Handling Missing Data in Epidemiology Research

Exploring Multiple Imputation as a Missing Data Strategy

Mackenzie Mueller

Department of Mathematics
Pacific Lutheran University

May 3, 2024

Background

Introduction

Why is missing data important?

- Poor reporting, analysis, and handling of missing data [5]
- Impacts the results of a study
- Occurs everywhere

Types of Missing Data

MCAR

Data is said to be missing completely at random if the probability of being missing is the same for all variables. [1]

MAR

Data is said to be missing at random if the probability of being missing is the same only within groups defined by the *observed* data. [1]

MNAR

Data is said to be missing not at random if the probability of being missing varies for unknown reasons and depends on unobserved values. [1]

Data Context

- National Health and Nutrition Examination Survey (public dataset)
- Around 5,000 participants ages 0-80+ every two years
- NHANES oversamples to obtain **nationally representative** data (people aged 60 and over, African Americans, and Hispanic-identifying individuals)
- Constantly evolving to meet emerging health/nutrition needs [2]

Variables

For the purpose of this project, relevant variables were selected from NHANES 2018 data:

Variable	Description
age	Patient age in years
sex	Sex of patient (M/F)
race_ethnicity	Self identified patient race
combined_education	Patient education level
reference_education	Reference person education level
fipr	Federal income poverty ratio (0-5)
tap_water	Amount of tap water drank, day 1, grams

Motivating Question

Can `fipr` help predict `tap_water` consumption?

Definition

fipr: A ratio of family income to poverty guidelines, where a 1 represents a family at the federal poverty line. (Range from 0-5) [2]

Data

- Sample size: 8704
- Complete cases: 4056
- Notice by ommitting NA's, we lose 53% of our data
- Missing fipr cases: 1070
- For the scope of this project, we only had time to focus on missing data in `fipr` variable

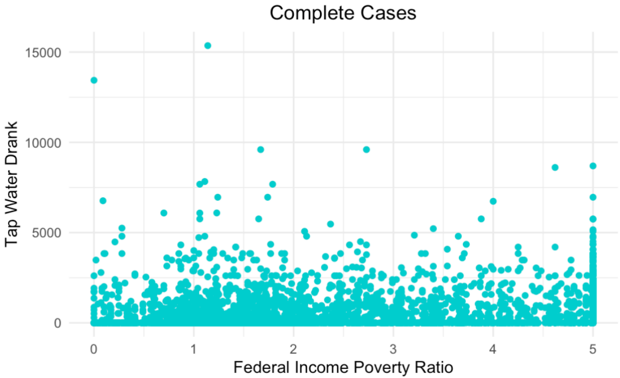
Data Cleaning

- Created combined_education variable
- Created new_race_ethnicity variable
- Recoded “don’t know” or “refused” as NA values [2]

Ad-hoc Solutions to Handling Missing Data

Complete Cases

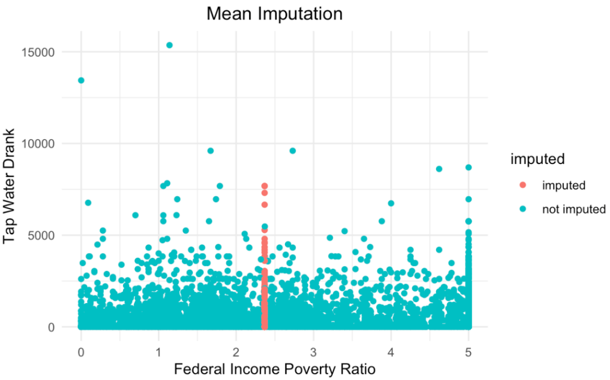
Also known as listwise deletion, the complete cases method omits all rows that contain a missing value in any category. [1]



Ad-hoc Solutions to Handling Missing Data

Mean Imputation

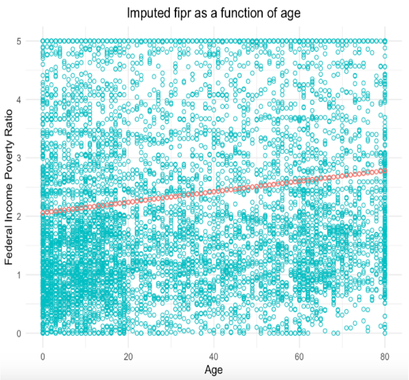
Replaces missing data values with the mean of that variable. [1]

