Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances

Arthur B. Kennickell*

November 26, 1997

Key words: Disclosure limitation, simulated data, multiple imputation.

Abstract

Recent developments in record linkage technology together with vast increases in the amount of personally identified information available in machine readable form raise serious concerns about the future of public use datasets. One possibility raised by Rubin [1993] is to release only simulated data created by multiple imputation techniques using the actual data. This paper uses the multiple imputation software developed for the Survey of Consumer Finances (Kennickell [1991]) to develop a series of experimental simulated versions of the 1995 survey data.

* Senior Economist and Project Director SCF. Please address correspondence to the author at Mail Stop 153, Board of Governors of the Federal Reserve System, Washington, DC 20551 U.S.A. Phone: 202-452-2247; fax: 202-452-5295; email: m1abk00@frb.gov . The views presented in this paper are those of the author alone and do not necessarily reflect those of the Board of Governors or the Federal Reserve System. The author wishes to thank Kevin Moore and Amy Stubbendick for a very high level of research assistance. The author is also grateful to Gerhard Fries, Barry Johnson, and R. Louise Woodburn for comments, and to Fritz Scheuren for encouragement in this project. Any errors remaining are the responsibility of the author alone.

Typically, in household surveys there is the possibility that information provided in confidence by respondents could be used to identify the respondent. This possibility imposes an ethical, and sometimes a legal, burden on those responsible for publishing the survey. Generally, it is necessary to review the data for items that could reveal the identity of individuals, and to filter the data made available to the public to minimize the degree of disclosure.[1] A recent issue of the *Journal of Official Statistics* (Vol. 9, no. 2, 1993) deals with many aspects of this problem.

The Survey of Consumer Finances (SCF), which is the focus of this paper, presents two particularly serious disclosure risks. First, the survey collects sensitive data on families' balance sheets and other aspects of their financial behavior. Second, the SCF oversamples wealthy families, who might be well-known, at least in their localities.

There is a growing belief that publicly available records, such as credit bureau files, real estate tax data, and similar files make it increasingly likely that an unscrupulous data user might identify survey respondents.[2] Several protective strategies have been proposed, but many of these proposals—truncation, simple averaging across cells, random reassignment of data, etc.—raise serious obstacles for many of the analyses for which the SCF is designed. The prospect of either being unable to release any information, or having to alter the data in ways that further restrict their usefulness makes it imperative that we explore alternative approaches to disclosure limitation.

Most disclosure limitation techniques attempt to release some transformation of the data that preserves what is deemed to be the important information. Taking this idea to one farsighted conclusion, Donald Rubin has suggested on several occasions creating an entirely synthetic dataset using techniques of multiple imputation (see, e.g., Rubin [1993]).[3] My impression is that most people have viewed the idea of completely simulated data with at least suspicion.[4] Such an exercise also presents considerable technical difficulties. However, even if it is not possible to create an ideal simulated dataset, we may learn something from the attempt to create one. This paper describes several experimental explorations in this direction.

Multiple imputation has played an important role in the creation of the public datasets for the SCF since 1989. In both the 1989 and 1992 surveys, a set of sensitive monetary variables was selected for a set of cases, the responses to those variables were treated as range responses (rather than exact dollar responses), and they were multiply-imputed using the FRITZ software developed for the SCF. The approach has been

---

[1] As Fienberg [1997] argues, releasing any information discloses something about the respondent, even if the probability of identification is minuscule.

[2] In his address to the Conference on Record Linkage held in Washington, DC in March 1997, Ivan Fellegi suggested if such problems continue to grow rapidly, we may no longer be able to create public use datasets as we know them now.

[3] For example, Rubin [1993] says "Under my proposal, no actual unit's confidential data would ever be released. Rather, all actual data would be used to create a multiply-imputed synthetic microdata set of artificial units..."

[4] However, Fienberg and Makov [1997] have proposed creating simulated data for the purpose of evaluating the degree of disclosure risk in a given dataset and Feinberg, Steele and Makov [1996] have examined the problem of simulating categorical data.

broadened in the 1995 survey based on the work reported here. In the experiments discussed in this paper, several approaches are taken to imputing all of the monetary values in the 1995 SCF.

The first section of the paper provides some general information on the content and sample design of the SCF and gives a review of the past approach to disclosure review. Because of the importance of imputation in the work reported here, the second section reviews the FRITZ imputation model. The third section discusses the special data used for the experiments, and presents some descriptive results. A final section summarizes the findings of the paper and points toward future work.

1 The 1995 Survey of Consumer Finances

The SCF is sponsored by the Board of Governors of the Federal Reserve System in cooperation with the Statistics of Income Division of the IRS (SOI). Data collection for the 1995 SCF was conducted in the second half of 1995 by the National Opinion Research Center (NORC) at the University of Chicago. The interviews, which were performed largely in person using computer-assisted personal interviewing (CAPI), required an average of 90 minutes—though some took considerably longer. The final dataset for 1995 contains information on 4,299 households.

Because the major focus of the survey is household finances, the SCF includes questions about all types of financial assets, tangible assets, and debts. To encourage accurate reporting and to serve the analytic goals of the survey, the questions are at a high level of detail. To provide adequate contextual variables for analysis, the SCF obtains data on the current and past jobs of respondents and their spouses or partners, their pension rights from current and past jobs, their marital history, their education, and other demographic characteristics. Data are also collected on past inheritances, future inheritances, charitable contributions, attitudes, and many other variables.

Although the combination of such a broad array of variables alone is sufficient cause to warrant intensive efforts to protect the privacy of the survey participants, an aspect of the SCF sample design introduces further potential disclosure problems. The survey is intended to be used for the analysis of financial variables that are widely distributed in the population—e.g., credit card debt and mortgages—and variables that are more narrowly distributed—e.g., personal businesses and corporate stock. To provide good coverage of both types of variables, the survey employs a dual-frame design (see Kennickell and Woodburn [1997]). In 1995, a standard multi-stage area-probability sample was selected from 100 primary sampling units across the United States (see Tourangeau, Johnson, Qian, Shin and Frankel [1993]). This sample provides good coverage of the broadly-distributed variables. A special list sample was designed to oversample wealthy households. Under an agreement between the Federal Reserve and SOI, data from the Individual Tax File (ITF), a sample of individual tax returns specially selected and processed by SOI, are made available for sampling.[5]

The area-probability design raises no particularly troubling issues beyond the need to protect geographic identifiers that is common to most surveys. However, the list sample raises two distinct problems. First, it increases the proportion of respondents who

[5]Use of the ITF for the SCF is strictly controlled to protect the privacy of taxpayers.

are wealthy. Such people are likely to be well-known, at least in their locality, and because of the relatively small number of such people, it is more likely that data users with malicious intent could match a respondent to external data if sufficient information were released in an unaltered form. Second, because SOI data have been used in the sample design, there is a legal requirement that SCF data made public be subjected to a disclosure review similar to that required for the release of the public version of the ITF.

Generally, the SCF data have been released to the public in very conservative stages, thus allowing time to deal with more complex disclosure issues. However, it is important to keep in mind that once a variable has been released, no amount of disclosure review can retrieve the information, and it can be much more difficult to add variables later because of the possible interactions of sensitive variables. However, this strategy has worked well in the past both in accommodating the most pressing needs of data users and in allowing users to build a case for including additional variables.

Table 1: Rounding of Continuous Variables

Data range	*Rounded to nearest*
>1 million	10,000
10,000 to 1 million	1,000
1,000 to 10,000	100
5 to 1,000	10
-5 to -1,000	10
-1,000 to -10,000	100
-10,000 to -1 million	1,000

Negative numbers smaller than -1 million truncated at -1 million
Negative numbers between -1 and -5 unaltered

For the 1992 SCF, the last survey for which final data had been released when this work was begun, the internal data were altered in the following ways for release.[6] First, geographic indicators, which were released at the level of the nine Census divisions, were altered: Observations were sorted and aligned by key characteristics, and location was swapped across similar cases. Second, unusual categories were combined with similar categories—e.g., among owners of miscellaneous vehicles, the categories "boat," "airplane," and "helicopter" were combined. Third, a set of cases with unusual wealth or income was chosen along with a random set of other cases. For these cases, key variables for which complete responses were originally given were multiply imputed subject to range constraints that ensured the outcomes would be close to the initially reported values. Fourth, a set of other unspecified operations were performed to increase more broadly the perceived uncertainty associated with all variables in every observation; these operations affected both actual data values and the "shadow" variables in the dataset that describe the original state of each variable.[7] As a final step, all continuous variables were rounded as shown in table 1. It is impossible to tell with certainty from the variables in the public dataset which variables may have been altered and how they were altered.

