Imputation of missing categorical values in survey research data

VESKA NONCHEVA
Plovdiv University, "Paisii Hilendarski"
Plovdiv, Bulgaria
wesnon@uni-plovdiv.bg, veskanoncheva@gmail.com

The problem of bias due to missing data has received a good deal of attention over the last 20 years and the correction of bias due to nonresponse remains an important problem for investigators using survey data. For data missing because of item nonresponse, imputation of the missing data is often the best solution. Many software packages will automatically remove cases with missing values from the analysis, greatly reducing the sample size, often causing a drastic loss of information. Additionally, if the data are not missing completely at random, removing cases with missing items will result in biased parameter estimates in subsequent analyses.

This study discusses a new method for imputation of missing survey data with a large number of categorical variables. Our method for direct ascription of missing categorical values needs both a tool for association and a tool for detecting which parts of the table are responsible for this association. We use the ubiquitous chi-square test for association in a cross-tabulation. However this test is not a tool detecting which parts of the contingency table are responsible for this association. Correspondence Analysis (see [1], [2]) is a tool that can fill this gap, allowing the machine learning algorithm to see the pattern of association in the data and to generate hypotheses for ascription of missing values that can be tested.

This method for direct ascription of missing categorical values is based on both the association between row and column points and the inertia.
The result can be visualized by plots. The graphical solution is restricted to two dimensions. A three-dimensional display can also be created. This type of display offers the advantage that one can zoom and navigate using the mouse.

We report the result of imputing the missing data from a survey dataset.

Acknowledgments. The research is supported by the Fund NPD, Plovdiv University Paisii Hilendarski, under Grant NI15-FMIIT-004.

REFERENCES

[1] Greenacre, M., *Correspondence Analysis in Practice*, Second Edition. London: Chapman & Hall / CRC, 2007

[2] Nenadic, O., Greenacre, M., *Correspondence Analysis in R, with Two- and Three-dimensional Graphics*: The ca Package, Journal of Statistical Software; Vol. 20, No. 3, 2007