# Counterfactual Framework of Causal Inference

Han Zhang

# 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 $\beta$; 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

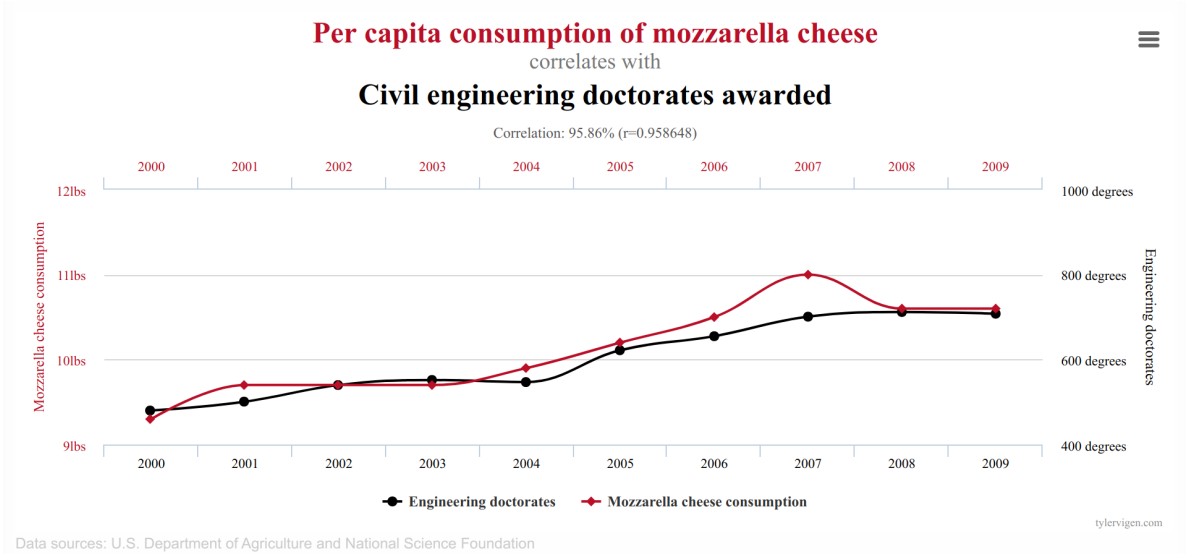
- Rubin, 1977, Multiple Imputation
  - Stochastic imputation for $m$ times; ending up with $m$ imputed datasets.
  - Analysis: Run your model (regression $Y$ on 10 $X$) $m$ times
  - Pooling: parameter estimates for $m$ different models can be used for estimation and inference:
    - The final parameter estimates of $\beta$ is the mean of $\beta$ across $m$ models
    - The standard error of final $\beta$ is more complex in math
    - basically it's the (within model standard error of $\beta$) + (between model standard error of $\beta$)
    - 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 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
  - E.g., regardless of whether I go to college or not, my potential outcomes are fixed
    - We just don't know what values they are, but they are fixed (stable)

# Violation of SUTVA

- A common violation of SUTVA is in the presence of network effect
  - 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
- It is a cutting edge research area in causal inference
- In everything we are going to learn in this class, we assume no interference; SUTVA holds

# Individual Level Treatment Effect

## Definition (Unit-Level Treatment Effect)

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

- After defining individual level treatment effect, we can rewrite observed outcome $Y_i$ as:
- $Y_i = Y_i^0 + D_i(Y_i^1 - Y_i^0)$
- This way of writing suggests observed $Y$ for a treated unit $i$ are the two sums:
  - potential outcome if $i$ were not treated
    - as a baseline: your wages if you did not go to college
  - 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 $\rho_i$ is unknown.

# Fundamental problem of causal inference

- Measure each unit-level treatment effect is extremely hard
  - But not impossible; this is the cutting edge research area
  - Intuition: use machine learning to predict each observation's potential outcome
- The easiest solution:
- Calculating average (expectation)

# 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; it is what we can easily calculate based on your 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

# Selection bias

- In general, $ATE \neq ATT \neq ATE_{naive}$
- $ATE_{naive}$ differs from $ATT$ by selection bias

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

- It's the (counterfactual outcome of the treated users) - (factual outcome) of control users
- If selection bias is 0, $ATE_{naive} = ATT$
- Imagine if there were no selection bias, then (the potential outcome of college educated if they did not go to college) should be equal to (earning of non-college educated)
- If there is positive selection bias, then those college educated, even if they did not go to college, would earn more than the non-college educated.

