ORIE 5355: People, Data, & Systems

Lecture 3: Survey weighting methods

Nikhil Garg

Course webpage: https://orie5355.github.io/Fall_2021/

Reminder: the task

- Each person j has an opinion, Y_j
- We want to measure $\bar{y} = E\left[Y_j\right]$, the population mean opinion on some issue
- Each person also has covariates, x_j^k (e.g., where they live)
- Sometimes, we also care about *conditional* means

$$E[Y_j \mid \text{lives in Roosevelt Island}]$$

Challenge 1: people don't give "true" opinion

People gave you $\widetilde{Y_j}$, instead of Y_j

$$\hat{y} = \frac{1}{N} \sum_j \widetilde{Y_j}$$

$\hat{y}$ does not converge to $\bar{y}$, *unless errors cancel out*

Challenge 2: Sample doesn't represent pop

- For each person j, let $A_j \in \{0,1\}$ be whether they answered
- You have $\boldsymbol{Y} = \{(A_j, Y_j)\}_{j=1}^{N}$, if called N people

  Where $Y_j = \emptyset$ if $A_j = 0$ (they did not answer)
- Again, you do

$$\hat{y} = \frac{1}{|\{j \mid A_j=1\}|} \sum_{j \in \{j \mid A_j=1\}} Y_j$$

where $\{j \mid A_j = 1\}$ denotes the set of people who answered

and so $|\{j \mid A_j = 1\}|$ is the number of people who answered

$\hat{y}$ does not converge to $\bar{y}$ *unless* Y_j and A_j are uncorrelated

Questions from last time?

Plan for today

Methods for tackle sample representation issues

- Stratifying sample *before* you poll
- Weighting techniques *after* you have responses

Differential response on *known* covariates

- Suppose we have a single binary covariate $x_j \in \{0,1\}$ indicating whether they graduated to college
  - Half the population went to college
- Suppose whether people answer is correlated with education

$$\Pr(A_j = 1) = \begin{cases} 0.1 \text{ if } x_j = 0 \\ 0.4 \text{ if } x_j = 1 \end{cases}$$

- Education also correlated with opinion Y_j in some unknown manner
- We want to measure $\bar{y} = E[Y_j]$, the population mean
- *No other correlations between whether they answer and opinion:*
  - Opinion Y_j *is independent of whether they respond* A_j*, conditional on* x_j

New notation

- Number of people *called*: N
- Population response rate for group ℓ: A^{ℓ}
- Population mean response for group ℓ: $\bar{y}^{\ell}$
- Population fraction for group ℓ: P^{ℓ}
- Corresponding ppl *called* values are: $\hat{A}^{\ell}$, $\hat{y}^{\ell}$, $\hat{P}^{\ell}$

(i.e., $\mathrm{N}\hat{P}^{\ell}\hat{A}^{\ell} = \left| \{j \mid A_j = 1, x_j \text{ in Group } \ell\} \right|$)

and so:

$$\bar{y} = \frac{P^0\bar{y}^0 + P^1\bar{y}^1}{P^0 + P^1} = P^0\bar{y}^0 + P^1\bar{y}^1$$ = $0.5\,\bar{y}^0 + 0.5\,\bar{y}^1$ in example

$$\hat{y}_{naive} = \frac{\hat{A}^0\hat{P}^0\hat{y}^0 + \hat{A}^1\hat{P}^1\hat{y}^1}{\hat{A}^0\hat{P}^0 + \hat{A}^1\hat{P}^1} \to \frac{A^0P^0\bar{y}^0 + A^1P^1\bar{y}^1}{A^0P^0 + A^1P^1}$$ = $0.2\,\bar{y}^0 + 0.8\,\bar{y}^1$

Naïve method in more detail

$$\hat{y}_{naive} = \frac{\left(\sum_{j\in\{j \mid A_j=1, x=0\}} Y_j + \sum_{j\in\{j \mid A_j=1, x=1\}} Y_j\right)}{|\{j \mid A_j = 1,\ x = 0\}| + |\{j \mid A_j = 1,\ x = 1\}|}$$

$$= \frac{\hat{A}^0\hat{P}^0\hat{y}^0 + \hat{A}^1\hat{P}^1\hat{y}^1}{\hat{A}^0\hat{P}^0 + \hat{A}^1\hat{P}^1} = \frac{\left(\#(Y_j{=}1)\text{ from Group 0 } + \#(Y_j{=}1)\text{from Group 1}\right)}{\text{Total Respondants}}$$

$$\rightarrow \frac{P^0A^0\bar{y}^0 + P^1A^1\bar{y}^1}{P^0\mathrm{A}^0 + P^1A^1} \neq \bar{y} \text{ unless } \mathrm{A}^0 = A^1$$

