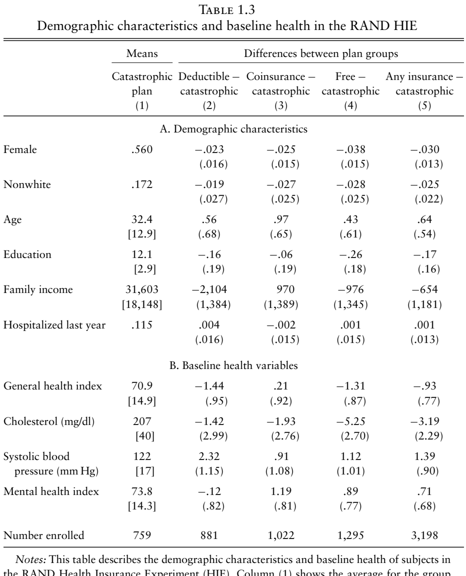

TABLE 1.3
Demographic characteristics and baseline health in the RAND HIE

	Means	Differences between plan groups			
	Catastrophic plan (1)	Deductible – catastrophic (2)	Coinsurance – catastrophic (3)	Free – catastrophic (4)	Any insurance – catastrophic (5)
A. Demographic characteristics					
Female	.560	−.023 (.016)	−.025 (.015)	−.038 (.015)	−.030 (.013)
Nonwhite	.172	−.019 (.027)	−.027 (.025)	−.028 (.025)	−.025 (.022)
Age	32.4 [12.9]	.56 (.68)	.97 (.65)	.43 (.61)	.64 (.54)
Education	12.1 [2.9]	−.16 (.19)	−.06 (.19)	−.26 (.18)	−.17 (.16)
Family income	31,603 [18,148]	−2,104 (1,384)	970 (1,389)	−976 (1,345)	−654 (1,181)
Hospitalized last year	.115	.004 (.016)	−.002 (.015)	.001 (.015)	.001 (.013)
B. Baseline health variables					
General health index	70.9 [14.9]	−1.44 (.95)	.21 (.92)	−1.31 (.87)	−.93 (.77)
Cholesterol (mg/dl)	207 [40]	−1.42 (2.99)	−1.93 (2.76)	−5.25 (2.70)	−3.19 (2.29)
Systolic blood pressure (mm Hg)	122 [17]	2.32 (1.15)	.91 (1.08)	1.12 (1.01)	1.39 (.90)
Mental health index	73.8 [14.3]	−.12 (.82)	1.19 (.81)	.89 (.77)	.71 (.68)
Number enrolled	759	881	1,022	1,295	3,198

Notes: This table describes the demographic characteristics and baseline health of subjects in the RAND Health Insurance Experiment (HIE). Column (1) shows the average for the group

- Differences in demographic characteristics & baseline health are statistically insignificant

- Assignment of health insurance plans is indeed random!

Results of RAND HIE

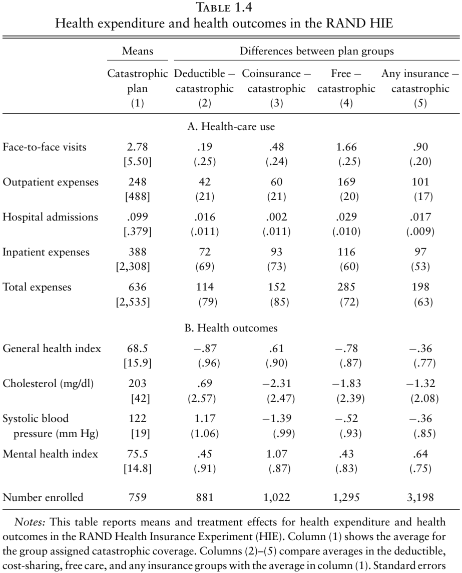

Table 1.4
Health expenditure and health outcomes in the RAND HIE

	Means	Differences between plan groups			
	Catastrophic plan (1)	Deductible – catastrophic (2)	Coinsurance – catastrophic (3)	Free – catastrophic (4)	Any insurance – catastrophic (5)
		A. Health-care use			
Face-to-face visits	2.78	.19	.48	1.66	.90
	[5.50]	(.25)	(.24)	(.25)	(.20)
Outpatient expenses	248	42	60	169	101
	[488]	(21)	(21)	(20)	(17)
Hospital admissions	.099	.016	.002	.029	.017
	[.379]	(.011)	(.011)	(.010)	(.009)
Inpatient expenses	388	72	93	116	97
	[2,308]	(69)	(73)	(60)	(53)
Total expenses	636	114	152	285	198
	[2,535]	(79)	(85)	(72)	(63)
		B. Health outcomes			
General health index	68.5	−.87	.61	−.78	−.36
	[15.9]	(.96)	(.90)	(.87)	(.77)
Cholesterol (mg/dl)	203	.69	−2.31	−1.83	−1.32
	[42]	(2.57)	(2.47)	(2.39)	(2.08)
Systolic blood	122	1.17	−1.39	−.52	−.36
pressure (mm Hg)	[19]	(1.06)	(.99)	(.93)	(.85)
Mental health index	75.5	.45	1.07	.43	.64
	[14.8]	(.91)	(.87)	(.83)	(.75)
Number enrolled	759	881	1,022	1,295	3,198

Notes: This table reports means and treatment effects for health expenditure and health outcomes in the RAND Health Insurance Experiment (HIE). Column (1) shows the average for the group assigned catastrophic coverage. Columns (2)–(5) compare averages in the deductible, cost-sharing, free care, and any insurance groups with the average in column (1). Standard errors

- HI increases health spending (Panel A)
- But, HI has no statistically significant effect on health outcomes

Matching

- Idea: Compare **individuals with the same characteristics** X across treatment and control groups

- Let X_i denote the observed characteristics: age, income, education, race, etc...

