

Missing Data? A Look at Two Imputation Methods

Anita Rocha, Center for Studies in Demography and Ecology
University of Washington, Seattle, WA

ABSTRACT

Statistical analysis is greatly hampered by missing data. It represents a loss of key information, but worse, it can introduce bias in the results of the analysis when the data is not missing at random. In the end can we make confident assertions? One way to help rectify the problem of missing data is to employ a sound method of imputation, a way to replace missing values with reasonable estimates. There are advantages and disadvantages to imputation. Here we look at two approaches. First, we consider the hot-deck method using SAS® macros developed by Stiller and Dalzell, sometimes described as the nearest-neighbor approach. Next, we explore the multiple imputation method using the SAS PROC MI. The examples herein were run on SAS 9.1 for Windows requiring Base SAS and SAS/STAT.

INTRODUCTION

Imputation is the replacement of missing values in data with estimates using any number of techniques and assumptions. The major problem that hampers useful application of imputation methods is bias, when an estimator's long-run average (expectation) differs from the quantity being estimated. "The danger comes from the possible difference between responders and nonresponders. When this difference is systematic, the results of analyses may be biased and false conclusions are easily drawn." (Huisman 2000) Ideally, an imputation method will be plausible and consistent, reduce bias while preserving the relationship between items within the data, and can be evaluated for bias and precision (Sande 1982). By way of illustration, two popular methods of imputation will be reviewed in this paper, the hot-deck method and the multiple imputation method.

HISTORY

With the general public's relatively strong trust in government in the 1940s and 1950s, the U.S. Census Bureau could garner high response rates, even percentages in the high 90s, for its Current Population Surveys (CPS). However, a tendency to want to analyze data without missing values was strong, so methods at the Bureau were devised to fill-in the missing values. The most famous during this era was the hot deck method, a simple approach given that computers were expensive and unreliable (Scheuren 2005). The hot deck involves sampling from a group of "like" records to fill-in missing values. For example, if the data had personal income missing for a working woman, age 38, married with 2 children, the imputed estimate for personal income would be randomly drawn from a group from the survey that share the same or similar characteristics. This estimate would then take the place of the missing value for personal income, and subsequent analyses would be performed using this "more complete" data.

Early on, assessments of precision for the imputed hot deck estimates were not routine although the variance properties were already developed (Hansen et al. 1953). As for bias, the assumption was implicitly made that the data were missing completely at random (MCAR) or conditionally at random (MAR), where the conditional variables were used to create the hot deck bins for the random sampling. However, there is yet another form of bias called nonignorable missing at random (NMAR), one that is likely impossible to account for accurately, even employing modeling augmentation, after the data has been collected. The early solution to avoid bias was to reduce the number of missing values by double sampling (Scheuren 2005). In double sampling, the respondents that didn't complete the survey the first time are asked a second time for a response. This technique, along with triple and quadruple sampling, is still in use today.

The hot deck method is simple to apply but found to underestimate variance, at least in its early applications. In the 1970s, another popular method took root called multiple imputation (Rubin 1977). Multiple imputation involves repeating the imputation three or more times. This method, unlike the hot deck, tends not to underestimate variance.

HOT-DECK

John Still and Donald R. Dalzell of the US Census Bureau published a paper in the 1998 *Proceedings SUGI 23* that outlined a hot-deck implementation strategy commonly used at the US Census Bureau, including highly efficient code for the necessary SAS macros. Below is the basic methodology and then the results from a practical application. (Please see their paper for more detail.)

Example 1: Hot-deck to impute missing income values from a survey

A researcher at the University of Washington wanted to use the hot-deck method to impute missing income values in the NLSY data (the

CONCLUSION

REFERENCES

Hansen, M.H., Hurwitz, W.N., and Madow, W. (1953) *Sample Survey Methods and Theory*, New York: Wiley.

Huisman, Mark (2000) Imputation of Missing Item Responses: Some Simple Techniques, *Quality & Quantity* 34: 331-351.

Rubin, D.B. (1977) Inference and Missing Data, *Biometrika* 63: 581-592.

Sande, I.G. (1982) Imputation in surveys: Coping with reality. *The American Statistician* 36: 145-152.

Scheuren, Fritz (2005) Multiple Imputation: How It Began and Continues. *The American Statistician* 59: 315-319.

Stiller, John and Dalzell, Donald R. (1998) Hot-deck Imputation with SAS® Arrays and Macros for Large Surveys, *Proceedings SUGI 23*, SAS Institute, Inc (<http://www2.sas.com/proceedings/sugi23/Stats/p246.pdf>).

ACKNOWLEDGMENTS

The author would like to acknowledge support from the Center for Studies in Demography and Ecology, University of Washington, Box 353412, Seattle, WA 98195-3412 (<http://csde.washington.edu>).

A good portion the author's work with the hot-deck method was sponsored by the grant "Give and take: Child agency in resource allocation", grant number NICHD R01HD045635-01, and principal investigator Jennifer Romich.

RECOMMENDED READING

Recommended reading lists go after your acknowledgments. This section is not required.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Anita Rocha
Center for Studies in Demography and Ecology
University of Washington
Box 353412
Seattle, WA 98195-3412
Work Phone: 206.616.6687
Fax: 206.616.8135
E-mail: alrocha@u.washington.edu
Web: <http://csde.washington.edu/>

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.