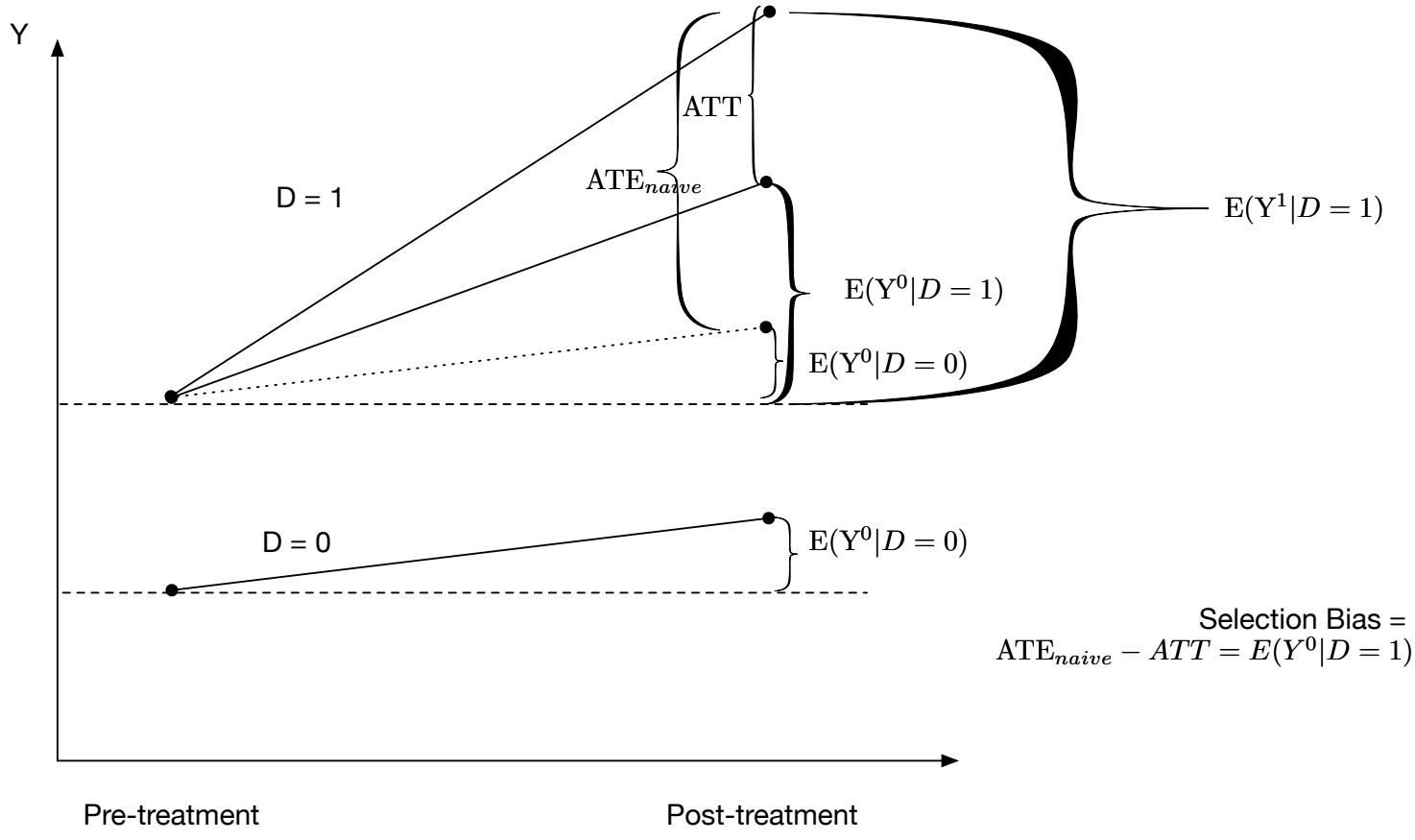
# Selection bias

- Alternatively speaking,

$$ATT = ATE_{naive} - \text{selection bias}$$

- So if you have positive selection bias, the real causal effect $ATT$ is smaller than naive estimates
- If there is positive selection bias, then those college educated, even if they did not go to college, would earn more than the non-college educated.
    - So the causal effect of education is really not as high as the naive calculation based on observed data shows

