Counterfactual Framework of Causal Inference

Han Zhang

March 8, 2022

Outline

Identification

- We have learned how to use samples to estimate and make statistical inference over some population quantity (e.g., $P(X)$ or $E(Y|X)$)
- What if we cannot observe some random variables?
- Statistical identification: use some observed random variables to infer properties about random variables that cannot be observed, or unobserved.
- To address identification problem, we need additional assumptions about our data

Missing Data

- Missing data is one common identification problem
- E.g., in a survey, people answer "Don't know"
- Let us work with the simplest case: we are interested in only one random variable Y.
- And we draw a sample of n points, $Y_1, \cdots, Y_n$ from population Y.
- Define R_i be an indicator for whether or not we observe Y_i
- General solutions:
  1. Bounds: possible ranges of Y
  2. Deletion: discard missing ones
  3. Imputation: predict missing Y

Missing data: bounds

- Assume we are interested in $E(Y)$
- How do we estimate $E(Y)$ in the presence of missing data?
- Suppose we see a data that looks like the below

Unit	Y_i	R_i
1	1	1
2	?	0
3	1	1
4	0	1
5	1	1
6	?	0

(1)

- And we know that Y can take values between $[0, 1]$ (Y can be continuous)
- What is the maximum possible value of $E(Y)$?

Missing data: bounds

- The largest value of Y is 1. We just fill in them, and calculate the largest possible value of $E(Y)$

Unit	Y_i	R_i
1	1	1
2	1	0
3	1	1
4	0	1
5	1	1
6	1	0

(2)

- The largest possible $E(Y)$ is $5/6$
- Likewise, we plug in the smallest value of Y
- The smallest possible value of $E(Y)$ is $3/6$
- We obtained bounds for $E(Y|X)$: $[3/6, 5/6]$; this is known as Manski bounds.
- Note that bounds are not confidence intervals. WHY?

Missing data: deletion

- Bounds can often be very wide, making them not that useful
- We can make stronger assumption to obtain more meaningful point estimation of $E(Y)$

Definition (MCAR: Missing Complete at Random, Rubin, 1976)

Y is missing completely at random if:

1. The missing $Y \perp\!\!\!\perp R$ (Response is independent of the missing Y we are interested in).
2. $P(R = 1) > 0$ (non-zero response probability)

Missing data: deletion

- MCAR assumption implies that

$$E(Y) = E(Y|R = 1) \tag{3}$$

- The right hand side is something we can estimate: the sample mean for those we can observe (apply plug-in principle)
- Practical implication: if MCAR holds, we can safely delete missing Y, and $E(Y|R = 1)$ is an unbiased estimates of $E(Y)$

Missing data: imputation

- We can also impute missing values to estimate $E(Y)$

Unit	Y_i	R_i
1	1	1
2	?	0
3	1	1
4	0	1
5	1	1
6	?	0

- Instead of deleting missing rows, we can fill in values

Imputation Method 1: Unconditional Mean Imputation

- Unconditional mean imputation fill in missing Y by the sample mean of observed Y

Unit	Y_i	R_i
1	1	1
2	$\hat{\mathrm{E}}[Y\|R=1]=\frac{3}{4}$	0
3	1	1
4	0	1
5	1	1
6	$\hat{\mathrm{E}}[Y\|R=1]=\frac{3}{4}$	0

- After unconditional mean imputation, the sample mean of imputed Y is an unbiased estimate of Y
  - Note: this is not the only way to make $E(Y)=E(Y|R=1)$
- Deletion and imputation all lead to unbiased estimate of $E(Y)$
- Their variance estimates are usually different!
  - $\hat{V}_{deletion}(Y)=0.25$
  - $\hat{V}_{imputation}(Y)=0.15$

MCAR in multivariate case

- When we have multiple variables, we can extend MCAR assumptions: each variable is independent of response.
- And with MCAR assumptions, we can perform listwise deletion by removing any row that has missing entries.

