Introduction to Missing Data and Imputation (Session 1)

Dale Barnhart, Robert Brenan, Simo Goshev

4 Oct 2018 , v0.01

The following is taken from Simo's tutorial on missing data

Missing data

What is missing data?

Generally we want data to be complete – that is we wish that all individuals we have sampled agree to be interviewed and subsequently provide valid responses to all applicable questions in the questionnaire.

However, more often than not data are incomplete. For example, sampled individuals may fall out of reach (i.e. move or simply ignore us), or decline to be interviewed. If we retained this type of respondents in the sample, our dataset would have observations that are subject to *unit non-response*. That is, we would not have any response data at all for these subjects.

Subjects can also decline to answer certain questions or simply skip questions. This is known as *item non-response* and it manifests as a "gappy" pattern in the dataset.

Unit and item non-response are the patterns of missingness commonly present in data used in social science research, and missing data, as you can tell already, is simply the full or partial absence of valid values or measurements for one or more subjects for one or more variables.

Why missing data may be a problem?

Missing data mechanisms

As we will see shortly, there are three distinct missing-data mechanisms talked about in the literature. To understand better these mechanisms and how they defer from each other, we first need to define two terms: *observed* and *unobserved* data.

Observed data are the values or measurements that a researcher collects. In our example dataset, these would be the recorded values (except for the missing values) of the eight variables.

Unobserved data are data that a researcher fails to measure. These data are represented by the missing values in our variables.

Missing completely at random (MCAR)

When we say that data are missing completely at random (MCAR), we mean that the missing-data mechanism is independent of both the observed and unobserved data. We can think of MCAR as a mechanism that introduces missigness to complete data by changing, on an entirely independent and random basis, the values of some variables to missing.

In theory, we would not expect incomplete records (or observations that contain missing values for some variables) to differ systematically from complete records.

When working with social surveys, very rarely can researchers claim/assume that missingness is MCAR. The reason is that usually there are underlying causes for subjects not to respond, causes which may or may not have been observed by the researcher (as part of the study).

Missing at random (MAR)

When we say that data are missing at random (MAR), we mean that the missing-data mechanism is dependent on the observed data but is independent of the unobserved data. We can think of this mechanism as one that introduces missigness to complete data by changing the values of some variables to missing based on the observed values of other variables belonging to the same record (subject).

In theory, we expect to see difference between complete and MAR data, specifically in the variables affected by missingness.

Since under MAR the mechanism of the missing data is dependent on the observed data only, we are able to both test (see Little 1988) for this missingness mechanism and address the missingness methodologically.

When working with social surveys, MAR is the type of mechanism that is commonly assumed. There are standard methods for handling it, some of which we will discuss shortly.

Missing not at random (MNAR)

When we say that data are missing not at random (MNAR), we mean that the missing-data mechanism is dependent on the unobserved data. We can think of this mechanism as one that introduces missigness to complete data by changing the values of some variables to missing based on either the unobserved values of

the same variable or the values of other, unobserved variables belonging to the same record.

With MNAR, we expect incomplete and complete cases to differ systematically. There are several classes of methods that can handle MNAR data of which sample selection models (with the Heckman selection model being the flagship model) is the most popular class in social science.

Why is this distinction important?

We conduct analyses to estimate quantities of interest and we want our estimates to be unbiased (or consistent) and efficient. Missingness can impact both of these characteristics in different ways depending on the underlying missing-data mechanism. The following table offers a summary of the effect of missingness on the characteristics of estimates (assuming again we are using complete cases only):

Missingness	Bias	Efficiency	Can be corrected?
MCAR	No	Lesser	Yes
MAR	? Yes	? Lesser	Yes
MNAR	? Yes	? Lesser	? Yes

Strategies for Handling (Anticipated) Missingness

We have seen so far that missingness may impact the estimates of quantities of interest. Data collectors and applied researchers utilize a variety of strategies and methods to minimize or mitigate the impact of missingness.

Prior to data collection

As the old saying goes, "Design trumps model". And it could not be more relevant to addressing missingness. To minimize missingness, the data collector has to maximize effort in designing a solid data collection plan (containing various provisions and fallbacks) and executing it with a high level of fidelity.