# Selection Bias




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

# 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 $\rho_i$, which is the individual-level treatment effect

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

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

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

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

# 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)

# Balance in covariates between treatment and control groups

- $\hat{\rho}_X$: difference-in-means of $X$ across treatment and control
  - If the assignment is truly random, then $\hat{\rho}_X$ should approach 0
  - This is often referred as balance between treatment and control groups
  - Or covariate balance
- If you are running experiments, you should always check this to make sure that the randomization is performed properly
- If $\hat{\rho}_X$ is significantly differnet from 0, then your randomization is probably wrong at some point

# Regression estimates of treatment effect with and without 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 with small-sized sample
  - Both are consistent
    - because when sample size goes to infinity, $\hat{\rho}_X$ approaches 0

-This formula shows that if your treatment and control groups are not balanced, the treatment effect estimates without and with covariates can differ considerably

# Additional Covariates

- Another classical justification to add covariates in regression is to reduce standard error estimates of regression coefficients
- 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 you have enough number of observations
- not only guarantees smaller standard error, but also detects treatment effect heterogeneity (next week)

# Stratified Randomized Experiment

- Sometimes the treatment assignment are not completely random
- Whether small class size improves students' test scores?
- Tennessee Project STAR experiment: whether class sizes impact test scores
- Within each school, we random select some classes to be in the treatment group (small class size), and other classes in regular class group (control group)
- This is known as cluster randomized, or stratified randomized experiments

# Estimate ATE under Stratified Randomized Experiment

- Neyman estimator:
- Estimate $ATE$ for each group
- $ATE$ is the weighted sum of group-specific $ATE$, with weights proportional to group size

$$\hat{ATE}_{neyman} = \sum_{j=1}^{J} \frac{\omega(j)}{\left(\sum_{j=1}^{J} \omega(j)\right)} \cdot \hat{ATE}_j \tag{23}$$

- Weights for each group is: $\omega(j) = \frac{N_j}{N}$
- $V(ATE)$; weight for each group is $\omega(j) = \frac{N_j}{N}^2$

# Star Example: Neyman estimator

| School | # Classes | Estimated Effect | $\widehat{\text{(s.e.)}}$ |
|---|---|---|---|
| 1 | 4 | 0.223 | (0.230) |
| 2 | 4 | −0.295 | (0.776) |
| 3 | 5 | 0.417 | (0.404) |
| 4 | 4 | 0.748 | (0.215) |
| 5 | 4 | −0.077 | (0.206) |
| 6 | 4 | 1.655 | (0.405) |
| 7 | 4 | −0.254 | (0.255) |
| 8 | 6 | 0.429 | (0.306) |
| 9 | 4 | −0.006 | (0.311) |
| 10 | 4 | −0.014 | (0.182) |
| 11 | 4 | −0.003 | (0.605) |
| 12 | 5 | 0.222 | (0.309) |
| 13 | 4 | 0.432 | (0.179) |
| 14 | 4 | 0.340 | (0.336) |
| 15 | 4 | 0.207 | (0.396) |
| 16 | 4 | −0.306 | (0.245) |
| overall | | 0.241 | (0.092) |

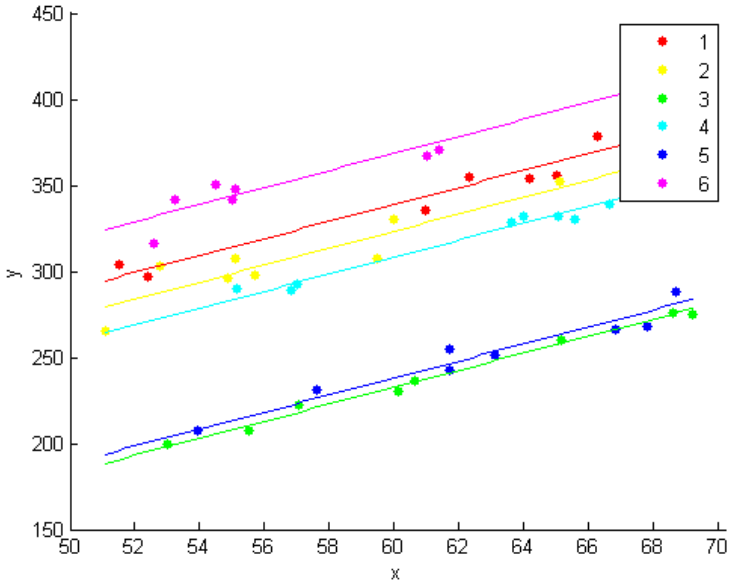
(24)

