Missing Data: Introduction

David S. Rosenberg

NYU: CDS

February 3, 2021

Contents

Missing Data Example

The Mayor's Survey: Setup

- A new mayor has grand plans to improve satisfaction level of residents.
- She'll try many interventions during her term to improve satisfaction.
- She needs a baseline estimate for satisfaction levels.
- She calls $n = 100$ randomly selected residents from the city and asks:
  - "Do you think the city government is doing a good job? (yes or no)"
- Many people don't answer, and many others hang up without responding (surprise!)...

Missing survey responses

- The mayor gets a 10% response rate to her survey.
- What can she do?
- Should she just take the average of the responses she gets?
- What about response bias?

Notation and Terminology

- Suppose every individual i has a response $Y_i \in \{0,1\}$.
- But we only observe this response Y_i for 10% of those called.
- Let $R_i = \mathbb{1}[i \text{ responded}]$ be an indicator that we observe Y_i.
- We can write our observation for i as $(R_i, R_i Y_i)$.
- We get $(0,0)$ if there's no response and $(1, Y_i)$ if there is a response.

Missing completely at random (MCAR)

Just taking the average

- Consider just taking the mean of the observed Y_i's:
$$\hat{\mu}_{\text{cc}} = \frac{\sum_{i=1}^{n} R_i Y_i}{\sum_{i=1}^{n} R_i}$$
- Seems reasonable if whether a person responds is independent of their opinion.
- We'll formalize these intuitions, but first...
- Quick math question: can we compute $\mathbb{E}\hat{\mu}_{\text{cc}}$?

2021-02-19

Just taking the average

- Consider just taking the mean of the observed Y_i's:

$$\hat{\mu}_{cc} = \frac{\sum_{i=1}^{n} R_i Y_i}{\sum_{i=1}^{n} R_i}$$

- Seems reasonable if whether a person responds is independent of their opinion.
- We'll formalize these intuitions, but first...
- Quick math question: can we compute $\mathbb{E}\hat{\mu}_{cc}$?

Hint: Is there some probability that the estimator is undefined? (e.g. is 0/0)?

Missing Completely at Random (MCAR)

- Response indicators: $R, R_1, \ldots, R_n \in \{0,1\}$ are i.i.d. with $\mathbb{P}(R=1)=\pi$.
- Satisfaction indicators: $Y, Y_1, \ldots, Y_n \in \{0,1\}$ are i.i.d. with $\mu = \mathbb{E}Y$.

Definition (Missing completely at random (MCAR))

We say $Y_1, \cdots, Y_n$ are **missing completely at random** if Y_i and R_i are independent for each i.

Definition (Complete cases)

We'll refer to the observations pairs $(R_i, R_i Y_i)$ for which $R_i = 1$ as **complete cases**.

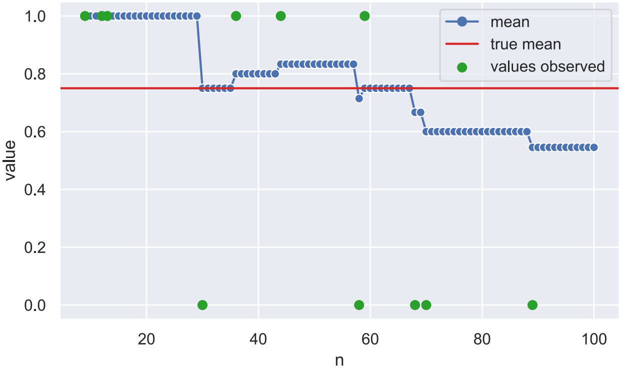
The complete case mean estimator

- The **complete case mean estimator** is defined as

$$\hat{\mu}_{cc} = \frac{\sum_{i=1}^{n} R_i Y_i}{\sum_{i=1}^{n} R_i}.$$

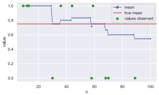
How does the complete case mean perform?

- Number of surveys: $n = 100$
- Response probability: $\mathbb{P}(R = 1) = 0.1$.
- True probability of satisfaction: $\mathbb{P}(Y = 1) = 0.75$.



How does the complete case mean perform?