---

[6]See Fries, Johnson and Woodburn [1997a] for an overview of the disclosure review for the 1995 SCF.

[7]For each final variable, the shadow variables contain information about whether the variable was inapplicable, reported completely or as one of a large number of types of range outcomes, missing for various reasons, or altered by one of a variety of editing processes.

A similar strategy is being followed for the 1995 SCF. The one significant change is in the imputation of data for the cases deemed "sensitive" and the random subset of cases described in step three. For the 1995 survey, all monetary data items in the selected cases will be imputed. Depending on the reception of the data by users, this approach may be extended in the 1998 SCF.

2 FRITZ Imputation Model

Because the principal evidence reported in this paper turns critically on the imputation of monetary variables, it is important to outline some of the more important characteristics of the FRITZ model, which was originally developed for the imputation of the 1989 SCF (see Kennickell [1991]) and has been updated for each round of the survey since then. This discussion focuses on the imputation of continuous variables.

Figure 1: Hypothetical Missing Data Patterns

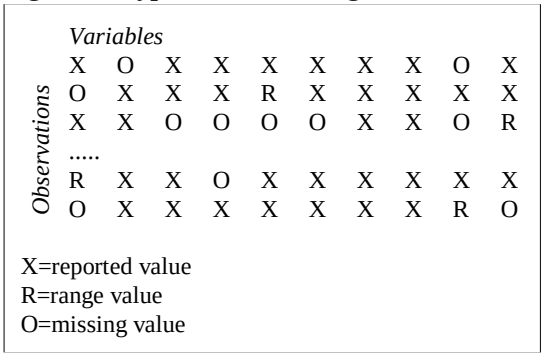

Figure 1 shows a hypothetical set of observations with various types of data given by respondents. In the figure, "X" represents complete responses, "R" symbolizes responses given as a type of range, and "O" indicates some type of missing value. In the SCF, there is a lengthy catalog of range and missing data responses, and this information is preserved in the shadow variables.[8] Data may be missing because the respondent did not know the answer, refused to answer, did not answer a question of a higher order in a sequence, because of recording errors, or other reasons.

The FRITZ system is an iterative multiple imputation model based on ideas of Gibbs sampling. The system acts on a variable-by-variable basis, rather than simultaneously drawing a vector of variables.[9] Within a given iteration, the most generally applied continuous variable routine is, in essence, a type of randomized regression, in which errors are assumed to be normally distributed.[10]

One factor that distinguishes the model from the usual description of randomized regression imputation models is the fact that the FRITZ model is tailored to the missing data pattern of each observation. In figure 1, all of these patterns shown are different, and they are not monotone (Little [1983]). For most continuous variables, the program generates a covariance matrix for a maximal set of variables that are determined to be relevant as possible conditioning variables. Once a variable has been imputed, its

[8]The collection of range data in the 1995 SCF is described in detail in Kennickell [1997].

[9]For an excellent example of a simultaneously determined system, see Schafer [1995]. Geman and Geman [1984] discuss another type of structure involving data "cliques."

[10]In general, continuous variables are assumed to follow a conditional lognormal distribution.

value is taken in later imputations as if it were originally reported by the respondent. For a given case, the model first determines whether a particular variable should be imputed. Given that the variable should be imputed, the FRITZ model computes a regression for the case using the variables in the maximal set that either are not originally missing or are already imputed within the particular iteration for the case. Finally, the model draws from the estimated conditional distribution until an outcome is found that satisfies any constraints that may apply. Constraints may take several forms. When a respondent has given a range response to a question, FRITZ uses the range to truncate the conditional distribution. Constraints may also involve cross-relationships with other variables, or simply prior knowledge about allowable outcomes. Specification of the constraints is very often the most complex mechanical part of the imputations.

In a given imputation, variables which were originally reported as a range but are not yet imputed within the iteration, are given special treatment that is particularly important for the experiments reported here. Range responses often contain substantial information on the location of the true value, and one would like to use this knowledge in imputation. In the ideal, it is not difficult to write down a general model that would incorporate many types of location indicators. However, in practice, simple models of this sort would quickly exhaust the degrees of freedom available in a modestly sized survey like the SCF. In practice, we adopt a compromise solution. Values reported originally as ranges are initialized at their midpoints, and these values are used as conditioning variables for other imputations until a value within the range is imputed.

The FRITZ model produces multiple imputations. For simplicity, each original observation is replicated five times, and each of these "implicates" is imputed separately. This arrangement allows users to apply standard software to the data.

The iteration process is fairly straightforward. In the first iteration, all the relevant population moments for the imputation models are computed using all available data. As imputations progress in that iteration, the covariance estimation is based on increasingly "complete" data. In the second iteration, all population moments are computed using the first iteration dataset, and a new copy of the dataset is progressively "filled in." In each successive iteration, the process is similar. Generally, the distribution of key imputations changes little after the first few iterations. Because the process is quite time-consuming, the model for the 1995 SCF was stopped after six iterations.

3 Experiments in Imputation for Disclosure Limitation

In this section, I report on three experiments using multiple imputation for disclosure avoidance (summarized in figure 2). In these experiments, every monetary variable for every observation in the survey was imputed.[11] In the first experiment, all complete reports of dollar values were imputed as if the respondent had originally reported ranges which ran from ten percent above the actual figures to ten percent below that figure. In keeping with our usual practice of using midpoints of ranges as proxies

[11]There are 480 monetary variables in the SCF, but it is not possible for a given respondent to be asked all of the underlying questions.

Figure 2: Design of Experiments

Experiment	*Range constraints*	*Use original value as initial location indicator*
1	±10%	Yes
2	None	Yes
3	None	No

for location indicators in imputation, the original values were retained until the variables were imputed. The second experiment also retained the reported value for conditioning, but imposed no range constraints on the allowed outcomes other than those required for cross-variable consistency. The third experiment treated the original values as if they were completely missing (that is, they were unavailable as conditioning variables) and, like the second experiment, imposed no prior bounds on the imputations; other monetary responses that were originally reported as ranges were also treated as completely missing values for purposes of conditioning, but their imputed values were constrained to lie within the reported ranges.

For several reasons, these experiments fall short of Rubin's ideal that one impute an entire dataset—possibly even starting by conditioning only on distributional data external to the actual sample. First, the experiments deal only with the dollar variables in the SCF. Second, all complete responses other than monetary responses are used as conditioning variables. Third, imputations of original range responses are constrained to lie within the reported ranges, even in experiment three. Finally—and probably most importantly—the results are specific to the particular specification of the FRITZ model. Inevitably there are deep compromises of theory made in implementing almost any empirical system. For imputation, such compromises may be less pressing when the proportion of missing data is relatively small, as is usually the case in the SCF. These compromises may cause larger distortions when much larger fractions of the data are imputed. A key question in evaluating the results here is how well the system performs under this more extreme condition. Because we also have the originally reported values, it is possible to make a direct evaluation of the performance of the model.

Despite the shortcomings of the three experiments, they seem very much in the sprit of Rubin's proposal. Because the experiments show the effects of progressively loosening the constraints on imputation, I believe the results should provide useful evidence in evaluating the desirability of going further in developing fully simulated data.

The mechanical implementation of these experiments was reasonably straightforward. In the first experiment, the shadow variables of all complete reports of dollar values were set to a value which would normally indicate to the FRITZ model that the respondent had provided a distinct dollar range, and the range values ten percent above and ten percent below the reported value were placed in an external file normally used for this purpose. In the second and third experiments, a special value was given to the shadow variable to indicate that there were no range constraints on the imputations other those that enforce cross-variable consistency. In experiments one and two, the initial values of complete responses were left in the dataset at the beginning of imputation; during the course of imputations, these values were used for conditioning until they were replaced by an imputed value, which was used to condition subsequent imputations. In experiment three, values originally reported completely were set to a missing value, and the usual midpoints of range responses were also set to a missing

value. Thus, no dollar variables in the third experiment were available for conditioning until they were imputed. In each of the experiments, the imputations were treated as if they were the seventh iteration of the SCF implementation of FRITZ. Thus, estimates of the population moments needed for the model were computed using the final results of the sixth iteration.

