MA4128

Kevin O'Brien

March 23, 2013

Abstract

Missing Data

1 Missing Data

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

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 completely 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.

1.1 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.

1.2 What is multiple imputation?

Imputation, the practice of ’filling in’ missing data with plausible values, is an attractive approach to analyzing incomplete data. It apparently solves the missing-data problem at the beginning of the analysis. However, a naive or unprincipled imputation method may create more problems than it solves, distorting estimates, standard errors and hypothesis tests, as documented by Little and Rubin (1987) and others.

The question of how to obtain valid inferences from imputed data was addressed by Rubin’s (1987) book on multiple imputation (MI). MI is a Monte Carlo technique in which the missing values are replaced by m¿1 simulated versions, where m is typically small (e.g. 3-10). In Rubin’s method for ‘repeated imputation’ inference, each of the simulated complete datasets is analyzed by standard methods, and the results are combined to produce estimates and confidence intervals that incorporate missing-data uncertainty. Rubin (1987) addresses potential uses of MI primarily for large public-use data files from sample surveys and censuses. With the advent of new computational methods and software for creating MI’s, however, the technique has become increasingly attractive for researchers in the biomedical, behavioral, and social sciences whose investigations are hindered by missing data. These methods are documented in a recent book by Schafer (1997) on incomplete multivariate data.