Using Graphical Analyses to Improve all Aspects of the Survey of Consumer Finances

Gerhard Fries, Federal Reserve Board, and R. Louise Woodburn, Internal Revenue Service
Gerhard Fries, FRB, Mail Stop 180 Washington, DC 20551, m1gxf00@frb.gov

Key Words: Complex Survey, Scatter Plot, Editing, Imputation, Weighting, Disclosure Avoidance

Introduction

Data providers and analysts are now equipped with sophisticated tools in a highly technical computing environment. Surveys are at the same time becoming more intricate resulting in data files which may be both quite complicated and potentially very large. As a result, traditional data processing tasks have become much more complex and often are very data dependent. In order to better understand the underlying data, it is always helpful and sometimes critical for the data analyst to "look at" the data using graphical techniques. The Survey of Consumer Finances (SCF) is a complex survey with data on all aspects of a household's financial characteristics. Additionally, the sample design is not simple. The use of graphics has become a key ingredient in producing high-quality data for the Survey of Consumer Finances. Different types of graphical plots have been used at all stages of processing. These stages include sample selection, imputation, editing, weighting, analyses, and disclosure review. The objective of this paper is to detail how the graphics are used, how their use has improved the survey processing, and how the use of graphics will be expanded for future surveys.

Including this introduction, this paper contains four sections. In the next section, we detail the Survey of Consumer Finances data, the sample design and other processing unique to the survey. The third section discusses the use of graphics in the areas listed above. The final section includes some closing comments to encourage the use of graphics. Examples are included to show how the implementation works in the survey.

The Survey of Consumer Finances

The SCF is a triennial household survey sponsored by the Federal Reserve Board with cooperation from the Statistics of Income (SOI) of the Internal Revenue Service. Data are collected on household finances, income, assets, debts, employment, demographics, and businesses. The interview averages about 75 minutes, but interviews of households with more complicated finances sometimes last several hours. An important objective of the SCF effort is to collect representative data to measure wealth. In order to accomplish this, the sample is selected from a dual frame that is composed of an area probability frame (AP) and a list frame (see Kennickell, A. B. and McManus, D. A., [1993] for details on the strengths and limitations of the sample design). The list frame is based on administrative records maintained by SOI. The list frame sample is stratified on an estimated wealth index with the higher indices selected at higher sampling rates. The 1989 sample was additionally complicated by the inclusion of a panel follow-up from 1983, a portion of which is also appropriately included in the 1989 cross section data set (see Heeringa, S. et al. [1994] for a description of the 1989 sample design). The 1992 and 1995 studies do not incorporate panel components. This paper is based on experiences from the 1989, 1992 and 1995 studies.

Due to the sensitive nature of the financial questions, both unit and item nonresponse are concerns in the SCF. The complex sample design and the use of frame information for estimation helps to address the unit nonresponse concern. For the item nonresponse, missing values are multiply imputed using a Gibbs sampling approach (see Kennickell [1991]). For the SCF, the respondent has three options for a given question, he can: 1) give a particular value, 2) answer with a refusal or a don't know, or 3) choose an interval from a range card provided by the interviewer. In the imputation procedure, refusals, don't knows, and range card values are imputed. The imputations for the range card responses are constrained by the range interval boundaries. The Gibbs sampling approach involves iteratively estimating a sequence of large randomized regression models to predict the missing values based on variables that are available for a given respondent. The result is an imputed dataset that preserves the distributions and relations found in the non-imputed data. A shadow variable is included that indicates the status of the original data, such as, whether or not the value is imputed, and what the range card interval was, if given. The imputation machinery is used in the disclosure avoidance preparation of the public use file.

Figure 1. Scatter Bi-Plot Comparing Different Stratification Variables to Financial Income