- Assumption 1:

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

  - Conditional on X_i, no selection bias.
  - Selection on observables assumption / ignorability

- Assumption 2: Overlap assumption

$$P(D_i = 1 | X_i = x) \in (0, 1) \; \forall x$$

  - Given x, we should be able to observe people from both control and treatment group.

Identification

- The assumption implies that

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

- The *ATT* for $X_i = x$ is given by

$$\begin{aligned} E[Y_{1i} - Y_{0i}|D_i = 1, X_i] &= E[Y_{1i}|D_i = 1, X_i] - E[Y_{0i}|D_i = 1, X_i] \\ &= E[Y_i|D_i = 1, X_i] - E[Y_{0i}|D_i = 0, X_i] \\ &= \underbrace{E[Y_i|D_i = 1, X_i]}_{avg\ \ with\ \ X_i\ \ in\ \ treatment} - \underbrace{E[Y_i|D_i = 0, X_i]}_{avg\ \ with\ \ X_i\ \ in\ \ control} \end{aligned}$$

- The components in the last line are identified (can be estimated).
- Intuition: Comparing the outcome across control and treatment groups after conditioning on X_i

ATT and ATE

- ATT is given by

$$\begin{aligned} ATT &= E[Y_{1i} - Y_{0i}|D_i = 1] \\ &= \int E[Y_{1i} - Y_{0i}|D_i = 1, X_i = x] f_{X_i}(x|D_i = 1) dx \\ &= E[Y_i|D_i = 1] - \int \left(E[Y_i|D_i = 0, X_i = x]\right) f_{X_i}(x|D_i = 1) \end{aligned}$$

- ATE is

$$\begin{aligned} ATE =& E[Y_{1i} - Y_{0i}] \\ &= \int E[Y_{1i} - Y_{0i}|X_i = x] f_{X_i}(x) dx \\ &= \int E[Y_i|D_i = 1, X_i = x] f_{X_i}(x) dx \\ &= + \int E[Y_i|D_i = 0, X_i = x] f_{X_i}(x) dx \end{aligned}$$

Estimation Methods

- We need to estimate $E[Y_i|D_i = 1, X_i = x]$ and $E[Y_i|D_i = 0, X_i = x]$
- Several ways to implement the above idea
- Regression: Nonparametric and Parametric
- Nearest neighborhood matching
- Propensity Score Matching: Skipped

Regression, or Analogue Approach

- Let $\hat{\mu}_k(x)$ be an estimator of $\mu_k(x) = E[Y_i|D_i = k, X_i = x]$ for $k \in \{0, 1\}$
- The analog estimators are

$$A\hat{T}E = \frac{1}{N}\sum_{i=1}^{N} \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$$

$$A\hat{T}T = \frac{N^{-1}\sum_{i=1}^{N} D_i(Y_i - \hat{\mu}_0(X_i))}{N^{-1}\sum_{i=1}^{N} D_i}$$

- How to estimate $\mu_k(x) = E[Y_i|D_i = k, X_i = x]$?

Nonparametric Estimation

- Suppose that $X_i \in \{x_1, \cdots, x_K\}$ is discrete with small K
  - Ex: two demographic characteristics (male/female, white/non-white). $K = 4$
- Then, a nonparametric binning estimator is

$$\hat{\mu}_k(x) = \frac{\sum_{i=1}^{N} \mathbf{1}\{D_i = k, X_i = x\} Y_i}{\sum_{i=1}^{N} \mathbf{1}\{D_i = k, X_i = x\}}$$

- Here, I do not put any parametric assumption on $\mu_k(x) = E[Y_i | D_i = k, X_i = x]$.
- Issue: Poor performance if K is large due to many covariates
  - **curse of dimensionality**
- If X can take continuum value, you can use kernel regression.

Parametric Estimation, or going back to linear regression

- If you put parametric assumption such as

$$E[Y_i|D_i = 0, X_i = x] = \beta' x_i$$
$$E[Y_i|D_i = 1, X_i = x] = \beta' x_i + \tau_0$$

  then, you will have a model

$$y_i = \beta' x_i + \tau D_i + \epsilon_i$$

- You can think the matching estimator as controlling for omitted variable bias by adding (many) covariates (control variables) x_i.

- This is one reason why matching estimator may not be preferred in empirical research.
  - Remember: Controlling for those covariates is of course important. This can be combined with other empirical strategies (IV, DID, etc).

$M-$Nearest Neighborhood Matching

- Fine the counterpart in other group that is close to me.

- Define $\hat{y}_i(0)$ and $\hat{y}_i(1)$ be the estimator for (hypothetical) outcomes when treated and not treated.

$$\hat{y}_i(0) = \begin{cases} y_i & if \;\; D_i = 0 \\ \frac{1}{M}\sum_{j \in L_M(i)} y_j & if \;\; D_i = 1 \end{cases}$$

- $L_M(i)$ is the set of M individuals in the opposite group who are "close" to individual i
  - Closeness is defined as the distance between X_i and X_j
  - There are several ways to define the distance. For example,

$$dist(X_i, X_j) = ||X_i - X_j||^2$$

- You need to choose (1) M and (2) the measure of distance to implement this.
- R has several packages for this.

Other Approaches

- Instrumental Variable: same idea.
    - IV estimation in program evaluation framework involves with the argument of local average treatment effect (LATE), which is beyond the scope of this course.
- Difference in differences (week 14)
- Regression discontinuity design (week 15)