Multiple Imputation in the Survey of Consumer Finances

Arthur B. Kennickell, Board of Governors of the Federal Reserve System

As Arnold Zellner remarked at a session on multiple imputation at the 1997 Joint Statistical Meetings, one should always try to get the actual data one needs rather than trying to create a proxy later. Missing information usually reflects failures in the information collection process, and we should not lose sight of this fact. However, it is regrettably the case that except in the most controlled experimental settings, it is unusual for studies of any human activity to generate complete information.

The Survey of Consumer Finances (SCF) focuses intensely on the details of households' finances. Owing to the perceived sensitivity of this topic to some people, unit and item nonresponse rates in the SCF are substantial. In addition, many of the assets and liabilities treated in the survey may have values that are not always precisely defined—for example, a closely held business may have a distribution of possible values and the market value might not be knowable without actually selling the business. Such assets can have relatively high rates of missing information. Multiple imputation (MI) has provided a means of providing a public dataset that is more informative overall than anything that could be constructed with the data available to the public. At the same time, the variability in imputations allows us to give a more honest picture of the limits of our knowledge about the missing data.

MI also plays a key role in the SCF in disclosure limitation. Fairly unusual observations in the population are relatively common in the SCF, and this aspect of the survey complicates the creation of public versions of the data. MI used as a data simulation tool has been a key part of a larger strategy to protect the identity of respondents while avoiding distortions in the data made available to the public.

The next section provides a brief description of the SCF and some of the actions that have been taken to reduce item nonresponse. The second section reviews the FRITZ imputation software originally constructed for the SCF and reviews some indicators of the performance of the model. The third section presents some evidence on the performance of the model in simulating data for values that were not originally missing, for use in the SCF program of disclosure limitation. The fourth section discusses how multiple imputation has been received by analysts, and the final section concludes and points toward further research.

■ The Survey of Consumer Finances

Background

The SCF is a triennial survey conducted by the Board of Governors of the Federal Reserve System in cooperation with the Statistics of Income Division (SOI) of the IRS.[1] Data for the survey were collected by the Survey Research Center at the University of Michigan from 1983 to 1989 and they have been collected by the National Opinion Research Center at the University of Chicago (NORC) since that time.

The survey is best known for the detailed data it collects on a broad array of assets and liabilities and related data on such items such as forms of account ownership, interest rates on loans, types of automobiles, etc. Although less well known, part of the survey maps out the network of relationships households have with the financial institutions that provide the financial services. In addition, the survey collects data on current and past employment, current and past pension rights, inheritances, health insurance coverage, marital history, demographics, attitudes, and a variety of other areas. Altogether there are about 2,700 main variables in the main instrument, of which 486 are from questions about amounts. Reflecting the skip patterns in the instrument and other logical constraints, the data have a complex tree structure.

Survey Sample and Unit Response

Many of the financial variables of interest in the analysis of the survey are extremely skewed in their distributions while many others are not. In order to sup-

port analysis of both types of variables, the survey is based on a dual frame sample design: a standard multi-stage national area-probability design (see Tourangeau *et al.* [1993]) and a list sample design intended to provide a disproportionate representation of wealthy households, who hold a large fraction of the skewed assets (see Kennickell [1998a]). The list sample is selected from a sample of tax returns developed by SOI, using a stratifier ("wealth index") that is correlated with household wealth.

By traditional standards, the unit response rate for the survey is low. In 1995, about 66 percent of eligible area-probability sample respondents participated. For the list sample, the rate varied from about 44 percent in the lowest wealth index stratum to about 13 percent in the highest stratum. Unlike the area cases, the list sample cases are given an opportunity to refuse participation in the survey by returning a postcard before they are approached by an interviewer. Overall, more than a fifth of the list sample availed themselves of this option in 1995. Over time, increasing efforts have been devoted to motivating both interviewers and respondents, though the net effect has been to maintain approximately constant unit response rates. Consequently, research on adjustments for unit nonresponse—particularly adjustments for the list sample for which at least some economic data are available in the frame—has been an important focus for the survey (see Kennickell [1997a] for references).

Inclusion of the list sample adds four important complications to the data problems faced in the survey. First, the SCF includes an unusually large number of values in the far right tail of the distributions of many asset and liability variables; such values make it difficult to distinguish between reporting errors and genuine outliers. Second, wealthy households are also more likely to hold a greater variety of assets and liabilities than other households, and as a consequence, they have more chances to report missing information (see Juster and Kuester [1991]). Third, wealthy households are also more likely to own complex assets for which the actual value is difficult to ascertain. Fourth, their unusual characteristics make them more challenging for the disclosure review of the data.

Data Quality

Because of the importance of accurate financial data values in the survey, SCF interviewer training has always had a strong focus on probing to collect the most accurate values possible. Moreover, because experience suggests that many reluctant or uncertain respondents can provide at least partial information for questions requesting dollar amounts, the collection of range information has also had an important role in the survey. Before the 1995 survey, use of a range card was a key tool in collecting partial information from reluctant or uncertain respondents. Other types of ranges (e.g., "more than a million dollars") were incorporated as a single amount using a coding convention. With the movement of the survey to CAPI in 1995, a great effort was devoted to constructing a mechanism for entering the types of partial information that interviewers had previously recorded on paper questionnaires. At the same time, the procedure served to automate interviewers' probing after the first level of probing. Analysis of the 1995 data suggests that the range routines have a particularly powerful effect in reducing the proportion of respondents who initially report that they do not know an answer (see Kennickell [1996]).[2] The gain appears to be largely one of efficiency. Experiments in imputing the cases with range reports suggests the distributions of values imputed with and without accounting for the range constraints do not differ markedly. If partial reporters are similar to fully missing reporters, this result provides some support for the usual ignorability assumptions.

Despite all efforts to the contrary, missing data remain a serious problem in the SCF. Table 1 provides a summary of item response for a selection of variables, broken down by the information content of the response. A stylized interpretation of these results is that almost everyone reveals whether they have an item, the great majority of cases provide either a complete response or a range response, very few respondents end with a "don't know" response, but a substantial minority still refuse to give an answer. The patterns of missing data across variables are sufficiently heterogeneous that it is a good approximation to assume that every observation has a unique pattern of missing data.

In summary, the survey contains a very large number of variables, there is substantial missing or partially missing (range) information, the patterns of missing information are highly heterogeneous, the distributions of many of the variables are highly skewed, and the data have a complex structure. Analysis of the survey in the absence of imputation would be a formidable task. Moreover, anyone using the public version of the dataset would lack key frame data that turn out to be important for understanding the distributions of the missing data. [3] Thus, even on pure efficiency grounds, there is a good case for imputing the missing data.

■ Imputation in the SCF

The Movement to MI in the SCF

Imputation for missing data has always played an important role in the survey. In the SCFs before 1989, missing data were singly imputed using a variety of techniques, including randomized regressions, hot deck, and other techniques (see Avery and Elliehausen [1985]). With a comprehensive redesign of the survey in 1989, the approach to imputation was reconsidered, and an MI model—FRITZ (Federal Reserve Imputation Technique Zeta)—was developed following the ideas of Rubin [1987] (see Kennickell [1991]). The structure was also influenced by ideas of Gibbs sampling as well as techniques developed in image processing (see Geman and Geman [1984]). The implementation of FRITZ for the SCF contains a highly structured set of constraints that embody the complex logic of the data. There were two principal motivations for the movement to MI: first, it was desired to have a structure that gives as honest a picture as possible about the limits of what is known, and second, it was desired to have an open structure that

Table 1: Reporting rates for various items, percent. Full sample for 1995 SCF, unweighted.

Item	Have item		Value reported by respondent, for those reporting having the item					
	Yes	Unknown	Number	Range response			DK	Other missing*
				Tree	Card	$ range		
Credit card balance	76.0	0.4	93.6	0.4	4.1	0.2	0.1	1.7
Principal residence	67.6	0.0	88.9	1.1	7.2	1.1	0.0	1.7
Borrowed on mortgage	42.9	0.3	89.6	1.5	5.9	0.2	0.3	2.6
Owe on mortgage	42.9	0.3	86.1	1.6	8.5	0.1	0.2	3.5
Mortgage payment	42.2	0.3	92.7	0.4	4.2	0.0	0.1	2.5
Rent	23.8	0.0	95.1	0.4	2.9	1.0	0.0	1.5
Other real estate	32.4	0.6	84.0	1.7	9.5	0.7	0.4	3.7
Business	26.8	0.4	61.9	5.7	18.6	1.0	1.2	11.5
Car loan payment	23.7	0.2	93.0	0.8	4.1	0.0	0.2	1.9
Checking account	88.7	0.3	80.1	1.9	10.4	0.5	0.4	6.7
Money market account	17.3	0.7	71.7	1.8	14.4	0.5	0.9	10.6
Savings account	33.6	0.7	80.2	1.7	11.1	0.1	0.1	6.8
Certificates of deposit	17.0	1.0	69.7	3.4	11.1	0.3	0.3	15.3
IRA/Keogh account	34.6	1.2	74.4	2.6	13.5	0.3	0.4	8.9
Savings bonds	24.0	0.7	76.1	3.2	13.0	0.2	0.8	6.8
Municipal bonds	8.1	1.2	59.8	2.9	15.2	0.9	1.2	20.1
Tax-free mutual funds	8.3	1.6	59.6	2.5	16.6	0.0	0.8	20.5
Stock	28.4	0.9	63.8	2.5	16.0	2.2	1.4	14.1
Trusts and annuities	7.2	0.6	65.9	3.9	16.1	0.6	0.0	13.5
Face value of whole life ins.	38.6	2.2	76.7	2.5	11.1	0.3	0.8	8.6
Cash value of whole life ins.	38.6	2.2	55.5	7.8	15.5	0.5	2.1	18.7
Wage income	73.6	1.0	72.8	1.5	16.7	0.2	0.3	8.4
Business income	20.6	1.5	68.5	2.4	12.8	0.3	0.5	15.6
Pension and Soc. Sec. inc.	26.5	1.2	73.3	1.9	11.1	0.0	0.4	13.3
Total income	100.0	0.0	69.1	1.5	16.8	0.1	0.5	12.1

* Almost entirely refusals. Also includes some missing values due to editing decisions.

is as statistically uniform as possible..