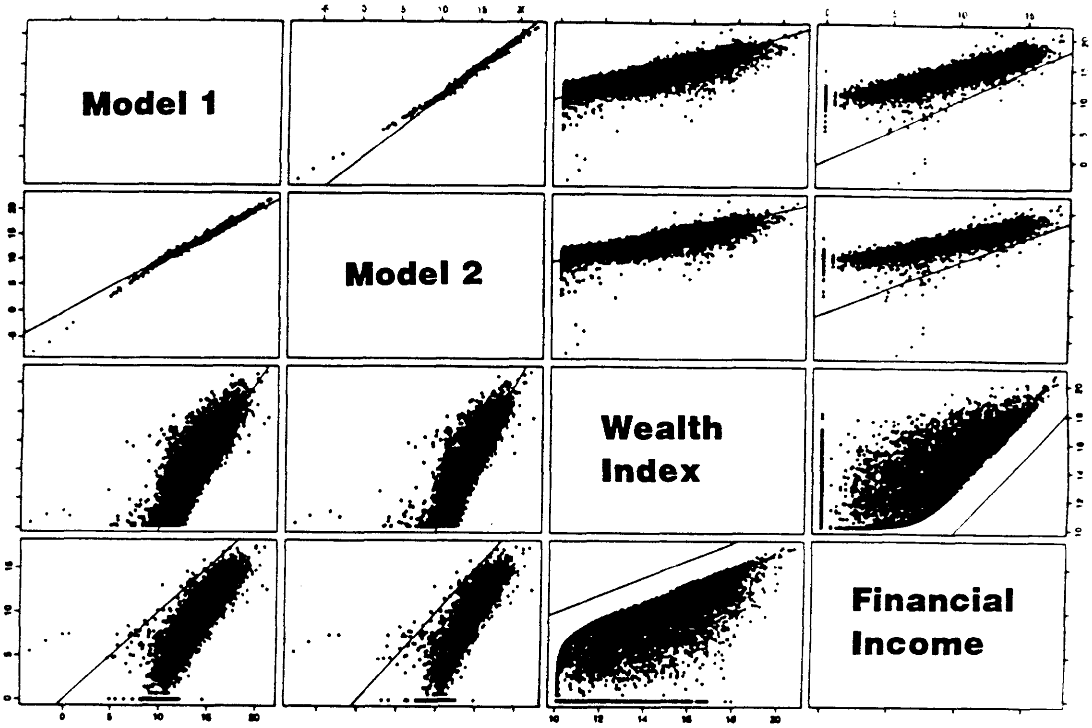
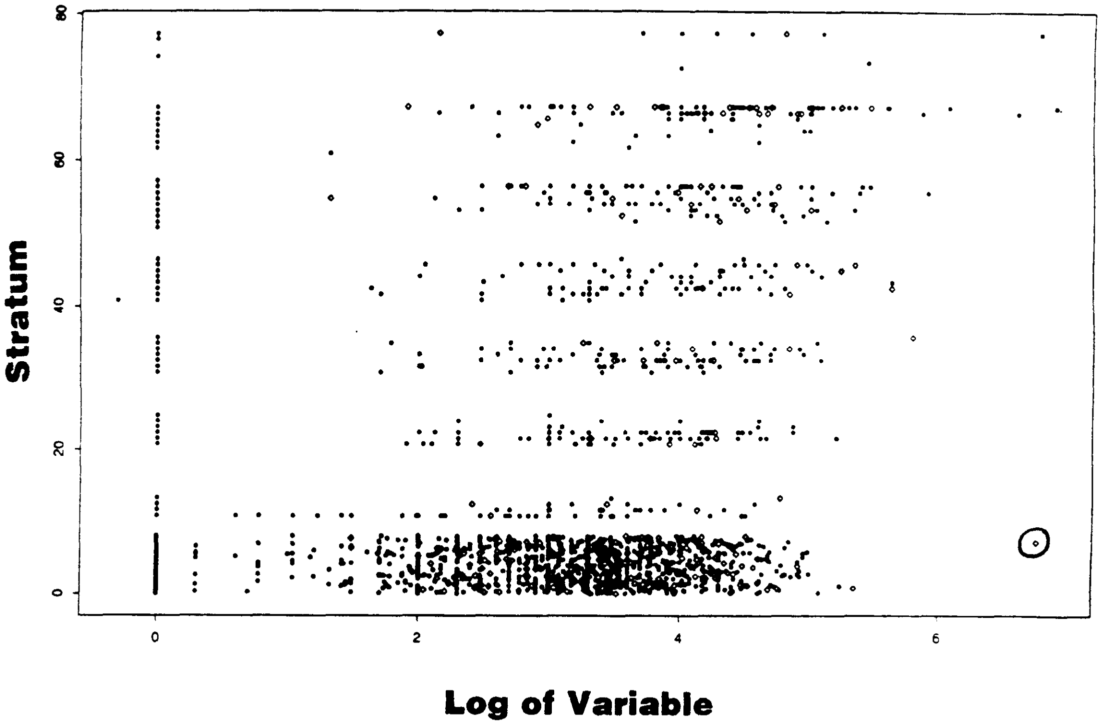


