MA4128

Kevin O'Brien

February 8, 2018

Abstract

Missing Data

1 Missing values and statistical procedures

Valid and missing cases// Automatic handling of missing values are one of the key features of any statistical package. To avoid stupid mistakes, it is essential to know on how many observations your current analysis is really based. Especially with multivariate procedures the automatic missing value deletion might reduce the number of valid observations drastically, if you are note careful. As always a preliminary diagnosis of your variables helps you avoid this, but still make sure to check with every procedure you run that the number of valid observations included in the analysis is sufficient.

SPSS excludes missing values, when accessing data for any analysis.

- If your analysis implies a single variable, e.g. display an average for a single variable, the average will be based on the valid values (valid n) for that variable.
- If your analysis implies two variables, e.g. produces a scatterplot or a crosstabulation of two variables, only observations that are non-missing on both variables will be plotted or counted.
- If your analysis implies several variables, e.g. a multiple regression with a dependent and five independent variables, it will be based only on the observations that are not missing on all these variables, i.e. even a single missing value on one of the variables will exclude that case.

- This mechanism is known as Listwise missing value deletion and is the default mechanism for all statistical procedures.

2 Missing Value Analysis

The Missing Value Analysis procedure performs three primary functions:

- Describes the pattern of missing data. Where are the missing values located? How extensive are they? Do pairs of variables tend to have values missing in multiple cases? Are data values extreme? Are values missing randomly?
- Estimates means, standard deviations, covariances, and correlations for different missing value methods: listwise, pairwise, regression, or EM (expectation-maximization). The pairwise method displays counts of pairwise complete cases.
- Fills in (imputes) missing values with estimated values using regression or EM methods.

Missing value analysis helps address several concerns caused by incomplete data. If cases with missing values are systematically different from cases without missing values, the results can be misleading. Also, missing data may reduce the precision of calculated statistics because there is less information than originally planned. Another concern is that the assumptions behind many statistical procedures are based on complete cases, and missing values can complicate the theory required.

3 Introduction to Missing Values

Cases with missing values pose an important challenge, because typical modeling procedures simply discard these cases from the analysis. When there are few missing values (very roughly, less than 5% of the total number of cases) and those values can be considered to be missing at random; that is, whether a value is missing does not depend upon other values, then the typical method of listwise deletion is relatively "safe". The Missing Values option can help you to determine whether listwise deletion is sufficient, and provides methods for handling missing values when it is not.

Missing Value Analysis versus Multiple Imputation procedures
The Missing Values option provides two sets of procedures for handling missing values:
The Multiple Imputation procedures provide analysis of patterns of missing data, geared toward eventual multiple imputation of missing values. That is, multiple versions of the dataset are produced, each containing its own set of imputed values. When statistical analyses are performed, the parameter estimates for all of the imputed datasets are pooled, providing estimates that are generally more accurate than they would be with only one imputation.
Missing Value Analysis provides a slightly different set of descriptive tools for analyzing missing data (most particularly Little's MCAR test), and includes a variety of single imputation methods. Note that multiple imputation is generally considered to be superior to single imputation.

3.1 Missing Values Tasks

You can get started with analysis of missing values by following these basic steps:

Examine missingness. Use Missing Value Analysis and Analyze Patterns to explore patterns of missing values in your data and determine whether multiple imputation is necessary.

Impute missing values. Use Impute Missing Data Values to multiply impute missing values.

Analyze "complete" data. Use any procedure that supports multiple imputation data. See Analyzing Multiple Imputation Data for information on analyzing multiple imputation datasets and a list of procedures which support these data.

4 Missing Data

- Missing at Random
- Missing Completely at Random
- Missing Not An Random

Displaying Patterns of Missing Values

You can choose to display various tables showing the patterns and extent of missing data. These tables can help you identify:

- Where missing values are located
- Whether pairs of variables tend to have missing values in individual cases
- Whether data values are extreme

Display

Three types of tables are available for displaying patterns of missing data.