The FRITZ Model

The FRITZ model is sequential in the sense that it follows a predetermined path through the survey variables imputing missing values one (occasionally, two) at a time. The model is also iterative in that it proceeds by filling in all missing values in the survey dataset, using that information as the basis for imputing a following round, and continuing the process until key estimates are stable. Five imputations are made for every missing value. For analytical convenience, the data are stored as five replicates ("implicates" to distinguish them from other replicates) of the dataset.

The FRITZ system performs three basic types of imputations: continuous variables, multinomial variables, and binary variables. The core of the model turns on two types of estimates, a covariance-based structure and a frequency-based structure. The latter is a type of hot deck, but because it can accommodate continuous variables as both target imputation variables and as conditioning variables, I prefer to think of it as two-dimensional nonparametric imputation with a nearest neighbor rounding rule. The majority of the implementation

of FRITZ for the SCF uses the covariance-based procedures. These covariance-based procedures are of two types: imputation of binary variables and imputation of nominally continuous variables.

Continuous Variable Imputation Technique

Because of the particular importance of dollar values in the SCF, I focus here on the part of FRITZ that is normally used for imputing continuous variables. For convenience, suppose the iterative imputation process has completed t-1 rounds and we are currently somewhere in round t, with a data structure as given in figure 1. In the figure, "y" indicates complete reports of the variable currently the subject of imputation, and "ψ" indicates round t-1 imputations of missing values of the variable; "x" indicates complete reports and "χ" indicates completed imputations in the past or current iteration of the set of other variables available to condition the imputation Variables that were originally reported as a range that are not yet imputed are represented by "r," and "●" represents values that were reported as completely missing values that are not yet imputed. Every "x" variable becomes a "y" variable at its place in the sequence of imputations within each iteration. Note that no missing values remain in the stylized t-1 dataset.

Figure 1: Data Structure for Imputations

Iteration t-1

Y	X →		
y_1	χ_{11}	x_{12}	x_{13}
ψ_2	χ_{21}	x_{22}	χ_{23}
y_3	x_{31}	x_{32}	x_{33}
...			
y_{n-2}	$x_{n-2,1}$	$x_{n-2,2}$	$x_{n-2,3}$
y_{n-1}	$x_{n-1,1}$	$\chi_{n-1,2}$	$x_{n-1,3}$
ψ_n	χ_{n1}	χ_{n2}	x_{n3}

Iteration t

Y	X →		
y_1	●	x_{12}	x_{13}
●	r_{21}	x_{22}	●
y_3	x_{31}	x_{32}	x_{33}
...			
r_{n-2}	$x_{n-2,1}$	$x_{n-2,2}$	$x_{n-2,3}$
y_{n-1}	$x_{n-1,1}$	$\chi_{n-1,2}$	$x_{n-1,3}$
●	r_{n1}	χ_{n2}	x_{n3}

In the ideal, we would like to condition every imputation on as many variables as possible, as well as on interactions and higher powers of those terms (see, e.g., Little and Raghunathan [1997]). However, there are practical limits to this strategy as a result of the relatively small number of degrees of freedom in a survey like the SCF. Some judgment must play a part in variable selection, but we have tried to span a broad set of possible relationships. Given that maximal set of allowed variables (denoted by X), not every element may be non-missing at a given stage of imputation.[4] For each variable to be imputed, the FRITZ model determines the non-missing variables among the maximal set for each observation (denoted X_i for observation i). Given the selection of available conditioning variables, the model essentially runs a regression of the target imputation variable on the subset of conditioning variables using the previous iteration of the model. In practice, this process is made more technically efficient by estimating a maximal normalized cross-product matrix for each variable to be imputed (denoted $\Sigma(X,Y)_{t-1}$), and then subsetting rows and columns corresponding to the non-missing conditioning variables for a given observation (denoted $\Sigma(X_i,Y)_{t-1}$).[5] The imputation for observation i in iteration t is given by

$$\psi_{it} = \beta_{it} X_{it} + e_{it}$$

where X_{it} is the subset of X that is available for i in iteration t, $\beta_{it} = \Sigma(X_iX_i)_{t-1}^{-1} \Sigma(X_iY)_{t-1}$, and e_{it} is a random error term. Once a value is imputed, its imputed value is used along with reported values in conditioning later imputations.

Error Term

In the past, the error term has been taken to be a draw from a truncated normal distribution. The truncations are usually determined by several factors. First, to avoid imputing extreme cases, the range of draws is usually restricted to the central 95 percent of the distribution. Second, logical conditions imposed by the structure of the data sometimes limit the range of the error. Finally, partial information in the form of respondent-provided ranges often provides a bound for the outcome.

The FRITZ system can also accommodate the use of empirical residuals, and I report some ongoing work in this area later in the paper.

The Role of Range Data

Range data play an important role in the imputations beyond their role in limiting the range of allowed outcomes. Range reports are broadly used, and this use appears more common for households with relatively complicated finances. The ranges contain information that would be useful in conditioning the imputations of other variables. Although it is theoretically possible to build a model that would account for the locational information in the ranges, this approach would quickly become unwieldy in light of the multiplicity of types of ranges that a respondent might report. The compromise adopted in the SCF implementation of FRITZ is to treat the midpoint of each range as a conditioning value like any other until that variable is imputed. As in the case of completely missing variables, once a range response is imputed, the imputed value is used as a conditioning variable in all subsequent imputations.

Variability of Imputations

Typically, the imputations of most dollar variables are reviewed graphically during each iteration to look for weaknesses in the imputation models or other unusual outcomes. As an example, figure 2 shows a plot of the values of total household income for each of the five implicates in the 1995 SCF, separated by sampling stratum. In this plot, the zero level stratum corresponds to the area-probability sample, and levels one through seven correspond to successively higher levels of implied wealth in the list sample.[6] In the figure the dots represent data that were originally reported completely, the crosses represent imputations based on range reports, and the diamonds are values that were imputed without any respondent-provided range data. Two points are noteworthy. First, with few exceptions, the outliers in every stratum are a mixture of complete reports and range reports. Second, within each stratum, the imputations of the completely missing values tend to be more centered.