Unit	Y_i	R_i	X_i
1	1	1	0
2	?	0	0
3	1	1	0
4	0	1	0
5	1	1	?
6	?	0	1

(4)

- Or taking the imputation perspective, we can perform unconditional mean imputation for each variables

MAR

- MCAR is often too strong in multivariate case
  - If there are many variables, we can delete a lot of observations
  - Often these variable are correlated with each other;
- One weaker assumption is MAR, also known as ignorability

Definition (MAR: Missing at Random, Rubin, 1976)

Y is missing at random if:

1. $Y \perp\!\!\!\perp R | X$ (Response is independent of Y, given some other variables X).
2. $P(R = 1) > 0$ (non-zero response probability)

- That is, Y is missing at random, once we condition on some control variables X.

Post-stratification estimator of sample mean

- Under MAR, we can estimate the mean of Y using post-stratification estimator

$$E(Y) = \sum_x E(Y|R = 1, X = x)p(X = x) \tag{5}$$

- In other words, we estimate $E(Y)$ as the weighted mean of the conditional expectation of Y given X in observed data, with weights $P(X = x)$
- Both terms on the right hand side can be estimated from samples (plug-in sample analog)
- Note: post-stratification estimator does not impute; directly estimate $E(Y)$

MAR vs MCAR

- Under MCAR: $\hat{\mathrm{E}}[\mathrm{Y}_i] = 3/4$

Unit	Y_i	R_i	X_i
1	1	1	0
2	?	0	0
3	1	1	0
4	0	1	0
5	1	1	1
6	?	0	1

- Under MAR, with stratification estimator, $\hat{\mathrm{E}}[\mathrm{Y}_i] = 7/9$

$$\hat{\mathrm{E}}[\mathrm{Y}] = \hat{\mathrm{E}}[Y|R=1,X=0]\,\widehat{\mathrm{P}}[X=0] + \hat{\mathrm{E}}[Y|R=1,X=1]\,\widehat{\mathrm{P}}[X=1]$$
$$= \frac{2}{3}\cdot\frac{4}{6} + 1\cdot\frac{2}{6} = \frac{7}{9}$$

- MCAR and MAR will yield different estimates of $E(Y)$
- Each estimate is unbiased estimate only if the corresponding assumption is true

Imputation method 2: Conditional Mean Imputation

- With MAR, we can also impute Y using conditional mean imputation: use the conditional mean of Y as our guesses of the missing Y
- $Y_i = \hat{E}(Y|R = 1, X = X_i)$

Unit	Y_i	R_i	X_i
1	1	1	0
2	$\hat{\mathrm{E}}[Y_i\|X_i = 0] = \frac{2}{3}$	0	0
3	1	1	0
4	0	1	0
5	1	1	1
6	$\hat{\mathrm{E}}[Y_i\|X_i = 1] = 1$	0	1

(6)

- Then we can calculate sample mean over imputed Y
- Under conditional mean imputation, $\hat{E}(Y)$ is again $7/9$
- The below two gives the same estimate of $E(Y)$:
  - conditional mean imputation of Y, and then take sample mean of imputed Y
  - post-stratification estimator

Conditional Mean Imputation using linear regression

- If we further assume all assumptions of linear regression are correct: $E(Y|R=1, X=x)$ is linear in X
- Then conditional mean imputation just uses predicted values of linear regression as imputed values

Unit	Y_i	R_i	$X_{[1]i}$	$X_{[2]i}$
1	1	1	0	3
2	?	0	0	7
3	1	1	0	9
4	0	1	0	5
5	1	1	1	4
6	?	0	1	3

(7)

Conditional Mean Imputation using regression

Unit	Y_i	R_i	$X_{[1]i}$	$X_{[2]i}$
1	1	1	0	3
2	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 7$	0	0	7
3	1	1	0	9
4	0	1	0	5
5	1	1	1	4
6	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 3$	0	1	3

(8)

Conditional Mean Imputation using other methods

- The interpretation advantage of linear regression is not relevant now; we do not care about interpreting β; we want our predictions of $E(Y|R=1, X=x)$ to be more precise
- So you can use GLM to predict $E(Y|R=1, X=x)$
  - GLM
- Or other more complex machine learning algorithms. It's a prediction problem!
- These options are all provided in `R` package `mice`

Imputation Method 3: hot-deck imputation

- Hot-deck imputation uses nearest-neighbor matching
- For unit i with missing Y_i, and non-missing X_i
  - Find the X_j that has the smallest distance to/is closet X_i
  - Use the Y_j associated with j as the imputed Y value for i