Figure 2. Plot to Investigate Editing and Imputation Processes



Use of Graphical Analyses

Graphical analyses for the SCF has been facilitated by the use of the software package S-Plus. The different graphs used are derived from scatter plots, bi-plots, cumulative distribution plots, and qq plots. For example, one plot used in all stages of processing is a scatter plot of the variable of interest versus an indication of sampling stratum. In this plot, we can discern list cases from area probability cases, and imputed data from reported data. The application of the graphical analyses are described below by processing task in order to show the benefits to each step.

Sample Design

Due to the data rich administrative records and the historical SCF data available, there were many opportunities to exploit the use of graphics in the sample design process. First, a series of scatter plots was used to view the results from the 1992 SCF. For the list sample, it was important to know how well the wealth index stratifier predicted the wealth reported on the survey. Based on this information, new stratifiers were created using regression models (see Frankel and Kennickell [1995].) In order to evaluate the possible stratifiers for use in the 1995 sample design, several types of plots were viewed. Included as Figure 1 is a scatter bi-plot that compares two regression models, the wealth index and financial income. In the 1992 SCF, financial income was highly correlated with nonresponse. Thus, it was comforting that the models appear to be an improvement over the wealth index, at least in the sense that they are more correlated with financial income.

Editing and Imputation

After initial data cleaning, much of the editing and imputation processing is done in parallel. Plots that are examined to find outlying imputations also may reveal problems related to data editing. The detection of outliers or strange patterns in the raw responses is extremely important. This may indicate that data editing, either at the vendor level or the in-house level, has not been totally effective. Perhaps, there is a problem with the wording of certain questions in the survey instrument, or there could be significant interviewer errors. Over time and subsequent surveys, all of these aspects of the survey are strengthened in part due to the graphical review of the data.

One of the main plots used in the editing and imputation process is a scatter plot of logged (base 10) continuous or discrete variables displaying bands which separate the list cases for 7 different wealth strata from the area-probability cases. These plots distinguish raw data points from imputed data points and range-card responses by the use of special symbols. Not only are these plots useful in spotting potentially erroneous data and "funny" imputations, but the distribution of imputed values is easily compared with the distribution of non-imputed values.

One typical such graph is shown in Figure 2. A box indicates a range-card response, a diamond indicates an imputed value, and a dot indicates a raw response value. The different strata are easily identifiable with AP cases occupying the lowest band in the plot and the wealthiest list cases the highest band in the plot. The wealth strata begin with the second band from the bottom. The list data points are further stratified vertically by financial income (highest income level appears highest within wealth strata.) The AP cases are adjusted randomly to produce a vertical effect.

Another type of graph used is a plot of one variable versus another variable where different symbols are used to show which variables have been imputed. In Figure 3, "<" indicates that the y-axis variable has been imputed, "V" indicates that the x-axis variable has been imputed, "v" indicates that both variables have been imputed, and "." indicates that neither variable has been imputed. These plots are useful in looking at imputation patterns in a bivariate setting.

Weighting and Analysis

During the process of computing analysis weights, the resulting weighted data needs to be carefully reviewed. After basic weights are constructed, graphs that show the influence of each case for a particular variable can be constructed. These types of plots can indicate whether a variable for a specific observation is contributing too much to the overall weighted total for that variable. If a problem is detected, then these plots might reveal whether the weight is too large and/or whether the value of the variable is too large (e.g. due to an editing or imputation error). Such a plot is shown in Figure 4. The log of the variable of interest is plotted versus the cumulative percent of the weight (cdf1), and the cumulative percent of the weighted value, i. e. the weight multiplied by the variable of interest (cdf2). The cdf2 plot shows a large gap. One case is contributing about 20 percent of the weighted total. Additionally, the cdf1 plot shows a significant gap, about 6 percent, indicating that the weight is quite large. These types of influence