Figure 2: Distribution of Log₁₀(Total Household Income), by Jittered Stratum Indicator and by Type of Original Reported Value, 1995 SCF
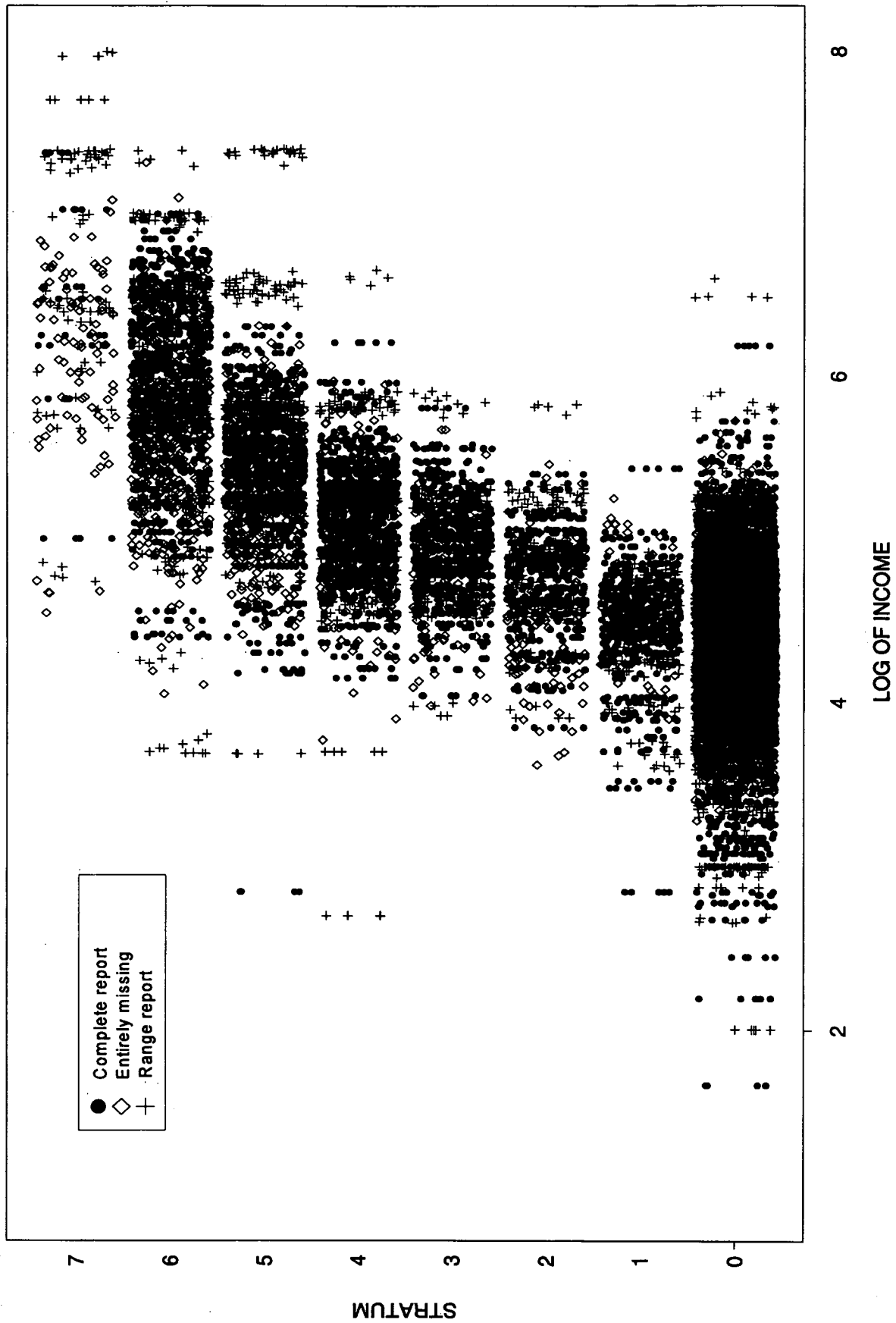
Table 2: Weighted coefficient of variation for mean and median values due to imputation, sampling and both sources, selected variables, percent; weighted percent of total value imputed, by type of original response; 1995 SCF.

| | Weighted coefficients of variation | | | | | | Wgtd. % of value imputed | | Memo: |
| | Imputation std. err. | | Sampling std. err. | | Combined std. err. | | Range | Missing | Wgtd % |
	Mean	Median	Mean	Median	Mean	Median			having
Credit card balance	0.6	0.0	3.9	17.7	3.9	17.7	10.4	1.7	66.6
Principal residence	0.6	0.0	2.5	2.8	2.6	2.8	7.6	1.1	59.9
Borrowed on mortgage	0.5	0.7	2.5	3.8	2.6	3.9	6.3	3.0	39.6
Owe on mortgage	0.5	0.1	2.7	3.6	2.7	3.6	7.5	2.5	39.6
Mortgage payment	0.5	0.0	2.5	2.2	2.5	2.2	5.7	3.1	39.6
Rent	0.6	0.6	2.1	2.5	2.2	2.6	5.4	2.3	31.1
Other real estate	4.8	2.7	29.0	9.5	29.5	9.9	12.7	3.0	17.2
Business	2.9	10.6	13.1	18.8	13.5	22.1	40.4	9.7	11.7
Car loan payment	0.3	0.5	2.5	1.5	2.6	1.6	4.5	2.3	27.7
Checking account	1.7	0.0	5.4	1.7	5.8	1.7	15.6	6.4	85.0
Money market account	4.7	2.8	10.2	14.4	11.5	14.7	22.0	7.9	9.9
Savings account	2.2	0.0	7.9	13.7	8.2	13.7	13.3	7.2	36.3
Certificates of deposit	3.5	2.2	8.0	8.8	8.9	9.2	11.8	19.5	14.7
IRA/Keogh account	1.6	0.0	7.7	6.5	7.9	6.5	19.9	11.5	22.6
Savings bonds	2.8	0.0	8.6	5.9	9.2	5.9	20.5	6.8	23.1
Municipal bonds	6.1	16.8	18.0	83.8	19.2	85.8	21.1	16.7	2.3
Tax-free mutual funds	6.8	15.5	16.4	21.0	18.0	27.0	18.8	28.6	4.0
Stock	2.9	2.0	12.0	11.2	12.4	11.5	31.6	14.2	13.7
Trusts and annuities	5.3	15.5	12.5	29.9	13.8	34.3	32.0	14.3	4.2
Face value of whole life ins.	2.7	4.4	5.6	11.6	6.4	12.5	15.0	13.3	32.9
Cash value of whole life ins.	2.7	4.5	6.9	7.4	7.5	8.9	18.2	23.2	32.9
Wage income	0.5	0.3	2.0	2.2	2.1	2.2	21.6	1.7	69.6
Business income	3.4	5.1	16.0	17.4	16.4	18.3	17.5	40.9	11.4
Pension and Soc. Sec. inc.	0.8	1.7	2.8	4.1	2.9	4.5	11.4	11.1	30.6
Total income	0.5	1.4	2.3	2.1	2.3	2.6	20.8	10.6	100.0

To get a general sense of how MI performs in the SCF, it is useful to start by looking at the variability of the imputations. Table 2 shows the weighted coefficients of variation with respect to imputation for the same dollar values reported in table 1. For comparison, standard errors with respect to sampling error, and the combined standard errors are also given. Examination of such information is often important in identifying questions that need work to help respondents provide better information or where we need to be concerned about increasing the sample size. For the examples in the table, the coefficient of variation with respect to imputation is substantially smaller than the sampling CV—one would expect this to be the case. In a few cases, the mean imputation is substantially less variable than the median. This result is driven by two factors, the highly skewed nature of the distribution of these variables, and the nature of the sample design. The great part of the total value of these items is held by a very small group, and the oversampling of wealthy households provides a relatively large number of observations to estimate the mean.

Holdings by less wealthy households of the sort that make up almost all of the area-probability sample are relatively rare and amounts held are quite variable. Because the weighted population in this group is very large relative to the group that holds the bulk of the dollar values, the median is less precisely estimated than the mean.

Table 2 also shows the proportion of the total value of the items imputed for original reports of ranges and of missing values. Clearly, because a large share of the imputed values is based on range reports, such information may be important at least in refining the precision of the imputations. In most cases, the fraction of the total value imputed for range reports is greater than that for completely missing variables. Regression analysis (not shown) of the coefficients of variation with respect to imputation for the mean and median suggests that these values are positively related to the proportion of the total value imputed, but, surprisingly, the model finds no independent effects for the proportion of the imputation

made from ranges and that from completely missing values.[7]

■ Using Multiple Imputation for Disclosure Limitation

As Ivan Fellegi [1997] has recently expressed more eloquently, the strong increase in the amount of individually-identified information available in the private sector poses a corresponding growing risk for the release of survey databases for research. In an effort to make information available to analysts in such an environment, various proposals have been made to license data users, to set up secure centers for using data, to require posting of bonds by users, and to implement other techniques to limit basic access to the data. In contrast, Rubin [1993] proposed creating an entirely simulated dataset that could be made available to everyone without any disclosure risk. In a world where researchers increasingly demand immediate access to information and the presumption is that everything should be available seamlessly on the Internet, this idea has great appeal.

Multiple imputation for data simulation has an important role in disclosure limitation in the creation of the public versions of the SCF. Beginning with the 1989 disclosure review, various sets of dollar values for selected cases were flagged and these values were multiply imputed subject to constraints intended to minimize distortion of key qualities of the data. The use of multiple imputation for disclosure limitation has progressed with each survey.

In the 1995 SCF, the process was carried to the point that all dollar values for the selected cases were simulated (see Kennickell [1998b]). Cases were selected on the basis of their having unusual levels of wealth or income given other characteristics, or other unusual combinations of responses. A further set of random cases was added to damage the ability of an intruder to identify even the set of cases we determined presented excessive disclosure risk. A new dataset was created for all the selected cases, and the shadow variables were set in such a way that the FRITZ model interprets the responses as range responses. The type of range mim-

icked was one where the respondent volunteered a dollar range. For the disclosure imputations, a dollar amount of $\pm p$ percent (where p is an undisclosable point somewhere between 10 and 20 percent) around the original value was stored in a dataset normally used to contain unique range reports. Subsequent evaluation (Fries et al.[1997]) indicates that while the imputations substantially masked individual cases, the effect on important distributional characteristics was minimal.

In an attempt to find how far we might be able to go with Rubin's larger proposal, I have done several additional experiments simulating all of the reported dollar values in the SCF with less severe constraints.[8] Insofar as these experiments deal only with dollar values, they fall considerably short of Rubin's proposal. Nonetheless, I believe the work provides insights into the usefulness and practicality of the creating simulated data.

