Missing Data Analysis

Matthew Brems

Goals

- Describe techniques for handling different types of missingness.
- Understand the differences between the different types of missingness.
  - MCAR, MAR, MNAR
- Assess the pros and cons of these techniques given the situation.

Introduction

- In almost any project that involves data, there will be some amount of "nonresponse" or "missingness."
  - Political Robo-Calls
  - Customer Satisfaction Surveys
  - U.S. Census
  - Data Corruption/Loss

- Because missingness is so prevalent, it is important to understand how to recognize when missingness is an issue and how to account for it.

Introduction

- Step 1: Prevent missing data.
  - Incentivize completing surveys/experiments.
  - Do work electronically so handwriting isn't an issue.
  - Try contacting subject multiple times.

Introduction

- Step 1: Prevent missing data.
  - Incentivize completing surveys/experiments.
  - Do work electronically so handwriting isn't an issue.
  - Try contacting subject multiple times.

- Step 2: Attempt to ignore missing data.
  - If you have 95% or more of your data, it *might* make sense to ignore it.
  - This treats observed sample as representative of entire sample.
  - This is usually the software default.

Introduction

- Step 1: Prevent missing data.
  - Incentivize completing surveys/experiments.
  - Do work electronically so handwriting isn't an issue.
  - Try contacting subject multiple times.

- Step 2: Attempt to ignore missing data.
  - If you have 95% or more of your data, it *might* make sense to ignore it.
  - This treats observed sample as representative of entire sample.
  - This is usually the software default.

- Step 3: Adjust for it.

Design Factors Affecting Missingness

- Survey Content
- Time of Survey
- Interviewers
- Data-Collection Method
- Questionnaire Design
- Respondent Burden
- Survey Introduction
- Incentives/Disincentives
- Follow-Up

Unit Nonresponse vs. Item Nonresponse

- Nonresponse generally will come in one of two forms: unit nonresponse or item nonresponse.

- Political Robocalls – I sample 1,000 voters to call:
  - 750 people never pick up the phone.
  - 75 answer Q1, then hang up.
  - 75 answer Q2, then hang up.
  - 100 complete the 3 question survey.

Unit Nonresponse vs. Item Nonresponse

- Political Robocalls:
  - "We saw a response rate of 25%."
  - "We saw a response rate of 18%."
  - "We saw a response rate of 10%."

- AAPOR (American Association for Public Opinion Research) has recommendations on how to report nonresponse.
  - This will ultimately vary on your use-case and your audience.

Addressing Unit vs. Item Nonresponse

- Unit Nonresponse:
  - *Almost* nothing can be done for that person, but can still get aggregate information.

- Item Nonresponse:
  - Can we ignore it?
  - If we can't ignore it, can we otherwise account for it?

Techniques for Addressing Unit Nonresponse

- Ignore It
  - Reduces precision of estimates or power of results.
    - Why does increasing the sample size not necessarily address this problem?
  - Known as "complete-case analysis" or "available-case analysis."

- Weight Class Adjustments
  - Reweight your observations so that your observed data reflects the population of interest.
  - I believe those who vote in 2016 will be 50% male and 50% female. However, 75% of my responses came from males and 25% came from females.
    - $w_{male} = \frac{proportion\ of\ responses}{true\ proportion} = \frac{0.25}{0.50} = \frac{1}{2}$
    - $w_{female} = \frac{proportion\ of\ responses}{true\ proportion} = \frac{0.75}{0.50} = \frac{3}{2}$

What this looks like…

- Ignore It
  - Find the proportion of people who will vote for Clinton.
    - $\hat{p} = \frac{\sum_i I(Clinton\ vote)_i}{N_{responses}}$
    - Recall that $I(\cdot) = 1$ if $\cdot$ is true and $I(\cdot) = 0$ if $\cdot$ is false.

- Weight Class Adjustments
    - $\hat{p} = \frac{\sum_i w_i \cdot I(Clinton\ vote)_i}{\sum_i w_i}$

Complete-Case Analysis vs. Available-Case Analysis

- 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.

Techniques for Addressing Item Nonresponse

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

Deductive Imputation

- Uses logical relations to fill in missing values.
  - Respondent mentions he was 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 he was not the victim of a crime, we know that the respondent was not the victim of a violent crime.
  - If a woman has 2 children in year 1, NA children in year 2, and 2 children in year 3, we can reasonably impute that she has 2 children in year 2.