In the absence of technical problems—far from the case with the work for this paper as the imputation system was subject to a massively larger than normal stress—each version of the experiment would require approximately three weeks to run through the entire dataset. A potentially much large amount of time would be required to debug the associated software. To make this range of experiments feasible, a compromise has been adopted here. The first of the eight modules of the SCF application of FRITZ was run for all of the experiments. This module deals largely with total household income and various financial assets.

Figures 3 through 6 show descriptive plots of data from the three experiments for the following four variables: total income, amount in the first savings account, the amount of directly held corporate stock, and the total value of financial assets.[12] The first three of these variables are intended to span a broad set of types of distributions; total financial assets, a variable constructed from many components, is included to show the effects of aggregating over the potentially large number of responses to questions about the underlying components. The impression from looking at a broader set of variables is similar. Each of the figures is divided into two sets of three panels. The top three panels show the distribution for experiments one through three, of the (base-10) logarithm of the originally reported values less the average across the five implicates of the logarithm of the corresponding imputed values ("bias"), where the distribution is estimated as an unweighted average shifted histogram (ASH). The bottom three panels are ASH plots for the three experiments, of the distribution over all cases of the standard deviation of the multiply-imputed values within observations.

Not surprisingly, the distribution of bias for experiment one has a mode at approximately zero for all the variables. In the case of income, savings balances, and stocks, the distribution of bias is fairly concentrated, with the 10th and 90th percentiles of the distribution corresponding to a bias of only about 5 percent (±0.02 on the scale shown) . The distributions of bias for savings accounts and stocks are relatively "lumpy,"

---

[12]The sets of observations underlying the charts include only respondents who originally gave a complete response for the variable, or, in the case of financial assets, who gave complete responses for all the components of financial assets. For many sub-models of the SCF implementation of FRITZ, general constraints are imposed for all imputations to ensure values that are reasonable (e.g., amounts owed on mortgage balloon payments must be less than or equal to the current amount owed); in the actual data, these constraints are occasionally violated for reasons that are unusual, but possible. When reimputing these values subject to dollar range constraints in experiment one, a small number of imputations violated the bounds imposed. To avoid major restructuring of the implementation of the FRITZ model for the experiments, these instances are excluded from the comparisons reported here. In each of the figures, the set of observations is the same across all six of the panels. For the income plots, households originally reporting negative income have been excluded.

largely reflecting the smaller samples used to estimate these distributions: about 1,100 observations were used for the savings account estimate and about 800 observations were used for the stock, but about 2,900 were used to estimate the distribution for total income. Reflecting an averaging over possibly many imputations for each of about 2,300 observations, the distribution of bias for total financial assets is quite smooth. In every case shown, there is some piling up of cases at the outer bounds corresponding to ±10 percent (about ±0.04 on the log scale). The FRITZ model is allowed to draw as many as 400 times from the predicted conditional distribution of the missing data before selecting the nearest endpoint of the constraint.[13] Thus, it is likely that these extreme observations are ones for which the models do not fit very well. Not surprisingly, examination of selected cases suggests that these observations are more likely to have unusual values for some of the conditioning variables in the imputation models. The median variability of the imputations within implicates shown by the ASH plots of the distributions of standard deviations, is about ±6 percent for income, savings accounts, and stocks. The variability within implicates is substantially lower for the sum of financial assets, reflecting offsetting errors in imputation.

In the second experiment, the relaxation of the simple range constraint in experiment one has the expected effect of increasing the variability of the bias, and increasing the standard deviation of imputations across implicates. In the case of total household income, the bias corresponding to the 90th percentile of the bias distribution jumps to about 25 percent. The effect is even larger for the other variables (the bias is nearly 300 percent at the 90th percentile for total financial assets). It is somewhat surprising just how much these values increase given that the imputations are potentially conditioned on a large number of reported values.[14]

In the third experiment with the removal of the reported values used for initial conditioning in experiment two, the range of the bias rises further. The 90th percentile of the bias distribution is about 140 percent for total income, and about 400 percent for total financial assets.

Because these results are reported on a logarithmic scale, it is possible that they could be unduly influenced by changes that are small in dollar amounts, but large on a logarithmic scale. The data do not provide strong support for this proposition. For income, scatterplots reveal that the logarithmic bias appears to be approximately equally spread at all levels of income for experiments one and two.[15] In the third experiment, the dominant relationship is similar, but there are two smaller groups that deviate from the pattern: a few dozen observations with actual incomes of less than a few thousand dollars are substantially over-imputed on average, and a somewhat larger number of observations

---

[13]By default, the range from which values are drawn is globally truncated at 1.96 standard errors above and below the conditional mean (the 95 percent confidence region under normality). Thus, the model is disallowed from generating relatively unusual outcomes.

[14]For example, total income is the first variable imputed, and all reported values (or midpoints of ranges) for variables included in the model for that variable are used to condition the imputation.

[15]For disclosure reasons, the scatterplot supporting this claim cannot be released.

with actual incomes of more than $100,000 are substantially under-imputed. The data suggest similar relationships across the experiments for the other variables as well.

To gauge the effects of the experiments on the overall univariate distributions of the four variables considered, figures 7-10 show quantile difference (QD) plots of the mean imputations across implicates and the reported values, both on a logarithmic (base-10) scale. A QD plot shows the difference in the values of two distributions at a common quantile point.[16] To focus on the changes in the actual observations, the plots here are unweighted; using weights does not substantially alter the impression one gains from the plots. Across these variables, the distributions are barely affected by experiment one. In the second experiment, in all cases there is an over-prediction of values at the bottom of the distributions. Results for the rest of each distribution are more mixed: outcomes are fairly close for income and financial assets, and there is a general under-prediction for savings accounts and stocks. In the third experiment, the correspondence between the two distributions deteriorates markedly, though the distribution of total financial assets appears the most resilient.

Univariate and simple bivariate statistics are important for many purposes, but for the SCF and many other surveys, the most important uses of the data over the long run are in modeling. Table 1 presents (unweighted) estimates of the coefficients of a simple linear regression of the logarithm of total household income on dummy variables for ownership of various financial assets and the log of the maximum of one and the value of the corresponding asset. This model has no particular importance as an economic or behavioral characterization. It is intended purely as a descriptive device designed to examine the effects of the variation across the experiments on the partial correlations of a set of variables imputed in all the experiments. The model is estimated using the final version of the data and the data from the three experiments. For each dataset, two subsets of observations are used: first, all observations in each dataset regardless of whether the variables included were originally reported completely by the respondent, and second, only cases for which every variable in the model was originally reported completely.

Relative to the estimates based on the final data, the estimates using data from the experiments perform fairly well, though there are some type one and type two errors in the classification of significance, and there is a curious (significant) change of sign for the coefficient on the ownership of whole life insurance with the data from experiments one and two. Interestingly, the estimates using data from experiment three appear to be closest overall to the base estimates. The R^2 of the regressions changes little except in the third experiment, where this value drops about 10 percent. While the experimental estimates overall are not markedly worse than the results from the original data, it would be very surprising if the outcome were otherwise. However, such regressions are only the

---

[16]These plots contain the same information as a quantile-quantile (Q-Q) plot. In essence, a QD plot is a Q-Q plot rotated by 45 degrees, with the horizontal axis relabeled with quantiles instead of the nominal level of the reference distribution. For comparison, an appendix contains the Q-Q plots corresponding to the QD plots in figures 7-10.

start of what many economists would consider applying to the data, and it is possible that more complex models or methods of estimation would give a different impression.

4 Summary and Future Research

By design, experiment one is virtually guaranteed to induce minimal univariate distortions, but it also leaves the outcomes near the original values. Unfortunately, simply knowing that an outcome is in a certain range may already be sufficient information to increase too much the probability of identifying some of the respondents in the SCF. My ex ante choice of contenders among the experiments was the second one, in which imputations condition on actual values, but there is no prior constraint on the outcomes that is connected to the original values. Ex post, I find the results relatively disappointing. The univariate outcomes reported for the third experiment look least attractive. Perhaps there may be practical ways of more globally constraining or aligning the outcomes of experiments two and three with the unadjusted data, but I suspect the choice of method would depend critically on a ranking of the importance of the types of analyses to be performed with the data. I hope that someone in the SCF group or elsewhere will be able to take the next step.

