Enhancing the SCF with Administrative and Alternative Survey Data

Gerhard Fries, Federal Reserve Board

The Survey of Consumer Finances (SCF) is well known as a major source for household data on finances, employment, and demographics (see Kennickell, Starr-McCluer, and Sundén, 1997). However, it is not generally known that the SCF is enhanced in many ways by data from a variety of other sources, including administrative data and alternative survey data. Some of these data are merged into the SCF and are useful for economic analyses, while others play an integral role in sampling, weight design, and imputation. It is important to note that there is no personally identifying information about survey respondents in the SCF dataset. This paper focuses on these issues for the 1995 SCF. The next section provides a brief summary of the SCF, while the following section describes the SCF data structure. The subsequent sections detail the use of administrative and alternative survey data in sample design, weight design, and imputation. The final section has some closing remarks.

■ Background on the SCF

The SCF is sponsored by the Board of Governors of the Federal Reserve System (FRB) with cooperation from the Statistics of Income Division (SOI) of the Internal Revenue Service. It is a household survey that is conducted every three years. Data for the 1995 survey were collected via laptop computers (CAPI) by the National Opinion Research Center (NORC) at the University of Chicago between the months of June and December. The average interview required 90 minutes, but complicated cases took significantly more time.

Perhaps best known for collecting information on household wealth, the SCF also includes questions on employment, demographics, credit card use, financial institutions, and attitudes about saving and spending. There is much diversity in the data, including variables that are not widely distributed (e.g., corporate stock or bonds). To provide adequate coverage of these variables, the sample is selected from a dual-frame design that is composed of a standard multistage area-probability (AP) frame and a list frame based on administrative records maintained by SOI.[1] Of the 4,299 interviews completed in the 1995 SCF, 2,780 families were from the AP sample, and 1,519 were from the list sample. The response rate for the survey was about 70 percent for the AP sample and around 34 percent for the list sample. Weighting adjustments are used to compensate for this type of unit nonresponse, while multiple imputation techniques are used to deal with item nonresponse.

■ 1995 SCF Data Structure

Main Data

The majority of the information in the SCF dataset derives from questions asked directly to respondents. Of course, the publicly released dataset may not have all of these data (see Fries, Johnson, and Woodburn, 1997).

Selected data from the following sources are merged into the main internal dataset:

Frame data

These data consist of variables relating to the sample design, including those for geography, sample selection probabilities, and weight adjustment factors. None of this information is included in the publicly released version of the 1995 SCF except geography for the four-level Census regions and nine-level Census divisions.

HEF, HCO, ROC, and Interviewer Survey

Additional information is collected by the survey interviewer in the Household Enumeration Folder (HEF). This is a folder of confidential materials and includes the Screener, the Household Contact Observation Form (HCO), and the Record of Calls (ROC). There is one version for the list sample and another one for AP households. The main difference is in the screeners, which are used to determine the respondent for the chosen

household. The criteria used for sample eligibility differ, depending on whether the household is from the list sample or from the AP sample. For AP cases, the main focus of the screener is to determine the "head of the household." Because the list sample consists of a set of names, the eligible person is restricted to the sample names or the spouses of sample names. In general, the person most knowledgeable about the household finances is interviewed in each sample. The HCO includes questions about the respondent's housing structure, as well as characteristics of the neighborhood surrounding the dwelling. Information is also collected relating to the first personal contact with the HEF informant. The ROC is a detailed record of all attempts, contacts, and discussions with informants, respondents, gatekeepers, etc. All data from these sources are purged of identifying information before they are given to FRB staff. The personal contact information in the HEF and the ROC data are useful in understanding unit nonresponse in the SCF (see Kennickell, 1997). One last set of data incorporates completed questionnaires of interviewers regarding their employment and educational background and certain attitudes. None of these data is released to the public.

Car Value Data

The CAPI instrument does not ask the respondents the value of their car(s), but information is collected on the car(s), including make/model and model year. Given this information, car values are estimated using N.A.D.A. official used car guides. Although the actual make and model of vehicles cannot be released, variables representing the value of each vehicle are included in the publicly released dataset.

CPS Data By Occupation Code

Certain employment statistics estimated from the March 1995 Current Population Survey (CPS) conducted by the Census Bureau are linked by a three-digit occupation code to the 1995 SCF. These data include average fraction of the last 52 weeks worked, average hours worked in 1995, fraction unemployed in 1995, and the mean wage for 1995.