- Each school has its own $ATE_j$
- $ATE$ is the weighted sum of school-specific $ATE_j$, with weights proportional to relative group size $(N_j/N)$
- Variance of ATE is weighted sum of school-specific variance with weights proportional to $(N_j/N)^2$

# Regression Estimator of ATE

- Regression estimator for overall ATE (still assume constant treatment effect)
- Regression $Y$ on $D$, with group fixed effects, that is, a separate intercept for each group $j$ (more on fixed effects later)
- $Y_{ij} = \alpha_j + \rho D_{ij} + \epsilon_{ij}$
- The regression coefficient for $D$ is the estimate of ATE

# Fixed effect




- Each group has its own intercepts
- But slope is the same
- More on this in next two weeks

# Use fixed effect regressions to estimate ATE

$$Y_{ij} = \alpha_j + \rho D_{ij} + \epsilon_{ij} \tag{25}$$

- $i$ is unit and $j$ is group
- $\alpha_J$ are group-level fixed effects
- Still, constant treatment effect assumption; $\rho$ is the overall ATE
- Estimated ATE is $\hat{\rho} = 0.238$, and its S.E. (0.103)
- Compare this with ATE estimated using Neyman estimator

# Neyman vs Regression

- In general, ATE estimates using Neyman mean estimator and regression estimator are different
- In fact, we can express $ATE_{OLS}$ as the following weighted mean:

$$\hat{ATE}_{OLS} = \hat{\rho} = \sum_{j=1}^{J} \frac{\omega(j)}{\left(\sum_{j=1}^{J} \omega(j)\right)} \cdot \hat{ATE}_j \tag{26}$$

- Weights for each group is proportional to:

$$\omega(j) = \frac{N_j}{N} \cdot P(D = 1|X = j) \cdot (1 - P(D = 1|X = j))$$

- $N_j/N$ is relative size
- $P(D = 1|X = j)$ is the proportion of treated units in group $j$, or put it differently, group-specific treatment propensity

# Neyman vs Regression

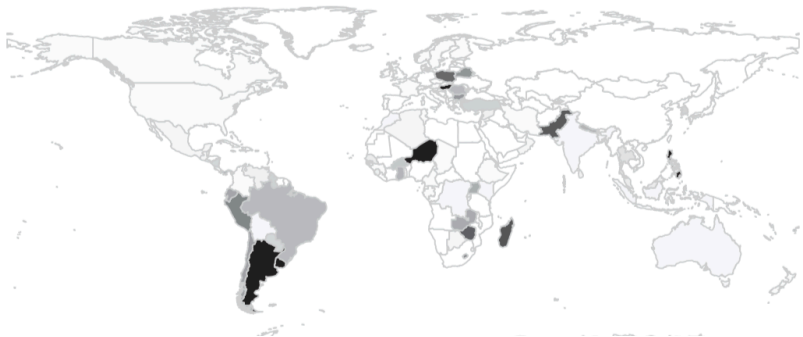
- $\hat{ATE}_{OLS}$ gives additional weights
  $P(D = 1|X = j) \cdot (1 - P(D = 1|X = j))$
- Regression estimator puts most emphasis on group whose $P(D = 1|X = j) = 1/2$, that is, group with the same number of treated and control units (Read MHE 3.3 for details)
- Note that we have not add covariates yet; adding covariates makes the weight to be
  $P(D = 1|X = j, othercovariates) \cdot (1 - P(D = 1|X = j, othercovariates))$
- In general: $\hat{ATE}_{OLS}$ is neither consistent nor unbiased
- $\hat{ATE}_{neyman}$ is consistent and unbiased
- But harder to add covariates

# Example

Peter M. Aronow and Cyrus Samii, *Does Regression Produce Representative Estimates of Causal Effects?*, American Journal of Political Science **60** (2016), no. 1, 250–267