Oversampling

At the stage of survey sampling design, it is almost always advisable to provision for oversampling. Oversampling means that we will sample more subjects that we think we would need. How many more subject you may ask? Well, some survey agencies sample 5% more, others sample 10% more. It all depends on the amount of missingness or non-response one might expect (and also of course on cost).

If in doubt, it is always recommended to ask a survey design specialist and/or someone very familiar with the population that will give rise to the sample.

Pilot studies

Researchers should always run pilot studies. Pilots can help in several imporant ways:

- Validate the survey instrument (questionnaire that is)
- Validate the rules and procedures of data collection, management and storage
- Gain preliminary inderstanding of expected unit and item non-response
- Test strategies for minimizing non-response (i.e. incentivisation)

Rigor during data collection

Once a respondent agrees to participate, we want to help them provide a complete response. This may mean working around respondent's constraints or employing some form of incentivisation. We always want to consult with the Institutional Review Board's (IRB) about acceptable forms of incentivising response.

After data collection

Despite our best effort during the planning and collection stages, our data may still be subject to missingness. What can we do ex-post?

Weighting

Weighting, or adjusting the relative contribution of respondents' data to estimated quantities of interest, is a popular way of correcting for unit and item non-response.

One of the most popular weighting techniques involves calculating probability of reponse weights for each subject in the sample and then using these weights in subsequent analyses. This technique is known as *inverse probability weighting* (IPW). In the context of a MAR dataset where subjects in the middle age-range skipped or declined questions on income, exercise and education (the process we had previously), IPW could be implemented in the following manner: we would predict the probability of a complete response for every subject as a function of age, and then use the inverse of the predicted probabilities as a probability weight in our complete-case regression of happiness on physical exercise.

What does IPW really do for us?

Suppose the likelihood of responding of a person of a specific age is 1. This means that everyone of this age would respond to the income, exercise and education questions and therefore everyone of this age for the purpose of our analysis would be counted once. Now, suppose the likelihood of responding to three questions for a person of a different age is 0.5. This means that we would

expect only half of the people of this age to respond and therefore everyone who actually responds will be given a weight of 2, or will be counted twice: once for themselves and once for a respondent of the same age who did not respond. With a likelihood of 0.1, everyone of the same age who responded will represent themselves and 9 other non-responders of the same age.

Two-phase estimation for missing data

This is a method that assumes the following process governing missingness:

Phase 1: Original sample selected and variables without missing values are measured
Phase 2: A subset of respondents is selected and the remaining variable are observed

Obvious limitations of this method? It is only useful when we can divide the observations into complete and incomplete

Imputation

Imputation methods are an alternative. Imputation generally refers to the process of filling in the "gaps" or missingness in our dataset with valid values in a principled manner. Imputation methods generally fall into two categories: methods for *single imputation* and methods for *multiple imputation*.

Single imputation

Methods that fall under the umbrella of *single imputation* include hot- and cold-deck imputation, mean substitution and regression.

Hot-deck imputation

The method involves replacing missing values with values from a random *similar* observation in the same dataset.

Cold-desk imputation

The method involves replacing missing values with values from a random *similar* observation from a different dataset.

Mean substitution

The method involves replacing missing values of a variable with the *mean* value of the variable computed over non-missing observations.

Regression

The method involves estimating a regression equation from all complete cases and then replacing the missing values with the respective predicted values from the regression.

There are many other single imputation techniques but ultimately all of them are subject to one common limitation. Could you name it?

Multiple imputation (MI)

Multiple imputation, a simulation-based technique proposed by Donald Rubin, overcomes the problems of variability that single imputation methods fail to address. (Yes, single imputation methods do not account for the uncertainty in the imputed values; that is their common limitation.) In a series of articles in the 1970's and 1980's, culminating with his seminal book "Multiple Imputation for Nonresponse in Surveys" published in 1987, Rubin developed the theoretical basis of MI and proposed the following general algorithm for its implementation:

1. Impute the missing values M times, thus generating M complete datasets
2. Estimate the quantities of interest from every m dataset in M
3. Combine the estimates from every dataset in M into a final set of estimates that can be used for inference.