Figure 3. Scatterplot of Two Variables to Investigate Imputation Process

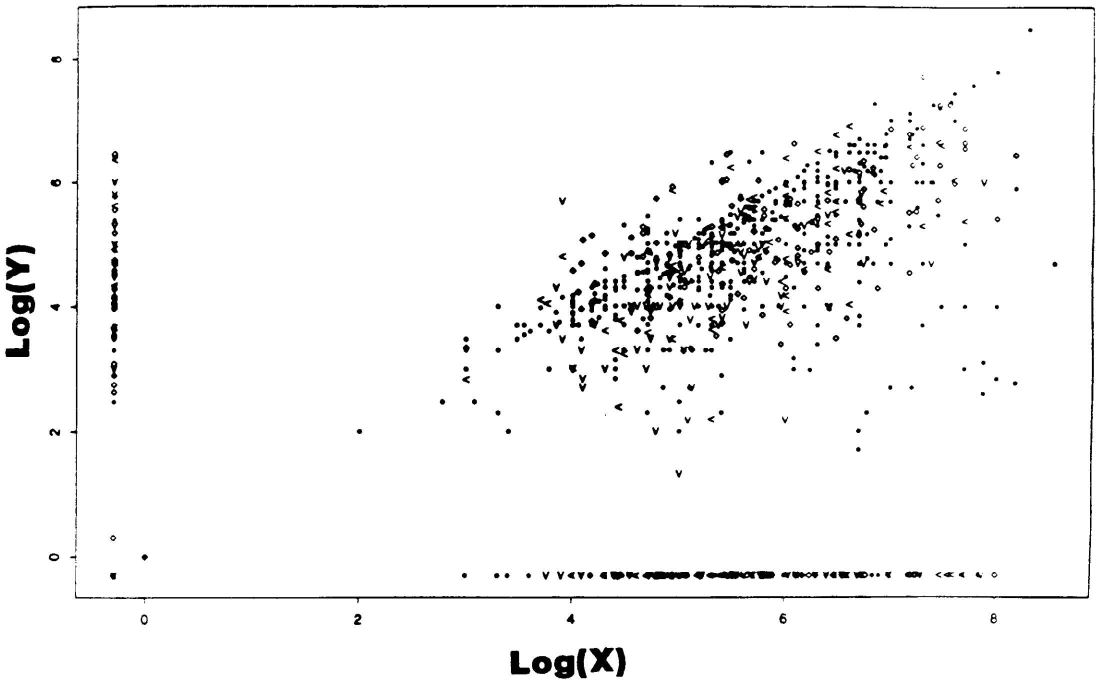


Figure 4. CDF Plots to Investigate the Influence of the Computed Weights