Ultimately, the experimental results reported in this paper say at least as much about the nature of the SCF imputations as they do about the possibility of creating a fully simulated dataset. Although the imputation models have been refined over three surveys, the results of experiments two and three, in particular, suggest that there is room for improvement. Indeed, a number of changes were instituted in the process of getting the experiments to produce meaningful data, and further changes will be implemented during the course of processing the 1998 SCF. Other options, including the possibility of using empirical residuals and looser constraints on the range of outcomes allowed, deserve further attention. However, I am not optimistic that there are many major feasible improvements in our ability to impute the SCF data to be discovered soon. There is a difference in what one can accept in imputing a relatively small fraction of the data and what is acceptable for the whole dataset. With fully simulated data, we are left with a difficult tradeoff of noise (however structured) and complex possibilities of bias against disclosure risk.

Disclosure limitation techniques have a Siamese twin in record linkage techniques. Close attention should be paid to advances in that area, to be aware of both new threats and new technologies that have a parallel use.

One technical question that appears potentially troublesome is how to estimate sampling error in a fully simulated dataset.[17] It is possible, in theory, to simulate records for the entire universe, but even in this case there would still be sampling variability in the imputations because the models were estimated on a sample. This variation may be a second order effect in most imputation problems, but we need to deal with the issue carefully if we expect to simulate all the data. Perhaps we could find an approximate solution in independently multiply imputing a set of observations using replicates of the actual data to generate the required sets of estimates of population moments.

---

[17]Fienberg, Steele and Makov [1996] also raise this question.

A large problem in planning all disclosure reviews is how to accommodate the needs (but not necessarily all the desires) of data users. I expect that users will express considerable resistance to the idea of completely simulated ("fake") data. Some statisticians may be troubled about how to address questions of estimating total error with such data. Among economists, there are substantial pockets of opposition to all types of imputation, and some researchers have raised carefully framed questions that need to be addressed equally carefully. However, given the choice between having no data or having data that are limited in some way, most analysts will likely opt for some information. Nonetheless, to avoid developing disclosure strategies that yield data that do not inform interesting questions, it may be important to engage users in the process where possible.

BIBLIOGRAPHY

Fienberg, Stephen E. [1997] "Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistics Research," working paper, Department of Statistics, Carneige Mellon University, Pittsburgh, PA.

_____ and Udi E. Makov [1997] "Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data," working paper, Department of Statistics, Carneige Mellon University, Pittsburgh, PA.

_____, Russell J. Steele, and Udi E. Makov [1996] "Statistical Notions of Data Disclosure Avoidance and their Relationship to Traditional Statistical Methodology: Data Swapping and Loglinear Models," *Proceedings of the 1996 Annual Research Conference and Technology Interchange*, U.S. Bureau of the Census, Washington, DC, pp. 87-105.

Fries, Gerhard, Barry W. Johnson, and R. Louise Woodburn [1997a] "Analyzing Disclosure Review Procedures for the Survey of Consumer Finances," paper for presentation at the 1997 Joint Statistical Meetings, Anaheim, CA.

_____, _____ and _____ [1997b] "Disclosure Review and Its Implications for the 1992 Survey of Consumer Finances," *Proceedings of the Section on Survey Research Methods*, 1996 Joint Statistical Meetings, Chicago, IL.

Geman, Stuart and Donald Geman [1984] "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-6, no. 6 (November), pp. 721-741.

Kennickell, Arthur B. [1991] "Imputation of the 1989 Survey of Consumer Finances," *Proceedings of the Section on Survey Research Methods*, 1990 Joint Statistical Meetings, Atlanta, GA.

_____ and R. Louise Woodburn [1997] "Consistent Weight Design for the 1989, 1992 and 1995 SCFs, and the Distribution of Wealth," working paper, Board of Governors of the Federal Reserve System, Washington, DC.

_____ [1997] "Using Range Techniques with CAPI in the 1995 Survey of Consumer Finances" *Proceedings of the Section on Survey Research Methods*, 1996 Joint Statistical Meetings, Chicago, IL.

Little, Roderick J.A. [1983] "The Nonignorable Case" in *Incomplete Data in Sample Surveys*, Academic Press, New York.

Rubin, Donald B. [1993] "Discussion of Statistical Disclosure Limitation," *Journal of Official Statistics*, vol. 9, no. 2, pp. 461-468.

Schafer, Joseph [1995] *Analysis of Incomplete Multivariate Data*, Chapman and Hall.

Tourangeau, Roger, Robert A. Johnson, Jiahe Qian, Hee-Choon Shin, and Martin R. Frankel [1993] "Selection of NORC's 1990 National Sample," working paper, National Opinion Research Center at the University of Chicago, Chicago, IL.

Figure 3a: ASH Plots of Distribution over Observations of Log_{10}(Actual Value)-Mean(Log_{10}(Imputation)) within Observations, Total Household Income, Experiments 1-3.

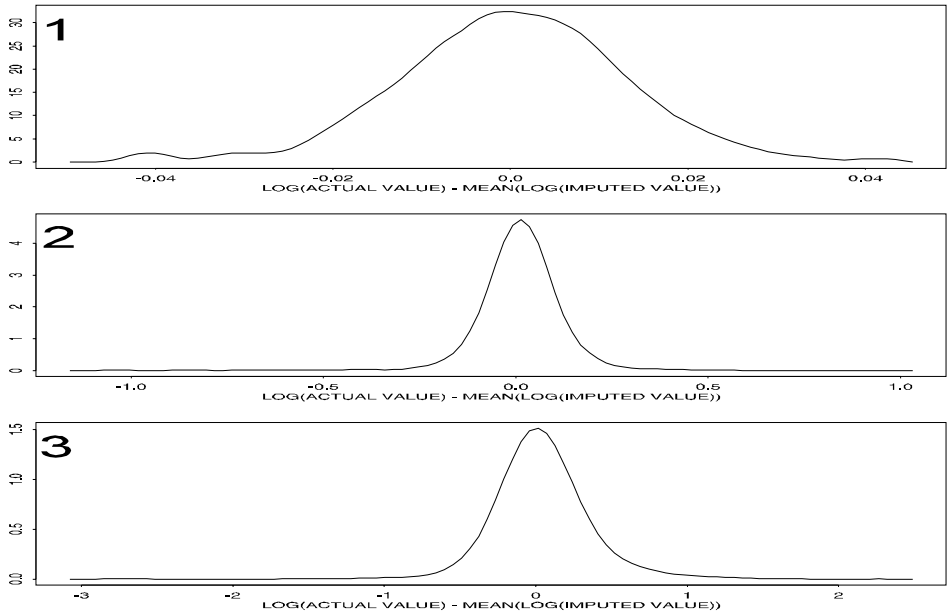

Figure 3b: ASH Plots of Distribution over Observations of the Standard Deviation of Log_{10}(Imputation) within Observations, Total Household Income, Experiments 1-3.

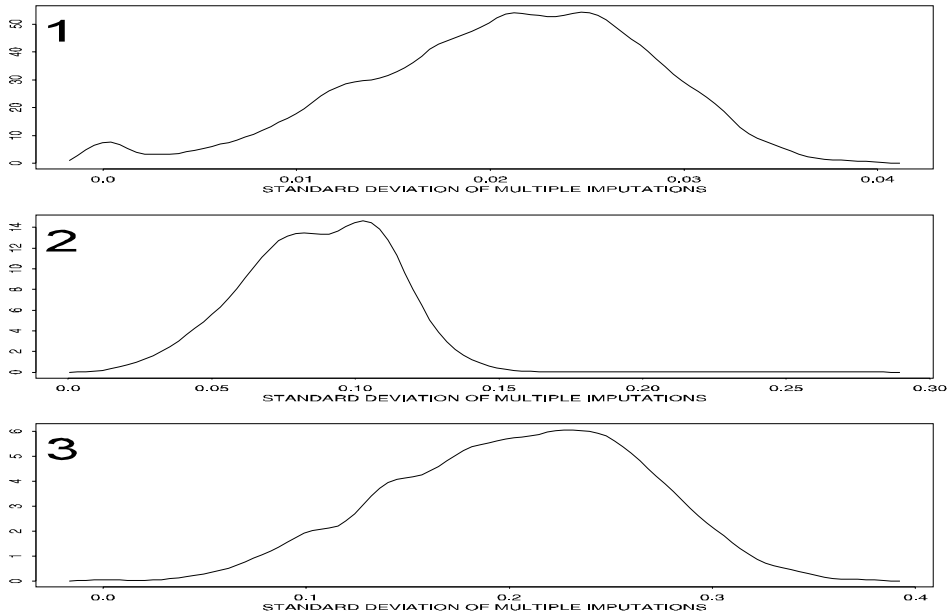

Figure 4a: ASH Plots of Distribution over Observations of Log_{10}(Actual Value)-Mean(Log_{10}(Imputation)) within Observations, Balance in First Savings Account, Experiments 1-3.