- Question: effect of regime types on foreign direct investment
- Data: time-series cross-sectional analysis of 114 countries from 1970 to 1997
- Jensen (2003)'s conclusion: "democratic regimes are associated with higher level of foreign direct investment"
- Group here: country
- So intuitively, if a country experiences changes in their democracy level, it will contribute more to the effect estimates

# Example



- "findings are driven primarily by the experiences of Latin American, Eastern European and African cases"

# Neyman vs Regression

| | regression | Neyman |
|---|---|---|
| consistency | inferior | superior |
| standard error | smaller (with covariate balance) | larger |
| practice (with ctrl) | easier | harder |

# Categorical Treatment

- When there are categorical treatment $D$ with more than 2 levels, simple
- Each group has its own treatment effect (defined with respect to the control group)
- You can still use Neyman estimator or linear regression estimator (treatments as dummies, and control group as the reference group)

# Continuous Treatment

- When there are ordinal or continuous treatment $D$:
- E.g., effect of years of schooling on future income
- One way to extend the Neyman-Rubin Causal Model

$$\begin{aligned} Y_i &= \begin{cases} Y_i^0 : D_i = 0 \\ Y_i^d : D_i = d \end{cases} \\ &= Y_i^0 + d(Y_i^d - Y_i^0) \end{aligned} \tag{27}$$

- This extension assumes that causal effects is linear in $D$ at the unit level
- Assume we manipulate the treatment from $d$ to $d'$.

$$ATE = E\left(\frac{Y_i^{d'} - Y_i^d}{d' - d}\right) \tag{28}$$

- If $d' - d' = 1$ (changes for one unit), this becomes to binary

# Categorical and Continuous Treatment with Regression

- You can use regression to estimate ATE in continuous treatments
- Assume $D$ is a continuous treatment, then
- $Y_i = \alpha + \rho D_i + \epsilon_i$ (for random experiments)
- $Y_{ij} = \alpha_j + \rho D_{ij} + \epsilon_{ij}$ (for stratified random experiments)
- $\rho$ will identify ATE, assuming causal effect is linear in treatment

# Example

Alan S. Gerber, Donald P. Green, and Christopher W. Larimer, *Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment*, American Political Science Review **102** (2008), no. 01, 33–48

- This paper wants to estimate the causal effect of social pressure on voter turnout (voters go to vote)
- A real stratified randomized experiment with categorical treatment
- 10,000 geographic cells in Michigan; sample 18 households within each cell
- 4 different treatment conditions; 1 control group
- 1/9 of households enter into each treatment group; the rest 5/9 enter into control group
- In total, 180,000 households (observations)

# Covariate balance

**TABLE 1. Relationship between Treatment Group Assignment and Covariates (Household-Level Data)**

| | Control | Civic Duty | Hawthorne | Self | Neighbors |
|---|---|---|---|---|---|
| | Mean | Mean | Mean | Mean | Mean |
| Household size | 1.91 | 1.91 | 1.91 | 1.91 | 1.91 |
| Nov 2002 | .83 | .84 | .84 | .84 | .84 |
| Nov 2000 | .87 | .87 | .87 | .86 | .87 |
| Aug 2004 | .42 | .42 | .42 | .42 | .42 |
| Aug 2002 | .41 | .41 | .41 | .41 | .41 |
| Aug 2000 | .26 | .27 | .26 | .26 | .26 |
| Female | .50 | .50 | .50 | .50 | .50 |
| Age (in years) | 51.98 | 51.85 | 51.87 | 51.91 | 52.01 |
| *N* = | 99,999 | 20,001 | 20,002 | 20,000 | 20,000 |

*Note*: Only registered voters who voted in November 2004 were selected for our sample. Although not included in the table, there were no significant differences between treatment group assignment and covariates measuring race and ethnicity.

