

Treatment of Missing Values

Missing Data

- Missing Data occurs in a data set when an observation is missing a value on a variable

Elimination

- If a missing value occurs on any of the p variables, eliminate the entire observation.
- This is the default method for most procedures.
- Consequence
 - A data set with even a modest amount of missing values scattered throughout can result in a substantial reduction in sample size.

Imputation

- Imputation describes the process of filling in the missing values of a variable

Imputation Methods

- Substitution with a measure of central tendency
- Distribution based Imputation
- Tree Imputation
- Regression Imputation
- Imputation using the EM algorithm
- Multiple Imputation

Substitution with a measure of central tendency

- Mean (the default in SAS EM)
- Median
 - 50th percentile
- Midrange
 - $(\text{Max} + \text{Min})/2$
- Mode

Substitution with a measure of central tendency

- Robust M-Estimators of Location
 - Tukey's biweight
 - Huber's
 - Andrew's wave
- These estimators limit the effect of extreme observations

Distribution based

- Replacement values are calculated based on the random percentiles of the variable's distribution
- Values are assigned based on the probability distribution of the non-missing observations
- This imputation method seeks to preserve the empirical distribution of the data

Tree Imputation

- For an input variable with missing values, replacement values are estimated by treating this input as a target, and using the remaining input variables (and other rejected variables) as predictors in a Decision Tree
- The predicted value obtained from the Decision Tree replaces the missing value

Regression Imputation

- For an input variable with missing values, replacement values are estimated by treating this input as a target, and using the remaining input variables (and other rejected variables) as predictors in a Regression Model
- The predicted value obtained from the Regression replaces the missing value

Imputation using the EM algorithm

- Conceptually similar to Regression Imputation
- Uses the Expectation-Maximization (EM) algorithm to develop maximum-likelihood estimates of the regression parameters and the variance-covariance matrix

Tree & Regression Imputation

- Because the imputed value is based on other input variables, these techniques may be more accurate than simply substituting a measure of central tendency

Multiple Imputation

1. Sets of plausible values for missing observations are created that reflect uncertainty about the non-responses. Each of these sets of plausible values is used to replace the missing values and create a completed data set.
2. Each of these completed data sets is analyzed.
3. The results of these analyses are then combined, which allows the uncertainty regarding the imputation to be taken into account.

Multiple Imputation

- An example of how plausible values can be created:
 - Use Regression Imputation to build a regression model. The parameters in this model are estimated, and have some distribution.
 - Create a series of regressions, where for each model, the parameter values are drawn from the distribution of the parameters

Availability of Imputation Algorithms

- SAS Enterprise Miner
 - Substitution with a measure of central tendency
 - Distribution based
 - Tree Imputation
- SAS Procedures MI & MIAnalyze
 - Multiple Imputation