In this paper, I report on the four experiments summarized in figure 3. In the first experiment, the data simulation is conditioned on other reported values until the reported values are replaced by the simulated values, and the error draw is taken from a normal distribution with variance tailored to the particular imputation. The second experiment excludes initial reports of dollar variables from the conditioning, but maintains the normal error structure; this experiment is intended to give a sense of the ability of the models to simulate data from more general inputs, as one might want to do in creating a fully simulated dataset. The third and fourth experiments parallel the conditioning structure of the first two experiments; however, the errors are random draws of empirical residuals where the residuals are computed and drawn separately for each of the seven list sample strata and for each of the four major geographic regions in the area-probability sample.[9] The empirical residuals are

Figure 3: Design of Experiments

Experiment	Type of error draw	Condition on reported dollar values?
1	Normal	Yes
2	Normal	No
3	Residual	Yes
4	Residual	No

subject to the same sorts of structural constraints as the normal draws.

For purposes of this paper, I examine the data simulated from the four experiments for only two dollar variables: total household income, and direct holdings of publicly-traded corporate stock (see Appendix). The choice is intended to indicate the performance of the model in different size samples of data available for estimation. By definition, all households have income, but only about 14 percent of the population is estimated to have direct stock holdings.

The results of the experiments are best described graphically. For each of the four experiments and each of the two variables, figures 4a and 4b show an estimate of the density of the mean over imputations for individual cases of the deviation of the simulated values from the reported values ("bias").[10] The bias distributions are all centered around zero, which is a minimal desirable property for data simulation—and a reasonable expectation given the structure of the model. However, partially as a consequence of the number of imputations and partially as a result of variations in the ability of the model to reproduce individual cases (degrees of model failure), there is also considerable dispersion of the bias. To some extent, this dispersion in the simulated data is desirable for disclosure reduction—otherwise a data intruder could always recover the exact value by averaging the imputations. As expected, the spread is greater for both the experiments two and four, which do not condition on the reported dollar values, and for the stock simulation where the model is based on a much smaller set of observations. The underlying models are noisier, so a larger number of imputations would be required to achieve the precision of experiments one and three. Differences in the bias distributions between the experiments with the normal draws and the corresponding ones with the empirical residuals are minor.

Figures 5a and 5b show the corresponding estimates of the density of the standard error of the simulated data over imputations for individual cases.[11] The information loss in not conditioning on reported dollar values is more clearly reflected in comparing experiments two and four with experiments one and three. As expected

from the truncations of the normal errors, the variability of imputations for experiments using empirical residuals is somewhat greater than those using the normal draws. The variability of imputations within observations reflects our ignorance about the structural relationships among variables. However, this variability in conjunction with the variability in the conditional mean outcome is good protection against disclosure. An important question to be answered is how much of each type of variability is "enough" to guarantee to a high degree that a data intruder could not identify individual observations.

Even if we can figure out the optimal amount of variability for disclosure limitation, simulated data must also reproduce a wide variety of characteristics of the original distributions if the data are to be useful in a broad range of applications. Figures 6a and 6b show unweighted quantile-difference (QD) plots for the actual and simulated distributions for each experiment and for the two variables.[12] To focus on the different effects of the data simulation, the plots for each variable are based only on the set of cases that initially gave a complete response.[13] For both income and stock, when the model conditions on other reported values, the distributions of the actual and simulated data differ little for most of their range for either type of error draw. For the bottom end of the stock distribution, there is sizable over-prediction on the logarithmic scale, but most of the dollar amounts are relatively small. In contrast, when the imputations do not condition on the reported dollar values, the actual and simulated distributions are not as well aligned. In the case of the income simulations, there is a notable under-prediction at the top of the range when normal error draws are used, while the alignment with empirical residuals is nearly as close as when the model conditions on the reported dollar values. For stocks, the fit deteriorates substantially with either type of error, though the fit in the experiment with the empirical residuals is notably better.

If these results carry over to other variables and to other surveys, data simulation may be a reasonable solution to at least one broad class of disclosure problems. The models appear unbiased in the conventional sense, and the degree of variability of some types of estimates

