Missing Data and Imputation Methods

Presented by Geoffrey Hubona

What is missing data?

- We define ***missing data*** as missing for some (but not all) variables and missing for some (but not all) cases.
- Missing data can affect the properties of the estimates obtained from the data:
  - Means, percentages, percentiles variances, ratios, regression parameters, and so forth.
- Missing data can also affect inferences:
  - Properties of tests and confidence intervals.

Missing mechanism

- Crucial factor to assess the extent of these biasing effects is the ***missing mechanism***: ***How does the probability of an item missing depend on other observed or non-observed variables as well as on its own value ?***
- We denote the collected data as $\boldsymbol{X}$ and partition the data into: $\boldsymbol{X} = \{\boldsymbol{Xo},\ \boldsymbol{Xm}\}$

  where $\boldsymbol{X_o}$ denotes the observed elements of $\boldsymbol{X}$

  and $\boldsymbol{X_m}$ represents the missing elements.

Missing mechanism

- Assume the missingness indicator matrix $\boldsymbol{R}$ corresponds to $\boldsymbol{X}$, such that each element of $\boldsymbol{R}$ is 1 if the corresponding element of $\boldsymbol{X}$ is missing, and 0 otherwise.
- In this way, we can define the ***missingness mechanism*** as the probability of $\boldsymbol{R}$ conditional on the values of the observed and missing elements of $\boldsymbol{X}$:

$$\mathbf{Pr}(\boldsymbol{R}|\boldsymbol{Xo}, \boldsymbol{Xm})$$

Missing by Design

- For example, where survey participants are excluded from the analysis because they are not part of the population under investigation.
  - "Cut-off sampling" in which organizational units above a certain size threshold are included with certainty and those below the threshold are excluded.
    - Is **unit non-response**, we do not consider these further.
  - Or "valid skips," when a question is not answered because it is not applicable to the given unit.

Missing Completely at Random (MCAR)

- Pattern of missing values is totally random and does not depend on any variable which may or may not be included in the analysis.
- The **MCAR** assumption can be expressed as:

$$\mathbf{Pr}(\boldsymbol{R}|\boldsymbol{X}) = \mathbf{Pr}(\boldsymbol{R})$$

which means that the probability of missingness in $\boldsymbol{X}$ depends neither on the observed values in any variable in $\boldsymbol{X}$ nor the unobserved part.

Missing at Random (MAR)

- **MAR** is a weaker assumption and can be expressed by:

$$\mathbf{Pr}(\boldsymbol{R}|\boldsymbol{X}) = \mathbf{Pr}(\boldsymbol{R}|\boldsymbol{Xo})$$

  which means that the missingness on $\boldsymbol{X}$ may depend on the observed part of $\boldsymbol{X}$, but it does not depend on the unobserved part itself.

- Missing data mechanism said to be ***ignorable*** if the data are **MAR** and the parameters governing the missing-data mechanism are distinct from the parameters in the model to be estimated.
  - **MAR** cannot be tested definitively

Missing Not at Random (MNAR)

- **MAR** assumption is violated, data is also said to be ‘not missing at random’ (**NMAR**) which means that some unknown process is generating the missing values.
- Classic example of MNAR is question about income where the high rate of missing values (20%-50%) is related to the value of the income itself (both high and low values).
- **MNAR** can occur when:
  - Missingness depends on unobserved predictors; or
  - Missingness depends on the missing value itself.

Traditional approaches to handle missing data

- List-wise deletion
- Pairwise deletion
- Non-response weighting
- Mean substitution
- Regression substitution
- Last value carried forward
- Using information from related observations
- Dummy variable adjustment
- Deterministic imputation

List-wise deletion

- Also called ‘case-wise deletion’ and ‘complete case analysis.’
- Any observation with missing data for a variable is removed.
- Analysis is performed with the remaining observations, the ‘complete cases.’
- Only justified if data generation mechanism is **MCAR**.
- In R, can omit observations with missing values using function `na.omit()` or extract complete observations using `complete.cases()`.

Pairwise deletion

- Uses all available cases to conduct the analysis.
- If data used in multivariate analyses to compute a covariance matrix, each two cases will be used for which the values of both corresponding variables are available.
  - Pairwise deletion can cause the computed covariance matrix to be non-positive definite.

Non-response weighting

- This involves adjusting for non-response by **weighting**.
- Non-response is problematic if the non-respondents are a random sample of the total sample which is seldom the case.
  - Achieved response rates usually follow a particular pattern:
    - Household surveys: the non-respondents are typically younger so respondents over 30 are usually overrepresented.
    - Men do not participate in surveys as readily as women.
    - Response rates in cities and deprived areas are lower on average.
  - One might use weights to bring the data set more in line with the sampled population.
    - But becomes complicated if more than one variable is missing.

Mean substitution

- Substitute **means** (or **medians** if skewed or **modes** if categorical) for the missing values.
- Preserves sample size so statistical power is not reduced.
- Can distort the distribution of the variable with missing values:
  - Underestimates the standard deviation
  - Pulls estimates of correlations towards zero.

Regression substitution

- Replaces missing values in an input variable by first treating missing input variable as a target and using remaining input variables as predictors in a regression model.
  - Predicted value of regression model substitutes for the missing value.
  - Better than simply substituting a measure of central tendency.
  - But, underestimates the variance of the predicted variable and consequently the standard error (because you have introduced multiple sources of error, or variability in the predicted target value).

Last value carried forward

- For longitudinal data (for example, repeated measures on same subjects), one could apply the ***Last Value Carried Forward***.
  - Last observed value fills in missing values in later observations as the 'best guess' for subsequent missing values.
  - Can introduce bias.
  - Function **`na.locf()`** in package **`zoo`** in R.

Using information from related observations

- Involves imputing missing values with a donor from the underlying data (***hot-deck imputation***) or with a donor from external data (***cold-deck imputation***).
  - Split data set into domains and order data in domains based on pre-defined ***sorting*** variables.
  - ***Sequential*** hot/cold-deck imputation selects the observation which is at the **top** of the sorted variables.
  - ***Random*** hot/cold-deck imputation **randomly** selects an observation in the domain.
  - Is in the package **`VIM`** which is the basis for **`VIMGUI`**.

Dummy variable adjustment

- Also called ***indicator variable adjustment***
- Sets the missing values of a variable to be equal to some arbitrary value, often the mean for non-missing cases, and creates a dummy variable indicating missingness which is then included in the regression model.
  - Ad hoc means to keep observations in the analysis but has no sound theoretical justification.
  - Can bias regression parameter estimates and standard errors.
  - Not a good choice overall.

Deterministic imputation

- Identifies cases in which there is only one possible solution, based on logical rules.
- We will look at several approaches inherent in **`VIMGUI`**:
  - ***Sequential*** and ***random*** **Hot-Deck** imputation;
  - Distance-based, **k-nearest neighbor** imputation;
  - Individual, **regression**-based imputation;
  - Iterative, model-based, stepwise regression imputation (the **IRMI** algorithm) with both standard and robust methods.