# Neyman estimator

- No explicit calculation of causal effect, but you can easily calculate that yourself

**TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election**

| | Experimental Group | | | | |
|---|---|---|---|---|---|
| | Control | Civic Duty | Hawthorne | Self | Neighbors |
| Percentage Voting | 29.7% | 31.5% | 32.2% | 34.5% | 37.8% |
| N of Individuals | 191,243 | 38,218 | 38,204 | 38,218 | 38,201 |

# Regression estimator

**TABLE 3. OLS Regression Estimates of the Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election**

| | Model Specifications | | |
|---|---|---|---|
| | (a) | (b) | (c) |
| Civic Duty Treatment (Robust cluster standard errors) | .018* (.003) | .018* (.003) | .018* (.003) |
| Hawthorne Treatment (Robust cluster standard errors) | .026* (.003) | .026* (.003) | .025* (.003) |
| Self-Treatment (Robust cluster standard errors) | .049* (.003) | .049* (.003) | .048* (.003) |
| Neighbors Treatment (Robust cluster standard errors) | .081* (.003) | .082* (.003) | .081* (.003) |
| N of individuals | 344,084 | 344,084 | 344,084 |
| Covariates** | No | No | Yes |
| Block-level fixed effects | No | Yes | Yes |

*Note*: Blocks refer to clusters of neighboring voters within which random assignment occurred. Robust cluster standard errors account for the clustering of individuals within household, which was the unit of random assignment.

* $p < .001$.

** Covariates are dummy variables for voting in general elections in November 2002 and 2000, primary elections in August 2004, 2002, and 2000.

# Heterogeneous treatment effect by subgroups

- Broadly speaking, heterogeneous treatment effect (HTE) just means that treatment effect varies
- For cluster-randomized experiment, Neyman estimator naturally gives heterogeneous treatment effect for each subgroup (e.g., each school)

# Heterogeneous treatment effect by subgroups

- Regression estimator:
- Fixed effect regression gives an overall ATE, but does not estimate heterogeneous treatment effect for each group
- To estimate group-specific ATE, we fit linear regression with interactions between group dummy and treatment $D$ (and no fixed effects)
- (Interaction coefficients + the coefficient on treatment ) captures the treatment effect for each subgroup
    - But this interaction model does not give us overall ATE

# Heterogeneous treatment effect by propensity score

- The second type of treatment effect heterogeneity is effect difference by treatment propensity score: $p(D = 1|X = x)$
- That is, we are comparing treatment effects among units that are more likely to be treated, and units that are less likely to be treated.

  Jennie E. Brand and Yu Xie, *Who Benefits Most from College?: Evidence for Negative Selection in Heterogeneous Economic Returns to Higher Education*, American Sociological Review **75** (2010), no. 2, 273–302

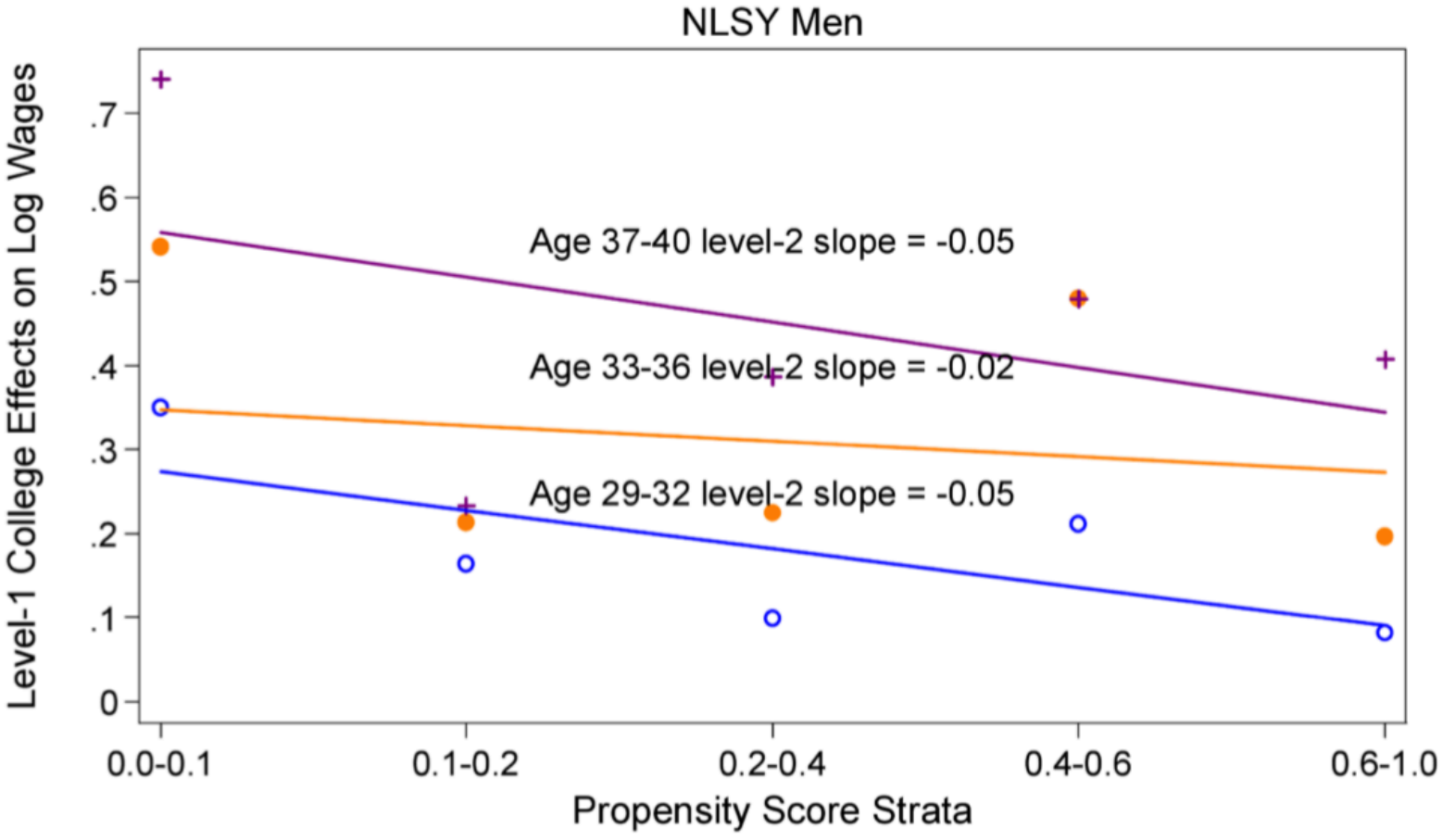
# Heterogeneous treatment effect by propensity score