Figure 4a: Unweighted Density Estimate of Log_{10}(Actual Value) Minus Mean(Log_{10}(Imputed Value), Total Household Income, Experiments 1-4.
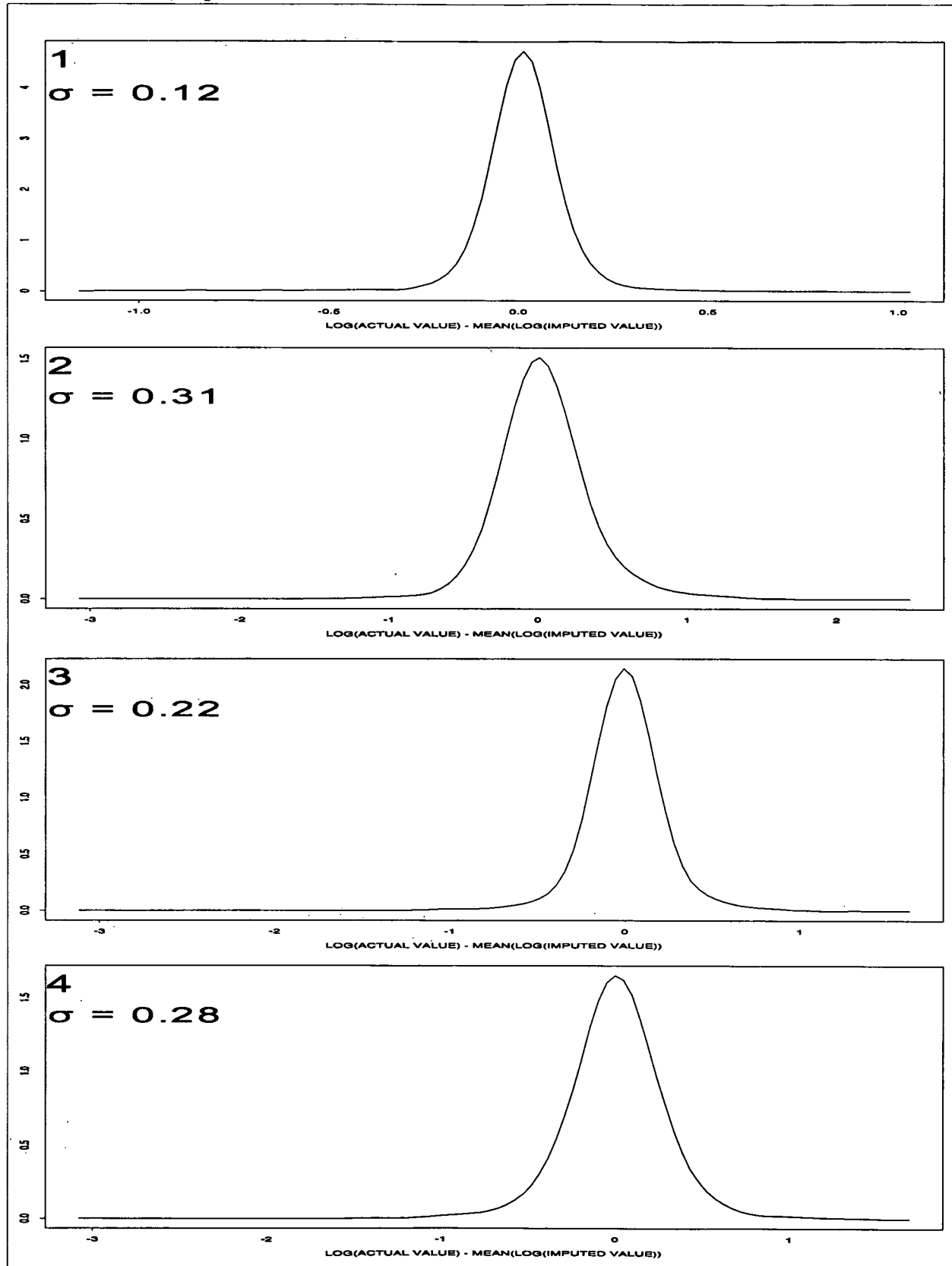
Figure 4b: Unweighted Density Estimate of Log_{10}(Actual Value) Minus Mean(Log_{10}(Imputed Value), Direct Holdings of Publicly-Traded Corporate Stock, Experiments 1-4.
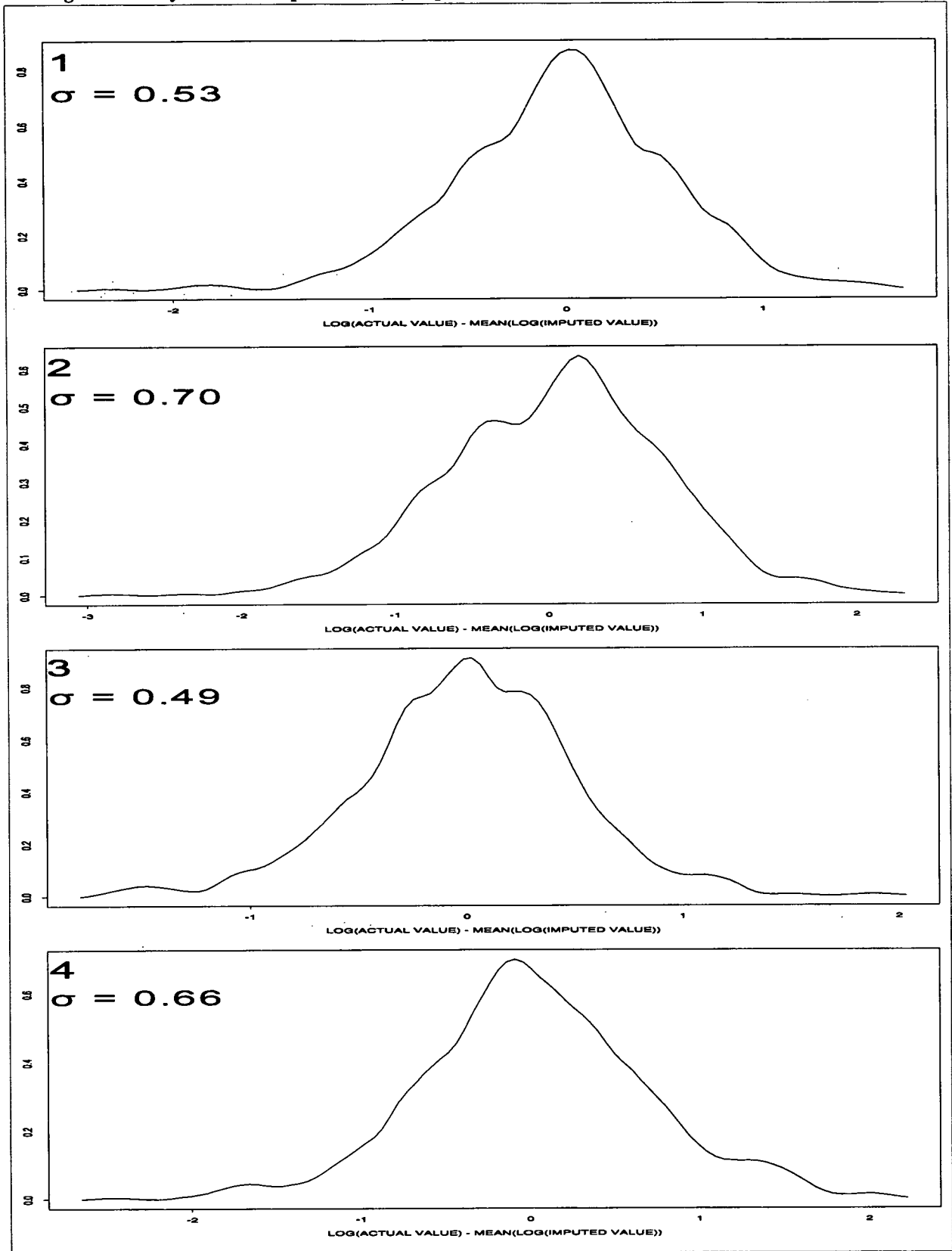
Figure 5a: Unweighted Density Estimate of Standard Deviation of Log₁₀(Imputed Values) Across Observations, Total Household Income, Experiments 1-4.
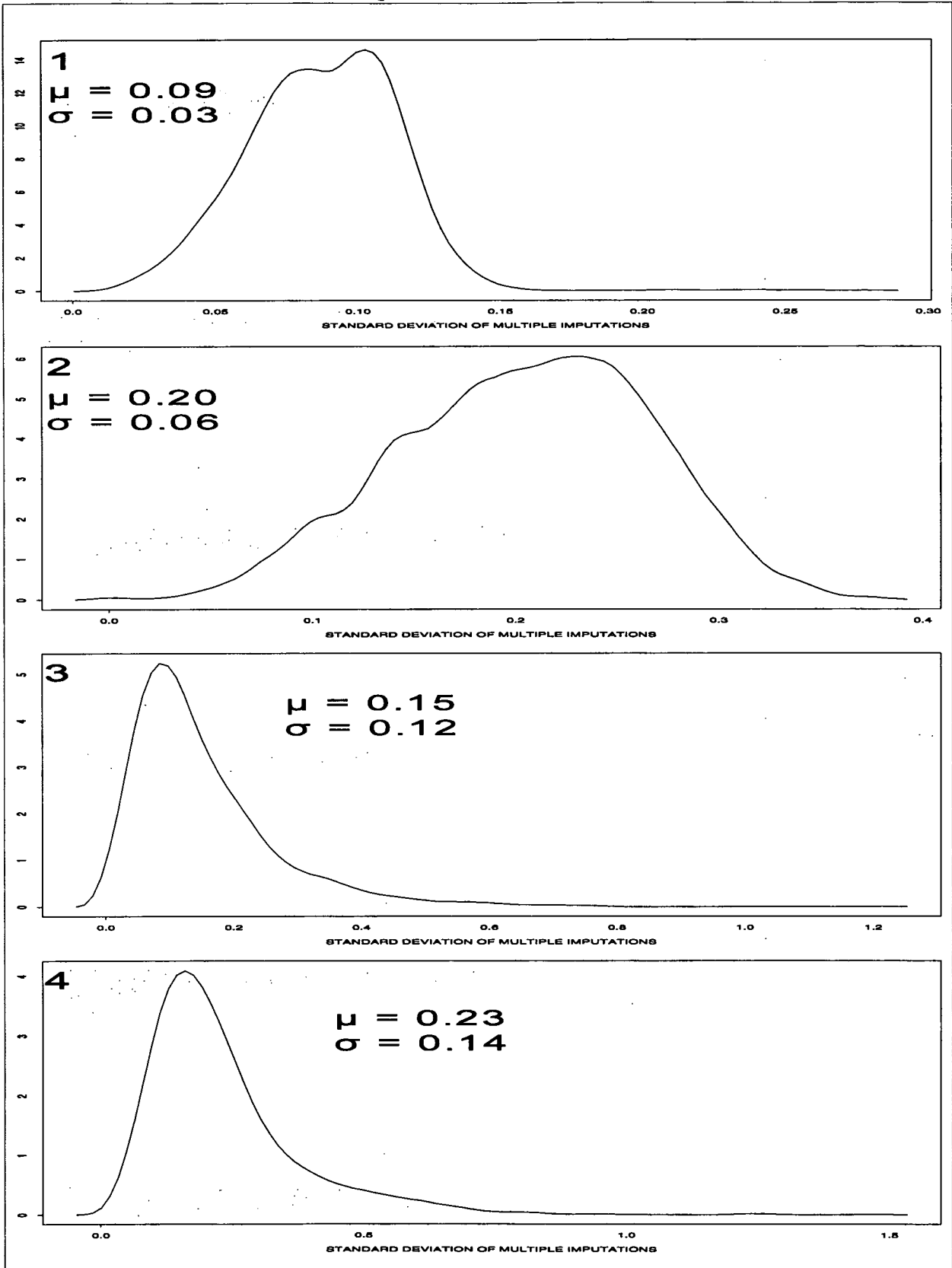
Figure 5b: Unweighted Density Estimate of Standard Deviation of Log$_{10}$(Imputed Values) Across Observations, Direct Holdings of Publicly-Traded Corporate Stock, Experiments 1-4.