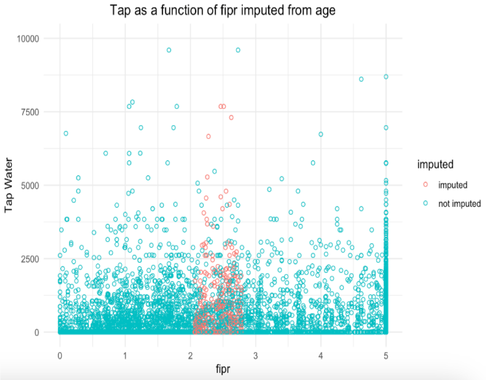


Ad-hoc Solutions to Handling Missing Data

Regression Imputation

A model is built from the observed data first, then predictions for missing values are calculated under the fitted model. The missing values are assigned the "most likely" values on the regression line. [1]





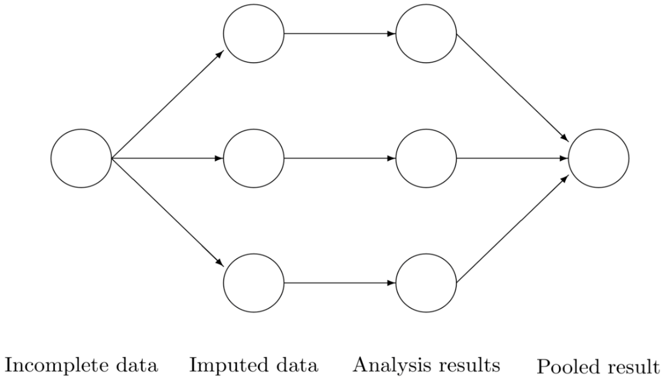
Research Question

How can multiple imputation be used to handle missing data?

Overview of Multiple Imputation

- MI deals with the uncertainty of the imputations themselves by imputing *several* values for each missing case (g between 5-10) [4]
- Imputed datasets are identical for observed data entries and differ in imputed values
- Imputed values are averaged into a pooled dataset

Visualization



Flexible Imputation of Missing Data, van Buuren, 2018.

Multiple Regression

We begin with a multiple regression model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + E$$

Where

- Y is first `fipr`, and later `tap_water`
- β_0 is the intercept, or constant
- $\beta_1, \beta_2, ..., \beta_k$ are the coefficients of our predictor variables
- $x_1, x_2, ...x_k$ are the predictor variables
- E is the random error term that represents the part of Y that cannot be explained by the predictor variables [3]

Model Prediction using R Statistical Software

We use multiple linear regression to predict a model for fipr based on the complete cases of the following variables:

```
fipr_model <- lm(fipr ~ reference_education +
combined_education + new_race_ethnicity + age + sex)

summary(fipr_model)
```

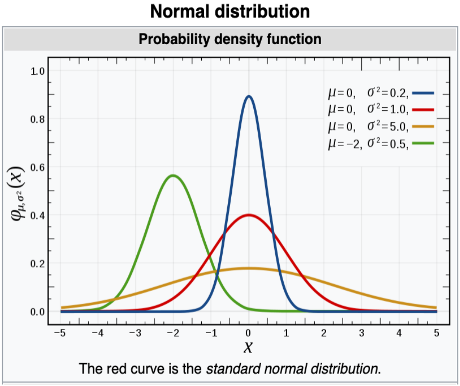
Random Noise

We impute our dataset g times, adding randomly generated noise E to Y based on the standard deviation of the residuals in `fipr_model` that assumes a normal distribution:

$$E \sim N\left(0, \sigma^2\right)$$

based on $n - m$ complete cases:

$$\sigma^2 = \frac{\sum_{i=m+1}^{n}\left(Y - \hat{Y}\right)^2}{n - m}$$



"Normal Distribution." Wikipedia.

fipr Models

We fit 5 models to impute the missing `fipr` values as follows:

- If `fipr` is missing, replace NA value with

$$\beta_0 + \beta_1 x_1 + ... + \beta_{11} x_{11} + E_i$$

  Where

  - $\beta_1, \beta_2, ..., \beta_{11}$ are the regression coefficients
  - $x_1, x_2, ... x_{11}$ are the indicator variables
  - E is the random noise

- Else, use given `fipr`

Note: the `pmin()pmax()` functions were used to adjust negative fipr values to 0, and fipr values > 5 to 5.