Additionally, CPS data were used to run regressions by occupation group separately for males and females for the log of annualized wages. Regressors included a constant, a spline on age (age, max(0, age-35), max(0,age-55)), a dummy variable for part-time employment, a dummy variable for self-employment, a dummy variable for race, and dummy variables for years of education. These coefficients were then employed to predict the annual level of expected income of respondents and their spouses or partners using the 1995 SCF data. The predicted wage and recodes of all of the CPS coefficients are then merged into the SCF dataset by household. None of these data is released to the public.

Data from 1990 Census of Population

To better characterize the locality or "neighborhood" in which a household resides, the 1995 SCF also contains tract and zip-code-level census information. Data were extracted from the Summary Tape Files (STF) of the 1990 Census of Population conducted by the Census Bureau. This information includes population, race, marital status, education, labor, housing information, commuting time, and other variables by zip code and census tract. Some county information was also selected. None of these data is released to the public.

Bank Market Data

The FRB maintains data bases on banking markets, and many such data items are available on a geographic basis. Data on bank market concentration, bank failures, and related items have been matched to the SCF at the State and county level. None of these data is released to the public.

■ Sample Design (Using Census Data and Administrative Files from SOI)

Area Probability Sample

As previously mentioned, the SCF has a dual-frame sample design. The AP sample is drawn from the NORC National Sample Frame based on administrative files relating to the 1990 Census. The first stage in the cre-

ation of NORC's National Sample involved dividing the U.S. into geographical units and then stratifying by urbanization, population size, and region. This resulted in 2,489 primary sampling units (PSU's) for the U.S. from which 100 were selected for the NORC National Sample. Nineteen of these represented large metropolitan areas such as Los Angeles or New York which were selected with probability one. The remaining PSU's were selected using systematic sampling with probability proportional to their size in housing units. In the next stage, smaller sampling units called "segments" were selected from within these 100 PSU's. Each segment in the U.S. had a probability of being selected into the sample proportional to its 1990 housing unit count. NORC field staff then listed dwelling units on a preselected subsample of segments. These dwelling units provide the source for the AP component of the SCF sample. Every unit had an equal probability of selection.

The unit of analysis for the SCF is the household, which is divided into the "primary economic unit" (PEU) and everyone else in the household. The PEU is defined to be the economically dominant single individual or pair of individuals within a household and all other persons financially dependent on that person or persons.

List Sample

The list sample is designed to provide adequate coverage for highly concentrated variables by oversampling wealthy households. Based on an agreement between the Federal Reserve and SOI, the list sample was selected from an annual sample of tax data from the 1993 Individual Tax File (ITF) containing 222,000 records. The ITF oversamples high-income taxpayers and is stratified by various income types. Additional agreements among the Federal Reserve, NORC, and SOI are designed to protect the privacy of individuals. These agreements pertain to both the type of information released to the public from the SCF and to the use of the ITF. For instance, the SCF staff at the Federal Reserve has access to the ITF file, but without names, and uses of the ITF file are limited to areas involving such issues as sampling and weighting.

For the 1995 SCF, the ITF was used to construct a "wealth index" (WI). In earlier SCF's, WI's were created by grossing up capital flows observed from the corresponding ITF files and using market rates of return. In 1995, the WI was defined as a combination of this earlier version and an index estimated from a model regressing gross assets on income and various tax variables. The WI was then used to create strata, and units were sampled at disproportionate rates, with sampling rates increasing for the wealthier strata. The highest stratum was not sampled. For cost reasons, only units living in PSU's selected from the AP sample were included. There are potential drawbacks to using the ITF for the SCF sample design (see Kennickell and McManus, 1993), including the fact that the unit of observation in the ITF is the taxpayer and not the household. However, it does appear that the technique outlined above is effective in capturing the wealth distribution for the U.S. population (see Kennickell, 1998).

■ Weight Design (Using SOI Data and Alternative Survey Data from the CPS)

Analysis Weights

Analysis weights are constructed for using the two samples jointly, and these weights are based on separate weights computed for each of the two sample frames. For the AP weights, the 1995 March CPS is used to compute control totals for both post-stratification and raking adjustments within age categories, age-homeownership categories, and household population figures by region. For the list weights, the ITF file is used for the selection probability[2], non-response adjustments, calculation of estimated control totals used for ratio adjustments, and raking procedures. The two sets of weights are then combined by a post-stratification technique. The two samples cannot be simply merged together because it is not possible to compute a joint probability of observation given the two frames. In general, one can assume that the list sample will provide the best estimates for the upper end of the wealth distribution and that the AP sample will provide the most reliable estimates for the lower end of the wealth distribution. For a very detailed look at the computation of the 1995 analysis weights, see Kennickell and Woodburn (1997).