Deductive Imputation

- Uses logical relations to fill in missing values.
  - Respondent mentions he was 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 he was not the victim of a crime, we know that the respondent was not the victim of a violent crime.
  - If a woman has 2 children in year 1, NA children in year 2, and 2 children in year 3, we can reasonably impute that she has 2 children in year 2.

- Pros: Requires no “inference,” true value can be assessed, <u>valid method</u>.
- 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, <u>invalid method</u>.

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 $\text{income} = \hat{\beta}_0 + \hat{\beta}_1\text{age} + \hat{\beta}_2\text{sex}$, 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 $\text{income} = \hat{\beta}_0 + \hat{\beta}_1 \text{age} + \hat{\beta}_2 \text{sex}$, 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, <u>invalid method</u>.

Stochastic Regression Imputation

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

Stochastic Regression Imputation

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

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

Multiply Stochastic Regression Imputation

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

Proper Multiply Stochastic Regression Imputation

- For any "NA" value in a given column, regression imputation replaces "NA" with a predicted value based on a regression line and random error.
  - i.e. Estimate $\text{income}_\text{i} = \hat{\beta}_0 + \hat{\beta}_1\text{age} + \hat{\beta}_2\text{sex} + \varepsilon_i$; $\hat{\beta}_{j,i} \sim N\left(\hat{\beta}_j, S\hat{E}(\hat{\beta}_j)\right)$ and $\varepsilon_i \sim N(0, \hat{\sigma})$, then impute missing income data using random draws from $N\left(\hat{\beta}_j, S\hat{E}(\hat{\beta}_j)\right)$ and $N(0, \hat{\sigma})$.
  - 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: Best version, valid method.
- Cons: Takes the most 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.

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 <u>with replacement</u>.
- Pros: Can use "Nearest-Neighbor Hot-Deck Imputation" based on respondents who are "close" to one another.
- Cons: If columns are imputed separately, multivariate relationships are not preserved. <u>Invalid method</u>.

Cold-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 <u>from another dataset</u> 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 <u>with replacement</u>.

- Pros: *lol*
- Cons: Requires multiple datasets. Worse than hot-deck imputation. Usually, multivariate relationships are not preserved. <u>Invalid method</u>.

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)
  - Cold-Deck Imputation (invalid)

Imputation: One Final Note

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

- 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 <u>missing not at random</u>.

Types of Missingness

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

- This type of missingness is called <u>missing at random</u>.

Types of Missingness

- Scenario 3: I am working in a lab, conducting an experiment with Petri dishes. At the end of the experiment, I want to record the amount of bacteria in each dish. However, I accidentally drop one of the Petri dishes and can't record the measurement.

Types of Missingness

- Scenario 3: I am working in a lab, conducting an experiment with Petri dishes. At the end of the experiment, I want to record the amount of bacteria in each dish. However, I accidentally drop one of the Petri dishes and can't record the measurement.

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

Types of Missingness

- Missing Not 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 (NMAR)

- The data of interest is systematically different for the respondents and nonrespondents. Whether or not an observation is missing depends on the unobserved data.
  - During the 2016 election, I administer a political robo-call, asking people who they plan to vote for in the upcoming Presidential election. People who truly will vote for Donald Trump are less likely to respond to the survey.
  - This is a case where whether or not a value is missing depends on the missing value itself!
  - There is a special type of NMAR data called "censoring," which is studied through survival analysis.

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

Missing at Random (MAR)

- <u>Conditional on data we have observed</u>, the data of interest is not systematically different between respondents and nonrespondents.
  - I administer a survey covering demographic and salary information. Every person responds with their sex but some people leave their salary blank. 70% of women report their salary and 50% of men report their salary.
  - In this case, we say that salary data is MAR conditional on sex.

Missing Completely at Random (MCAR)

- The data of interest is not systematically different between respondents and nonrespondents.
  - I administer a five question survey to 20 students on paper. The papers are handed in as students leave. I enter them into a computer but knock my coffee over onto the stack, obscuring Q5 on the bottom three surveys.

- 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. $H_A: MAR$
  - (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

- If our data are MCAR, then:
  - 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

- If our data are MAR, then:
  - We cannot use these methods:
    - 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 regression.

Methods for NMAR

- If our data are NMAR, then:
  - We cannot 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.
    - Recommendations:
      - Deductive Imputation

Resources

- http://www.stat.columbia.edu/~gelman/arm/missing.pdf
- https://liberalarts.utexas.edu/prc/_files/cs/Missing-Data.pdf
- http://scikit-learn.org/stable/auto_examples/missing_values.html#sphx-glr-auto-examples-missing-values-py