$P^0A^0/\ (P^0\mathrm{A}^0 + P^1A^1)$ is limit fraction of respondents from Group 0

Bias (even with $N \rightarrow \infty$): Limit fraction does not match the population fraction

Variance (with finite N): Sample values do not match limit values

Stratified sampling

Main idea for stratification

- Suppose you have L mutually exclusive demographic groups:
  - A population that is heterogeneous *across* groups
  - Relatively homogenous *within* groups
  - (Exactly the setup we have)

Y_j is independent of A_j, conditional on x_j

- Then, instead of calling N completely random people
  - Call N^{ℓ} people from group ℓ
  - Where N^{ℓ} is determined by how likely each group is to respond
- Even if each group responds at same frequency, this leads to *lower variance* estimates
- With differential response rates, can also correct the *bias* in *mean*

Why does it work?

- Even without differential response rates, just differential opinion:

  There are two sources of variance in estimation:

    Which groups are over- and under- sampled due to noise

    What the opinion of each person is

  Stratification mitigates the first source of variance

- With differential response rate: we can “cancel out” the differential response rate by just calling more people from that group

Why does it work? (Mathematically)

Now
$N^{\ell}\hat{A}^{\ell}$ instead of
$\mathrm{N}\hat{P}^{\ell}\hat{A}^{\ell}$

$$\hat{y} = \frac{\left(\sum_{j\in\{j \mid A_j=1, x=0\}} Y_j + \sum_{j\in\{j \mid A_j=1, x=1\}} Y_j\right)}{|\{j \mid A_j = 1,\ x = 0\}| + |\{j \mid A_j = 1,\ x = 1\}|}$$

$$= \frac{(\#1 \text{ from group } 0\ \ + \#1 \text{ from group } 1)}{\text{Total Respondants}}$$

$$\rightarrow \frac{N^0 A^0 \bar{y}^0 + N^1 A^1 \bar{y}^1}{N^0 \mathrm{A}^0 + N^1 A^1} = \bar{y} \text{ if } \frac{N^0 \mathrm{A}^0}{P^0} = \frac{N^1 \mathrm{A}^1}{P^1}$$

With stratification, cancel out the bias *because* you simply asked more people from the group with lower response rate

It also reduces variance, even if $\mathrm{A}^0 = A^1$ (and $\mathrm{N}^0 = N^1$)

Stratification in practice

- You often don't know group specific response rates A^{ℓ}
  - Define groups and then keep sampling until you have enough samples
  - Weighting after sampling (covered next)
- How many groups/what groups do you choose?
  - Our example had a binary covariate we called "education"
  - What about stratifying ethnicity, or intersectional groups (ethnicity x gender)?
  - Why stop there? Why not ethnicity x gender x education x age …?
  - As number of groups increase, number of people in each group goes down
- Remember the rule: create groups such that the response rates is not correlated with whether they answer, *within each group*

Response Y_j *is independent of whether they respond* A_j*, within each group* x_j

Questions?

Weighting

Main idea for weighting

- In stratified sampling, we balanced out the groups according to their population percentage *before* we called people
- With weighting, we try to do the same thing, but *after* we call people and know how many from each group responded
- Why?
  - You might not know response rates per group
  - You might not know a person's demographics until you call them
  - Can run *sensitivity analyses:* "what would the estimate be if this demographic group only composes x% of the population instead of y%?"
- Comes at a cost: doesn't have the same variance reduction properties as does stratified sampling

Main idea, 2 steps:

Step 1: Use the responses to estimate the mean response for each group ℓ, i.e., get an estimate $\hat{y}^{\ell}$ of the true opinion $\bar{y}^{\ell}$