Unit	Y_i	R_i	X_i
1	1	1	4
2	?	0	8
3	1	1	1
4	0	1	12
5	1	1	20
6	?	0	3

- Example: unit 6's X is closest to unit 1's X. So we impute Y_6 as $Y_1 = 1$

Hot-deck imputation using propensity scores

- When we have multivariate X, it is not easy to calculate their distances
- Instead, it is popular to estimate propensity score of response

$$P(R = 1|X) \tag{9}$$

- Propensity score of response provides an single-number summary of multivariate X
- Hot-deck imputation based on nearest propensity score, not based on original distances between X
  - In other words, you want to match units whose response propensity are similar
- Estimation of propensity scores
  - Logistic regression is the default choice
  - But apparently other machine learning methods are acceptable

Hot-deck example

Unit	Y_i	R_i	$X_{[1]i}$	$X_{[2]i}$	$P(R_i = 1 \mid X_i)$
1	2	1	0	3	0.33
2	?	0	0	7	0.14
3	3	1	0	9	0.73
4	10	1	0	5	0.35
5	12	1	1	4	0.78
6	?	0	1	3	0.70

(10)

- Unit 6's propensity score of response is closest to unit 3's propensity score. Thus Y_6 is imputed as $Y_3 = 3$

Deletion vs Imputation

- In practice, assume we want to run a regression based on Y and 10 predictors X
- Solution 1: Listwise deletion
  - Both R and Stata uses this strategy by default
  - Pros: simple; unbiased if MAR is true
  - Cons: large standard errors (since you will drop many cases)
- Solution 2: mean imputation (unconditional or conditional)
  - Pros:
    - give you more cases to work with
    - also unbiased if MCAR/MAR is true
  - Cons: small standard errors. Why?
    - Artificially fix the missing Y to its mean.
- Solution 3: hot-deck imputation
  - Pros: preserve the support of original data
    - Bear similarity to propensity score matching
  - Cons: how to estimate propensity scores?

Stochastic Imputation

- Problem of mean imputation: small standard error issues when we use sample mean for imputation
- A workaround—stochastic imputation—add some random noise to the sample mean
  - Say, we still use regression to impute Y, but add some random noises to your predicted Y
  - If we are working with complex machine learning models, there may be some inherent stochastic component (results are not the same every time)
- Problem: stochastic imputation also have some uncertainty, based on what noises you use
  - These random noises are not added to your final analysis, thus still producing small standard error estimates

Multiple Imputation

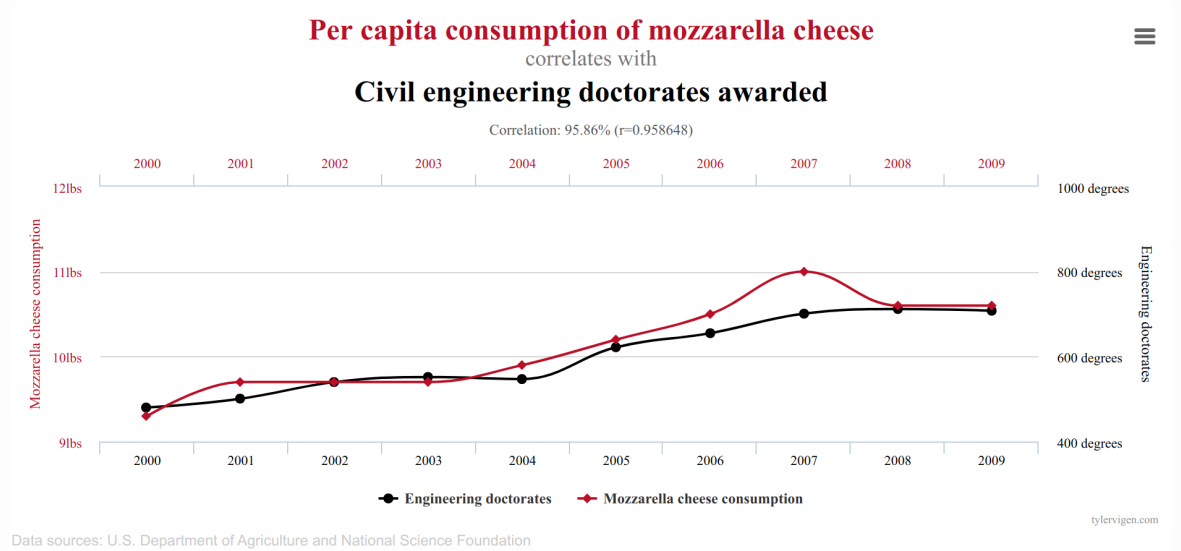
- Rubin, 1977, Multiple Imputation
  - Stochastic imputation for m times; ending up with m imputed datasets.
  - Analysis: Run your model (regression Y on 10 X) m times
  - Pooling: parameter estimates for m different models can be used for estimation and inference:
    - The final parameter estimates of β is the mean of β across m models
    - The standard error of final β is more complex in math
    - basically it's the (within model standard error of β) + (between model standard error of β)
    - Or use bootstrap if m is large enough