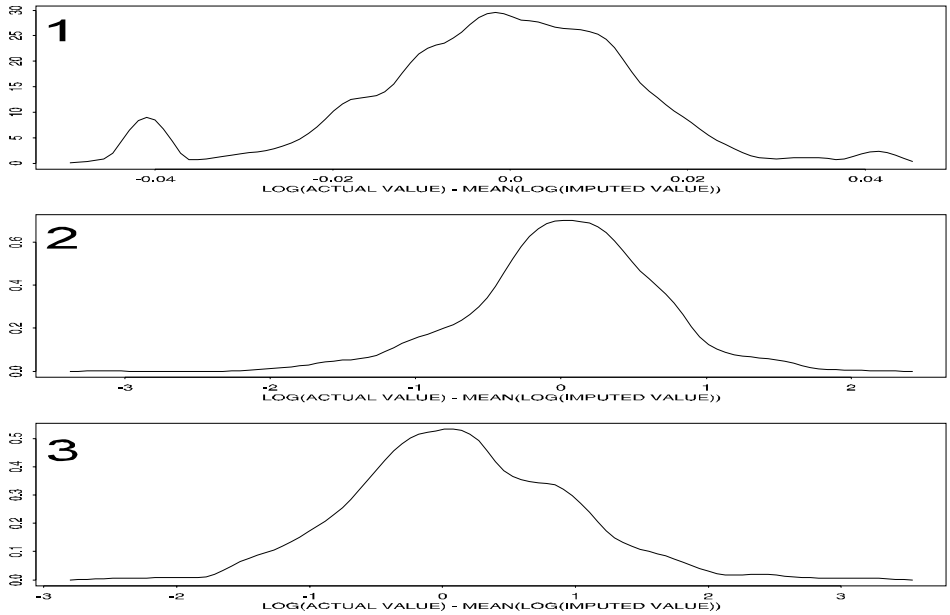

Figure 4b: ASH Plots of Distribution over Observations of the Standard Deviation of Log_{10}(Imputation) within Observations, Balance in First Savings Account, Experiments 1-3.

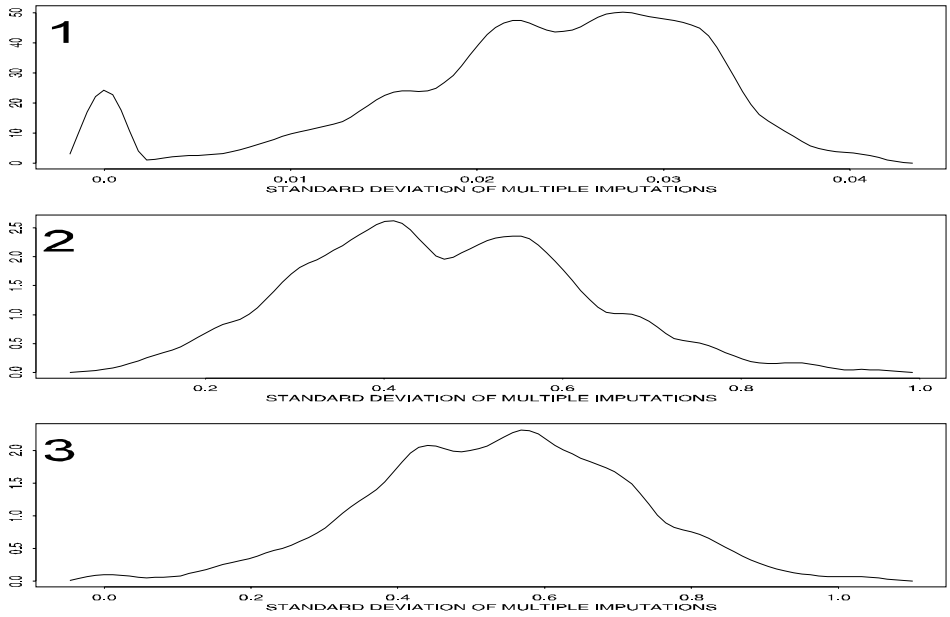

Figure 5a: ASH Plots of Distribution over Observations of Log_{10}(Actual Value)-Mean(Log_{10}(Imputation)) within Observations, Publicly Traded Stock, Experiments 1-3.

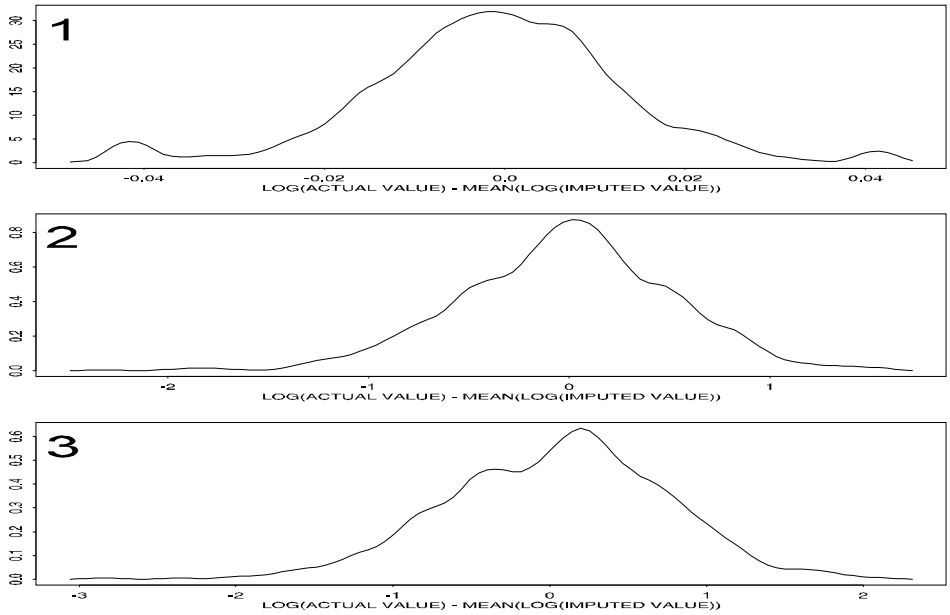

Figure 5b: ASH Plots of Distribution over Observations of the Standard Deviation of Log_{10}(Imputation) within Observations, Publicly Traded Stock, Experiments 1-3.

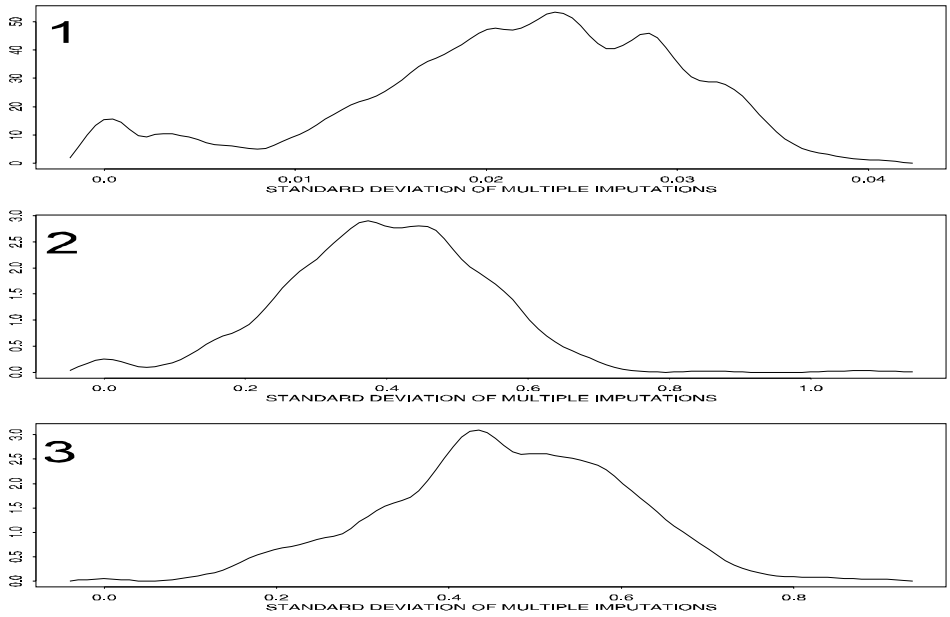

Figure 6a: ASH Plots of Distribution over Observations of Log_{10}(Actual Value)-Mean(Log_{10}(Imputation)) within Obervations, Total Financial Assets, Experiments 1-3.