Tap Models

Now we have 5 complete datasets across our education, race, age, sex, *and* fipr variables. We use these predictor variables to model tap water consumption.

```
tap1 <- lm(reference_education + combined_education +
       new_race_ethnicity + age + sex + fipr1)

                         ⋮

tap5 <- lm(reference_education + combined_education +
       new_race_ethnicity + age + sex + fipr5)
```

Coefficients and Standard Errors

From these models we aquire:

- $\beta_0^\ell, \beta_1^\ell, \beta_2^\ell, ..., \beta_k^\ell$ for $\ell = 1...g$
- $SE(\beta_0^\ell), SE(\beta_1^\ell), ..., SE(\beta_k^\ell)$ for $\ell = 1...g$

  Where

  - k represents each of our 11 predictor variables, and
  - ℓ represents each of our 5 imputed datasets.

Regression Coefficients

Averaging across our 5 imputed datasets produces point estimates of our regression coefficients...

$$\tilde{\beta}_j \equiv \frac{\sum_{\ell=1}^{g} \beta_j^{\ell}}{g}$$

Standard Errors

... Standard errors of estimated coefficients are obtained by combining info about *within* and *between* imputation variation:

$$\tilde{SE}\left(\tilde{\beta}_j\right) \equiv \sqrt{V_j^{(W)} + \frac{g+1}{g} V_j^{(B)}}$$

- Within imputations:

$$V_j^{(W)} \equiv \frac{\sum_{\ell=1}^{g} SE^2\left(\beta_j^{\ell}\right)}{g}$$

- Between imputations:

$$V_j^{(B)} \equiv \frac{\sum_{\ell=1}^{g}\left(\beta_j^{\ell} - \tilde{\beta}_j\right)^2}{g-1}$$

Final Model

... And we are left with a final model:

$$\hat{tap} = \texttt{570.1 + (-86.3)ref\_some\_college + (-62.9)ref\_college\_grad + (-164.9)com\_some\_college + (-361.9)com\_less\_hs + (184.1)race\_asian + (17.2)race\_black + (286.9)race\_white + (254.0)race\_other + (-0.8)age + (-12.9)sex\_female + (38.8)fipr}$$

Note that there is uncertainty around each of our β values, for example:

- The standard error for `fipr` is 19.29, with a 95% confidence interval between 0.19 and 75.8.
- The standard error for `sex_female` is 23.44, with a 95% confidence interval between -58.78 and 33.08.

Example with Toy Data

	fipr	race	age	tap_water
1	0.5	white	44	400
2	4.0	black	37	200
3	2.5	white	74	150
4	3.0	white	20	600
5	*NA*	black	25	550
6	*NA*	white	61	500
7	4.7	black	40	1000

Figure: Before

Imputations

imputed_fipr1	imputed_fipr2	imputed_fipr3	imputed_fipr4	imputed_fipr5
0.5	0.5	0.5	0.5	0.5
4.0	4.0	4.0	4.0	4.0
2.5	2.5	2.5	2.5	2.5
3.0	3.0	3.0	3.0	3.0
5.0	4.3	5.0	2.5	5.0
2.6	1.7	3.4	0.0	3.5
4.7	4.7	4.7	4.7	4.7

Figure: During

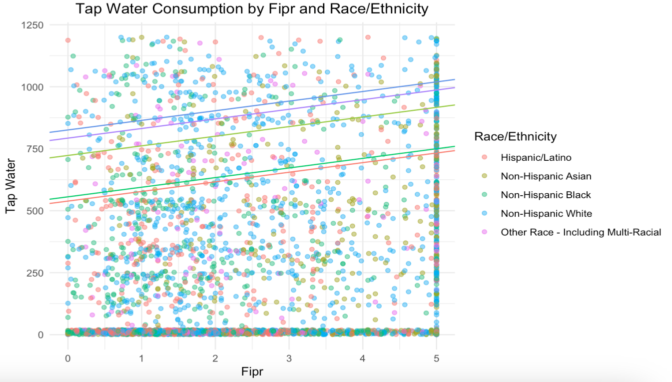
Average Imputed Values

	imputed_fipr_avg	race	age	tap_water
1	0.5	white	44	400
2	4.0	black	37	200
3	2.5	white	74	150
4	3.0	white	20	600
5	4.3	black	25	550
6	2.2	white	61	500
7	4.7	black	40	1000