Tabulated cases. The missing value patterns in the analysis variables are tabulated, with frequencies shown for each pattern. Use Sort variables by missing value pattern to specify whether counts and variables are sorted by similarity of patterns. Use Omit patterns with less than n

Cases with missing values. Each case with a missing or extreme value is tabulated for each analysis variable. Use Sort variables by missing value pattern to specify whether counts and variables are sorted by similarity of patterns.

All cases. Each case is tabulated, and missing and extreme values are indicated for each variable. Cases are listed in the order they appear in the data file, unless a variable is specified in Sort by.

In the tables that display individual cases, the following symbols are used:

```
+ Extremely high value
-Extremely low value
S System-missing value
A First type of user-missing value
B Second type of user-missing value
C Third type of user-missing value
Variables
```

You can display additional information for the variables that are included in the analysis. The variables that you add to Additional Information for are displayed individually in the missing patterns table. For quantitative (scale) variables, the mean is displayed; for categorical variables, the number of cases having the pattern in each category is displayed.

Sort by. Cases are listed according to the ascending or descending order of the values of the specified variable. Available only for All cases.

Estimating Statistics and Imputing Missing Values

You can choose to estimate means, standard deviations, covariances, and correlations using listwise (complete cases only), pairwise, EM (expectation-maximization), and/or regression methods. You can also choose to impute the missing values (estimate replacement values). Note that Multiple Imputation is generally considered to be superior to single imputation for solving the problem of missing values. Little's MCAR test is still useful for determining whether imputation is necessary.

Listwise Method

This method uses only complete cases. If any of the analysis variables have missing values, the case is omitted from the computations.

Pairwise Method

This method looks at pairs of analysis variables and uses a case only if it has nonmissing values for both of the variables. Frequencies, means, and standard deviations are computed separately for each pair. Because other missing values in the case are ignored, correlations and covariances for two variables do not depend on values missing in any other variables.

EM Method

This method assumes a distribution for the partially missing data and bases inferences on the likelihood under that distribution. Each iteration consists of an E step and an M step. The E step finds the conditional expectation of the "missing" data, given the observed values and current estimates of the parameters. These expectations are then substituted for the "missing" data. In the M step, maximum likelihood estimates of the parameters are computed as though the missing data had been filled in. "Missing" is enclosed in quotation marks because the missing values are not being directly filled in. Instead, functions of them are used in the log-likelihood.

Roderick J. A. Little's chi-square statistic for testing whether values are missing completely at random (MCAR) is printed as a footnote to the EM matrices. For this test, the null hypothesis is that the data are missing com-

pletely at random, and the p value is significant at the 0.05 level. If the value is less than 0.05, the data are not missing completely at random. The data may be missing at random (MAR) or not missing at random (NMAR). You cannot assume one or the other and need to analyze the data to determine how the data are missing.

Regression Method

This method computes multiple linear regression estimates and has options for augmenting the estimates with random components. To each predicted value, the procedure can add a residual from a randomly selected complete case, a random normal deviate, or a random deviate (scaled by the square root of the residual mean square) from the t distribution.

4.1 Multiple Imputation

Multiple imputation is a simulation-based approach to the statistical analysis of incomplete data. In multiple imputation, each missing datum is replaced by m¿1 simulated values. The resulting m versions of the complete data can then be analyzed by standard complete-data methods, and the results combined to produce inferential statements (e.g. interval estimates or p-values) that incorporate missing-data uncertainty.

4.2 Little's MCAR test

The results of Littles MCAR test appear in footnotes to each EM estimate table. The null hypothesis for Littles MCAR test is that the data are missing completely at random (MCAR). Data are MCAR when the pattern of missing values does not depend on the data values. Because the significance value is less than 0.05 in our example, we can conclude that the data are not missing completely at random. This confirms the conclusion we drew from the descriptive statistics and tabulated patterns.