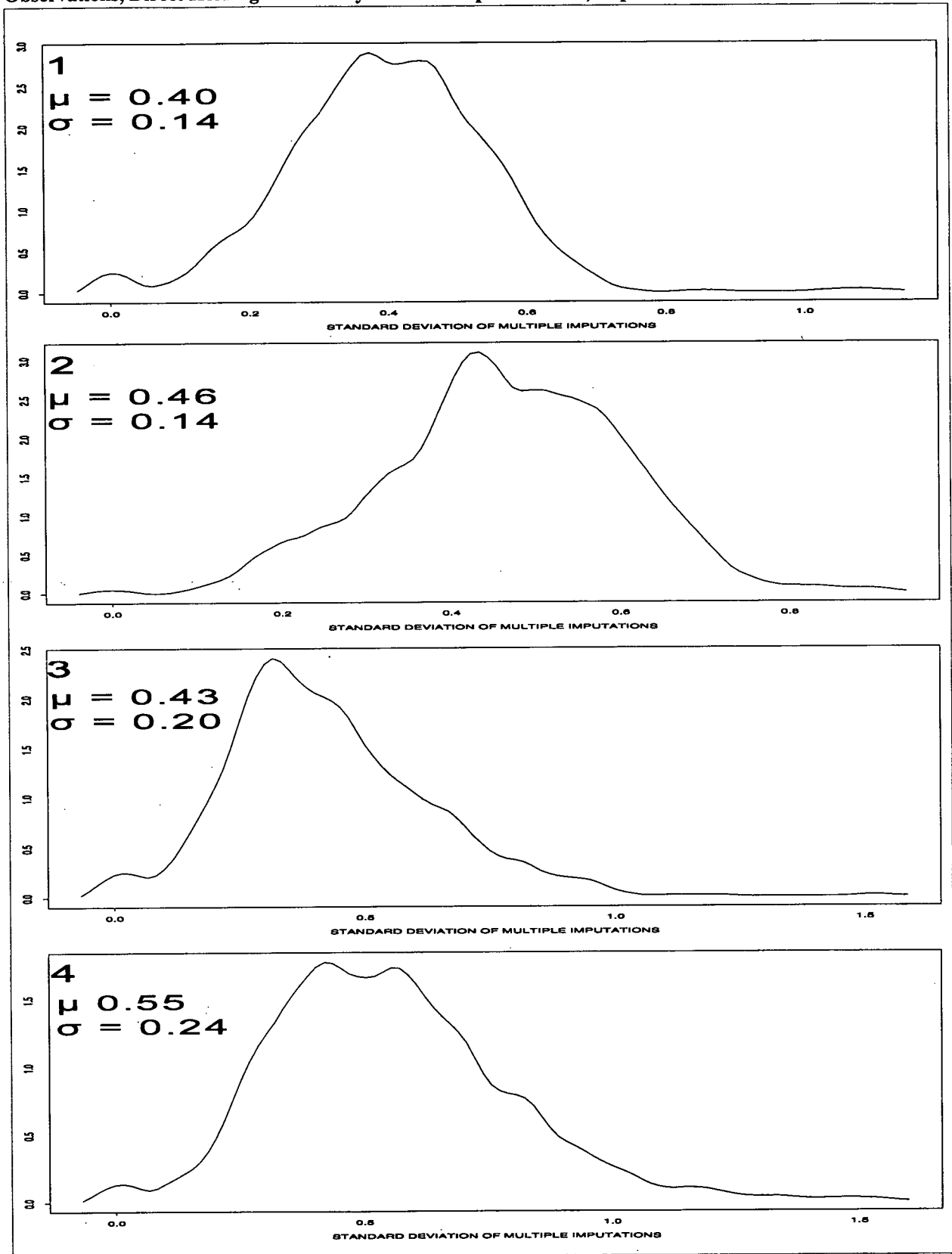
Figure 6a: Unweighted Quantile Difference Plots: Log$_{10}$(Actual Value) Minus Log$_{10}$(Imputed Value), Total Household Income, Experiments 1-4.
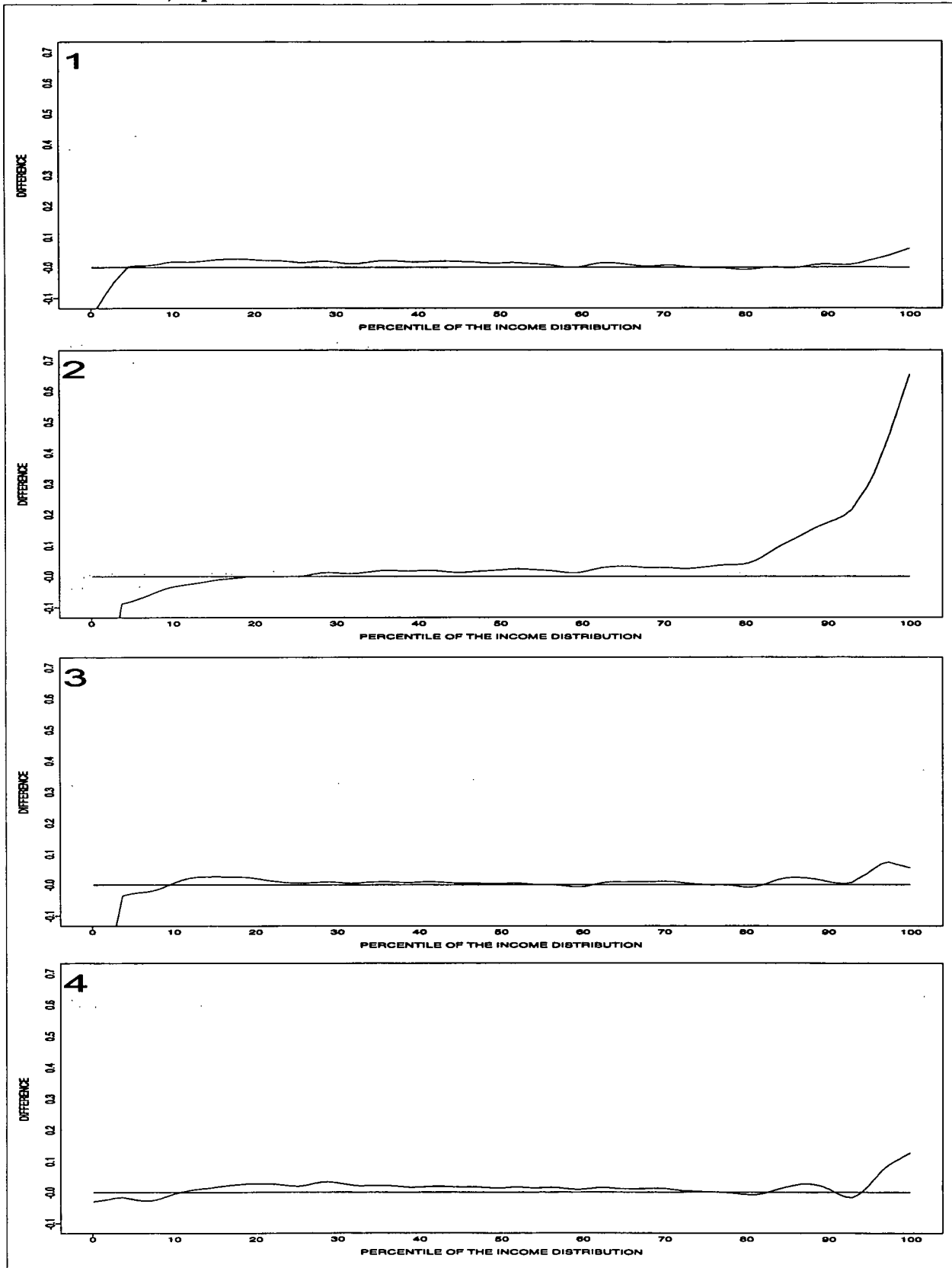
Figure 6b: Unweighted Quantile Difference Plots: Log₁₀(Actual Value) Minus Log₁₀(Imputed Value), Direct Holdings of Publicly-Traded Corporate Stock, Experiments 1-4.
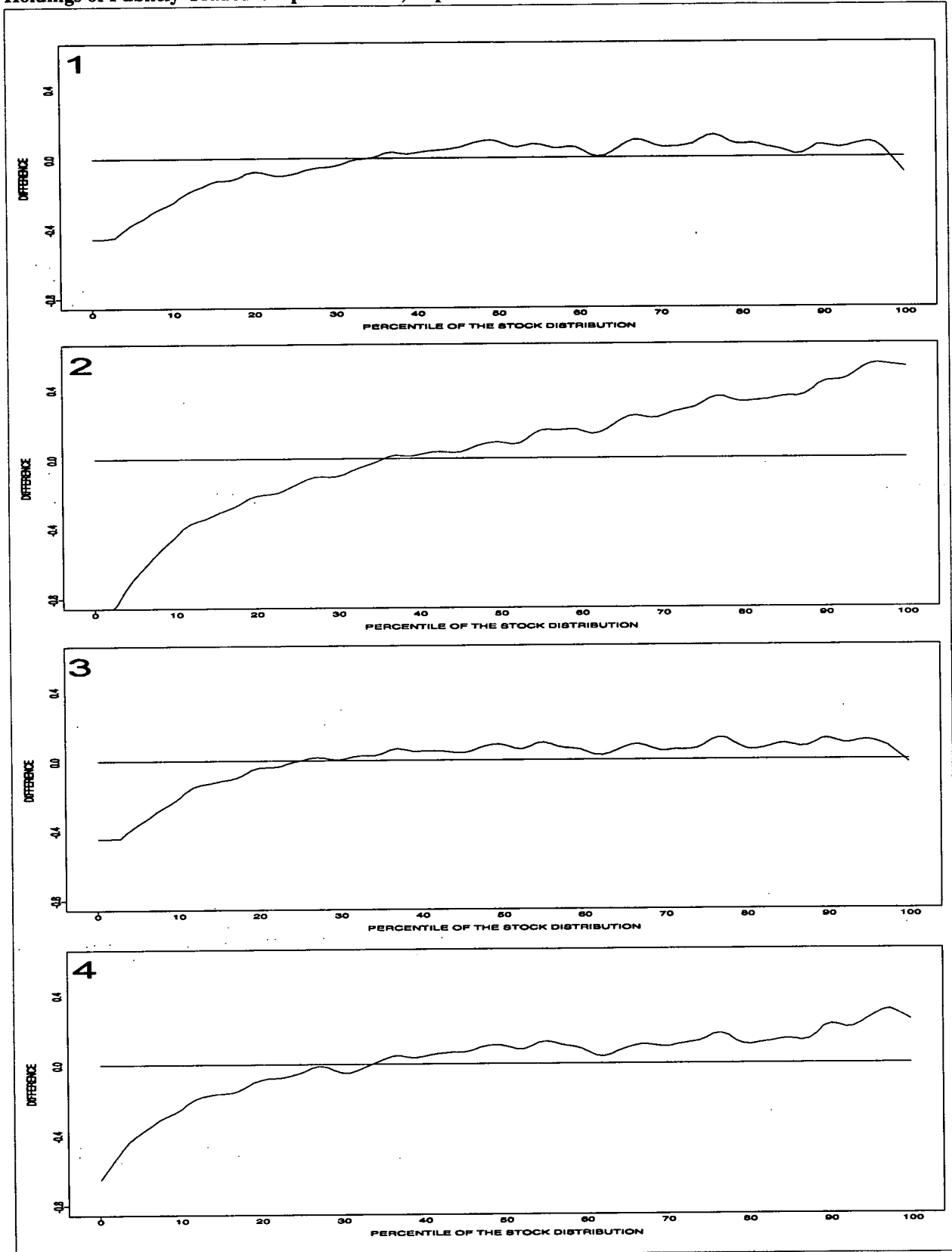
could be "tuned" by choosing different number of imputations. Although the results at least with empirical residuals mimic the reported distributions fairly well, more work needs to be done in a multivariate context—particularly the effects of data simulation on regressions.