Missing data by Chained Equations

- In practice, more than one X can have missing values
- Assume we have 5 X; we use $1, 2, 3, 4$ to impute the 5th variable, and then use $1, 2, 3, 5$ to impute the 4th variable, and so on and so forth
    - Imputed values are allowed to use in the next step

Prediction vs Causation

- Correlation $\neq$ causation
- We can use X to predict Y, and use Y to predict X
- $Y = g(X) \iff X = g^{-1}(Y)$
- This does not capture the intuitive idea that X causes Y



https://www.tylervigen.com/spurious-correlations

Confounder

- Does college education lead to higher wages?
- Observed (Factorial): on average, college graduates indeed earn more than people with only high school education
- Critique:
  - People who can go to college have higher ability
  - Even if they did not go to education, they could still earn more
  - Ability is a confounder, which produces the correlation between education and wages

Counterfactual thinking

- How can we say that education indeed has a causal effect on income?
- Guess the (counterfactual) earning of college graduates, if they did not go to college
  - In other word, if a college graduate did not go to college, what his/her wage would be?
- If the counterfactual earning equals to the factual earning, then college education does not matter; there is no causal effect
- If the factual earning is higher than counterfactual earning, then college education indeed matters
  - the difference between factual and counterfactual earning is the causal effect due to education

Neyman-Rubin Causal Model: potential outcomes

- Neyman-Rubin Causal Model formally write down the counterfactual idea
- We have a binary treatment D; $D = 1$ if treated and 0 otherwise
- For a person i in the population, her outcome Y_i is assumed to be:

Definition (Neyman-Rubin model)

$$Y_i = \begin{cases} Y_i^0 : D_i = 0 \\ Y_i^1 : D_i = 1 \end{cases}$$
$$= Y_i^0 + D_i(Y_i^1 - Y_i^0)$$

- Y_i is observed outcome
- Y_i^0 is the potential outcome if i is not treated
- Y_i^1 is the potential outcome if i is treated

Counterfactual Outcome as Missing Data Problem

Unit	Y_i	D_i
1	1	1
2	1	0
3	1	1
4	0	1
5	1	1
6	0	0

(11)

Unit	Y_i^0	Y_i^1	D_i
1	?	1	1
2	1	?	0
3	?	1	1
4	?	0	1
5	?	1	1
6	0	?	0

(12)

SUTVA

- In the content of this class, Neyman-Rubin model often adds another assumption: stable unit treatment value assumption (SUTVA)
- "the potential outcome observation on one unit should be unaffected by the particular assignment of treatments to the other units"
- What does this mean?
  - Being in the treatment or control group does not impact the underlying potential outcomes
  - Regardless of whether I go to college or not, my potential outcome (if I did not go to college) is fixed
  - We just don't know what values they are, but they are fixed

Violation of SUTVA

- A common violation of SUTVA is in the presence of social networks
  - E.g., my potential outcome, if I did not go to college, depends on my friends' choice
  - (all my friends are college educated) versus (all my friends are not college educated)
  - My potential outcome might be larger in the former case than the latter case
- This kind of violation of SUTVA is often called interference
- And is a cutting edge research area in causal inference
- In everything we are going to learn in this class, we assume no interference and no violation of SUTVA

Individual Level Treatment Effect

Definition (Unit-Level Treatment Effect)

For i in the population, the causal effect of treatment for unit i is : $\rho_i = Y_i^1 - Y_i^0$

- $Y_i = Y_i^0 + D_i(Y_i^1 - Y_i^0)$
- So the Neyman Rubin model suggests that the observed Y for a treated unit i are the two sums:
  - counterfactual outcome if i were not treated
  - individual level treatment effect

Fundamental Problem of Causal Inference

Definition (Fundamental Problem of Causal Inference, Holland, 1986)

At any given time, we only observe one of the potential outcomes for unit i — either Y_i^1 or Y_i^0 —but not both. Thus unit-level treatment effect ρ_i is unknown.

