Missing Data

Tim Kelly
George Montag
Rashid Ahmed

December 9 2021

Why Missing Data is an Issue?

- Missing data is a common problem
  - Survey - respondent didn't answer all questions
  - Time Series - data not available at desired intervals
  - Panel - attrition from study
- Absence of data reduces the statistical power
- Lost data can cause bias in the estimation of parameters
- Can reduce the representativeness of the sample.
- With a large dataset, you could discard the observations that are missing data. Mean imputation is a valid technique if you don't have outliers in the data set. Alternatively, you could use median or mode imputation if the dataset has outliers.

Missing data mechanisms

- Missing Completely at Random (MCAR).
- Missing at random (MAR).
- Not missing at random (NMAR)

Missing Completely at Random (MCAR)

- Suppose variable Y has some missing values. We will say that these values are MCAR if the probability of missing data on Y is unrelated to the value of Y itself or to the values of any other variable in the data set. However, it does allow for the possibility that "missingness" on Y is related to the "missingness" on some other variable X. (Briggs et al., 2003) (Allison, 2001)
- *Example:* We want to assess which are the main determinants of income (such as age). The MCAR assumption would be violated if people who did not report their income were, on average, younger than people who reported it.

Missing Completely at Random (MCAR)

- missing values not related to the specific variable which is to be obtained, or the set of observed responses
- Most benign case
- Missing data is a random subset of the observations
- No systematic difference between missing and present data values
- Power may be lost in the design, but the estimated parameters are not biased by the absence of the data

Missing at random (MAR)

- The probability of missing data on Y is unrelated to the value of Y after controlling for other variables in the analysis (say X).
- For Example: The MAR assumption would be satisfied if the probability of missing data on income depended on a person's age, but within age groups the probability of missing income was unrelated to income.

Missing at Random (MAR)

- MAR is the intermediate case in which there is information about the missing data contained in the complete observations that can be used to improve inference about the model.
- missing but the rate depends on same other variable in the data
- Consider a questionnaire where gender is known for each respondent and, for example, we find that missing data could be explained by a persons gender

Missing at Random (MAR)

- Could be a combination of data/factors that can explain the missing data
- Can break up the data by category or gender, and see if the missing rate is the same or different. If you break up your data and find that that your missing rate is the same, then you are, you are probably in a situation where your data is missing at random.
- Data are regarded to be MAR when the probability that the responses are missing depends on the set of observed responses, but is not related to the specific missing values which are expected to be obtained.
- Cannot just ignore data that is not missing at random

Not Missing at Random

- Not missing at random is a case where the gaps in the data set are not benign but are systematically related to the phenomenon being modeled. Happens in surveys when the data is self-selected or self-reported (Chapter 19 sample selection has more information).
- In short, missing value do depend on unobserved values.
- The cases of MNAR data are problematic. The only way to obtain an unbiased estimate of the parameters in such a case is to model the missing data.

Ways to address and fix missing data - Eliminating rows (Listwise Deletion)

- If not concerned with efficiency, simply delete the incomplete observation. Only issue is what possibly helpful information could be salvaged from the incomplete observations.
- This approach is known as the complete case (or available case) analysis or listwise deletion.
- Stata does this with missing values
- Eliminating data can produce a sample that is not representative of the population because the missing observation is not random. Eliminating an entire row could leave you with too few observations. For a survey, you might be getting rid of responses to other questions because the respondent didn't answer one question.
- Pairwise deletion

Use Sample Means (Mean Imputation)

- Use sample means- If the entire row of the X matrix is missing you use the sample means for each cell, this is no different than entirely eliminating the observation. If missing values are systematically related to X, the sample mean may not be a representative estimate of the true value of X. This was an issue with getting rid of the row entirely, which sounds like it still isn't fixed by using mean imputation.
- A con is that you are reducing the variablity.

Dummy Variable Approach w/Pros and Cons

- *Here, we create a dummy variable for each variable in the data set, allowing us to use the available data without deleting any values*
- **Pros:**
- -still a relatively simple method
- -keeps original data set intact; no deleting values unnecessarily
- **Cons:**
- -still assumes that data are missing completely at random; if not the case, then the estimates will be biased

Deck Imputations:

Hot deck imputation:

- Despite being used extensively in practice, the theory is not as well developed as that of other imputation methods
- involves replacing missing values of one or more variables for a non-respondent (called the recipient) with observed values from a respondent (the donor) that is similar to the non-respondent with respect to the characteristics observed in both cases.

Cold deck imputation:

- Same as hot deck except that the data is found in a previously conducted similar.

Regression imputation

- Existing variables are used to make a prediction, and then the predicted value is substituted as if an actual obtained value
- the imputation retains a great deal of data over the listwise or pairwise deletion and avoids significantly altering the standard deviation or the shape of the distribution

Regression Imputation

- Suppose we are estimating a regression model with multiple independent variables
- One of them, X, has missing values.
- We select those cases with complete information and regress X on all other independent variables
- Then, we use the estimated equation to predict X for those cases it is missing.

Regression Imputation

Stochastic Regression Imputation

- To add uncertainty back to the imputed variable values, we can add some normally distributed noise with a mean of zero and the variance equal to the standard error of the regression estimates.
- This method is called as Random Imputation or Stochastic Regression Imputation.

Regression Imputation

Limitations of single imputation techniques in general:

- They lead to an underestimation of standard errors and, thus, overestimation of test statistics.
- The main reason is that the imputed values are completely determined by a model applied to the observed data, in other words, they contain no error.
- Single value for the missing data point. Might be biased, not representative of what that value truly is.

Multiple Imputation

- Instead of substituting a single value for each missing data, the missing values are replaced with a set of plausible values which contain the natural variability and uncertainty of the right values.
- The idea was first proposed by Rubin (1987).
- The imputed values are drawn from a distribution, so they inherently contain some variation.
- Solves the limitations of single imputation by introducing an additional form of error based on the variation in the parameter estimates across the imputation, which is called "between imputation error".
- Is a simulation-based procedure. Its purpose is not to re-create the individual missing values as close as possible to the true ones, but to handle missing data to achieve valid statistical inference

Multiple Imputation Procedures

- Impute the missing values by using an appropriate model which incorporates random variation.
- Repeat the first step several times.
- Perform the desired analysis on each data set by using standard, complete data methods.
- Average the values of the parameter estimates across the missing value samples to obtain a single point estimate.
- Calculate the standard errors by averaging the squared standard errors of the missing value estimates.

After this, the researcher must calculate the variance of the missing value parameter across the samples. Finally, the researcher must combine the two quantities in multiple imputation for missing data to calculate the standard errors.

Multiple Imputations

Advantages:

- It has the same optimal properties as ML, and it removes some of its limitations.
- Multiple Imputation can be used with any kind of data and model with conventional software.
- When the data is MAR, multiple imputations can lead to consistent, asymptotically efficient, and asymptotically normal estimates.

Multiple Imputations

Limitations:

- It is a bit challenging to successfully use it.
- It produces different estimates every time you use it, which can lead to situations where different researchers get different numbers from the same data using the same method.

Maximum Likelihood

- This method uses all of the available data in the data set to find the variance-covariance matrix, then use this matrix to estimate a regression model for the data
- Two main types of ML estimation:
  - Direct Maximum Likelihood: deals with directly maximizing a normal likelihood function for the linear model
    - Pro: gives the right standard errors and efficient estimates
    - Con: computationally challenging
  - Expectation-Maximization Algorithm: requires large sample size and data must be missing at random. It first shows estimates for the mean and the covariance matrix. These are then used to estimate parameters of the model.
    - Only used for linear or log-linear models

Conclusion

- The best solution to the missing data is to maximize the data collection when the study protocol is designed and the data collected.
- Researchers should seek to understand the reasons for the missing data.
- Distinguishing what should and should not be imputed is usually not possible using a single code for every type of missing value

Conclusion

"The only really good solution to the missing data problem is not to have any. So in the design and execution of research projects, it is essential to put great effort into minimizing the occurrence of missing data. Statistical adjustments can never make up for sloppy research" (Paul D. Allison, 2001)