Help! I have missing data. How do I fix it (the right way)?

Matt Brems

- Global Lead Data Science Instructor, General Assembly
- https://github.com/matthewbrems/ODSC-missing-data-may-18
- Twitter: @matthewbrems
- LinkedIn: https://www.linkedin.com/in/matthewbrems/
- Site: www.argmatt.com

Outline

- Introduction
- What is missing data?
- Three strategies for tackling missing data.
- Comparing unit and item nonresponse.
- Discussing three types of missingness.
- Conclusion

Index	Age	Sex	Income
1			
2			
3			
4			
⋮			
10000			

Index	Age	Sex	Income
1			
2			
3			
4			
⋮			
10000			

Administer survey.

Index	Age	Sex	Income
1	65	M	70k
2	39	M	75k
3	42	F	100k
4	28	F	50k
⋮			
10000	18	F	10k

Index	Age	Sex	Income
1			
2			
3			
4			
⋮			
10000			

Administer survey.

Index	Age	Sex	Income
1	65	M	70k
2	39	M	75k
3	42	F	100k
4	28	F	50k
⋮			
10000	18	F	10k

What actually happens.

Index	Age	Sex	Income
1	NA	M	NA
2	39	NA	75k
3	NA	NA	NA
4	28	F	50k
⋮			
10000	18	F	NA

Why does this missing data occur?

- Accessibility issues.
- Sleepy undergraduates recording data.
- Maybe one sex is less likely to talk about income.
- Maybe people who earn less money are less likely to talk about income.

Let's shift our mindset.

- Don't use the phrase "missing data analysis."
- Instead, use the phrase "data science with missing data" or "statistical analysis with missing data."

In most cases, we aren't "fixing" the missing data. We're just learning how to cope with it.

Methods of Handling Missing Data

1. We can **avoid** it.

2. We can **ignore** it.

3. We can **account** for it.

Why talk about avoiding missingness?

- Avoiding missing data altogether allows us to reduce our uncertainty.
- It's usually cheaper for us to spend time avoiding missing data than to make guesses about the best way to fill it in.
- If we can avoid missing data, we make what comes afterward so much easier.

Unit Nonresponse vs. Item Nonresponse

- **Unit nonresponse** is where no values from an observation are observed.

- Suppose that I randomly generate 1,000 valid U.S. phone numbers in order to administer some survey. I write my survey in English and I call someone who doesn't speak English. This person didn't answer any of our questions, so this qualifies as "unit nonresponse."

Unit Nonresponse vs. Item Nonresponse

- **Item nonresponse** is where some, but not all, values from an observation are observed.

- Suppose that I randomly generate 1,000 valid U.S. phone numbers in order to administer some survey. I call all 1,000 individuals and ask them what their age, sex, and income are. 50 of them tell me their age and sex, but do not tell me their income.

Unit Nonresponse vs. Item Nonresponse

- In observation 3, we observe unit nonresponse.

- In observations 1, 2, and 10,000, we observe item nonresponse.

Index	Age	Sex	Income
1	NA	M	NA
2	39	NA	75k
3	NA	NA	NA
4	28	F	50k
⋮			
10000	18	F	NA

Unit vs. Item Nonresponse

- Unit Nonresponse
  - Avoid it.
  - Ignore it.
  - Account for it.
- Item Nonresponse
  - Avoid it.
  - Ignore it.
  - Account for it.

How do we avoid unit nonresponse?

- Respondent burden.
- Method of data collection.
- Accessibility.
- Time of survey.

How do we ignore unit nonresponse?

- This is straightforward – just assume that your sample of respondents is close enough to the sample of respondents and nonrespondents.
  - If your proportion of nonrespondents is low, you might feel comfortable making this assumption.

- This is usually the software default!

How do we account for unit nonresponse?

- Suppose I want to estimate the proportion of people who support candidate X in an upcoming election.

- Ignore it: $\hat{p} = \frac{\sum_i I(vote\ for\ X)_i}{N_{responses}}$

- Account for it: $\hat{p} = \frac{\sum_i w_i \cdot I(vote\ for\ X)_i}{\sum_i w_i}$

Weight Class Adjustments

- Reweight so your data reflect the population of interest.
- I believe those who vote will be 50% male and 50% female. However, 75% of my responses came from males and 25% came from females.

