Incomplete Data Analysis

Vanda Inácio

University of Edinburgh

Semester 1, 2018/2019

Multiple imputation

MI in practice – the `mice` package

- ↪ `R` provides several useful packages for MI. The most commonly used include `Amelia`, `mice`, and `MI`.
- ↪ They basically differ in the specific algorithm they implement for the imputation model (step 1).
- ↪ We will learn how the `mice` package works for conducting MI analyses.

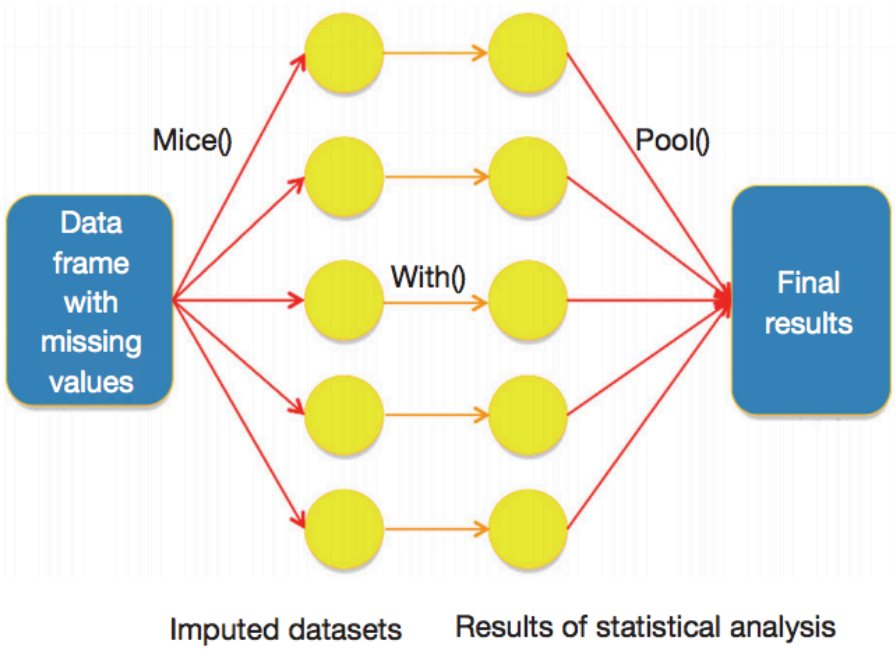
Multiple imputation

MI in practice – the `mice` package

- `mice` contains several popular methods for imputation under the normal linear model. Specifically:
  - Regression imputation, as we have seen it in lecture 3. The method is available as `norm.predict`.
  - Stochastic regression imputation. This method is available as `norm.nob`.
  - Bayesian multiple imputation that implements the method described in slide 30 of lecture 9. This method is available as `norm`.
  - The bootstrap alternative to the Bayesian multiple imputation, slide 33 of lecture 9, is also implemented and is available as `norm.boot`.

Multiple imputation

MI in practice – the `mice` package



Schematic illustration of how `mice` package works with a dataset with missing values. Note the sequential use of `mice()`,`with()`, and `pool()` functions. Figure from Annals of Translational Medicine, **4**, 2016.

Multiple imputation

MI in practice – the `mice` package

- We start with the function `mice()` which will perform the imputation step (step 1). We need to specify the method we want to use to impute the missing values and the number of imputations. By default this function creates 5 copies of the dataset, i.e., $M = 5$. We can also set a seed within this function, so results are reproducible.

- Now having the complete datasets at hand, the conventional statistical analysis that would have been conducted had the data been complete can be performed. This is done via the function `with()` and we need to feed this function with the imputed data from the previous step and with the model we want to fit (e.g., through the use of `lm` for linear regression and `glm` for generalised linear models).

- Lastly, we use the function `pool()` to combine the results of the M analyses (obtained in the previous step using `with()`).

- See the supplementary file on Learn for a concrete application.

Multiple imputation

Imputation under non-normal distributions

- ↪ The imputation methods discussed so far produce imputations draws from a normal distribution.
- ↪ In practice, the data could be skewed, heavy tailed, non negative, bimodal or rounded to name only some deviations from normality.
- ↪ This creates an obvious mismatch between observed and imputed data which could adversely affect the estimates of interest.
- ↪ The effect of non normality is general small for measures that rely on the center of the distribution, like means, but it could be substantial for estimates like a variance or a percentile (van Buuren, 2012).
- ↪ A sensible approach is to transform the data toward normality before imputation and back-transform after imputation.
- ↪ A beneficial side effect of transformation is that the relation between the missing and observed variables may become closer to a linear relation.

Multiple imputation

Imputation under non-normal distributions

- ↪ Sometimes applying a simple function to the data, like the logarithmic is all that is needed. More generally, the transformation could be made to depend on known covariates like, for instance, age and gender.

- ↪ It is also possible to directly draw imputations from non-normal distributions. For instance, Liu (1995) proposed methods for drawing imputations under the *t* distribution instead of the normal (the *t* distribution has heavier tails than the normal distribution).

- ↪ He and Raghunathan (2006) created imputations by drawing from Tukey's gh distribution, which can take many different shapes.