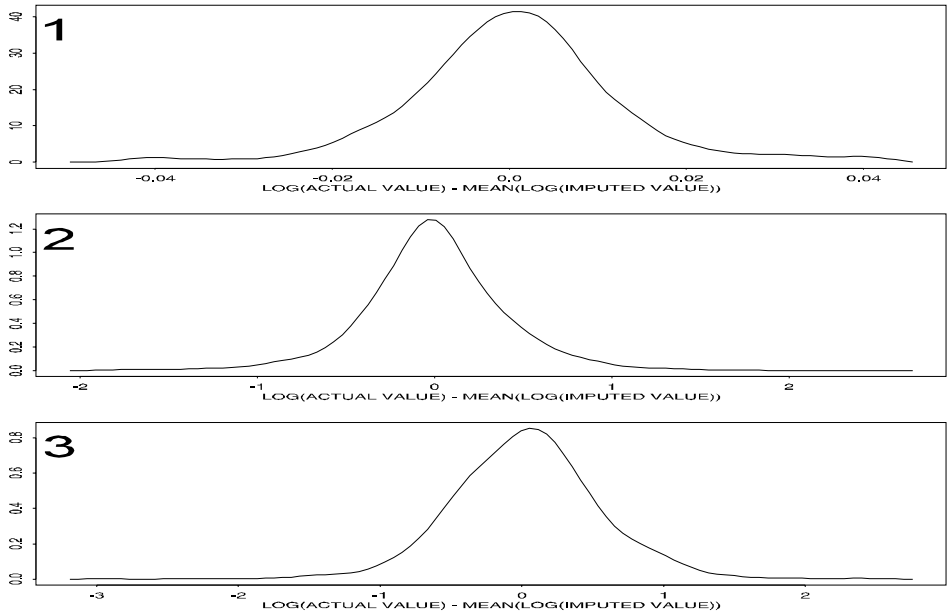

Figure 6b: ASH Plots of Distribution over Observations of the Standard Deviation of Log_{10}(Imputation) within Observations, Total Financial Assets, Experiments 1-3.

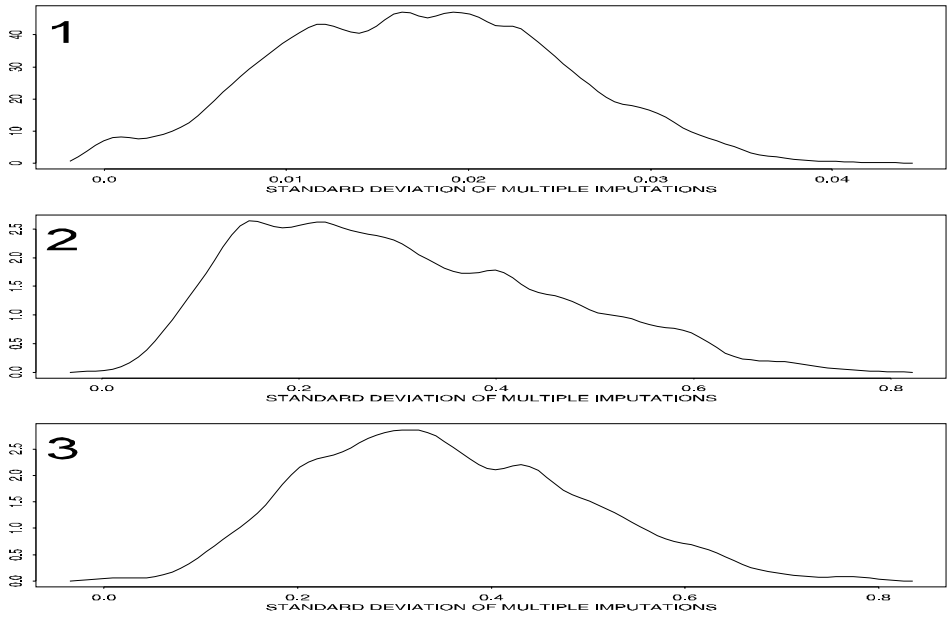

Figure 7: QD Plots of Actual Values Less Imputed Values, Total Household Income, Experiments 1-3.

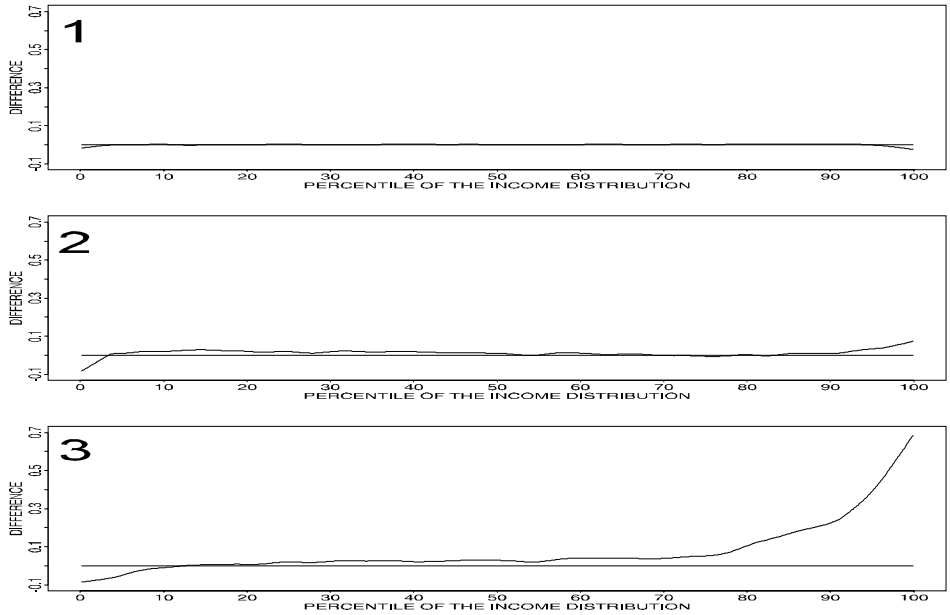

Figure 8: QD Plots of Actual Values Less Imputed Values, Balance in 1st Savings Account, Experiments 1-3.

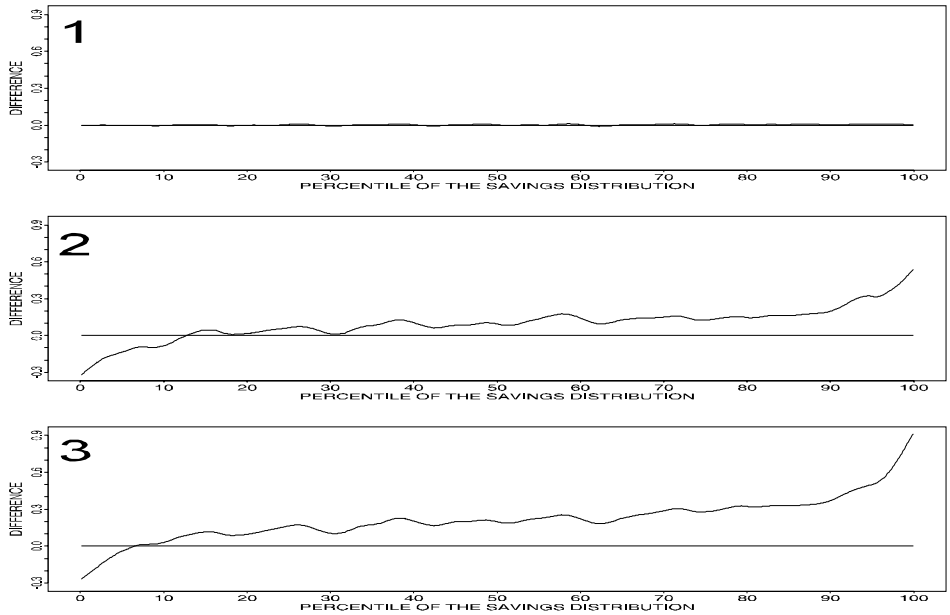

Figure 9: QD Plots of Actual Values Less Imputed Values, Publicly Traded Stock, Experiments 1-3.