Step 2: Do a weighted average of $\hat{y}^{\ell}$; each group is given weight W^{ℓ}

$$\hat{y} = \sum_{\ell} W^{\ell}\, \hat{y}^{\ell}$$

If $W^{\ell} = P^{\ell}$ and $\hat{y}^{\ell} \to \bar{y}^{\ell}$, then $\hat{y} \to \bar{y}$

Details differ in how to construct estimate $\hat{y}^{\ell}$, how to calculate weight W^{ℓ}, and what groups ℓ to consider

Naïve Weighting

Step 1: Use the mean response for each group ℓ separately, i.e.

$$\hat{y}^{\ell} = \frac{\sum_{j \in \{j \mid A_j = 1, x = \ell\}} Y_j}{|\{j \mid A_j = 1,\ x = \ell\}|}$$

Step 2: Weight W^{ℓ} is our best guess of true population fraction P^{ℓ} for group ℓ

Complication: How many groups/which ones?

- If group too broad (e.g., group ℓ just gender), then break cardinal rule:

  Need: Opinion Y_j *is independent of whether they respond* A_j*, conditional on group* ℓ

- If group is too specific (*ethnicity* x *gender* x *education* x *age*), then:

  Problem 1: Estimate $\hat{y}^{\ell} = \frac{\sum_{j \in \{j \mid A_j = 1, x = \ell\}} Y_j}{|\{j \mid A_j = 1, x = \ell\}|}$ might be really bad

  Too few respondents in a group → high variance (1 person might determine entire average)

  Problem 2: We might not know population fraction P^{ℓ}

Tackling Problem 2: Population weights

- Suppose very specific group (*ethnicity* x *gender* x *education* x *age*)
- Naïve: try to figure out true population fraction ("joint distribution")

  "W^{ℓ} = P^{ℓ} fraction of pop is college educated white women age 35-44"

- Easier: Use "marginal" distribution for each covariate

  "a fraction of population is women"

  "b fraction of population is college educated"

  "c fraction of population is white"

  "d fraction of population is age 35-44"

  $\Rightarrow$Pretend "W^{ℓ} = *abcd fraction of pop is college educated white women age 35-44"*

- *Not covered -- "raking": match marginal distribution for each covariate without assuming that marginal distributions make up joint distribution*

The homework

- In the homework, first we define groups just based on a single covariate, for example gender, ethnicity/race, political party, etc.
  - (e.g., group ℓ just based on gender); we give you P^{ℓ}
- Then we define groups based on 2 covariates; we give you P^{ℓ}
- Then we define groups based on 2 covariates and ask you to construct P^{ℓ} based on marginal distributions

Tackling Problem 1: MRP

Problem 1: Estimate $\hat{y}^{\ell} = \frac{\sum_{j \in \{j \mid A_j = 1, x = \ell\}} Y_j}{|\{j \mid A_j = 1, x = \ell\}|}$ might be really bad

Too few respondents in a group → high variance (1 person might determine entire average)

- Somehow this seems wrong: presumably, the estimate for a group should be very close to that of a "neighboring" group
- "Multi-level regression with post-stratification" (MRP)

Main idea: Train a (Bayesian) regression model to get estimate $\hat{y}^{\ell}$ for each set of covariates. Then, "post-stratify" by weighting $\hat{y}^{\ell}$ by population fraction P^{ℓ}

For groups with many samples, estimate $\hat{y}^{\ell}$ just based on that group; otherwise, based on "neighboring" groups

Parting thoughts on weighting

- Where do the population percentages come from? In political polling, you need to define a universe of “likely voters”
- Methods not covered here: *Inverse Propensity Scoring*, and *Matching*
- Note, can only weight when you observe the covariates for each respondent!
- What if sampling bias is correlated with a feature you don’t observe?
  - Next time!

Announcements

- Homework 1 posted
- My office hours: 2-3pm today, in Bloomberg 201 + Zoom
  - Potentially will add Fridays depending on demand
- TA office hours: Fridays 1:30 – 2:30 (Over zoom)
  - This week: Introduction to Google Colaboratory (~15-20 minutes)
  - Potentially 1:30 – 3:30 depending on demand
- Increased course capacity to 75; waitlist should be clearing soon
- Please make sure you have access to EdStem and are receiving announcement notifications

Questions?