- Number of surveys: $n = 100$
- Response probability: $\mathbb{P}(R = 1) = 0.1$.
- True probability of satisfaction: $\mathbb{P}(Y = 1) = 0.75$.



- The green dots represent observed values of Y_i.
- The blue dots show the value of $\hat{\mu}_{cc}$ as n increases.
- The blue dots don't start at 0, but rather at the first green dot, since the estimator isn't defined until we have at least one observation.
- Note that the estimate remains unchanged between observations.
- The horizontal line shows the true expected value of the Y_i's.

Complete Case Mean, MCAR: Properties

- The "complete case" mean estimator is defined as

$$\hat{\mu}_{cc} = \frac{\sum_{i=1}^{n} R_i Y_i}{\sum_{i=1}^{n} R_i} = \frac{\frac{1}{n}\sum_{i=1}^{n} R_i Y_i}{\frac{1}{n}\sum_{i=1}^{n} R_i}.$$

- Let $\mu = \mathbb{E}Y$ and $\pi = \mathbb{P}(R = 1)$.
- By the LLN, the numerator converges to $\mathbb{E}[RY] = \pi\mu$, by MCAR.
- By the LLN, the denominator converges to π.
- Thus $\hat{\mu}_{cc} \xrightarrow{P} \mu$.

Complete Case Mean, MCAR

- The "complete case" mean estimator is defined as

$$\hat{\mu}_{\text{cc}} = \frac{\sum_{i=1}^{n} R_i Y_i}{\sum_{i=1}^{n} R_i}$$

  and it **has a few oddities**.
- When everything is missing, the estimator is $0/0$, which is not defined.
- We can't even talk about whether it's biased, much less its variance.
- We could just define $\hat{\mu}_{\text{cc}} = 0$ when $R_1 = \cdots = R_n = 0$.
- Exercise: Show that doing this yields a biased estimator when $n = 1$.
- Exercise: Show that $\mathbb{E}\left[\hat{\mu}_{\text{cc}} \mid \sum_{i=1}^{n} R_i > 0\right] = \mathbb{E}Y$, so the complete case mean is at least conditionally unbiased.

Missing at random (MAR)

Missing at random (MAR)

- MCAR is a very strong assumption – often blatantly not true.
- Most commonly we make an assumption called **missing at random**.
- Usually more defensible than MCAR.
- Requires introduction of a covariate X into the picture.

Missing at random (MAR)

- Assume we have additional information X_i about each individual i.
- Also assume that X_i is **never missing**.

Definition (Missing at random (MAR))

$Y_1, \ldots, Y_n$ are missing at random if, after observing X_i, R_i has no additional information about Y_i. More formally, R_i and Y_i are conditionally independent given X_i, which we'll denote by

$$R_i \perp\!\!\!\perp Y_i \mid X_i \quad \forall i = 1, \ldots, n.$$

Can't check it...

- There is no way to verify this MAR assumption, at least not without full data
- Nevertheless, this is the assumption that is most commonly made.

Missing at random (MAR)

- Assume we have additional information X_i about each individual i.
- Also assume that X_i is **never missing**.

Definition (Missing at random (MAR))

$Y_1, \ldots, Y_n$ are missing at random if, after observing X_i, R_i has no additional information about Y_i. More formally, R_i and Y_i are conditionally independent given X_i, which we'll denote by

$$R_i \perp\!\!\!\perp Y_i \mid X_i \quad \forall i = 1, \ldots, n.$$

Can't check it. . .

- There is no way to verify this MAR assumption, at least not without full data
- Nevertheless, this is the assumption that is most commonly made.

- Note that if X is independent of Y (i.e.X is fairly useless covariate), then we're back in the MCAR case.
- Full data, we'll learn on the next slide, is data with nothing missing.