The following formulas for aggregation used in Step 3 are known as the *Rubin's rules*:

$$\widehat{\beta} = \frac{1}{m} \sum_{m=1}^{M} \widehat{\beta}_m$$

$$V_\beta = W + (1 + \frac{1}{m})B$$

where

$$W = \frac{1}{m} \sum_{m=1}^{M} s_m^2$$

$$B = \frac{1}{m-1} \sum_{m=1}^{M} (\widehat{\beta}_m - \widehat{\beta})^2$$

Nowadays, multiple imputation is the recommended technique for handling missing data.

Full-likelihood methods (FLM)

FLM are based on the joint distribution of the variables affected by missingness together with the outcome of the regression model we wish to estimate. We specify a likelihood function which we then estimate to retrieve the most likely regression paratemers. A main limitation of this method is that all variables need to be continuous and the regression model has to be linear.

Doubly robust methods (DRM)

While with IPW weighting methods we model the probability of complete cases and mulitple imputation models the distribution of missing variables, DRM

combine the two approaches. Generally, DRM use two models: one predicts the missing values and the other predicts the missing probabilities (which are used as weights). There are both parametric and semi-parametric implemetations of DRM but they have remained somewhat on the sidelines in social science research.

Multiple Imputation

Multiple imputation methods fall under two broad categories: *univariate* MI methods and *multivariate* MI methods. We use univariate methods for multiple imputation if missingness is MAR and we wish to impute the missing values of a single variable. In comparison, multivariate mulitple imputation (MMI) methods are helpful if (missingness is MAR and) we wish to impute multiple variables with missing values simultaneously and in relation to one other.

This ability to allow for relationships among the variables (missing or non-missing) in a dataset is what makes the MMI methods so popular in social science research. We will focus on three distinct flavors of MMI in order of specification flexibility they provide: imputation using multivariate normal regression, imputation in monotone data, and chained imputation.

Imputation using multivariate normal regression

This method of imputation can be used when imputing one or more continuous variables. It uses multivariate normal regression to model the mean function and Markov Chain Monte Carlo to impute the missing values.

Oh wait, what is a Markov chain?

A Markov chain is a mathematical model of a stochastic system. It is defined by a set of states and a set of transition probabilities for traversing among these states. Intuitively, we can think of Markov chains as random walks on graphs. *This* website has several interactive examples of Markov chains that can help us understand them better.

And what is a Markov Chain Monte Carlo (MCMC)?

MCMC is a class of algorithms that allow us to sample from complex probability distributions. The Monte Carlo part of MCMC refers to the fact that we are simulating a distribution and the Markov Chain part refers to the fact that we are using a Markov chain to do the sampling.

And how does MCMC impute the missing values?

Under specific conditions, a Markov chain will have a *stationary distribution*. This means that for a sufficiently long random walk, the set of transition probabilities will converge to some fixed quantities and will not depend on the starting state of the walk.

Now, if we set the stationary distribution of the Markov chain to the distribution that we wish to sample from (i.e. the *target distribution*), for a sufficiently long random walk, the Markov chain will reproduce empirically this distribution and will help us sample from it. In imputation problems (based on regression), we normally pick a target distribution and a mean function and let the developers of statistical software choose the specific MCMC algorithm that would be most efficient for sampling from this distribution.

The imputation algorithm has two steps: an *imputation* step and a *posterior* step. In the imputation step, the missing values are replaced by draws from our target distribution – a multivariate normal distribution, conditional on the observed data and current values of the model parameters. In the posterior step, new values of the model parameters and variance are drawn from their respective distributions, conditional on the observed data and the imputed values from the previous imputation step. This procedure is repeated until a pre-specified number of iterations is reached.

An obvious disadvantage of this method is that it only works for imputing continuous variables (that need to follow a normal distribution). This is a pretty significant limitation considering the variety of variable types that are present in social science data.