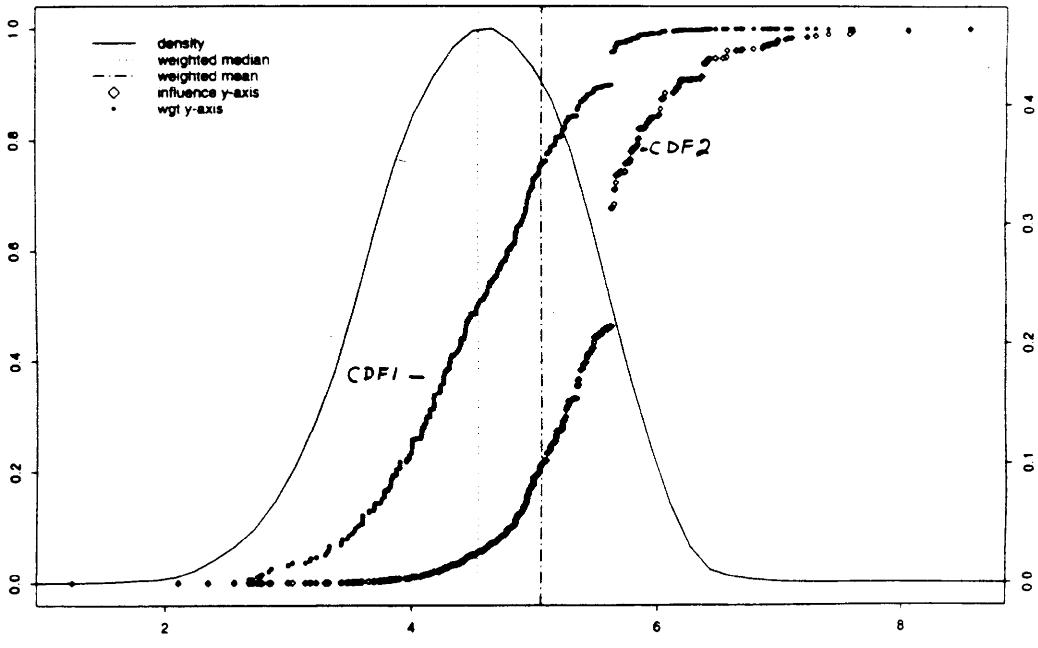


Figure 5. Scatter Plot to Investigate Weighting Process

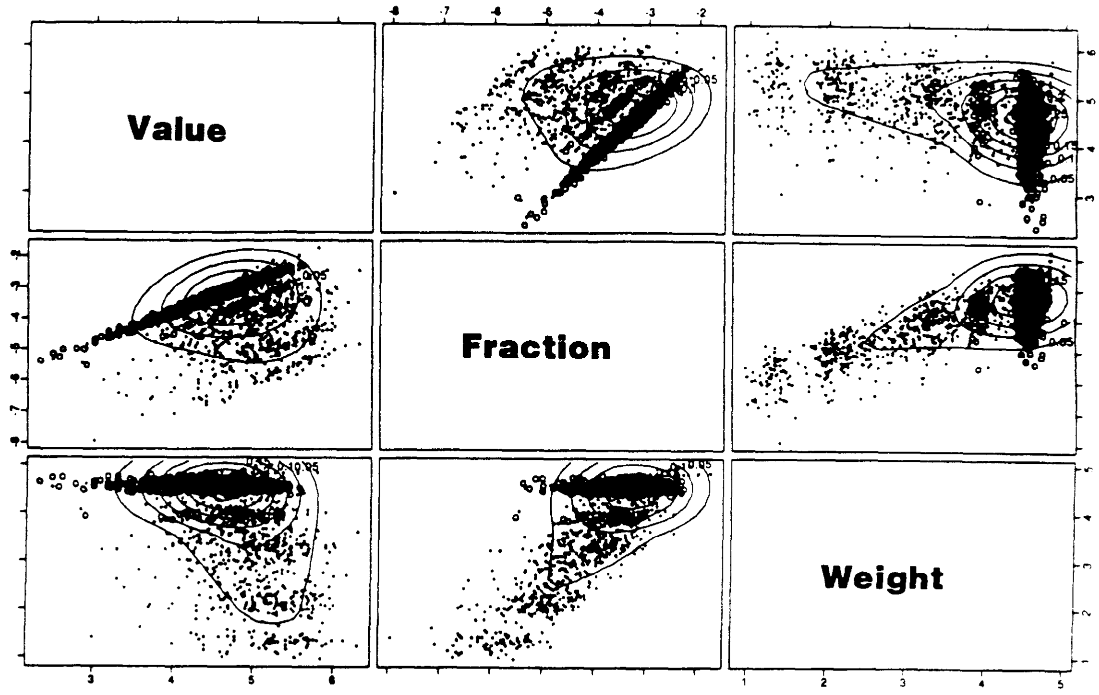
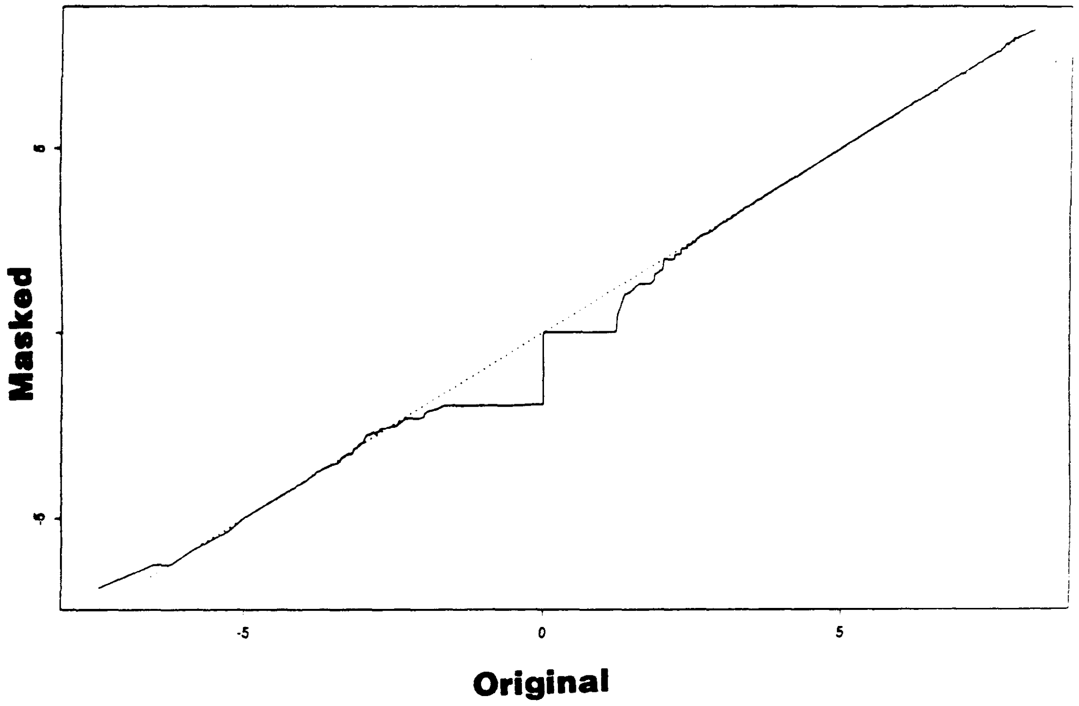


