Handling Missing Data in Survey

Cheng Peng

Contents

1 Introduction

Missing data are questions without answers or variables without observations. They are a common and significant challenge in the survey. They often influence the selection of a statistical method of analysis, and, depending on their severity, can undermine the confidence of analysis, or even result in wrong conclusions.

2 Sources of Missing Data

Missing data occur in survey research due to various reasons.

- **Total Non-response**: a sampled subject does not participate in the survey. Total non-response results from refusals to participate in the survey, non-contacts (not-at-homes), and other reasons such as a language barrier, deafness, or being too ill to participate.
- **Non-coverage**: a subject in the target population is not included in the survey's sampling frame. It occurs when some subjects in the population of inference for the survey are not included in the survey's sampling frame. These missing subjects have no chance of selection for the sample and hence go unrepresented.
- **Item Non-response**: a responding sampled subject fails to provide acceptable responses to one or more of the survey items. Item non-response may arise because
  - a respondent refuses to answer an item because it is too sensitive, does not know the answer to the item, gives an answer that is inconsistent with answers to other items and hence is deleted in editing;

  - the interviewer fails to ask the question or record the answer.

- **Partial Non-response**: Partial non-response falls between total and item non-response. Whereas total non-response relates to a failure to obtain any responses from a sampled subject and item non-response usually implies the failure to obtain responses for only a small number of survey items. Partial non-response involves a substantial number of item non-responses. It can occur, for instance,
  - when a respondent cuts off the interview in the middle,
  - when a respondent in a panel survey fails to provide data for one or more of the waves of the panel, or
  - when a respondent in a multi-phase survey provides data for some but not all phases of data collection.

3 Patterns and Mechanisms of Missing Data

Before choosing appropriate methods to handle missing data, we need to know the patterns and mechanisms of missing data.

3.1 Missing Data Patterns

Missing data can be grouped according to the missing data pattern, which describes which values are observed and which values are missing in the data matrix. In general, missing data patterns can be roughly classified into a variety of groups, such as univariate, multivariate, monotone, non-monotone, and file matching (Little and Rubin, 2002).

A univariate missing pattern indicates a situation where missing data occur only in a single variable. As an extension of the univariate case, the multivariate missing pattern refers to missing data in a set of variables, either for the entire unit or for particular items in a questionnaire.

- If a variable is missing for a particular subject not only at a specific time point but also at all subsequent time occasions, the missing data pattern for this individual is said to be a monotone missing pattern.
- If a case is missing at a given time point and then returns at a later follow-up investigation, then the missing data pattern for this subject is referred to as the non-monotone missing data.

In longitudinal data analysis, the non-monotone missing pattern can cause more problems than the monotone pattern, and thus deserves close attention. There are more missing data patterns such as the latent-factor patterns with variables that are never observed. For each missing data pattern, there are corresponding statistical techniques to handle its impact on the quality of data analysis.

3.2 Types of Missing Data Mechanisms

Missing data mechanisms concern the relationship between missing data and the values of variables in the data matrix. Given this focus, missing data mechanisms can be categorized into three classes: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).

- **Missing Completely at Random (MCAR)**: If missing data are unrelated to both the missing responses and the set of observed responses, the observed values are representative of the entire sample without missing values. This missing data mechanism is referred to as MCAR. This is the best we can hope for since we can simply ignore the records with missing components (list-wise deletion) without introducing bias to the data.
- **Missing at Random (MAR)**: If missing data depend on the set of observed responses but are unrelated to the missing values, the missing data are said to be MAR. For example, a registry examining depression may encounter data that are MAR if male participants are less likely to complete a survey about depression severity than female participants. That is, if the probability of completion of the survey is related to their sex (which is fully observed) but not the severity of their depression, then the data may be regarded as MAR. Since there is a systematic relationship between the propensity of

missing values and the observed data, but not the missing data. What it means, is that the missing data can be predicted by other variables in the data set.

- **Missing Not at Random (MNAR)**: There is a pattern in the missing data that affect the primary dependent variables. We can extend the above "depression" example if participants with severe depression are more likely to refuse to complete the survey about depression severity. Since the sources of missing data are themselves unmeasured means that (in general) this issue cannot be addressed in analysis, and the estimate of effect will likely be biased. Missing not at random is the worst-case scenario. We should use MNAR data with caution since these missing data cannot be "recovered".

4 Methods of Handling Missing Data

Methods of handling missing data are dependent on the missing data mechanisms. We have briefly discussed three major types of missing mechanisms in the previous sections. Here is a summary of these mechanisms.

- In **MCAR**, the missing mechanism is independent of characteristics of either the observed data or the unobserved values in the data set.
- In **MAR**, the missing mechanism is entirely explained by the observed data, that is, after observed values are accounted for, missingness is randomly distributed.
- In **MNAR**, missing observations are dependent upon unobserved values; missingness cannot be accounted for by controlling for observed data.