■ Reception of MI Among Analysts

In implementing MI for the SCF, there have been many problems, but not many of them have been more difficult than that of getting analysts to accept the data and to use them correctly. As far as I am aware, the 1989 SCF was the first large-scale survey to provide multiply imputed data for all variables. At that point in social science research, which is where SCF data are mainly used, there was very little professional attention paid to issues of missing information. One notable exception is Lillard et al. [1986] which examined imputation in the 1980 Census. In addition, the structure of many datasets—no imputations, no flags for imputation of missing data, or the presence of just one imputation—made it easy to ignore the issue. With the release of the SCF data and its five separate implicates for each original observation, even the most casual user of the survey needed to be aware of the multiple imputations.

Many analysts experienced initial difficulties in the mechanics of handling the data. The most common such problem was that some naive researchers who had not read the data codebook ran regressions on all five of the implicates together and often found delightfully (at first) inflated t-statistics. Others computing population totals using the survey weights (which sum to the full population for each implicate) found estimates about five times as large as expected.

A deeper and continuing issue is that having been forced to deal with imputations more directly, some analysts have questioned the entire practice of imputation. There is much misunderstanding among analysts about what they should do, and much suspicion about what data producers actually do. Most notably, some researchers have expressed concern about the effects of imputation on "structural modeling" as practiced by economists and others—particularly in models with individual fixed effects. David Brownstone and others have addressed this question in parts, but it is clear that much more ef-

fort is needed. For many other researchers who have at least implicitly accepted the idea of imputation, it has been a struggle to find a practical means of working with the data. Although the software needed to analyze multiply-imputed data is not particularly complicated, poor grounding in the conceptual framework has been a large impediment to analysts. Many non-statisticians have complained about the lack of "intermediate" level material. Simple software for regressions and other simple statistics has been made available (Montalto and Sung [1996] and Kennickell [1997b]), yet still there is often confusion about how such software should be applied. One sign of progress is that among consumer economists using the SCF, correct use of the multiple imputations is a standard for publication. Unfortunately, among other types of economists and other researchers, we still see people publishing articles using only one of the imputations or making other such errors in the use of the data. The conclusion I draw from what I have seen is that the need for education is great.

■ Conclusions and Future Research

It was difficult to implement MI for the SCF, but only marginally more difficult than setting up a system for single imputations. Computing time and storage scale up approximately linearly, but with current technology, this is no longer a serious consideration except, perhaps, for the most enormous datasets—even then, there are technical efficiencies that can reduce the processing burden. A more serious issue is the reception of MI by data analysts. Historically, many people have seen statistics as a "barrier to saying what we want to say." Clearly, an important part of statistics is aimed at determining the proper basis for inference, and this may often have the effect of limiting the claims that can be made with data. However, serious researchers should be united on the validity of this point, however irritating it may sometimes be in practice. In the particular case of MI, there appears to be a great deal of confusion about its statistical properties and how it should be used. Most publicly available datasets are singly-imputed, if they are imputed at all. Many of these are very fine data production efforts—for example, the Current Population Survey. Analysts look at such survey data, which appear to be accepted by statisticians, and wonder why they should have to worry about imputations, much less

multiple imputations. A serious effort needs to be made to reach such people and help them to see that MI is just another in the set of tools created to further the scientific basis of research. All these tools are imperfect, but they do move us along.

Overall, multiple imputation over three (and going on four) triennial waves of the SCF has been a success. Looking at imputation variance has helped us to understand which are the weak questions that need work— this has prompted important questionnaire and training revisions. MI has also become a routine part of our analytical work. At least within some groups of our external data users, using multiply-imputed data—or explaining why not—has become close to the standard for publication. MI has become an essential tool in creating the public version of the SCF.

There are three areas of imputation where I believe it is important for the SCF to work in the future. First, I believe that data simulation is a promising approach. I am not confident that licensing and controlled data centers will work sufficiently well to deal with the SCF disclosure problems, and I expect that many other surveys will be in a similar position. Second, the experiments with empirical residuals are promising, and I would like to make more general use of this technique outside of data simulation for disclosure limitation. Finally, it is very important to continue to try to reach data users by building analysis software for use with multiply imputed data and by striving for targeted clear explanations of why it is important to take imputation and multiple imputation seriously.

■ Acknowledgment

Footnotes

1   See Kennickell *et al.* [1997] for an overview of the survey.

2   The effect on initial refusals appears, at best, much more limited.

3   Such information includes frame data for the list sample, data obtained by matching characteristics of Census tracts in the SCF sample, data from the Current Population Survey matched by three-digit industry and occupation codes, and other more minor collections of data.

4   In a problem where the pattern of missing data has a certain type of hierarchy ("monotone"), there are other possible solutions (see Little and Rubin [1987]). Unfortunately, the highly assorted patterns of missingness noted earlier render such an approach infeasible for the SCF.

5   In the first iteration, the cross-products matrix is computed using all non-missing pairs of variables.

6   To make the distributions somewhat easier to see, the stratum indicators have been randomly altered within broad bands about the value of the original indicator.

7   As expected, the data show a strong negative relationship between the proportion of the population having the items and the coefficients of variation with respect to imputation and to sampling of the mean and median.

8   I have reported on earlier stages of these experiments at the 1997 Record Linkage Workshop in Washington, and the 1998 Statistical Data Protection conference in Lisbon.

9   Another distinction between the two types of error draws is that the error terms in the first two experiments are drawn from the central 95 percent region of the appropriate normal distributions, while the residuals in the second two experiments

are not similarly truncated. However, earlier investigations suggest that removing the truncations of the normal draws would make little difference in the results reported here.

[10] For reference, the plots also show the standard deviations of the bias distributions.

[11] For reference, the plot also shows the mean and standard deviation of the standard deviations of the individual observations.

[12] The vertical axis of the QD plot shows the difference in values of two distributions at their common percentile points shown on the horizontal axis.

[13] The quantile scales in figure 5a and 5b reflect the unweighted distribution of the observations included. In the case of total income, the following are the decile points 10 to 90 of the distribution in thousands of dollars: 9.9, 17.2, 24.5, 32.0, 41.8, 53.0, 73.0, 110.0, 270.0. The corresponding figures for directly held stock are: 1.2, 5.0, 10.0, 22.0, 50.0, 100.0, 231.7, 567.0, 1975.0. In contrast, the weighted figures for income for the same populations are: 7.4, 13.0, 18.5, 24.0, 30.0, 37.0, 45.4, 58.5, 80.0. The weighted figures for stock are: 0.5, 1.0, 2.5, 5.0, 9.0, 15.0, 22.0, 45.0, 100.0.

■ Bibliography

Avery, Robert B. and Gregory E. Elliehausen [1985] "1983 Survey of Consumer Finances: Technical Manual and Codebook," mimeo, Board of Governors of the Federal Reserve System, Washington, DC.

Fellegi, Ivan [1997] "Record Linkage and Public Policy — A Dynamic Evolution," in *Record Linkage Techniques — 1997*, Wendy Alvey and Betty Jamerson (eds.), Federal Committee on Statistical Methodology, Office of Management and Budget, Washington, DC, pp. 3-12.

Fries, Gerhard, Barry W. Johnson, and R. Louise

Woodburn [1997] "Analyzing Disclosure Review Procedures for the Survey of Consumer Finances," presented at the 1997 Joint Statistical Meetings, Anaheim, CA.

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.

Juster, F. Thomas and Kathleen A. Kuester [1991] "Differences in the Measurement of Wealth, Wealth Inequality and Wealth Composition Obtained from Alternative U.S. Wealth Surveys," *Review of Income and Wealth*, 37(1), pp. 33-62.

Kennickell, Arthur B. [1998a] "List Sample Design for the 1998 Survey of Consumer Finances," mimeo, Board of Governors of the Federal Reserve System.

_____ [1998b] "Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances," presented at SDP '98, Lisbon, Portugal.

_____ [1997a] "Analysis of Nonresponse Effects in the 1995 Survey of Consumer Finances," *Proceedings of the Section on Survey Research Methods*, 1997 Joint Statistical Meetings, Anaheim, CA.

_____ [1997b] "Codebook for the 1995 Survey of Consumer Finances," on the Internet at http://www.bog.frb.fed.us/pubs/oss/oss2/95/scf95home.html.

