

Overview of Multiple Imputation and Repeated-Imputation Inference (RII) Techniques

Catherine Phillips Montalto, The Ohio State University¹
Jaimie Sung, The Ohio State University²

Multiple imputation is a method of handling missing values in survey data. Multiple imputation uses stochastic multivariate methods to replace each missing value with two or more values to simulate the sampling distribution of the missing values.³ In the resulting multiple complete data sets ("implicates") nonmissing values are identical across the separate implicates but the values imputed for missing information are expected to vary since they are imputed using stochastic methods. Multiple imputation techniques (1) produce data sets with no missing data which increases efficiency of estimation by allowing the researcher to use all available data; (2) can incorporate information in an effort to reduce nonresponse bias; and (3) provide information which can be used to estimate the extra variability in the data due to missing values. As a result, multiply imputed data sets provide a basis for more valid inference and tests of significance. Since 1989 the triennial Surveys of Consumer Finances have used multiple imputation.

The relevant question for the empirical researcher using any SCF since the 1989 survey is how to generate the best point estimate and estimates of variance for parameters of interest given the existence of multiple complete data sets. In general this is achieved by simply combining results across the separate implicates. This method of inference, based on multiple complete data sets, is referred to as "repeated-imputation inference" (RII). Montalto and Sung (1996) provide an overview of multiple imputation in the 1992 SCF and the use of RII techniques.

The value of using the RII technique in empirical research is that it averages over the variability between the individual implicates to produce the best estimate of what the results would have been if the missing data had been observed. Additionally, RII techniques incorporate the variability due to missing values in the variance estimates. In general, this will increase the estimate of variance compared to estimates which ignore this variability resulting in more stringent tests for statistical significance.

The risk of using only a single implete for

empirical research is that point estimates may be biased and the variance of these estimates will be underestimated. Since estimates of variance which ignore the variability due to missing values are downward biased, statistical significance of relationships will be overstated.

Researchers should use RII techniques when the precision of estimated parameters or the statistical significance of relationships is of importance, for example in making policy recommendations or in program evaluation. Future research should carefully study and document the practical importance of imputation variability. The implications of high proportions of imputed values for variables of interest, both as dependent and independent variables in multivariate frameworks, need to be better understood.⁴

References

- Montalto, C. P. & Sung, J. (1996). Multiple imputation in the 1992 Survey of Consumer Finances. *Financial Counseling and Planning*, 7(1), 133-146.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley Sons.

Endnotes

1. Assistant Professor, Consumer and Textile Sciences Department, 1787 Neil Avenue, Columbus, OH 43210-1295. Ph 614-292-4571. E-mail: montalto.2@osu.edu.
2. Doctoral Candidate, Consumer and Textile Sciences Department.
3. Rubin (1987) provides an extensive discussion of both the theory and practice of multiple imputation.
4. A full version of this paper is available from the authors.