# Practical Issues and Challenges from the Users' Perspective

**Catherine P. Montalto, The Ohio State University[1]**

The Survey of Consumer Finances is a very complex data set due to the sample design, multiple imputation of missing data, and issues related to confidentiality and disclosure. Since the 1989 survey, the questionnaire, sample design and imputation techniques have changed only marginally. This provides important "economies" to researchers in that knowledge and skill are highly transferable from one SCF to another from 1989 forward.

The survey employs a dual frame sample design. One frame is a standard multi-stage area probability sample of households in the U.S. The second frame uses a special list sample drawn from a sample of tax records that is intended to oversample households that are more likely to be wealthy. This complex sample design provides a more legitimate basis for estimates of narrowly-held assets and highly concentrated wealth in the U.S. However, this over sampling of households more likely to be wealthy means that data must be weighted to generate valid estimates for the U.S. population. Additionally, the data also include some highly influential observations that present challenges in empirical research.

Multiple imputation is used to handle data missing in the Surveys of Consumer Finances due to item non-response. The technique used by the Federal Reserve Board uses stochastic multivariate methods to replicate each original observation multiple times. Beginning with the 1992 SCF, the public data sets have contained five complete data sets, referred to as "implicates." The benefit to researchers is that the public data sets contain no missing values. The cost is that researchers must learn how to analyze data appropriately in the presence of five complete data sets.

Confidentiality and disclosure issues are related to protecting the identity of individual respondents. These issues are extremely relevant to the SCF because (1) the survey collects sensitive data on household assets, liabilities, and financial behavior, and (2) the survey oversamples households that are more likely to be wealthy. The related disclosure issues include determining what information will and will not be released in the public data set, as well as techniques used to adjust potentially identifying information. Because the SCF is sample survey data, estimates derived from the SCF data can contain error due to sampling. Due to the aforementioned confidentiality and disclosure concerns, standard methods of calculating sampling error cannot be used.

Now I would like to take these issues and place them in the context of "practical issue from the users' perspective". From the standpoint of empirical research, once we understand the theoretical and statistical issues, I think our questions focus on the "how to" and the "practical significance" of these issues, specifically determining how much our empirical results change when we account for these issues in our estimation.

Let me begin with sampling error. As previously mentioned, legal and ethical confidentiality issues prevent the Federal Reserve Board from releasing information related to the sampling frame that users need to implement any of the classical resampling approaches to estimation of sampling error. As an alternative, the public data set contains a set of replicate weights that can be used to derive estimates of error due to sampling. The Codebook for the 1995 Survey of Consumer Finances contains SAS code that can be used to derive these estimates.

With respect to multiple imputation and imputation error, the "how to" appropriately analyze data in the presence of five complete data sets is fairly straightforward. This method of inference, based on multiple complete data sets, is referred to as "repeated-imputation inference" (RII). RII techniques incorporate the variability due to missing values, or imputation error, in the variance estimates. SAS code for estimating imputation error of point estimates is provided in the Codebook for the 1995 Survey of Consumer Finances, and code for the use of RII in an OLS regression is available from the SCF Users Group Web page. For an easily understandable discussion of multiple imputation in the SCF from a user's point of view, refer to the Montalto and Sung paper that appeared in Financial Counseling and Planning in 1996.

Members of the SCF Users Group had a lively discussion about the appropriateness of RII techniques for nonlinear models. In order to resolve the multiple points of view within our group, we contacted Professor Donald B. Rubin at Harvard University, and author of the eminent book on multiple imputation. Professor Rubin confirmed that RII techniques are applicable to both linear and nonlinear models. The criteria for determining whether RII techniques are appropriate is independent of the functional form of the estimating equation. RII techniques are appropriate whenever the complete-data analysis inferences are based on estimates and standard errors. These estimates can include population means, variances, correlations, factor loadings, and regression coefficients.

The "how to" estimate sampling error and imputation error is fairly straightforward. The practical significance of these issues is also a very important question. The use of the replicate weights to estimate sampling

error and the use of RII techniques to estimate imputation error are computationally intensive methods. These techniques require fairly large computer memory, and familiarity with computer matrix language. We do have some evidence of the relative importance of sampling error and imputation error in the context of population estimates, and of the importance of imputation error in multivariate analysis. Refer to the full version of this paper for more details.[2] Future research should carefully study and document the practical importance of imputation error in multivariate analyses.

The third empirical issue is related to weighting — specifically when to weight and when not to weight — and to influential observations. Because the SCF sample is not an equal-probability design, the final nonresponse-adjusted sampling weights should be used to produce point estimates and descriptive statistics that are generalizable to the U.S. population. However, even after weighting, influential observations often inflate estimates of means and standard errors of the mean. Researchers at the Federal Reserve routinely review calculations for the presence of overly-influential outliers, and apply robust techniques when appropriate. As researchers we need to increase our understanding of these techniques, including weight trimming and graphical analyses, and use these techniques appropriately in our analyses.

There seems to be less consensus on the role of weights in multivariate analyses. One camp argues that weights should be used in multivariate analyses, and the other camp argues that weights should not be used. The rationale for not weighting multivariate analyses is that if the strata defining variables used to construct the weights are controlled for in the multivariate analysis, it is not necessary, and possibly not appropriate, to weight. There are also issues related to whether the strata defining variables are endogenous or exogenous. If the weights are endogenous, weighted regression coefficients will suffer from simultaneity bias. From a users point of view, we need to understand what it means to "control for the strata defining variables" and how the strata defining variables should enter our models -- should they affect the intercept term only, or the slope coefficients as well? Clearly an important strata defining variable in the SCF is the wealth index strata used in the selection of high-income households.

Researchers can always think of variables or information they would like to have. I have constructed a "Wish List" for consideration by the Federal Reserve Board. (1) Disaggregate the race/ethnicity variable. The SCF collects information on the race and ethnicity of the respondent and spouse. Responses are recorded separately for six groups. However, in the public data set, Asian or Pacific Islander, Native American/Eskimo/Aleut, and Other are combined into one category. Thus, in empirical work it is impossible to separate Asian and Pacific Islanders from Native American/Eskimo/Aleut. This is disappointing, since many behaviors differ substantially between these two groups. (2) Ask individuals if they are covered by Social Security and include this information in the public data set. This will improve the accuracy with which we can estimate retirement income. While 95% of workers in the U.S. are covered by Social Security, 5% are not. And the retirement benefits available through State retirement programs are very different from those available through the Social Security program. Assuming that all persons are covered by Social Security will underestimate the retirement resources of persons covered by "generous" state programs, and in some cases the extent of this underestimation will be large. (3) If it is possible to simultaneously honor legal and ethical confidentiality issues, and to provide an indicator variable for rural or nonmetropolitan residence, analysis of the SCF data could enhance our understanding of the circumstances and needs of rural and small town communities.

The Survey of Consumer Finances is a rich source of information on assets and liabilities of U.S. households; it is also a very complex data set. The complexity arises from the sample design, multiple imputation of missing data, and issues related to confidentiality and disclosure. These complexities present abundant challenges to users of the data. Fortunately there is abundant information available to help researchers correctly and respectfully utilize the data to address important research questions.

## References

Montalto, C. P. & Sung, J. (1996). Multiple imputation in the 1992 Survey of Consumer Finances. Financial Counseling and Planning, Vol. 7, pp. 133-46. URL http:/www.hec.ohio-state.edu/hanna/mltimp.htm

Endnotes

1. Assistant Professor, Consumer and Textile Sciences Department.
2. The full version of this paper is available at http://www.hec.ohio-state.edu/scf/cmacci98.htm