Missing Data: Introduction

David S. Rosenberg

NYU: CDS

January 23, 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 $R_i = \mathbb{1}[i \text{ revealed } Y_i]$ 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.

"MCAR" and the complete case mean estimator

Just taking the average

- Let's consider the estimator that is the mean of the observed Y_i's:

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

- This seems reasonable if the probability a person responds is independent of their opinion.
- Let's formalize these intuitions.
- Quick math question: does this estimator have an expected value?
- 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

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

- By the LLN, the numerator converges to $\mathbb{E}[RY] = \pi\mu$, by MCAR.
- By the LLN, the denominator converges to π.
- Thus $\hat{\mu}_{\text{cc}} \xrightarrow{P} \mu$, as desired.

Complete Case Mean, MCAR

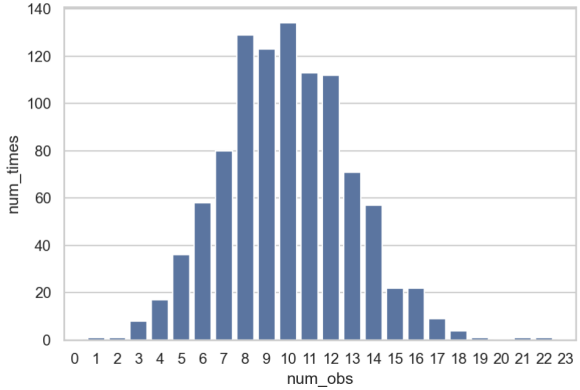
- The "complete case" mean estimator is defined as

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

  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 $n = 1$.)

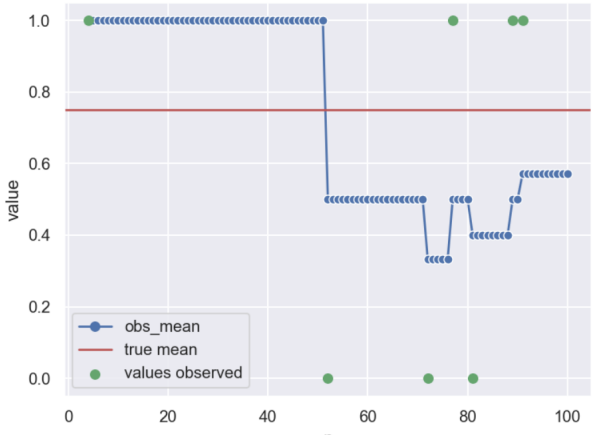
Probability review: how many responses will we get?

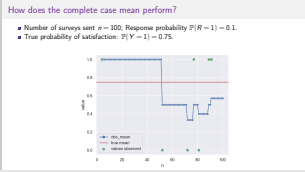
- Missingness indicators: $R, R_1, \ldots, R_n \in \{0, 1\}$ are i.i.d. with $\mathbb{P}(R = 1) = 0.1$.
- Number of observations: $N = \sum_{i=1}^{n} R_i$. *What's the distribution of N?*
- Expected number of responses is $0.1n$ and $N \sim \text{Binom}(n, p = 0.1)$.
- Histogram of N from 1000 simulations of our setup with $n = 100$:



How does the complete case mean perform?

- Number of surveys sent $n = 100$; 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 $\hat{\mu}_{\text{cc}}$ as n increases. Note that 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 also that the estimate remains unchanged between observations. The horizontal line shows the true expected value of the Y_i's.

Missing not at random (MNAR)

Missing not at random (MNAR)

- Suppose people are more likely to respond if they're satisfied.
  - $\implies$ NOT MCAR
- More generally, suppose $\mathbb{P}(R_i = 1 \mid Y_i) = \pi(Y_i)$ for some function $\pi(y)$.
- Nothing changes with our LLN argument,
  - estimator still converges to $\mathbb{E}(RY)/\mathbb{E}(R)$.
- However, now we can only say that

$$\frac{\mathbb{E}[RY]}{\mathbb{E}[R]} = \frac{\mathbb{E}[\mathbb{E}(RY \mid Y)]}{\mathbb{E}[\mathbb{E}(R \mid Y)]} = \frac{\mathbb{E}[Y\pi(Y)]}{\mathbb{E}\pi(Y)},$$

  which does not equal $\mathbb{E}Y$, at least in general.
- Is this a real issue?