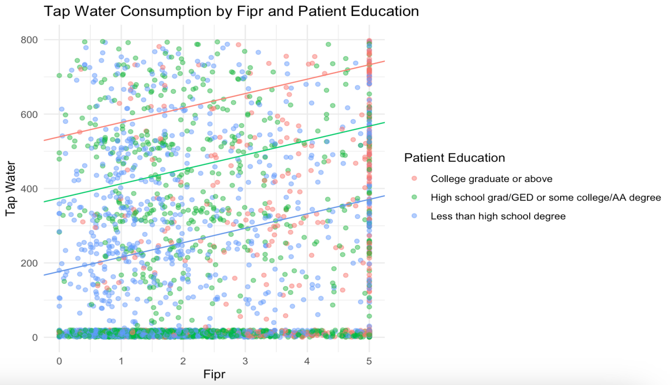
Figure: After

```r
##multiple regression model:
example_model <- lm(fipr ~ race + age, data = example_data)

##impute missing fipr values 5 times
set.seed(16)
imputed_fipr1 <- ifelse(
  is.na(example_data$fipr),
  pmin(pmax(
    4.558427 + (-2.309397) * example_data$race_ind + (-0.005414) * example_data$age +
      rnorm(1, mean = 0, sd = 1.36), 0), 5),
  example_data$fipr)
...
imputed_fipr5 <- ifelse(
  is.na(example_data$fipr),
  pmin(pmax(
    4.558427 + (-2.309397) * example_data$race_ind + (-0.005414) * example_data$age +
      rnorm(1, mean = 0, sd = 1.36), 0), 5),
  example_data$fipr)

##create a matrix containing imputed fipr values:
imputed_fiprs <- cbind(imputed_fipr1, imputed_fipr2, imputed_fipr3, imputed_fipr4, imputed_fipr5)

##average 5 imputed datasets:
average_fiprs <- rowMeans(imputed_fiprs, na.rm = TRUE)

##replace missing fipr values with the imputed values
example_data$imputed_fipr_avg <- example_data$fipr
example_data$imputed_fipr_avg[is.na(example_data$fipr)] <- average_fiprs[is.na(example_data$fipr)]
```

Visualizations



Visualizations



Limitations

- We only worked with missing fipr values, rather than missing values across our whole dataset
- new_race_ethnicity decision during data cleaning
- We simplified the MI process, by only adding random error to the regression outputs (we didn't also add random error to the coefficients)
- We looked at all `tap_water` cases, we did not remove 0's. This could be a direction of future research...

Takeaways

- Multiple imputation as a missing data strategy provides a more sophisticated way of handling missing data that accounts for randomness and provides plausible imputed values.
- In response to our motivating question, we can conclude that fipr is a statistically significant predictor of tap water consumption.

Acknowledgements

- Dr. Justice
- Dr. Simic-Muller
- Dr. Edgar

References

Stef van Buuren. *Flexible Imputation of Missing Data, Second Edition*. CRC Press, 2018. ISBN: 9780429960352, 0429960352.

National Center for Health Statistics. *About the National Health and Nutrition Examination Survey*. URL: https://www.cdc.gov/nchs/nhanes/about_nhanes.htm. (accessed: 03.09.2024).

Ramzi W. Nahhas. *Introduction to Regression Methods for Public Health Using R*. bookdown, 2024. ISBN: 9781483386478. URL: https://www.bookdown.org/rwnahhas/RMPH/#preface.

Mailman School of Public Health. "Missing Data and Multiple Imputation". In: (2024). URL: https://www.publichealth.columbia.edu/research/population-health-methods/missing-data-and-multiple-imputation.

Hairui Yu. "What is Missing in Missing Data Handling?" In: *Journal of Statistics and Data Science Education* 32.1 (2023), pp. 3–10. DOI: https://doi.org/10.1080/26939169.2023.2177214.

Questions?