More terminology and formalization

- The full data is the dataset we would observe if nothing were missing.
  - Denote that by $(X_1, Y_1), \ldots, (X_n, Y_n)$
- What we actually observe:

$$(X_1, R_1, R_1 Y_1), \ldots, (X_n, R_n, R_n Y_n)$$

- The complete data are the cases with observed Y (i.e. $R = 1$)
  - Explains the terminology "complete case estimator"
- The incomplete data cases are cases with missing Y (i.e. $R = 0$)

The propensity score

- Key piece in the MAR setting is the model for missingness:

$$\mathbb{P}(R=1 \mid X=x, Y=y) = \mathbb{P}(R=1 \mid X=x) = \pi(x).$$

- This model can be fit in the usual way, using tools from statistics and ML.
- Logistic regression is a common approach.
- For most of this course, it will be reasonable to assume we know $\pi(x)$,
  - or can estimate it relatively well.
- The model for missingness goes by different names in different contexts.
- We will generally refer to it as the **propensity score**. [RR83]

How can the mayor use the propensity scores?

- Suppose the mayor has a probability of response for each individual.
  - e.g. She has built a model using historical response data.
- Each individual i potentially has probability $\pi(X_i)$ to respond.
- Is our previous complete case mean still a reasonable estimator?
- It gives too much weight to individuals who are more likely to respond.

3 basic approaches to the MAR problem

Likelihood methods missing data are latent variables, find or estimate MLE

Imputation methods use X to impute Y, then proceed as with full data

Inverse propensity weighting (IPW) just use complete cases, but weight by propensity

- Likelihood methods are general and elegant, but often difficult to apply
- We will focus on the imputation and IPW methods
- We will also look into "doubly robust" methods, which combine IPW and imputation

Assumption: nonzero propensity scores

- Suppose $\pi(x) = 0 \ \forall x \in \mathcal{X}_0 \subset \mathcal{X}$.
- This means we never observe any Y values correspond to $x \in \mathcal{X}_0$.
- This is potentially a serious blind spot in our data collection.
- If the distribution of $Y \mid X \in \mathcal{X}_0$ is the same as $Y \mid X \in \mathcal{X}$, then we're still OK.
- If $\mathbb{P}(X \in \mathcal{X}_0)$ is very small, and the distributions aren't too different, then maybe we're OK.
- But generally, if there's a subset $\mathcal{X}_0$ with 0 response probability, we can have serious bias in our estimators, with no fix without additional assumptions.

Assumption

Unless otherwise noted, we will always assume that propensity scores are strictly positive: $\pi(x) > 0$.

Assumption: nonzero propensity scores

- Suppose $\pi(x) = 0\ \forall x \in \mathcal{X}_0 \subset \mathcal{X}$.
- This means we never observe any Y values correspond to $x \in \mathcal{X}_0$.
- This is potentially a serious blind spot in our data collection.
- If the distribution of $Y \mid X \in \mathcal{X}_0$ is the same as $Y \mid X \in \mathcal{X}$, then we're still OK.
- If $\mathbb{P}(X \in \mathcal{X}_0)$ is very small, and the distributions aren't too different, then maybe we're OK.
- But generally, if there's a subset $\mathcal{X}_0$ with 0 response probability, we can have serious bias in our estimators, with no fix without additional assumptions.

Assumption

Unless otherwise noted, we will always assume that propensity scores are strictly positive: $\pi(x) > 0$.

- For example, suppose we have a list of all the potential voters in an election, along with their phone number.
- We select randomly from the list and call to ask who they'll vote for.
- Different people have different propensities to response (e.g. retirement-age individuals tend to respond at a much higher rate than college age individuals).
- But consider the subset of voters $\mathcal{X}_0 = \{\text{missing or incorrect phone number}\}$.
- Unless we have another contact mechanism besides a phone call, the propensity score for any individual in $\mathcal{X}_0$ will be 0.
- If individuals with missing or incorrect phone numbers are different in the voting preferences, we will have a sample bias that we can't fix without additional assumptions.
- Along the same lines, the propensity score will probably be for individuals who do not speak a common language with whoever calls them with the survey question.

References

Resources

- Chapter 6 in Tsiatis's book *Semiparametric theory and missing data* gives a nice overview of the missing data problem. [Tsi06, Ch. 6].

References I

[RR83] Paul R. Rosenbaum and Donald B. Rubin, *The central role of the propensity score in observational studies for causal effects*, Biometrika **70** (1983), no. 1, 41–55.

[Tsi06] Anastasios A. Tsiatis, *Semiparametric theory and missing data*, Springer, 2006.