Multiple Imputation using R

Lisa McQuarrie

What is multiple imputation

Imputation is a method for handling missing data that involves replacing a missing observation with a prediction of its true value. A simple example of imputation is replacing a missing value on a variable X by the mean of the observed values of X. Another example of imputation is using a linear regression to predict the missing value of X. The method used to generate the imputed value is called the imputation model.

Single imputation produces one imputed value for each missing value in a data set. Then, in the analysis, the data set with the imputed values is treated as though all its values were observed. In other words, the analysis is carried out on the data set with the imputed values in the same way it would have been carried out if the data set had been fully observed in the first place. The problem with single imputation is that it treats imputed observations, which are predictions and have error, as though they were observed and have no error. If a regression is carried out on a data set that underwent single imputation, then the standard error of the estimated regression coefficients will be underestimated; this then affects confidence intervals or hypothesis testing.

Multiple imputation corrects this problem by essentially conducting single imputation many times. The output of multiple imputation is m data sets containing imputed values, where m is the number of imputations and is specified by the analyst. The analysis (ex. regression) is then carried out on all of the m data sets individually and the m sets of results (ex. estimated regression coefficients) are pooled together to produce one final set of results. The differences between the estimated coefficients on the m data sets provides an estimate of the error of the imputed values. Consequently, the imputation error is incorporated into the standard error of the regression coefficient estimates and confidence intervals or hypothesis testing is valid. Multiple imputation is generally preferred over single imputation.

There are three steps in the process of multiple imputation:

1. Generate m imputed datasets
2. Analyse each dataset separately
3. Pool the results of each analysis. If the analysis is a regression, then the pooling would produce one point estimate and one standard error for each regression coefficient

When to use multiple imputation

When deciding whether multiple imputation is a suitable approach to handling the missing observations in a data set it is important to consider what caused the observations to be missing. This is called the missing data mechanism, and there are 3 types.

- Missing completely at random (MCAR): The reason the data is missing is unrelated to any observed or unobserved data. For example, if, due to lack of resources, half of the participants in a study were chosen at random to provide measurements of a variable that is expensive to record.
- Missing at random (MAR): The reason the data is missing is related to observed data but not the unobserved data. For example, if in a study of learning outcomes, males were more

likely than females to skip school on the day of a test. If gender is recorded and included in the analyses, the missing test scores are MAR.

- Missing not at random (MNAR): The reason the data is missing is related to the value of the missing observation. For example, in a study of learning outcomes, students with lower grades skip school the day of a test. Had they taken the test, the unobserved students would have scored lower, therefore, their test scores are MNAR.

One of the best ways to determine the missing data mechanism is simply to consider what data is missing and why it might be missing based on knowledge of the data and the subjects. In addition, a hypothesis test called Little's test can be used to identify if the missing data are MCAR. If the null hypothesis of Little's test is not rejected then the data can be assumed to be MCAR, otherwise the data are either MAR or MNAR. However, this test is somewhat controversial and experts disagree as to its effectiveness. Additionally, there is no test to distinguish between MAR and MNAR data, which is why considering the underlying cause of the missing observations is important.

Multiple imputation assumes that the missing data mechanism is MCAR or MAR and will provide valid results under these two mechanisms. If the missing data mechanism is MNAR then the results from multiple imputation may not be accurate.

How to conduct multiple imputation

Before conducting multiple imputation, four decisions must be made.

1. Identify that the missing data mechanism is either MCAR or MAR.
2. Determine the appropriate form of the imputation model. For each variable that will be imputed, a method must be chosen. The choice depends mostly on the type of the variable that will be imputed (continuous, binary, ordinal…). The table below provides recommendations for each variable type with the corresponding method for the R mice function from the mice package.

Table 1: Recommended methods for use in the R mice function, according to the variable type. Note that the options listed are based on recommendations from experts, and do not represent all possible methods.

Variable type	Method
Continuous, approximately normal	norm (Bayesian linear regression)
Continuous, not normal	pmm (predictive mean matching)
Binary	logreg (logistic regression)
Ordinal	polr (proportional odds model)
Nominal	polyreg (polytomous logistic regression)
Sparse categorical	pmm (predictive mean matching)

3. Choose the set of variables to be included as predictors in the imputation model. If you are imputing multiple variables, different predictors can be used for each variable.
   a. Generally, you should include as many variables as possible in the imputation model, but in small to moderate sized data sets you may have too many variables to include them all. In that case, 15 – 20 variables has been shown to be enough.
   b. All the variables that will be included in the analysis model should be included as predictors in the imputation model. This includes the response variable from the analysis model and any interaction terms or functions of variables that you plan to include in the analysis model.

    c. Also include any variables that have moderate to strong correlation with the variable to be imputed as predictors in the imputation model.
    d. Include any variables that are related to the missing data mechanism (ie. include any variables that may be related to why the observations are missing). These variables can be identified using knowledge about the study and/or calculating associations between variables and a missing observation indicator variable (a variable that takes the value 1 if the observation is missing and 0 if the observation is observed).
    e. If there are other variables with missing data that you are not imputing, then, in general, do not include them in the imputation model. For example, suppose X_1 and X_2 both have missing values on observation 10, and X_1 will be imputed but X_2 will not. In that case, using X_2 in the imputation model for X_1 will result in an ‘NA’ value for observation 10 (ie. observation 10 cannot be imputed).

4. Choose the number of imputations, m. One rule of thumb is to set m equal to the percentage of missing observations in the data set. Ex. if there are 100 total observations on variable X_1 and 20 are missing then set m = 20. Choosing a larger m will result in larger power in hypothesis tests (or smaller confidence intervals), so if higher power is required you can choose a larger m. The only added cost for using a larger m is computation time.

After these decisions are made, multiple imputation can be implemented using the ‘mice’ package in R. The key steps are

1. Use the ‘mice’ function to generate the imputed data sets.
2. Use the ‘with’ function to apply the analysis model to the m imputed data sets.
3. Use the ‘pool’ function to pool the model results across the m imputed data sets and produce the final model estimates.

References

van Buuren, Stef. Flexible Imputation of Missing Data. 2012. Chapman & Hall/CRC. Available: https://stefvanbuuren.name/fimd/

van Buuren, Stef. Mice Package documentation. 2021. Available: mice.pdf (r-project.org)