- Similar to missing data problems, we have to make assumptions (here, assumptions about potential outcomes) to allow identification.
  - In particular, identification of the average of treatment effect ρ_i because identifying the effect for every unit can be extremely hard
  - But calculating average is easier

ATE and ATT

Definition (Average Treatment Effect (ATE))

$$ATE = E(\rho) = E(Y^1 - Y^0) = E(Y^1) - E(Y^0)$$

- ATE is the mean of unit-level treatment effect
- ATE is the difference between the mean of two potential outcomes

Definition (Average Treatment Effect on the Teated (ATT))

$$ATT = E(\rho|D = 1) = E(Y^1 - Y^0|D = 1) = E(Y^1|D = 1) - E(Y^0|D = 1)$$

- ATT is the mean of unit-level treatment effect for treated units

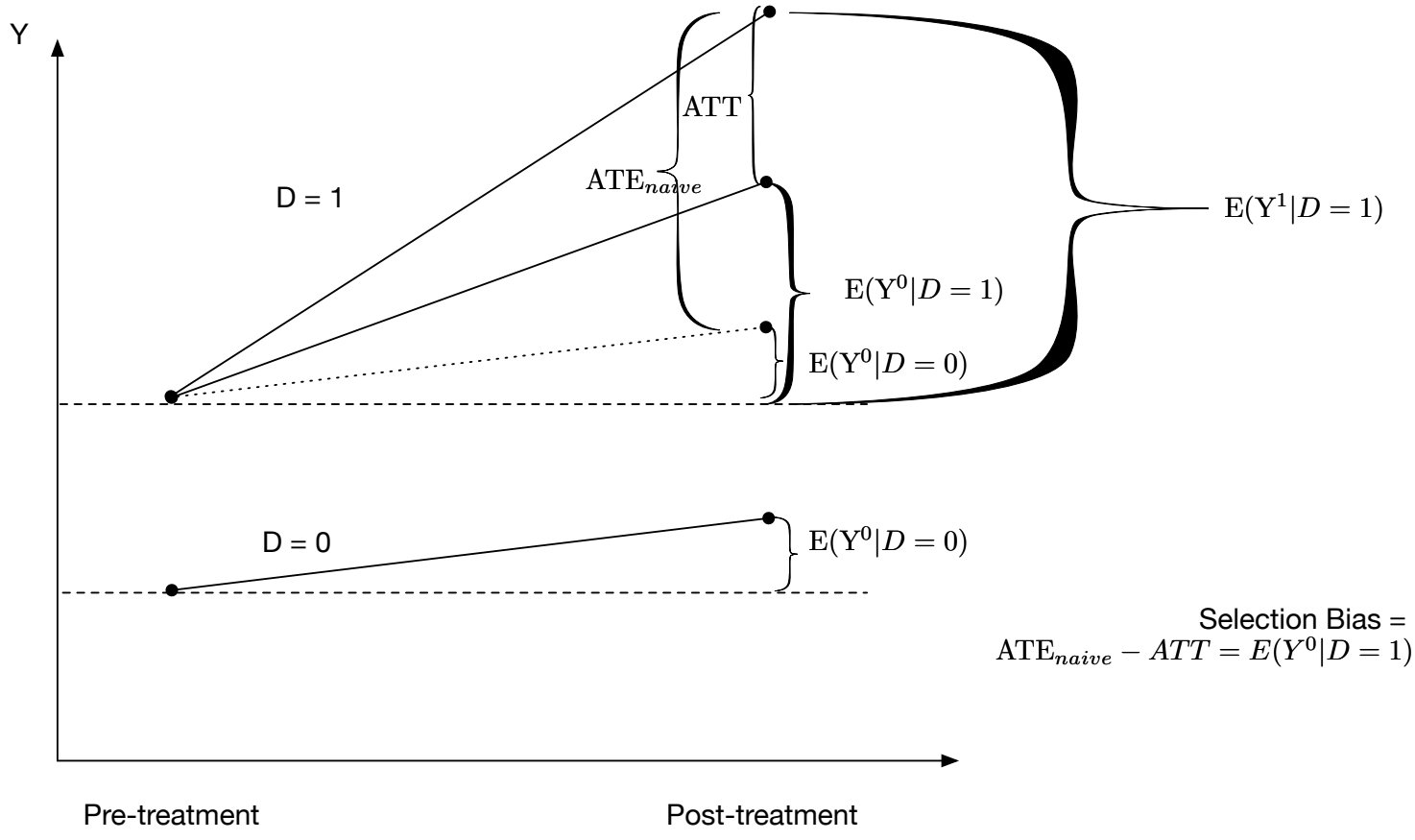
Naive estimate of ATE

- Naive estimate of ATE is just the difference in means of observed data

$$ATE_{naive} = E\left[\mathrm{Y}|\mathrm{D}=1\right] - E\left[\mathrm{Y}|\mathrm{D}=0\right] \tag{13}$$

- For instance, $D = 1$ for college education and $D = 0$ for less than college education
- ATE_{naive} is the mean earning of college education - mean earning of non-college educated
- In general, $ATE \neq ATT \neq ATE_{naive}$; what is their connection?

Selection Bias



Selection Bias =
$\text{ATE}_{naive} - ATT = E(Y^0|D=1) - E(Y^0|D=0)$

Selection bias

- ATE_{naive} neither estimates ATE nor ATT
- ATE_{naive} differs from ATT by selection bias

$$\text{selection bias} = ATE_{naive} - ATT = E(Y^0|D=1) - E(Y^0|D=0)$$

- Intuitively, this is the counterfactual earning of college educated if they did not go to college, minus the factual earning of non-college educated
  - This selection bias could be caused by ability, for example
- If selection bias is 0, $ATE_{naive} = ATT$

Random Assignment