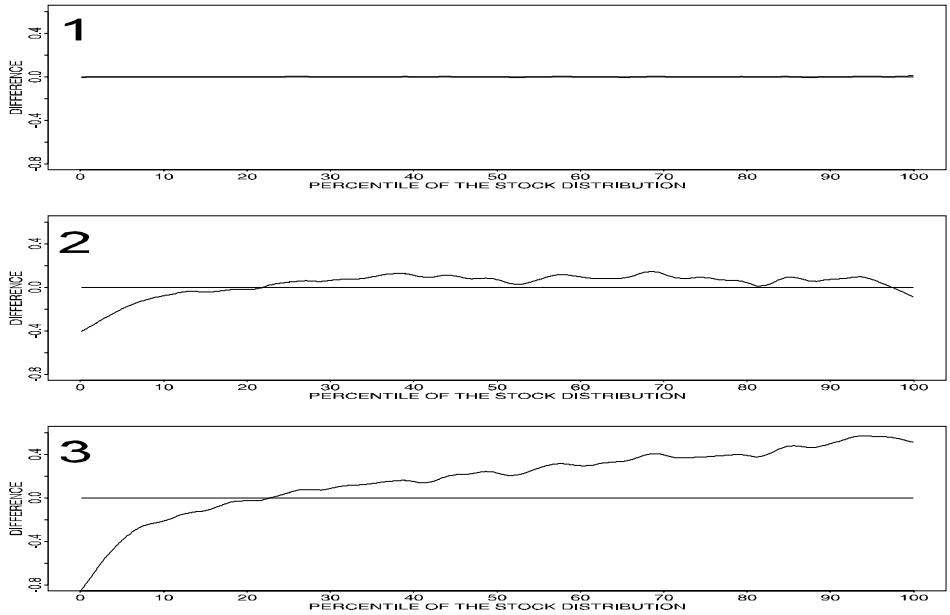

Figure 10: QD Plots of Actual Values Less Imputed Values, Total Financial Assets, Experiments 1-3

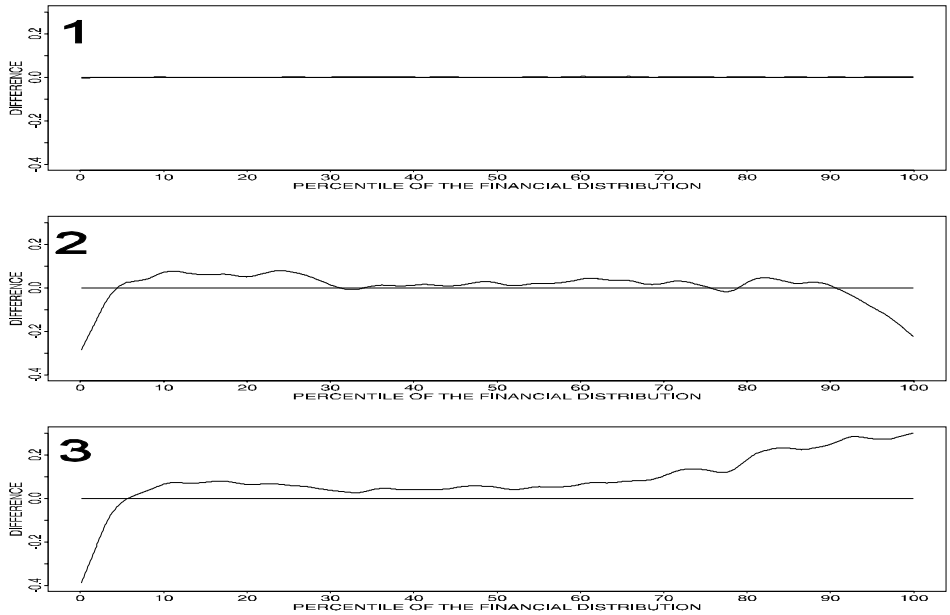

Table 3: Regression of Logarithm of Total Household Income on Various Variables, Original Data and Experiments 1-3, Using all Observations and Using Only Observations Originally Giving Complete Responses to all Variables in the Model.

	All observations included				Only complete responders included			
	Orig.	Exp. 1	Exp. 2	Exp. 3	Orig.	Exp. 1	Exp. 2	Exp. 3
Intercept	2.54+	2.56+	3.75+	2.64+	2.83+	3.42+	6.60+	2.83+
	0.76	*0.87*	*0.87*	*0.76*	*1.09*	*1.07*	*1.30*	*1.09*
Have checking	0.19+	0.25+	0.21+	0.18+	0.18+	0.19+	0.15+	0.17+
	0.03	*0.03*	*0.03*	*0.03*	*0.04*	*0.04*	*0.04*	*0.04*
Ln($ checking)	0.26+	0.30+	0.26+	0.25+	0.27+	0.27+	0.24+	0.26+
	0.01	*0.01*	*0.01*	*0.01*	*0.02*	*0.02*	*0.02*	*0.02*
Have IRA/Keogh	0.17+	0.18+	0.17+	0.16+	0.06	0.12	0.08	0.07
	0.05	*0.05*	*0.06*	*0.05*	*0.07*	*0.07*	*0.08*	*0.07*
Ln($ IRA/Keogh)	0.10+	0.11+	0.10+	0.10+	0.07+	0.10+	0.08*	0.07+
	0.02	*0.02*	*0.02*	*0.02*	*0.03*	*0.03*	*0.03*	*0.03*
Have savings acct.	0.01	0.01	0.01	0.01	-0.03	-0.02	-0.03	-0.03
	0.04	*0.04*	*0.04*	*0.04*	*0.05*	*0.05*	*0.05*	*0.04*
Ln($ savings acct)	0.03	0.03	0.04	0.03	0.00	0.01	0.01	0.00
	0.02	*0.02*	*0.02*	*0.02*	*0.02*	*0.03*	*0.03*	*0.02*
Have money mkt acct.	0.03	0.03	0.03	0.03	0.04	0.04	0.04	0.04
	0.07	*0.08*	*0.08*	*0.07*	*0.10*	*0.12*	*0.11*	*0.09*
Ln($ money mkt acct.)	0.03	0.00	-0.02	0.03	0.05	0.01	-0.02	0.05
	0.03	*0.03*	*0.03*	*0.03*	*0.04*	*0.05*	*0.05*	*0.04*
Have CDS	0.22+	0.31+	0.27+	0.24+	0.23*	0.27*	0.23	0.22*
	0.08	*0.09*	*0.10*	*0.08*	*0.11*	*0.11*	*0.13*	*0.10*
Ln($ CDS)	0.06	0.09*	0.08	0.07*	0.07	0.09*	0.07	0.07
	0.03	*0.04*	*0.04*	*0.03*	*0.04*	*0.04*	*0.05*	*0.04*
Have savings bonds	-0.02	-0.05	-0.09*	-0.02	-0.10	-0.12	-0.13*	-0.10
	0.04	*0.06*	*0.04*	*0.05*	*0.06*	*0.07*	*0.06*	*0.06*
Ln($ savings bonds)	0.02	0.00	-0.02	0.02	-0.03	-0.05	-0.04	-0.03
	0.02	*0.03*	*0.02*	*0.02*	*0.03*	*0.04*	*0.03*	*0.03*
Have other bonds	0.63+	0.51+	0.62+	0.62+	0.67+	0.54+	0.34+	0.68+
	0.09	*0.10*	*0.09*	*0.09*	*0.14*	*0.13*	*0.16*	*0.14*
Ln($ other bonds)	0.26+	0.22+	0.25+	0.26+	0.26+	0.22+	0.15+	0.27+
	0.03	*0.03*	*0.03*	*0.03*	*0.05*	*0.04*	*0.05*	*0.05*
Have mutual funds	0.06	0.09	-0.02	0.06	0.17	0.21*	-0.00	0.18
	0.07	*0.07*	*0.05*	*0.07*	*0.09*	*0.10*	*0.07*	*0.09*
Ln($ mutual funds)	0.04	0.05*	0.01	0.04	0.09+	0.10+	0.03	0.10+
	0.02	*0.02*	*0.02*	*0.02*	*0.03*	*0.04*	*0.03*	*0.03*
Have annuity/trust	0.02	0.03	0.01	0.02	-0.04	-0.07	-0.07	-0.04
	0.04	*0.04*	*0.02*	*0.04*	*0.05*	*0.06*	*0.06*	*0.05*
Ln($ annuity/trust)	0.04*	0.04*	0.02	0.04*	0.01	0.01	-0.29	0.01
	0.01	*0.01*	*0.02*	*0.01*	*0.02*	*0.02*	*0.27*	*0.02*
Have whole life ins	-0.69+	0.14+	0.19+	-0.70+	-0.63*	0.17*	0.20+	-0.61*
	0.18	*0.05*	*0.06*	*0.18*	*0.26*	*0.07*	*0.07*	*0.26*
Ln($ cash val life ins.)	0.10+	0.02*	0.01	0.10+	0.09+	0.03	0.02	0.09
	0.02	*0.01*	*0.01*	*0.02*	*0.03*	*0.04*	*0.04*	*0.03*
R^2	0.40	0.39	0.40	0.37	0.43	0.43	0.42	0.36