- $w_{male} = \frac{true\ proportion}{proportion\ of\ responses} = \frac{0.50}{0.75} = \frac{2}{3}$

- $w_{female} = \frac{true\ proportion}{proportion\ of\ responses} = \frac{0.50}{0.25} = 2$

Caution: Weight Class Adjustments

- Postweighting requires that we know what the true population distribution is.
  - In the example, we needed to know what percentage of voters will be female.
  - This is often unrealistic.

- Depending on how much weighting we do, we might see that variance of estimates increase dramatically.

When to use postweighting for unit nonresponse?

- If you're willing to assume what the population distribution will look like for the variables on which you're weighting.

- If you're willing to minimize bias at the expense of an increase in variance.

- There is a similar procedure called "weighted least squares regression" if you want to apply postweighting to a more complicated model.

Unit vs. Item Nonresponse

- Unit Nonresponse
  - Avoid it.
  - Ignore it.
  - Account for it.
- Item Nonresponse
  - Avoid it.
  - Ignore it.
  - Account for it.

How do we avoid item nonresponse?

- Questionnaire length.
- Questionnaire design.
- Survey content.

How do we ignore item nonresponse?

- Complete-Case Analysis
  - Drops any observation with any missing value.
    - Pros: Results will be well-behaved, simplest, usually software default.
    - Cons: Drops some collected data, loses “information” and precision.

How do we ignore item nonresponse?

- Complete-Case Analysis
  - Drops any observation with any missing value.
    - Pros: Results will be well-behaved, simplest, usually software default.
    - Cons: Drops some collected data, loses "information" and precision.

- Available-Case Analysis
  - Drops no observations and calculates results based on available data.
    - Pros: Uses all data available.
    - Cons: Can get "not well-behaved results," i.e. invalid covariance matrices.

How do we account for item nonresponse?

- Imputation
  - Deductive Imputation
  - Mean/Median/Mode Imputation
  - Regression Imputation
  - Stochastic Regression Imputation
  - Multiple Stochastic Regression Imputation
  - Proper Imputation
  - Hot-Deck Imputation

Deductive Imputation

- Uses logical relations to fill in missing values.
  - Respondent mentions they were not the victim of a crime, so the column for "victim of a crime" contains a 0. However, an "NA" exists in the column for "victim of a violent crime." Because the respondent mentioned they were not the victim of a crime, we know that the respondent was not the victim of a violent crime.
  - If someone has 2 children in year 1, NA children in year 2, and 2 children in year 3, we can probably impute that they have 2 children in year 2.

Deductive Imputation

- Uses logical relations to fill in missing values.
  - Respondent mentions they were not the victim of a crime, so the column for "victim of a crime" contains a 0. However, an "NA" exists in the column for "victim of a violent crime." Because the respondent mentioned they were not the victim of a crime, we know that the respondent was not the victim of a violent crime.
  - If someone has 2 children in year 1, NA children in year 2, and 2 children in year 3, we can probably impute that they have 2 children in year 2.

- Pros: Requires no "inference," true value can be assessed, valid method.
- Cons: Can be time consuming or requires specific coding.

Mean/Median/Mode Imputation

- For any “NA” value in a given column, mean imputation replaces “NA” with the mean of that column. (Same for median and mode imputation.)

Mean/Median/Mode Imputation

- For any "NA" value in a given column, mean imputation replaces "NA" with the mean of that column. (Same for median and mode imputation.)

- Pros: Easy to implement and comprehend. *Seems* reasonable.
- Cons: Significantly distorts histogram, underestimates variance, mean and median imputation will give very different results for asymmetric data, invalid method.

Regression Imputation

- For any "NA" value in a given column, regression imputation replaces "NA" with a predicted value based on a regression line.
  - i.e. Given observed demographic data, estimate $\mathbf{income} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{age} + \hat{\beta}_2 \mathbf{sex}$, then use observed age and sex as predictors to impute missing income data.