- In randomized controlled experiments, we randomly assign subjects into treatment and control groups; we have random assignment

Definition ((Completely) Random Assignment)

- $Y_i^0, Y_i^1 \perp\!\!\!\perp D_i$ (Potential outcome is independent of treatment assignment)
- $P(D = 1) > 0$ (non-zero treatment probability)

- Using the education and income example, imagine the God randomly assign people to go to college or not, so we do not know beforehand what potential outcome i gets

Random Assignment Solves the Identification Problem

- Under random assignment of D, we have:

$$ATE_{naive} = E\left[\mathrm{Y}|\mathrm{D}=1\right] - E\left[\mathrm{Y}|\mathrm{D}=0\right] = ATE \tag{14}$$

- Proof (the first line to second line is due to independence between D and Y^0, Y^1)

$$\begin{aligned} E\left[\mathrm{Y}|\mathrm{D}=1\right] - E\left[\mathrm{Y}|\mathrm{D}=0\right] &= E\left[\mathrm{Y}^1|\mathrm{D}=1\right] - E\left[\mathrm{Y}^0|\mathrm{D}=0\right] \\ &= E\left[\mathrm{Y}^1|\mathrm{D}=1\right] - E\left[\mathrm{Y}^0|\mathrm{D}=1\right] \\ &= E\left[\mathrm{Y}^1 - \mathrm{Y}^0|\mathrm{D}=1\right] \\ &= E\left[\mathrm{Y}^1 - \mathrm{Y}^0\right] \\ &= E\left[\mathrm{Y}^1\right] - E\left[\mathrm{Y}^0\right] \end{aligned} \tag{15}$$

Non-parametric estimator: difference-in-means

- With random assignment, estimating ATE is very simple: ATE_{naive}, which is just the difference in mean outcome of the treatment and the control group
- This is a non-parametric estimator
- Another important observation: $ATT = ATE$ for randomized experiments

Experiment as Imputation

Unit	Y_i^0	Y_i^1	D_i
1	?	1	1
2	1	?	0
3	?	1	1
4	?	0	1
5	?	1	1
6	0	?	0

(16)

- Random assignment implies that we can impute the missing values using observed sample mean; similar to the MCAR assumption in missing data
  - But here, random assignment is a fact, not an assumption