- ↪ Demirtas and Hedeker (2008) investigated the behaviour of methods for drawing imputations from the Beta and Weibull distributions.

Multiple imputation

Other types of variables

- When the variable to be imputed has a binary form, a sensible imputation model is a logistic regression model. In `mice` the procedure is implemented as `mice.impute.logreg`.

- If the variable to be imputed is categorical with K unordered categories then a reasonable model is a multinomial logit model, and this is implemented in `mice` as `mice.impute.polyreg`.

- A categorical variable with K ordered categories is imputed by the ordered logit model or proportional odds model. Implementation is done in `mice.impute.polr` function.

Multiple imputation

Multivariate multiple imputation

- In general, a dataset may have many variables with missing values.
- To make good imputations we should use a multivariate analysis of all variables within a single 'platform'.
- There are several methods for dealing with missingness in several variables, but all them require a quite advanced knowledge of Bayesian methods.
- The package `mice` also accommodates multivariate multiple imputations, with some pre-defined procedures.
- I will very briefly describe the main idea of how things would work in general.

Multiple imputation

Multivariate multiple imputation

- The following is from Schafer and Graham (2002).
- Consider a hypothetical data set with three variables Y_1, Y_2, and Y_3, which we assume to be jointly normally distributed.
- Suppose that one group of participants (Group A) has measurements for all three variables, another group (Group B) has measurements for Y_1 and Y_2 but missing values for Y_3, and a third group (Group C) has measurements for Y_3 but missing values for Y_1 and Y_2.
- The parameters of the trivariate normal model—three means, three variances and three correlations—are not known and should be estimated from all three groups.

Multiple imputation

Multivariate multiple imputation

- If the parameters were known, MIs could be drawn in the following way.
- Group A requires no imputation.
- For Group B, we would need to compute the linear regression of Y_3 on Y_1 and Y_2. Then, for each participant in Group B, we would use his or her own values of Y_1 and Y_2 to predict the unknown value of Y_3 and impute the predicted value $\widehat{Y_3}$ plus random noise drawn from a normal distribution with the appropriate residual variance.
- For Group C, we would compute the bivariate regression of Y_1 and Y_2 on Y_3, obtain the joint prediction $(\widehat{Y_1}, \widehat{Y_2})$ for each participant, and add random noise drawn from a bivariate normal distribution with the appropriate residual variance-covariance matrix.

Multiple imputation

Multivariate multiple imputation

- A crucial feature of MI is that the missing values for each participant are predicted from his or her own observed values, with random noise added to preserve a correct amount of variability in the imputed data.

- Another feature is that the joint relationships among the variables Y_1, Y_2, and Y_3 must be estimated from all available data in Groups A, B, and C.

- Maximum likelihood estimates of the parameters could be computed using an EM algorithm, but proper MI requires that we reflect uncertainty about these parameters from one imputation to the next.

- Therefore, instead of using maximum likelihood estimates, we need to draw random values of the parameters from a posterior distribution based on the observed data likelihood and a prior.

Multiple imputation

Multivariate multiple imputation

- The form of this posterior distribution is not easy to describe, but it can be sampled from by a variety of techniques; data augmentation (Schafer, 1997) is straightforward, but one could also use importance resampling.

- The effect of drawing parameters from a posterior distribution, rather than using maximum likelihood estimates, means that for Group B the regression of Y_3 on Y_1 and Y_2 will be randomly perturbed from one set of imputations to the next; similarly, in Group C the joint regression of (Y_1, Y_2) on Y_3 will also be perturbed.

Multiple imputation

Summary

- ↪ Multiple imputation overcomes the main problem associated with single imputation (underestimating variability) and is one of the state-of-the-art technique that researchers/methodologists currently recommend.
- ↪ As all the other techniques we have learned, MI assumes MAR data.
- ↪ A multiple imputation analysis consists of three distinct steps: the imputation phase, the analysis phase, and the pooling phase.
- ↪ The imputation phase creates multiple copies of the dataset, each of which contains different estimates of the missing values.
- ↪ The imputation phase, i.e., the specification of an appropriate imputation model, is the crux of MI, since if such model is misspecified there is the potential for bias.
- ↪ The purpose of the analysis phase is to analyse the filled-in dataset. This step applied the same statistical procedures that one would have used had the data been complete. Procedurally, the only difference is that one performs each analysis M times, once for each imputed dataset. The analysis phase yields several sets of parameter estimates and standard errors.

Multiple imputation

Summary

- ↪ Finally, the pooling phase uses Rubin's rules to combine the *M* sets of estimates and standard errors into a single set of results.
- ↪ The pooled parameter estimate is simply the arithmetic average of the estimates from the analysis phase.
- ↪ Combining standard errors is somewhat more complex because it involves two sources of sampling variation.
- ↪ The within imputation variance is the arithmetic average of the *M* sampling variances, and the between imputation variance quantifies the variability of an estimate across the *M* imputations.
- ↪ The within imputation variance estimates the sampling fluctuation that would have resulted had there been no missing data, and the between imputation variance captures the increase in sampling error due to missing data.