Regression Imputation

- For any "NA" value in a given column, regression imputation replaces "NA" with a predicted value based on a regression line.
  - i.e. Given observed demographic data, estimate $\mathbf{income} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{age} + \hat{\beta}_2 \mathbf{sex}$, then use observed age and sex as predictors to impute missing income data.

- Pros: Easy to comprehend, seems logical, better than mean/median/mode imputation.
- Cons: Still distorts histogram and underestimates variance, invalid method.

Stochastic Regression Imputation

- For any "NA" value in a given column, stochastic regression imputation replaces "NA" with a predicted value based on a regression line and random error.
  - i.e. Estimate $\mathbf{income}_i = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{age} + \hat{\beta}_2 \mathbf{sex} + \varepsilon_i$ and $\varepsilon_i \sim N(0, s)$, then use observed age and sex as predictors to impute missing income data, plus random draw from $N(0, s)$.

Stochastic Regression Imputation

- For any "NA" value in a given column, stochastic regression imputation replaces "NA" with a predicted value based on a regression line and random error.
  - i.e. Estimate $\text{income}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{age} + \hat{\beta}_2 \text{sex} + \varepsilon_i$ and $\varepsilon_i \sim N(0, s)$, then use observed age and sex as predictors to impute missing income data, plus random draw from $N(0, s)$.

- Pros: Easy to comprehend, better than regression imputation, allows for much better estimation of true variance.
- Cons: Still underestimates variance, invalid method.

Multiply Stochastic Regression Imputation

- For any "NA" value in a given column, multiply stochastic regression imputation replaces "NA" with a predicted value based on a regression line and random error.
  - i.e. Estimate $\text{income}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{age} + \hat{\beta}_2 \text{sex} + \varepsilon_i$ and $\varepsilon_i \sim N(0, s)$, then use observed age and sex as predictors to impute missing income data, plus random draw from $N(0, s)$.
  - Do this p times so that you create p imputed ("complete") datasets. Analyze results in each of p datasets. Aggregate or pool results across datasets by reporting mean, variance, and confidence interval.

- Pros: Better than singly-stochastic regression imputation, allows for much better estimation of true variance.
- Cons: Takes a bit of effort to implement, invalid method.

Proper Multiply Stochastic Regression Imputation

- For any "NA" value in a given column, proper regression imputation replaces "NA" with a predicted value based on a regression line and random error.
  - i.e. Estimate $\text{income}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{age} + \hat{\beta}_2 \text{sex} + \varepsilon_i; \hat{\beta}_{j,i} \sim N\left(\hat{\beta}_j, S\hat{E}\left(\hat{\beta}_j\right)\right)$ and $\varepsilon_i \sim N(0, \hat{\sigma})$, then impute missing income data using random draws from $N\left(\hat{\beta}_j, S\hat{E}\left(\hat{\beta}_j\right)\right)$ and $N(0, s)$.
  - Do this p times so that you create p imputed ("complete") datasets. Analyze results in each of p datasets. Aggregate or pool results across datasets by reporting mean, variance, and confidence interval.

- Pros: Very good version, valid method.
- Cons: Takes more effort to implement.

Hot-Deck Imputation

- Divide sample units into classes (i.e. based on age and sex). For any "NA" value in a given class, randomly select the value of one of the observed values in that class and impute that value for the missing value.
  - i.e. Among 18-34 year old women, there are 20 observed values and 3 missing values. For each missing value, pick one observed value at random and fill in the missing value with that observed value. You will select three observed values with replacement.
- Pros: You're using existing data.
- Cons: If columns are imputed separately, multivariate relationships are not preserved. Invalid method.

Techniques for Addressing Item Nonresponse

- Imputation
  - Deductive Imputation (valid)
  - Mean/Median/Mode Imputation (invalid)
  - Regression Imputation (invalid)
  - Stochastic Regression Imputation (invalid)
  - Multiple Stochastic Regression Imputation (invalid)
  - Proper Imputation (valid)
  - Hot-Deck Imputation (invalid)

Two comments about imputation!

- Assuming that you're using a valid method of imputation, you are not making up data.
  - You are conducting analyses with proper estimation of variance, which allows us to express the true amount of uncertainty we have in our results.