Imputation in monotone data

To understand this method, first we need to define the term "monotone missing pattern". Data exhibit a monotone missing pattern if missingness in variable Y_j for subject i means that variables Y_k, where $k > j$ for subject i are also missing.

The presence of a monotone missing pattern in data simplifies the imputation task by reducing it to a sequence of independent univariate (possibly conditional) imputation tasks.

The imputation algorithm has the following general logic:

$$\begin{aligned} X_1^* &\leftarrow Z \\ X_2^* &\leftarrow X_1^*, Z \\ X_3^* &\leftarrow X_1^*, X_2^*, Z \end{aligned} \tag{1}$$

The advantages of this method are that not only can we simplify the imputation task but we can also specify univariate models that correspond to the types of variables we are imputing. This is a rather big improvement over the method of multivariate normal regression. The disadvantage is that it is quite uncommon for data to have a monotone missing pattern.

Multiple imputation using chained equations

Multiple imputation using chained equations (MICE) is the most popular method for imputation in social science research. Its popularity is based primarily on its ability to impute multiple variables of different types simultaneously and conditional on one another.

One iteration of the imputation algorithm has the following general logic:

$$\begin{aligned} X_1 &\leftarrow \boldsymbol{X_{-1}}, Z \\ X_2 &\leftarrow \boldsymbol{X_{-2}}, Z \\ X_3 &\leftarrow \boldsymbol{X_{-3}}, Z \end{aligned} \tag{2}$$

In essence, MICE is based on a series of univariate imputation models, each of which is selected to correspond to the type of variable (e.g. continuous, categorical, proportion, etc.) that is being imputed. It uses chained equations, meaning that an outcome in one imputation equation by default (and this can be customized) is a predictor in all other equations. MICE is similar in logic to MCMC in that it builds a chain, iterates until the chain attains its stationary distribution and samples from it to replace the missing observations.

Some 'ground rules' of imputation

Multiple imputation is an extremely useful technique that can help us tackle missing data problems in a principled and statistically sound manner. However, practitioners should remain vigilent and not fall into some common pitfalls associated with the procedure.

To avoid trouble:

- We need to understand our data very well. We need to know whether all conditions for imputation are met, variables it makes sense to impute and variables that we should not impute for certain groups (e.g. wages for one-year-olds).
- We should not rely on statistical software to know how to handle our data properly. Indiscriminately throwing variables into an imputation routine and expecting the software to know their meaning and make decisions for us is a recipe for disaster. Likewise, cherry picking variables to include in imputation models is nothing short of bad research practice.
- We must pay attention to attrition and selection. Imputation algorithms do not know anything about this aspect of our missing data problem and would assume it away. Borrowing information from non-comparable subjects could severely affect the quality of our imputation and inference.
- The amount of missingness and sample size matter for the performance of imputation routines! The rule of thumb is that missingness of about 10% is manageable, and missingness of 25% is high and likely problematic, especially in samples of smaller size.

Imputation in panel data

Imputing panel data is similar to imputing cross-sectional data. Before imputation however, we would reshape the data from long to wide form. This is really important as imputation routines do not normally know the ways in which observations are be related and if the data are imputed while still in long form, in essence, the software would be making the assumption that observations of the same individual for different time periods are uncorrelated; and that would be a silly thing to assume!

References

Mack C, Su Z, Westreich D. Rockville (MD): Agency for Healthcare Research and Quality (US); 2018 Feb.

Little, R. J. A. 1988. A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association 83: 1198–1202.

Cheng Li, Little's test of missing completely at random The Stata Journal (2013) 13, Number 4, pp. 795–809

Rubin, Multiple Imputation After 18+ Years,Journal of the American Statistical Association, Vol. 91, No. 434 (Jun., 1996), pp.473-489

Lumley, T., 2010, Complex Surveys. A guide to analysis in R.

van Buuren, S., Groothuis-Oudshoorn, K., 2011, mice: Multivariate Imputation by Chained Equations in R, Journal of Statistical Software

Stata, 2017, Stata Multiple-Imputation Reference Manual, Stata Crop LLC, College Station, Texas