In the subsequent sub-sections, we will outline the major methods used to handle missing data with different types of missing mechanisms.

4.1 Listwise Deletion

Delete all data from any participant with missing values. If the sample is large enough, we likely can drop data without substantial loss of statistical power. Be sure that the values are missing at random and that you are not inadvertently removing a class of participants.

This method is used for MCAR and, occasionally, MAR. It is the easiest and simplest method among all other available methods.

The disadvantages are

- loss of valuable information;
- potential contribution to bias;
- loss of statistical power;

4.2 Mean Replacement

For variable "X" with missing values, take the mean of all included observations. Substitute the mean of "X" for missing values of "X". This is a type of simple imputation method.

It is valid for MCAR. The advantages are

- preserving the mean of the data set;
- simple;
- allowing use of all observations

The disadvantages are

- artificially reducing the standard deviation of the data set,
- distorting relationships between variables;
- yielding potentially biased estimates;
- producing results that are highly statistically significant, but inaccurate;

4.3 Regression Replacement

This is another method of single imputation. It estimates the distribution of the missing variable(s), given covariates, takes a random draw from this distribution for each value then performs analysis as usual.

The regression replacement is valid for both MCAR and MAR. The advantages are * avoidance of bias in estimating; * simpler than multiple imputations;

The disadvantages are

- misrepresenting uncertainty of estimates;
- more complicated than list-wise deletion or mean replacement
- reducing confidence intervals of estimates although theoretically unbiased.

4.4 Multiple Imputation

Multiple Imputation is the most sophisticated and, currently, most popular approach to take the regression idea further and take advantage of correlations between responses.

It estimates the distribution (Bayesian posterior distribution) of the missing variable, given covariates; takes random draws from this distribution to produce multiple versions (usually 3–10) of an imputed data set; performs analysis on each imputed data set and pools the results. It is a simulation-based Bayesian method. Most statistical computer programs can do multiple imputations.

It is valid for MCAR or MAR. The advantages are

- accounting for the extra uncertainty produced by imputing data;
- producing better estimates of missing values.

The disadvantages are

- requiring complicated statistical methods or complicated software;
- harder to understand;
- taking extra steps;
- because the method accounts for extra uncertainty, results can be interpreted as if data were not missing.

5 R Packages for Imputation

Several R packages have functions to perform imputation for missing values. This section lists a few commonly used R libraries. We only explain the concepts with no coding.

5.1 MICE Package

MICE (Multivariate Imputation via Chained Equations) is one of the commonly used packages by R users. Creating multiple imputations as compared to a single imputation (such as the mean) takes care of uncertainty in missing values.

MICE assumes that the missing data are Missing at Random (MAR), which means that the probability that a value is missing depends only on the observed value and can be predicted using them. It imputes data on a variable-by-variable basis by specifying an imputation model per variable.

As an example, we consider a data set with variables $X_1, X_2, \cdots, X_k$. If X_1 has missing values, then it will be regressed on other variables X_2 to X_k. The missing values in X_1 will be then replaced by predictive values obtained. Similarly, if X_2 has missing values, then X_1, X_3 to X_k variables will be used in the prediction model as independent variables. Later, missing values will be replaced with predicted values.

By default, linear regression is used to predict continuous missing values. Logistic regression is used for categorical missing values. Once this cycle is complete, multiple data sets are generated. These data sets

differ only in imputed missing values. Generally, it's considered to be a good practice to build models on these data sets separately and combine their results.

Precisely, the methods used by this package can be found in the document of the package at https://cran.r-project.org/web/packages/mice/mice.pdf.

5.2 Hmisc Package

Hmisc is a multiple-purpose package useful for data analysis, high–level graphics, imputing missing values, advanced table making, model fitting & diagnostics (linear regression, logistic regression & cox regression), etc. Amidst, the wide range of functions contained in this package, it offers 2 powerful functions for imputing missing values. These are impute() and aregImpute(). Though it also has transcan() function, aregImpute() is better to use.

impute() function simply imputes missing values using a user-defined statistical method (mean, max, mean). Its default is median. On the other hand, aregImpute() allows the mean imputation using additive regression, bootstrapping, and predictive mean matching.

In bootstrapping, different bootstrap resamples are used for each of the multiple imputations. Then, a flexible additive model (non-parametric regression method) is fitted on samples taken with replacements from original data and missing values (acts as dependent variable) are predicted using non-missing values (independent variable).

Then, it uses predictive mean matching (default) to impute missing values. Predictive mean matching works well for continuous and categorical (binary & multi-level) without the need for computing residuals and maximum likelihood fit.

More functions and updates in Hmisc package can be found at https://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf