Missing values

If there are missing data they are assumed to be “ignorable”, i.e., MAR or MCAR.

In the case of MAR and MCAR, multiple imputation i

MCAR (Missingness completely at random): A variable is missing completely at random if the probability of missingness is the same for all units: For example, if for each subject, we decide whether to collect the diabetes status by rolling a die and refusing to answer if a “6” shows up. Another example is a weighing scale that ran out of batteries. Some of the data will be missing simply because of bad luck. Another example is when we take a random sample of a population, where each member has the same chance of being included in the sample. The (unobserved) data of members in the population that were not included in the sample are MCAR. MCAR causes enlarged standard errors due to the reduced sample size, but does not cause bias (‘systematic error’ that is overestimation of benefits and underestimation of harms)

MAR (Missingness at random): Missingness that depends only on observed predictors - is that the probability a variable is missing depends only on available information. Here, we would have to be willing to assume that the probability of nonresponse to diabetes depends only on the other, fully recorded variables in the data. It is often reasonable to model this process as a logistic regression, where the outcome variable equals 1 for observed cases and 0 for missing. When an outcome variable is missing at random, it is acceptable to exclude the missing cases (that is, to treat them as NA), as long as the regression controls for all the variables that affect the probability of missingness.

MAR allows prediction of the missing values based on the participants with complete data.

MNAR (Missingness not at random): Missingness that depends on unobserved predictors. Missingness is no longer “at random” if it depends on information that has not been recorded and this information also predicts the missing values. If a particular treatment causes discomfort, a patient is more likely to drop out of the study. This missingness is not at random (unless “discomfort” is measured and observed for all patients). If missingness is not at random, it must be explicitly modeled, or else you must accept some bias in your inferences.

Missingness that depends on the missing value itself. Finally, a particularly difficult situation arises when the probability of missingness depends on the (potentially missing) variable itself. For example, suppose that people with higher earnings are less likely to reveal them. Essentially, MNAR are referred to collectively as non-random missingness, and cause more trouble for us than MCAR and MAR.

MCAR is generally unrealistic. Individuals who skip a question or refuse to be measured may be very different than those who comply and they may have systematically different outcome and/or predictor values. MAR is at least somewhat plausible – it allows for systematic differences, as long as those differences are predictable based on observed information. Under MAR (and another condition called ignorability which is beyond the scope of this text), MI leads to consistent estimates with correct standard errors.

MNAR is often plausible, as well, but is more complicated to handle since the distinction between MAR and MNAR depends on unknown information. As a result, a common approach is to assume the data are MAR and use MI to handle the missing data. The plausibility of MAR is improved by including in the imputation model any variable that could be related to the chance of missingness. Evaluating the potential impact of a violation of the MAR assumption (in other words, MNAR) involves using advanced methods to posit a missing data model that depends on the unknown information. This model, by necessity, requires strong assumptions, so MNAR analyses typically include a sensitivity analysis in which the assumptions are varied, resulting in a range of possible conclusions.

When the data are MCAR, a complete case analysis will yield unbiased estimates, although MI will be more efficient in that the effective sample size will increase since no cases need be discarded. Under MAR and MNAR, however, a complete case analysis will typically yield biased estimates. MI, however, can provide consistent estimates under MAR (assuming the imputation model does not leave out important variables and is correctly specified), but not under MNAR.

To summarize, in case of missing data, we observe a paired observation (y_{ij}, R_{ij}), where R_{ij} is a binary indicator variable denoting missingness. When $R_{ij} = 1$, y_{ij} is observed. But when $R_{ij} = 0$, y_{ij} is not observed. Now, our basic model is for $P(\mathbf{y}_i, \mathbf{R}_i \mid \mathbf{x}_i)$, where $\mathbf{x}_i$ is the set of predictors. Due to method of composition, we have $P(\mathbf{y}_i, \mathbf{R}_i \mid \mathbf{x}_i) = P(\mathbf{R}_i \mid \mathbf{y}_i, \mathbf{x}_i)P(\mathbf{y}_i \mid \mathbf{x}_i)$. The crux of all three missingness assumptions relates to the assumed structure for $P(\mathbf{R}_i \mid \mathbf{y}_i, \mathbf{x}_i)$. Here $\mathbf{y}_i = \{y_{ij}\}_{j=1}^{n_i}$, of which some are observed and some are not. Thus we can re-write it as $\mathbf{y}_i = \{\mathbf{y}_{iO}, \mathbf{y}_{iM}\}$ partitioning the observed and missing entries.

- If the data is MCAR: $P(\mathbf{R}_i \mid \mathbf{y}_i, \mathbf{x}_i) = P(\mathbf{R}_i \mid \mathbf{x}_i)$

  Complete case analysis is valid. Mulitple imputation or any other imputation method is valid.

- If the data is MAR: $P(\mathbf{R}_i \mid \mathbf{y}_i, \mathbf{x}_i) = P(\mathbf{R}_i \mid \mathbf{y}_{iO}, \mathbf{x}_i)$

  Some complete cases analyses are valid under weaker assumptions than MCAR. E.g. linear regression is unbiased if missingness is independent of the response, conditional on the predictors. Multiple imputation is valid (it is biased, but the bias is negligible).

- If the data is MNAR: $P(\mathbf{R}_i \mid \mathbf{y}_i, \mathbf{x}_i) = P(\mathbf{R}_i \mid \mathbf{y}_{iO}, \mathbf{y}_{iM}, \mathbf{x}_i)$

  You must model the missingness explicitly; jointly modeling the response and missingness. In some specific cases (e.g. survival analysis), MNAR data (e.g. censored data) is handled appropriately. Generally, we assume MAR whenever possible just to avoid this situation.

While setting up a model under MNAR, one may follow the general framework from `https://academic.oup.com/biomet/article/88/2/551/264970`. `mice` also has some functions for MNAR. But before using those functions, it is important to check the probability models considered for $\mathbf{R}_i$'s and $\mathbf{y}_i$'s by those packages.

Likelihood-based methods In these approaches, missing responses are implicitly imputed by modeling joint distribution of $P(\mathbf{y}_i \mid \mathbf{x}_i)$. In cases of MAR and MCAR, likelihood-based methods can be used based solely on the marginal distribution of the observed data. However, in some approaches like EM, the missing values are validly predicted by the observed data via the model for the conditional mean of the missing responses given the observed responses (and covariates). R package `mlmi` allows running likelihood-based imputation. However, they only allow unsupervised likelihood. If the likelihood is based on a supervised model, there is no standard package for that. It is better to code the EM algorithm directly. When the sample size is large enough, multiple imputation (MI) can be applied instead. Likelihood-based approaches require model for $P(\mathbf{y}_i \mid \mathbf{x}_i)$ must be correctly specified.

Why imputation?

The 'all available data analysis' is only appropriate when the missingness assumption is MCAR, but not MAR. And complete case analysis throws away a lot of data. Hence, imputation is an alternative technique that does not throw any data and some imputation methods also relax the missingness assumption to MAR.

Due to missingness, inherently balanced data may become unbalanced and thus covariance models such as AR or MA may become difficult to implement. Hence, the imputation methods help to impute those missing values and allows us to perform such analyses as well.

Monotone missing One can apply `mice` in case of monotone missingness setting `visit="monotone"`.

Single Imputation In single imputation analyses, NA values are estimated/replaced one time with one particular data value for the purpose of obtaining more complete samples, at the expense of creating some potential bias in the eventual conclusions or obtaining slightly less accurate estimates than would be available if there were no missing values in the data.

A single imputation can be just a replacement with the mean or median (for a quantity) or the mode (for a categorical variable). However, such an approach, though easy to understand, underestimates variance and ignores the relationship of missing values to other variables. Single imputation can also be done using a variety of models to try to capture information about the NA values that are available in other variables within the data set. The `simputation` package can help us execute single imputations using a wide variety of techniques, within the pipe approach used by the tidyverse. Another approach I have used in the past is the mice package, which can also perform single imputations.

Multiple imputation Methods that rely on just a single imputation fail to acknowledge the uncertainty inherent in the imputation of the unobserved responses.

Multivariate imputation by chained equations (MICE) is a particular multiple imputation technique and the most popular one. The chained equation process can be broken down into four general steps:

- Step 1: A simple imputation, such as imputing the mean, is performed for every missing value in the dataset. These mean imputations can be thought of as "place holders."
- Step 2: The "place holder" mean imputations for one variable ("var") are set back to missing.
- Step 3: The observed values from the variable "var" in Step 2 are regressed on the other

variables in the imputation model, which may or may not consist of all of the variables in the dataset. In other words, "var" is the dependent variable in a regression model and all the other variables are independent variables in the regression model. These regression models operate under the same assumptions that one would make when performing linear, logistic, or Poison regression models outside of the context of imputing missing data.

- Step 4: The missing values for "var" are then replaced with predictions (imputations) from the regression model. When "var" is subsequently used as an independent variable in the regression models for other variables, both the observed and these imputed values will be used.
- Step 5: Steps 2–4 are then repeated for each variable that has missing data. The cycling through each of the variables constitutes one iteration or "cycle". At the end of one cycle all of the missing values have been replaced with predictions from regressions that reflect the relationships observed in the data.
- Step 6: Steps 2–4 are repeated for a number of cycles, with the imputations being updated at each cycle.

Sensitivity analysis `https://www.gerkovink.com/miceVignettes/Sensitivity_analysis/Sensitivity_analysis.html` provides details to use `mice` package along with sensitivity analyses. The main purpose of the sensitivity analysis is to check the change in the estimates of the target model if the imputed data is perturbed slightly. `https://cran.r-project.org/web/packages/smdi/vignettes/d_narfcs_sensitivity_analysis.html` has a sensitivity analysis for Cox survival analysis model.

One of the most popular sensitivity analysis approaches is the δ-adjustment approach. `mice` implicitly assumes MAR which automatically covers the MCAR case. Now the only remaining thing is to check MNAR. δ-adjustment is a simple technique to do that, as described paragraph 7.2.3 in [1]. It is based on creating imputations under nonignorable models. We do so by simply adding and subtracting some amount from the imputations. Then run the analysis for these different datasets.

Balanced case In this case, we can run `mice` in wide format and may or may not add other covariates.

Unbalanced case In this case all available data analysis is possible, but under the MCAR assumption. We can still apply mice, but in long format and by adding the other covariates with no missingness. But it ignores within-subject correlation. The most appropriate approach would be based on the target model for the outcome and take EM-based approaches.

Last Value Carried Forward of Last Observation Carried Forward In clinical trials and also in hospital data, this method is widely used. It makes very strong assumptions that the response did not change during the missing period. It may lead to unrealistic conclusions. In some cases, however, such assumptions are reasonable.

References

[1] Stef Van Buuren. *Flexible imputation of missing data.* CRC press, 2018.