- Treatment: college education
- Outcome: future earnings
- HTE: corresponding to two different hypothesis
- individuals who are most likely to select into college benefit the most from college
  - Treatment effect is high among units whose treatment propensity $P(D = 1|X = x)$ is high
- individuals who are least likely to obtain a college education benefit the most from college
  - Treatment effect is high among units whose treatment propensity $P(D = 1|X = x)$ is low

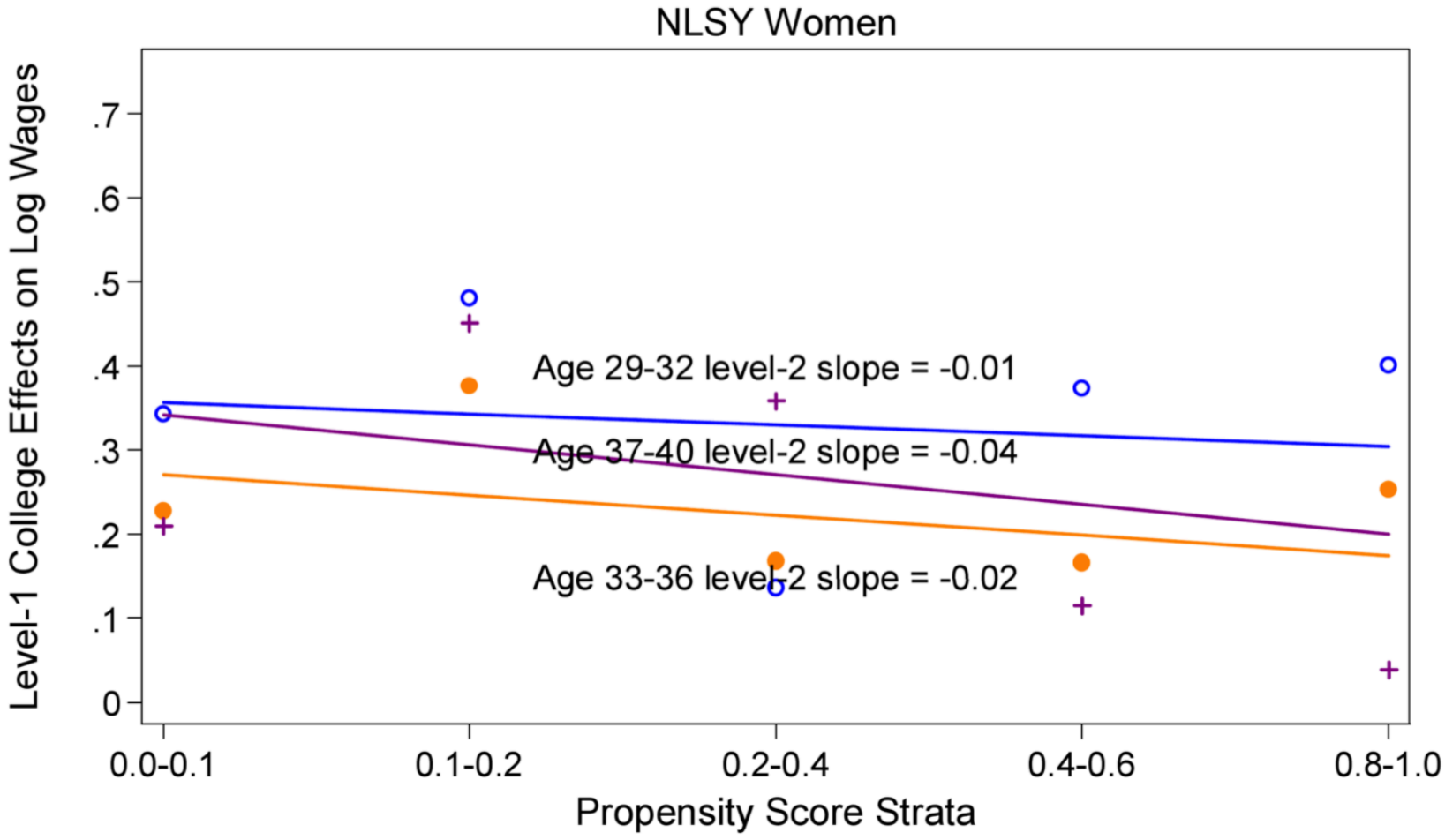
# Heterogeneous treatment effect by propensity score

