Proceedings of Statistics Canada Symposium 2002
Modelling Survey Data for Social and Economic Research

MULTIPLE IMPUTATION OF MISSING INCOME DATA AT THE INDIVIDUAL AND FAMILY LEVELS USING SEQUENTIAL REGRESSION IMPUTATION: APPLICATION TO THE NATIONAL HEALTH INTERVIEW SURVEY

Trevellore E. Raghunathan, Nathaniel Schenker, Pei-Lu Chiu, and Diane Makuc[1]

ABSTRACT

Analysts of health data are often interested in studying relationships between income and health. The National Health Interview Survey, conducted by the National Center for Health Statistics of the U.S. Centers for Disease Control and Prevention, provides a rich source of data for studying such relationships. The nonresponse rates on two key income items, an individual's earned income and a family's total income, are over 20%, however. Moreover, these nonresponse rates appear to be increasing over time. A project is currently underway to multiply impute individual earnings and family income along with some other covariates for the National Health Interview Survey in 1997 and subsequent years.

There are many challenges in developing appropriate multiple imputations for such large-scale surveys. First, these are many variables of different types, with different skip patterns and logical relationships. Second, it is not known what types of associations will be investigated by the analysts of multiple imputed data. Finally, some variables such as family income, are collected at the family level and others, such as earned income, are collected at the individual level. To make the imputations for both the family- and individual level variable conditions on as many predictors as possible, and to simplify modeling, were are using a modified version of the sequential regression imputation method described in Raghunathan et al. (2001, Survey Methodology). We are creating imputations in two stages to account for two different levels of analytic units, the individual and the family. In the first step, we impute individual-level covariates and then create summaries at the family level. Next, we impute the family-level variables conditional on these summaries. To build interdependence between the family- and individual-level variables, we iterate between two steps where the imputed variables at the family (individual) level in the previous iteration are used as predictors in the current iteration for imputing variables at the individual (family) level.

Besides issues related to the hierarchical nature of the imputations just described, there are other methodological issues of interest, such as the use of transformations of the income variables, the imposition of restrictions on the values of variables, the general validity of sequential regression imputation, and even more generally the validity of multiple-imputation inferences for surveys with complex sample designs.

This paper will describe the multiple imputation of income in the National Health Interview Survey and discuss the methodological issues involved. In addition, the paper will present empirical summaries of the imputations as well as results of a Monte Carlo evaluation of inferences based on multiply imputed income items.

[1] University of Michigan, and NCHS, USA, teraghu@umich.edu