StatNews #52

Recommended Solutions to Missing Data[1]
May 2002
Revised 2012

There are two methods for dealing with missing data that have become available in mainstream statistical software in the last few years. These two methods are vast improvements over traditional approaches, as described in StatNews #47. This newsletter outlines these two methods.

Both of the methods discussed here require that the missing data mechanism is ignorable, that is, not related to the missing values (see StatNews #46). If the mechanism is ignorable, resulting estimates (i.e., regression parameters and standard errors) will be unbiased with no loss of power.

The first method is Multiple Imputation (MI). Just like the imputation methods discussed in StatNews #47, Multiple Imputation fills in estimates for the missing data. However, to capture the uncertainty in those estimates, MI imputes the values multiple times. Because it uses an imputation method with error built in, the multiple estimates should be similar, but not identical. The result is multiple data sets with identical values for all of the non-missing values and slightly different values for the imputed values in each data set. The statistical analysis of interest, such as ANOVA or logistic regression, is performed separately on each data set, and the results are then combined. Because of the variation in the imputed values, there should also be variation in the parameter estimates, leading to appropriate estimates of standard errors and appropriate p-values.

Multiple Imputation is available in SAS (after v8.1), S-Plus and Solas. In SAS, PROC MI creates the multiple data sets, which can then be easily analyzed separately using standard statistical procedures. PROC MIANALYZE will then combine the results from these separate analyses. Joe Schafer at Penn State has developed four S-Plus libraries for multiple imputing normal, categorical, mixed, and panel data. He has made the library for normal data available as a free stand-alone package called NORM. Multiple Imputation is also available in Solas, but its algorithms have been questioned as inappropriate, and we cannot recommend its use at this time.

The second method is to analyze the full, incomplete data set using maximum likelihood estimation. This method does not impute any data, but rather uses all data observed for each case to compute maximum likelihood estimates. The maximum likelihood estimate of a parameter is the value of the parameter that is most likely to have resulted in the observed data. (For a more detailed explanation, see StatNews #50). When data are missing, we can factor the likelihood function. The likelihood is computed separately for those cases with complete data on some variables and those with complete data on all variables. These two likelihoods are then maximized together to find the estimates. Like

[1] Stata (version 11 and subsequent version) offers a multiple imputation command, the MI command. For more information, refer to our Stat Happenings #10. 2012, May

multiple imputation, this method gives unbiased parameter estimates and standard errors. One advantage is that it does not require the careful selection of variables used to impute values that Multiple Imputation requires. It is, however, limited to linear models.

Analysis of the full, incomplete data set using maximum likelihood estimation is available in AMOS. AMOS is a structural equation modeling package, but it can run multiple linear regression models. We have developed a step-by-step handout for running regression models in AMOS. AMOS is easy to use and directly reads SPSS files, but it will not produce residual plots, influence statistics, and other typical output from regression packages. The missing value analysis package in SPSS will do some very limited maximum likelihood estimates for means and correlations only.

If you would like help choosing a method to missing data and the software to implement it, please contact Karen Grace-Martin in the Office of Statistical Consulting.

References:

Schafer, J. Software for Multiple Imputation
Hox, J.J. (1999) A Review of Current Software for Handling Missing Data, Kwantitatieve Methoden, 62, 123-138.
Allison, P. (2000). Multiple Imputation for Missing Data: A Cautionary Tale, Sociological Methods and Research, 28, 301-309.

Author: Karen Grace-Martin

Back to StatNews Table of Contents