* = significant at the 95% level of confidence. + = significant at the 99% level of confidence. Standard errors are given in italics below each estimate; these estimates account for multiple imputation of the data.

APPENDIX

Quantile-Quantile Plots Corresponding to Figures 7-10

Figure 7a: Q-Q Plots of Imputed Distribution vs. Actual Distribution, Total Household Income, Experiments 1-3.

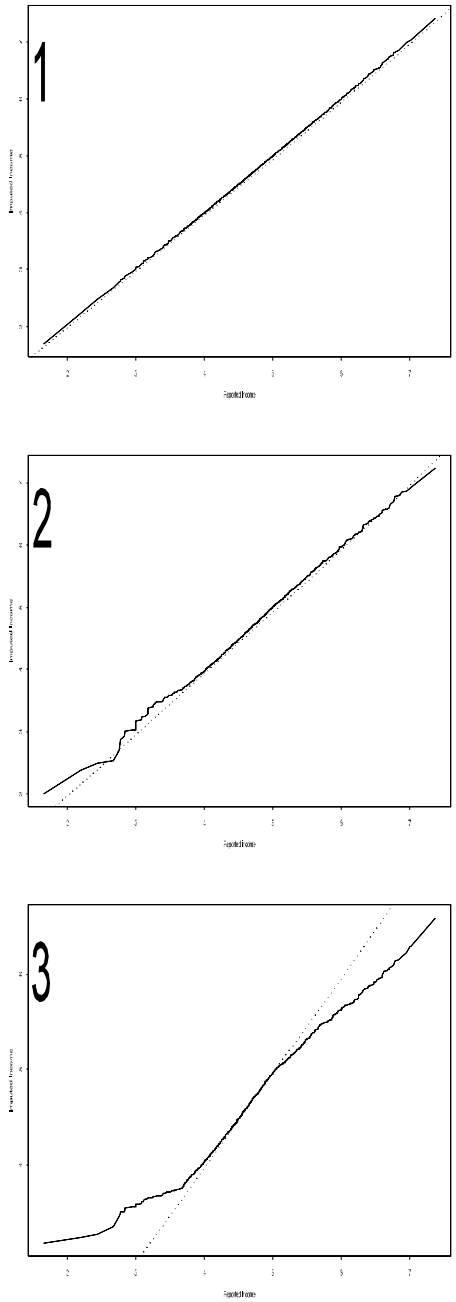

Figure 8a: Q-Q Plots of Imputed Distribution vs. Actual Distributrion, Balance in 1st Savings Account, Experiments 1-3.

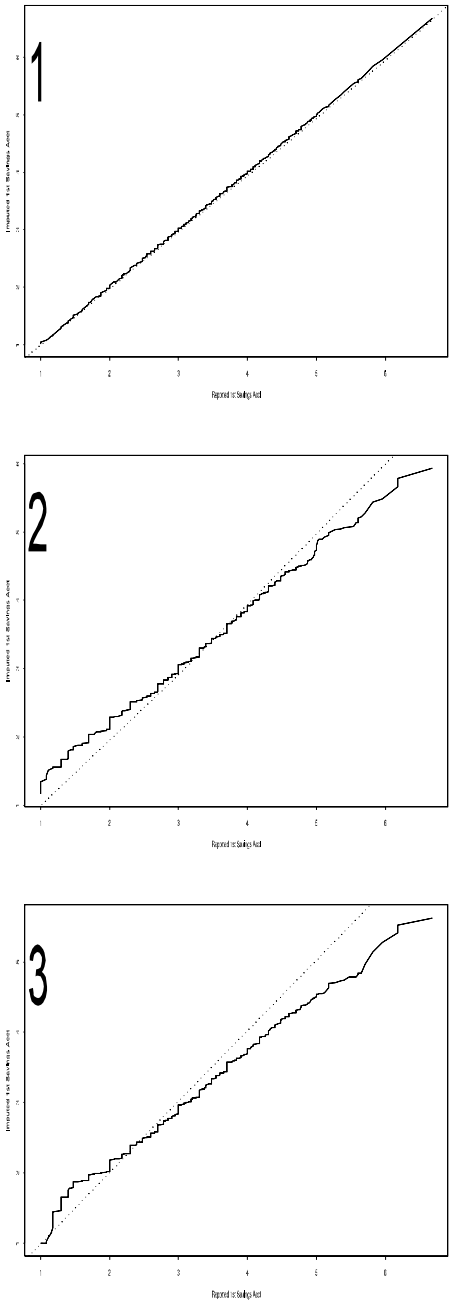

Figure 9a: Q-Q Plots of Imputed Distribution vs. Actual Distribution, Face Value of T-Bills, Experiments 1-3.

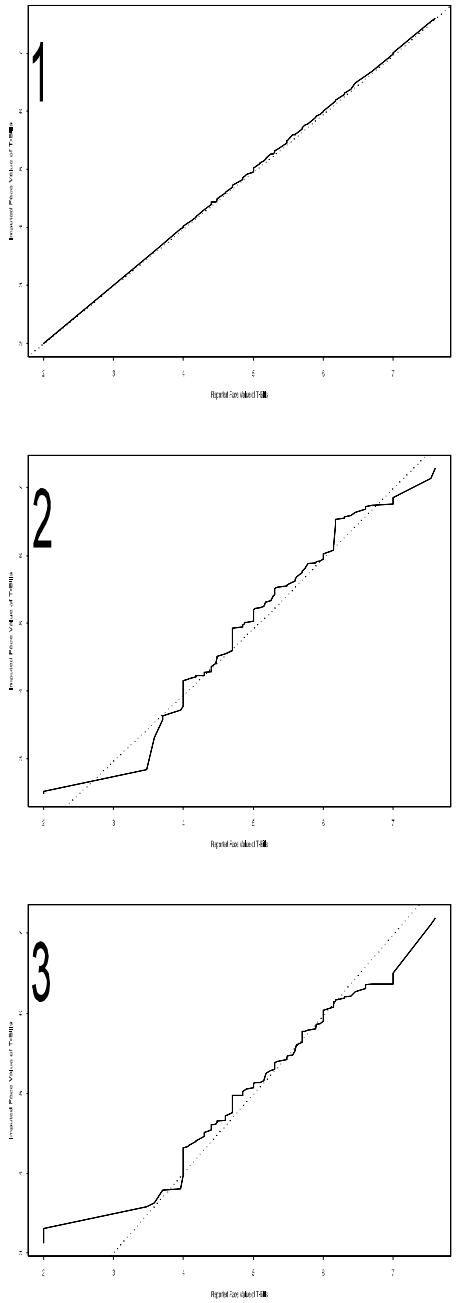

Figure 10a: Q-Q Plots of Imputed Distribution vs. Actual Distribution, Total Financial Assets, Experiments 1-3.