- Divide respondents into groups based on their propensity score
- Use regression to estimate treatment effect for each group
- Plot the effects by propensity score
- Caution: this article is not a randomized experiment; it uses a survey dataset
- rigorously what it estimates is not ATE, unless we add strong assumptions
- More on the assumptions next week
- I am using it to illustrate the idea of HTE

# Heterogeneous treatment effect by propensity score



# Heterogeneous treatment effect by propensity score



# Heterogeneous treatment effect by covariates

- The third, and probably the most common type of HTE you will see in research articles: treatment heterogeneity by covariates
- Formal notation: $\tau(x) = E(Y^1 - Y^0|X = x)$
- $\tau(x)$ is usually referred as conditional average treatment effect (CATE)
- In Project STAR example:
- We think the important variation of treatment heterogeneity come from whether teacher is more experienced or not

# Using interaction model to capture CATE

- To capture heterogeneity by covariate (CATE), the easiest approach is to add interaction between treatment and covariate:

$$Y_{ij} = \alpha_j + D_{ij}\rho + \text{experienced}_i\beta + D_{ij}\text{experienced}_i\gamma + \epsilon_{ij}$$

- $\gamma$ and $\rho$ together captures the treatment effect heterogeneity
- $\rho$: *CATE* for classes with inexperienced teacher
- $\gamma + \rho$: *CATE* for treated classes with experienced teacher

# Using interaction model to capture CATE: shortcomings

- Using interaction model to capture CATE is the dominant approach in applied literature now
- It also has shortcomings
- The interaction model assumes constant effect within covariate levels
- It slightly relax the overall constant effect assumption, but still can be unrealistic
- e.g., within class taught by experienced teacher, gender ratio still matters.
- [Question]: what regression model you should use if you believe that teacher experience and gender ratio all matters?
- A common situation is that treatment effect vary by certain covariate, but you forget to add that into the regression

# Using interaction model to capture CATE

- If the covariate is beyond binary and has more levels, or it is continuous variable, we are implicitly adding another assumption: linear interaction effect
- That is, CATE grows linearly in the covariate
- Is this realistic? certainly not.
- If time permits, at the end of semester, we can talk some advances of estimating heterogeneous treatment effects using machine learning
- The basic idea is to predict each individual's counterfactual outcomes
- So the need to split population into subgroups by propensity score or by covariates is decreased