Missing not at random: concrete example

- Suppose response probability is higher for satisfied people:

$$\pi(y) = \begin{cases} 0.2 & \text{when } y = 1 \\ 0.1 & \text{when } y = 0. \end{cases}$$

- Then our complete case mean converges to

$$\begin{aligned} \frac{\mathbb{E}[Y\pi(Y)]}{\mathbb{E}\pi(Y)} &= \frac{1 \cdot \mathbb{P}(Y=1)0.2 + 0 \cdot \mathbb{P}(Y=0)\,0.1}{\mathbb{P}(Y=1)0.2 + \mathbb{P}(Y=0)\,0.1} \\ &= \frac{0.2\mu}{0.2\mu + 0.1\,(1-\mu)} \end{aligned}$$

- If $\mu = \mathbb{P}(Y=1) = 0.5$, then we get $\hat{\mu}_{\text{cc}} \xrightarrow{P} \frac{\mathbb{E}[Y\pi(Y)]}{\mathbb{E}\pi(Y)} = \frac{2}{3}$.
- If $\mu = \mathbb{P}(Y=1) = 0.1$, then we get $\hat{\mu}_{\text{cc}} \xrightarrow{P} \frac{\mathbb{E}[Y\pi(Y)]}{\mathbb{E}\pi(Y)} \approx .18$.
- Generally speaking (exercise), our estimate will converge to an overestimate of the actual parameter $\mu = \mathbb{E}Y = \mathbb{P}(Y=1)$.

Missing not at random: what can we do?

- TL;DR: Not much, without some additional assumptions.
- But suppose somehow we know $\pi(y)$
  - i.e. the probability that somebody will respond given their opinion
- Then we can form a maximum likelihood estimator for μ. (Homework)
- Unfortunately, impossible to estimate $\pi(y)$ with our data.
  - We have no observations of Y when $R = 0$.

Missing not at random: make missings into observed?

- Suppose we could spend a lot of time and money and track down a random sample of nonresponders and hound them until they give their values of Y.
- If we assume the persistence and hounding doesn't change their answers,
  - we could use this data to estimate $\pi(y)$.
  - We could use the maximum likelihood approach suggested previously.
- But this may be very difficult in practice.
- Also not at all clear that plugging this estimate for $\pi(y)$ into the maximum likelihood approach would yield a better estimate than just looking at the complete data we managed to collect to estimate $\pi(y)$.

Missing at random (MAR)

Missing at random (MAR)

- MCAR is a very strong assumption – often blatantly not true.
- With MNAR... there's not so much we can do.
- Most commonly we make an assumption called "missing at random"
  - Which is usually more defensible than MCAR
  - There's a lot more we can do with that assumption compared to MNAR

Missing at random (MAR)

- Assume we have additional information X_i about each individual i.
  - X_i is never missing

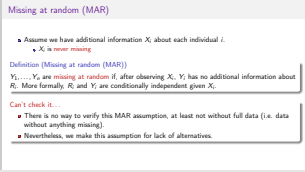
Definition (Missing at random (MAR))

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

Can't check it...

- There is no way to verify this MAR assumption, at least not without full data (i.e. data without anything missing).
- Nevertheless, we make this assumption for lack of alternatives.

└Missing at random (MAR)



Note that if X is independent of Y (i.e. a fairly useless covariate), then we're back in the MCAR case.

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:

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

- Note that there is only a dependency on x on the RHS.
- 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 $\pi(y)$,
  - 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.

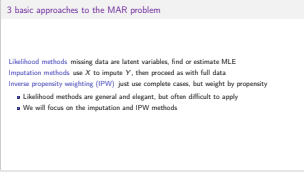
3 basic approaches to the MAR problem

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



Sutton and Barto ran the ε-greedy algorithms with 3 settings of epsilon on the 2000 bandit problems, each for 1000 steps. The average reward achieved at step t across the 2000bandit problems is plotted over time. Note that ε-greedy with $\varepsilon = 0.1$ will never choose the optimal action more than about 90% of the time. Or $.9 + \varepsilon/k = .91$ of the time, to be more precise. The pure greedy strategy ($\varepsilon = 0$)

References

Resources

- Terminology was based on [CFV17].

References I

[CFV17] Victor Chernozhukov and Iván Fernández-Val, *Treatment effects*, Econometrics—MIT Course 14.382, Cambridge MA, 2017, MIT OpenCourseWare.

[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.