Missing data

very short notes

Index

1 Missing data

Why do missing values exist?

- Faulty equipment, incorrect measurements, missing cells in a manual data entry, censored/anonymous data.
- Review scores for movies, books, etc.
- Very frequent in questionnaires for medical scenarios.

Missing values

- Frequently indicated by out-of-range entries or specific characters (max/min, NA, n/a ,..).
- Missing values may have a significance in itself, for instance, missing test in a medical examination. In this case it should be coded as a separate value.
- Most schemes assume this is not the case and *missing* may needed to be code as an additional value.

Types of missing values

- Missing Completely at Random (MCAR): the probability that an observation is missing (x_i) is not related to iyts value or to the value of any other variable. Any piece of data is just as likely to be missing as any other piece of data.
- Missing at Random (MAR): missingness is correlated with other variables that are included in the analysis. For instance, respondents in service occupations can be less likely to report their income.
- Not Missing at Random (NMAR): the distribution of an example having a missing value for an attribute depends on the missing value. For example, respondents with high incomes are less likely to report income. We'd better go back to the source of the data to obtain more information.

Dealing with missing values

- Ignore them (entirely or selectivity). Deletion method (listwise deletion, pairwise deletion).
- Single imputation methods: mean/mode substitution, regression substitution.
- Model-based methods (maximum Likelihood, multiple imputation).

Single imputation methods

- Mean/mode substitution (most common value)
  - Replace missing value with sample mean or mode.
  - Run analysis as if all were complete cases.
  - Advantages: you can use complete case analysis.
  - Disadvantages: it reduces variability.
- Dummy variable control
  - Create an indicator for missing value.
  - Impute missing values to a constant (for instance, the mean).
  - Include missing indicator in the algorithm.
  - Advantages: it uses all the available information about missing observations.
  - Disadvantages: results in biased estimates, not theoretically driven.
- Regression Imputation
  - It replaces missing values with predicted score from a regression equation.

Multiple imputation methods