Unit	Y^0	Y^1	D
1	$\hat{\mathrm{E}}[Y\|D=0]=\frac{1}{2}$	1	1
2	1	$\hat{\mathrm{E}}[Y\|D=1]=\frac{3}{4}$	0
3	$\hat{\mathrm{E}}[Y\|D=0]=\frac{1}{2}$	1	1
4	$\hat{\mathrm{E}}[Y\|D=0]=\frac{1}{2}$	0	1
5	$\hat{\mathrm{E}}[Y\|D=0]=\frac{1}{2}$	1	1
6	0	$\hat{\mathrm{E}}[Y\|D=1]=\frac{3}{4}$	0

(17)

Regression estimator of ATE

- We can rewrite Y_i in the following way (MHE, 2.3.1)

$$\begin{aligned} \mathrm{Y}_i &= E\left(\mathrm{Y}_i^0\right) &&+ \left(\mathrm{Y}_i^1 - \mathrm{Y}_i^0\right) D_i &&+ \quad \mathrm{Y}_i^0 - E\left(\mathrm{Y}_i^0\right) \\ &= \alpha &&+ \rho_i \mathrm{D}_i + &&\eta_i \end{aligned} \tag{18}$$

- This equation looks like linear regression! But each individual has its own regression coefficient ρ_i, which is the individual-level treatment effect

- Constant treatment assumption: assume that ρ_i is the same for every one, ρ, $ATE = E(\rho_i) = \rho$

$$\begin{aligned} E\left[\mathrm{Y}_i|\mathrm{D}_i = 1\right] &= \alpha + \rho + E\left[\eta_i|\mathrm{D}_i = 1\right] \\ E\left[\mathrm{Y}_i|\mathrm{D}_i = 0\right] &= \alpha + E\left[\eta_i|\mathrm{D}_i = 0\right] \end{aligned} \tag{19}$$

$$ATE_{naive} = E\left[\mathrm{Y}_i|\mathrm{D}_i = 1\right] - E\left[\mathrm{Y}_i|\mathrm{D}_i = 0\right]$$

$$= \underbrace{\rho}_{\text{ATE}} + \underbrace{E\left[\eta_i|\mathrm{D}_i = 1\right] - E\left[\eta_i|\mathrm{D}_i = 0\right]}_{\text{selection bias}} \tag{20}$$

Regression estimator of ATE

- Selection bias is 0, since $Y^0 \perp\!\!\!\perp D$ under random assignment
- Therefore,

$$ATE_{naive} = E\left[\mathrm{Y}_i|\mathrm{D}_i = 1\right] - E\left[\mathrm{Y}_i|\mathrm{D}_i = 0\right] = \underbrace{\rho}_{\text{ATE}}$$

- Therefore, if you are running a random experiment,
  - Non-parametric estimator: the difference in mean outcome of treated and control units
  - Parametric estimator:
    - assume constant treatment effect
    - run a regression of observed outcome on treatment D, and use coefficient of D as the estimate of ATE

Regression as Imputation

- The regression estimator of ATE is implicitly making counterfactual imputation using linear regression:

Unit	Y_i^0	Y_i^1	D_i	$X_{[1]i}$	$X_{[2]i}$
1	?	2	1	1	7
2	5	?	0	8	2
3	?	3	1	9	3
4	?	10	1	3	1
5	?	2	1	5	2
6	0	?	0	7	0

Regression as Imputation

- Fit a regression $Y = \beta_0 + \beta_1 D_i + \beta_2 X_{[1]i} + \beta_3 X_{[2]i}$, and impute counterfactual outcome using the linear regression:

Unit	Y_i^0	Y_i^1	D_i	$X_{[1]i}$	$X_{[2]i}$
1	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 1 + \hat{\beta}_3 \cdot 7$	2	1	1	7
2	5	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 8 + \hat{\beta}_3 \cdot 2$	0	8	2
3	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 9 + \hat{\beta}_3 \cdot 3$	3	1	9	3
4	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 3 + \hat{\beta}_3 \cdot 1$	10	1	3	1
5	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 5 + \hat{\beta}_3 \cdot 2$	2	1	5	2
6	0	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 7 + \hat{\beta}_3 \cdot 0$	0	7	0

(21)

- Then ATE and ATT can be easily calculated as the difference in means of Y^1 and Y^0