Replicate Weights

It is often important to estimate variances for statistics calculated from the SCF. Sampling variance is a major component of this total variance, but cannot be computed by classical means (e.g., balanced repeated replication) from the SCF due to the intricate nature of both the sample design and the weighting methodology. Instead, 999 bootstrap sample replicates have been constructed from the final dataset using detailed sample design information from each of the sample frames. Analysis weights are computed for each replicate using the same procedure as in calculating the main analysis weight. Bootstrap techniques can then be used to estimate variances due to sampling (see Sitter, 1992). A good example of how one can use the multiply imputed SCF's and replicate weights for variance estimation of survey estimates can be found in Kennickell and Woodburn (1997).

■ Imputation (Using Outside Data from N.A.D.A. and CPS)

To account for item nonresponse, the final versions of the SCF datasets since 1989 are all multiply-imputed (see Kennickell, 1991 and Rubin, 1987). The imputation approach mainly involves iteratively estimating a sequence of regression models. Values to replace missing data are drawn, based on available information for a given respondent household.[3]

Often, right-hand-side variables for the above mentioned regressions include data that are not directly in the SCF, but come from other sources. One example includes variables that contain the value of the household's cars (see Section II.). These variables are estimated using an outside data source--N.A.D.A. used-car guides. Additionally, a variable corresponding to annualized expected income (wage) is used as an independent variable for many of the labor-variable imputations (e.g., head of household/spouse wage). This variable, as well as others (see Section II) including variables formed from regression coefficients by Census occupation code of annualized wage, is derived from using the March 1995 CPS. Strata indicator variables defined from the list sample wealth index are also used in the same manner, allowing some of the imputations to condition on the sample design itself.

■ Closing Remarks

The SCF is a rich set of household data, but it is also a collection of various other information gathered from a variety of other sources, including other surveys and administrative data files. As explained earlier, some of these data are necessary for sampling, weighting, and the imputation process. In the future, there may be additional ways to enhance the SCF using such data. For example, the 1990 Census tract data might be employed in the sampling or weighting process.

The final payoff, of course, is that the SCF data are a major source for both academic and policy research in the areas of household saving, household portfolios, borrowing and liquidity constraints, and wealth inequality (see Fries, Starr-McCluer, and Sundén, 1998).

■ Acknowledgments

The author would like to express gratitude to all of the SCF staff for their support and a special thanks to Arthur Kennickell for his guidance and comments.

■ Endnotes

This paper will be available on the FRB SCF web site:

http://www.bog.frb.fed.us/pubs/oss/oss2/scfindex.html

■ Footnotes

[1] Sample selection procedures are such that only NORC staff ever see the names of the survey participants.

[2] The base input for the list weight is the inverse of the probability of selection, which is the product of the probability of selection into the SOI sample, the probability of selection of the areas of the AP sample, and the sampling rate within the list sample strata.

[3] These imputations are stored in successive replicates (five for 1995), called implicates, of each data record. A separate analysis weight is constructed for each of these implicates.

References

Fries, G.; Johnson, B.W.; and Woodburn, R.L. [1997], "Analyzing the Disclosure Review Procedures for the 1995 Survey of Consumer Finances," *Proceedings of the Section on Survey Research Methods,* 1997 Annual Meeting of the American Statistical Association, Anaheim, CA.

Fries, G.; Starr-McCluer, M.; and Sundén, A. [1998], "The Measurement of Household Wealth Using Survey Data: An Overview of the Survey of Consumer Finances," working paper, Board of Governors of the Federal Reserve System, Washington, DC.

Kennickell, A.B. [1998], "List Sample Design for the 1998 Survey of Consumer Finances," working paper, Board of Governors of the Federal Reserve System, Washington, DC.

Kennickell, A.B.; Starr-McCluer, M.; and Sundén, A. [1997], "Changes in Family Finances, 1992 to 1995: Evidence from the Survey of Consumer Finances," *Federal Reserve Bulletin*, Volume 83 (January), pp. 1-24.

Kennickell, A.B. and Woodburn, R.L. [1997], "Consistent Weight Design for the 1989, 1992, and 1995 SCF's, and the Distribution of Wealth," working paper, Board of Governors of the Federal Reserve System, Washington, DC.

Kennickell, A.B. [1991], "Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation," *Proceedings of the Section on Survey Research Methods*, 1991 Annual Meeting of the American Statistical Association, Atlanta, GA.

Rubin, D.B. [1987], *Multiple Imputation for Nonresponse in Surveys*, John Wiley and Sons, Inc.

Sitter, R.R. [1992], "A Resampling Procedure for Complex Survey Data," Journal of the American Statistical Association, 87(419), pp. 755-765.