Two comments about imputation!

- If you're simply imputing data in order to have a "complete" data set for further analysis (i.e. not doing multiple imputations, then multiple analyses, then pooling results), be careful.
  - After constructing this data set, nobody will know the difference between observed data and imputed data.

Types of Missingness

- Scenario 1: I administer a survey that includes a question about someone's income. Those with low incomes are significantly less likely to respond to that question.

Types of Missingness

- Scenario 1: I administer a survey that includes a question about someone's income. Those with low incomes are significantly less likely to respond to that question.

- This type of missingness is called not missing at random.

Types of Missingness

- Scenario 2: I administer a survey that includes a question about someone's income. Those who are female are less likely to respond to the question about income.

Types of Missingness

- Scenario 2: I administer a survey that includes a question about someone's income. Those who are female are less likely to respond to the question about income.

- This type of missingness is called missing at random.

Types of Missingness

- Scenario 3: I am a very sleepy graduate student who accidentally knocks coffee on some of the written surveys we've collected, so we lose all of the data that we otherwise would have had.

Types of Missingness

- Scenario 3: I am a very sleepy graduate student who accidentally knocks coffee on some of the written surveys we've collected, so we lose all of the data that we otherwise would have had.

- This type of missingness is called missing completely at random.

Types of Missingness

- Not Missing at Random (NMAR, pronounced “N-marr”)
- Missing at Random (MAR, pronounced “marr” or “M-A-R”)
- Missing Completely at Random (MCAR, pronounced “M-car”)

Not Missing at Random (MNAR)

- The data of interest is systematically different for the respondents and nonrespondents.
  - Whether or not an observation is missing depends on the value of the unobserved data itself!

- MNAR is the most difficult type of missingness to address.

Missing at Random (MAR)

- Conditional on data we have observed, the data of interest is not systematically different between respondents and nonrespondents.

Missing Completely at Random (MCAR)

- The data of interest is not systematically different between respondents and nonrespondents.

- MCAR is not usually the case, but if MCAR is a reasonable assumption, then there are a lot of convenient methods for handling missing data.

Which missingness do I have?

- 1. Little's Test for MCAR
  - Hypothesis test available in software packages. $H_0: MCAR$ vs. $H_A: not\ MCAR$
  - (No empirical test possible to establish NMAR!)

- 2. Partition data into "observed" and "unobserved" results and compare two datasets. (Are certain summaries significantly different?)

- 3. Think about missing data process - can you come up with reasonable answer based on how missing data came about?

Methods for MCAR

- We can use any of the methods we previously discussed with their respective caveats.
  - Recommendations:
    - Deductive Imputation
    - Proper Imputation
    - Multiply Stochastic Regression Imputation
    - Stochastic Regression Imputation
    - Hot-Deck Imputation
    - Complete-Case Analysis
      - Will be unbiased, but will underestimate variance.

Methods for MAR

- We cannot use complete-case analysis.
- We can use any of these methods we previously discussed with their respective caveats.
  - Recommendations:
    - Deductive Imputation
    - Proper Imputation
    - Multiply Stochastic Regression Imputation
    - Stochastic Regression Imputation
    - Hot-Deck Imputation
  - This assumes we include the MAR variables in our modeling of missing values.

Methods for NMAR

- We cannot *(should not)* use these methods:
  - Complete-Case Analysis
  - Proper Imputation
  - Multiply Stochastic Regression Imputation
  - Stochastic Regression Imputation
  - Hot-Deck Imputation

- We can use any of these methods we previously discussed with their respective caveats.
  - At this point, I only recommend deductive imputation (from a stats perspective).

What is my workflow?

1. I evaluate how much missing data I have during EDA. Is it worth my time to try to address it?

2. For each variable, can I estimate what type of missingness I have?

3. What is the best method of imputation I can use given my constraints? (time, money)

Thank You!

- Matt Brems
  - Global Lead Data Science Instructor, General Assembly
  - https://github.com/matthewbrems/
  - Twitter: @matthewbrems
  - LinkedIn: https://www.linkedin.com/in/matthewbrems/
  - Site: www.argmatt.com