Inference: Neyman Variance Estimator

- If we are running experiments, ATE can be estimated easily by taking the differences between $E(Y|D=1)$ and $E(Y|D=0)$
- What about statistical inference?
- Neyman Variance Estimator
  1. assume constant treatment effect
  2. Then
$$V(ATE) = \frac{V_t}{N_t} + \frac{V_c}{N_c}$$
  3. V_t is variance of Y for treated users, and N_t is number of treated users
- If treatment effect is not constant, true variance is usually smaller than $\frac{V_t}{N_t} + \frac{V_c}{N_c}$

Inference: Linear regression

- If we regress Y on D ($Y = \alpha + \rho D$), it can be shown that the regression estimates of the standard error of regression coefficient is exactly same as the Neyman Variance estimator

$$V(\rho) = V(ATE) = \frac{V_t}{N_t} + \frac{V_c}{N_c}$$

- See Imbens and Rubin (chapter 7) for proof

Inference: Randomization Test

- Null distribution: D has no causal effect on Y
  - then if we shuffle the outcome, $E(Y|D=1) - E(Y|D=0) = 0$
- Randomization test
  - Calculate ATE based on experimental data
  - Shuffle your observed Y, and recalculate $ATE_{shuffle}$ based on the shuffled data
  - Say you shuffled 1000 times, and have 1000 $ATE_{shuffle}$.
  - Then you can easily calculate 95% confidence interval/standard errors of ATE estimates
  - The p value for observing ATE is just the probability that your shuffled $ATE_{shuffle}$ is larger than ATE: p -value $= P(ATE_{shuffle} > ATE)$
- Pros: do not need to assume constant treatment effect
- Cons: time consuming

Additional Covariates

- Researchers often collect some additional covariates (i.e., pre-treatment variables)
- With additional variables, it is easier to work with regression estimator

$$Y_i = \alpha + \rho D_i + \beta X_i + \epsilon_i \tag{22}$$

- $\hat{\rho}^{adj}$:covariate-adjusted estimate of treatment effect
- $\hat{\rho}$: difference-in-means of outcome variables across treatment and control (or regression coefficient by regressing Y on D without covariates)
- $\hat{\rho}_X$: difference-in-means of X across treatment and control

Additional Covariates

- It can be shown that (Li and Ding, 2019, J. R. Stat. Soc, or Imbens and Rubins, Chapter 7):

$$\hat{\rho}^{adj} = \hat{\rho} - \hat{\beta}^T \hat{\rho}_X$$

- $\hat{\rho}_X$: difference-in-means of X across treatment and control
  - With completely randomized experiments
    - $\hat{\rho}^{adj}$ is biased; $\hat{\rho}$ is unbiased
    - $\hat{\rho}_X$ are usually not exactly 0, especially when the data size is not that large
  - Both are consistent
    - because with more and more data, $\hat{\rho}_X$ approaches 0; this is called covariate balance
- Be careful if your treatment and control groups are not balanced; in that case, the treatment effect estimates without and with covariates can differ a lot in finite sample

Additional Covariates

- Another classical justification to add covariates in regression is that $\hat{\rho}^{adj}$ has smaller standard error than $\hat{\rho}$
- For instance, MHE (p. 23): "Inclusion of the variable X... generate more precise estimates of the causal effect)"
- David A. Freedman, *On regression adjustments to experimental data*, Advances in Applied Mathematics **40** (2008), no. 2, 180–193
- It is not necessarily true!

Additional Covariates

Winston Lin, *Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique*, The Annals of Applied Statistics **7** (2013), no. 1, 295–318. MR3086420

- adding covariate is guaranteed to lead to smaller standard error estimates, if

1. full interaction is added; and
2. robust standard errors are used

$$Y_i = \alpha + \rho D_i + \beta X_i + \gamma D_i X_i + \epsilon_i$$

- Note that condition 1 is not easy to follow in practice; if you have 10 covariates, you have to add 10 interaction terms

Recommended practice

- David A. Freedman, *On regression adjustments to experimental data*, Advances in Applied Mathematics **40** (2008), no. 2, 180–193
- Always present two treatment effects: without and with covariates
- "Regression estimates. . . should be deferred until rates and averages have been presented"
- Always check pre-treatment covariate balance
- Add interactions if covariance-balance is passed
- not only guarantees smaller standard error, but also detects treatment effect heterogeneity (next week)