Effects of Multiple Imputation: Empirical Results Using Linear Regression to Predict the Level of Total Household Income

Catherine Phillips Montalto, The Ohio State University[1]
Jaimie Sung, The Ohio State University[2]

While theory recommends the use of repeated-imputation inference (RII) techniques for multiply-imputed data sets (Rubin, 1987; Montalto & Sung, 1996), the practical significance of the technique may best be illustrated by studying empirical results. We use the 1992 SCF to estimate OLS regression equations for total household income separately on each of the five implicates, and using RII techniques. The dependent variable is the natural logarithm of total household income. The independent variables are selected to measure characteristics of the respondent and household which are associated with variation in total household income.

We examine differences between the implicates and the RII results in the magnitude, sign and significance of estimated coefficients. The separate results from the five implicates are similar, each explaining about 38% of the variance. Thirty-six of the 38 estimated coefficients are similar in magnitude and statistical significance, and of the same sign across the five implicates.

Estimates derived using the RII techniques confirm results which were consistent across the five separate implicates. The magnitude of the RII coefficient is always in the middle of the range produced from the separate implicates since the RII best point estimate of a regression coefficient is the average of the five separate regression coefficients. Coefficients which were statistically significant across the five separate implicates remain statistically significant in the RII results, although the levels of significance are generally more stringent in the RII results, as is to be expected.

To further analyze variation, these equations are used to predict total household income. Predictions from the five separate implicates are compared with each other and to results obtained using the RII results. Two types of comparisons are made. First, actual levels of predicted total household income for given sets of characteristics are compared to identify which implicates over predict and which implicates under predict income relative to the RII predictions. Second, to control for differences in the levels of the predicted total household income between different sets of characteristics, percent errors in the level of total household income predicted from each implicate relative to the RII predictions are calculated.

The degree of the prediction error is small (0.23% to 4.78%). That these differences are small is not surprising since we use variables which are known to have very low proportions of nonresponse in the survey data. As a result, a low proportion of the data values are imputed and the amount of variability in the data values between the five implicates is small. In a sense, we set ourselves up to find "no important differences." In this light, it is interesting that we do find differences, and that the prediction error gets close to five percent in some cases. We are also able to illustrate that the differences between results derived separately from each implicate vary in a nonsystematic pattern; specifically in our case, the tendency to over predict or to under predict, and the relative size of the prediction error varies in a nonsystematic pattern across the five separate implicates.[3]

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, 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. A full version of this paper is available from the authors.