Figure 6. QQ Plot Comparing the Net Worth Distributions Using Masked Data vs Original Data



plots were also useful when doing data analysis by different groups or family cohorts.

Another graph used to evaluate the weights is a scatter bi-plot of the value, weight and percent contribution to the weighted total (fraction). This plot, shown in Figure 5, is useful in viewing the distribution of the weights for a chosen variable. Also, outlying weights can be detected.

Disclosure Adjustments

The main objective of the disclosure avoidance strategy of the SCF is to protect the respondent's identity while preserving the usefulness and integrity of the microdata. This leads to two main uses of graphics. First, scatterplots and many of the plots described above are used to search for 'unique' data points, overall and in various subgroups. Second, qq plots and other comparative plots are used to compare distributions computed using the original data versus those distributions computed using the masked data. Figure 6 shows such a qq plot of net worth. The aberrations around zero are due to the use of the log scale and do not represent significant differences. The upper tails of the distributions are remarkably similar. The lower tails differ due to some bounding of negative values. These plots are described in detail in Fries and Woodburn [1994].

Closing Comments

We have found graphics to be quite effective in helping us to understand and evaluate the processes used in the SCF. They provide a very thorough means of viewing the data. How many survey organizations can claim that they have looked at every variable to see how well the editing and imputation processes have worked? How often are plots used at the sample design stage? What better way to evaluate sampling weights than by looking at their distribution? How often is the effect of disclosure adjustments reported, not only in terms of means and totals, but considering the entire distribution? Overall, we strongly encourage survey practitioners to use graphics in all aspects of survey processing.

Acknowledgements

The authors would like to express their gratitude to all of the Survey of Consumer Finances staff for their support. A special thanks to Arthur Kennickell for his guidance and comments. Many thanks to Doug McManus, formerly of the SCF staff, who developed the S-Plus programs used to create the graphs shown here.

Bibliography

FRIES, G., and WOODUBRN, R. L. [1994], "The Challenges of Preparing Sensitive Data for Public Release," *Proceedings of the Section of Survey Research Methods, ASA.*

FRANKEL, M., and KENNICKELL, A.B., [1995], "Toward the Development of an Optimal Stratification Paradigm for the Survey of Consumer Finances," *Proceedings of the Section of Survey Research Methods, ASA.*

INTERNAL REVENUE SERVICE [1990], *Individual Income Tax Returns 1987,* Department of the Treasury, pp. 13-17.

HEERINGA, S., CONNOR, J. and WOODBURN, R. L., [1994], "The 1989 Survey of Consumer Finances, Sample Design Documentation," Working Paper, ISR, University of Michigan.

HOAGLIN, D. C. et. al. [1985]. *Exploring Data Tables, Trends, and Shape,* John Wiley and Sons, INc., pp. 432-442.

KENNICKELL, A.B. [1991]. "Imputation of the 1989 Survey of consumer Finances: Stochastic Relaxation and Multiple Imputation.: *Proceedings of the Section of Survey Research Methods, ASA.*

KENNICKELL, A.B., and MCMANUS, D.A., [1993]. "Sampling for Household Financial Characteristics Using Frame Information on Past Income," *Proceedings of the Section of Survey Research Methods, ASA.*

WILSON, O., and SMITH, W. J. Jr., (1983), "Access to Tax Records for Statistical Purposes," *Proceedings of the Section of Survey Research Methods,* American Statistical Association, pp. 591-601.