_____ [1996] "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.

_____ [1991] "Imputation of the 1989 Survey of Consumer Finances," *Proceedings of the Section on Survey Research Methods*, 1990 Joint

Statistical Meetings, Atlanta, GA.

_____ Martha Starr-McCluer; and Annika Sundén [1997] "Family Finances in the U.S.: Recent Evidence from the Survey of Consumer Finances," *Federal Reserve Bulletin*, vol. 83 (January), pp. 1-24.

Lillard, Lee; James P. Smith; and Finis Welch [1986] "What Do We Really Know about Wages? The Importance of Nonreporting and Census Imputation," *Journal of Political Economy*, v. 94, no. 3, pp. 489-506.

Little, Roderick J.A. and Trivellore Raghunathan [1997] "Should Imputation of Missing Data Condition on All Observed Variables?" *Proceedings of the Section on Survey Research Methods*, 1997 Joint Statistical Meetings, Anaheim, CA.

_____ and Donald B. Rubin [1987] *Statistical*

Analysis with Missing Data, John Wiley and Sons, New York.

Montalto, Catherine and Jaimie Sung [1996] "Multiple Imputation in the 1992 Survey of Consumer Finances," *Financial Counseling and Planning*, v. 7, pp. 133-146.

Rubin, Donald B. [1987] *Multiple Imputation for Nonresponse in Surveys*, John Wiley and Sons, New York.

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

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.

■ Appendix

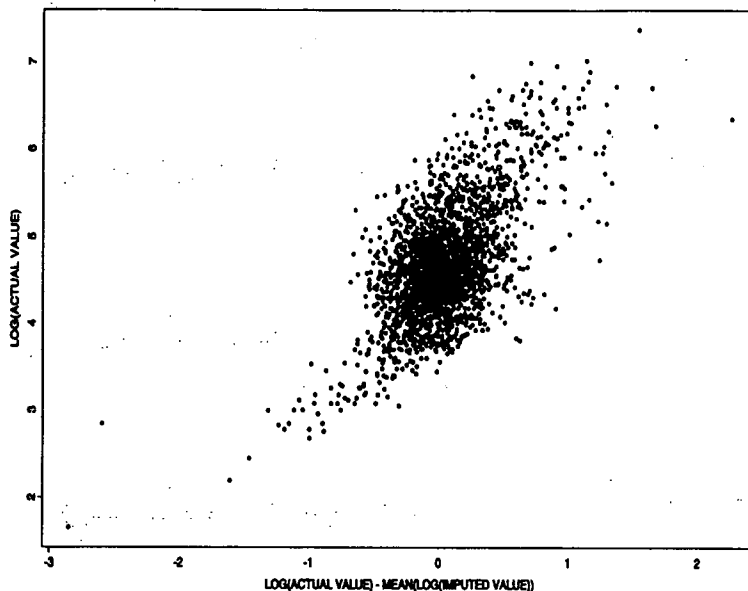**Scatterplot of Actual Values vs. "Bias," Total Household Income, Experiment 1**
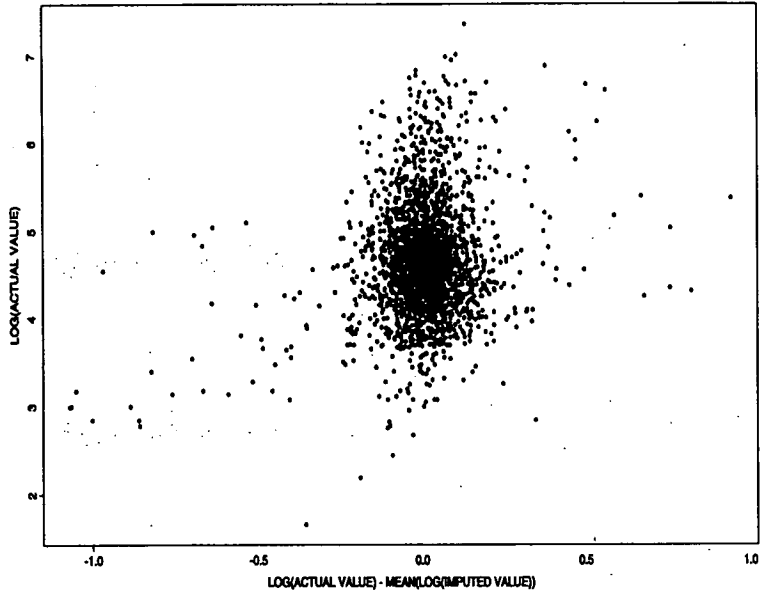
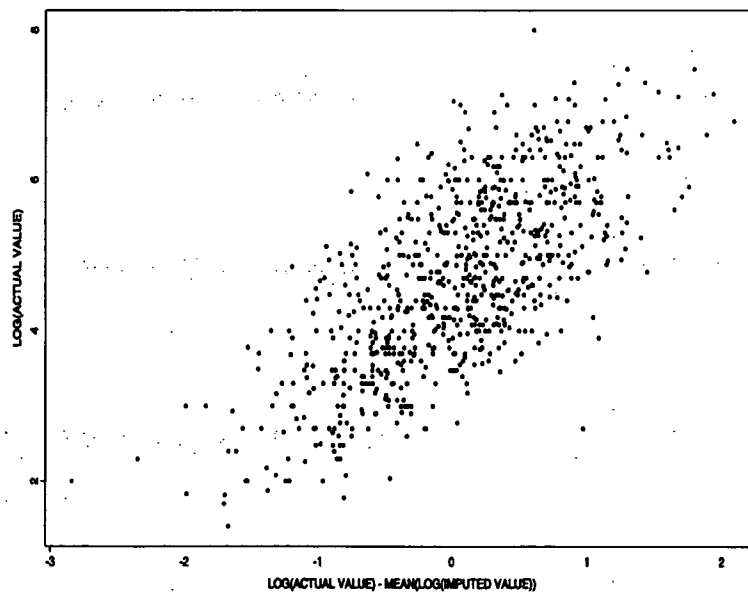**Scatterplot of Actual Values vs. "Bias," Publicly-Traded Stock, Experiment 1**
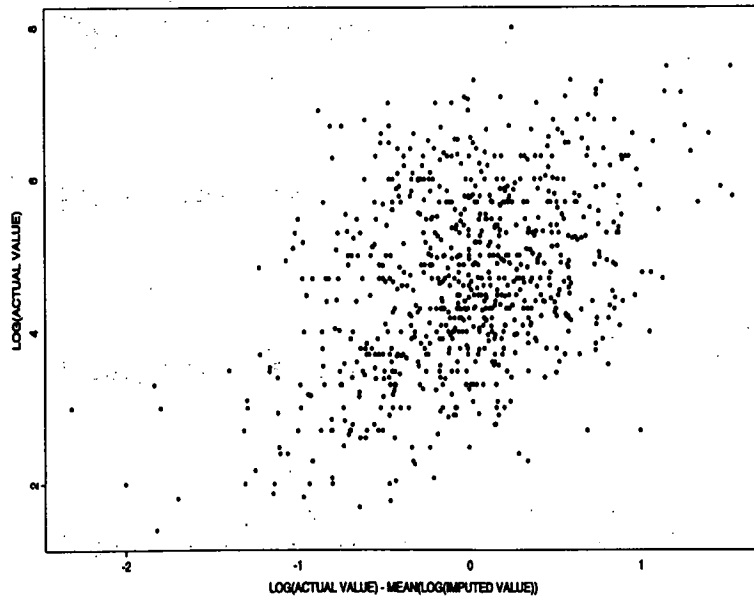
■ Appendix--continued

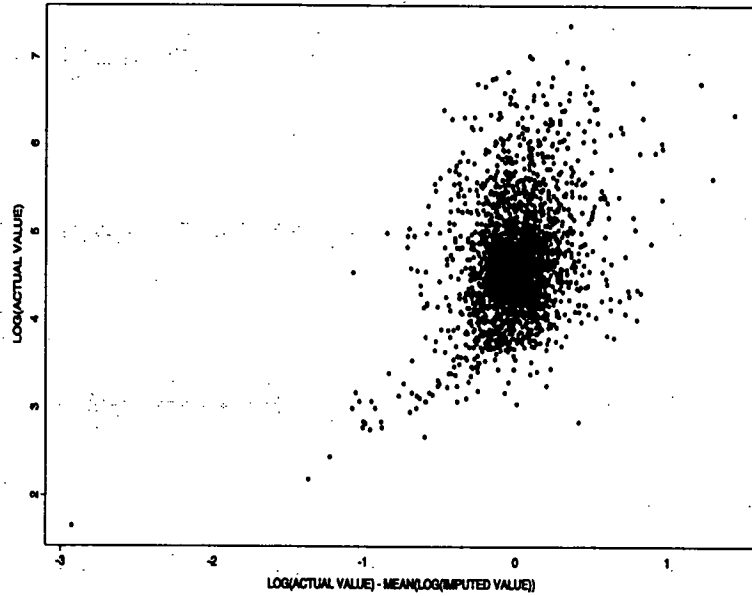**Scatterplot of Actual Values vs. "Bias," Total Household Income, Experiment 2**

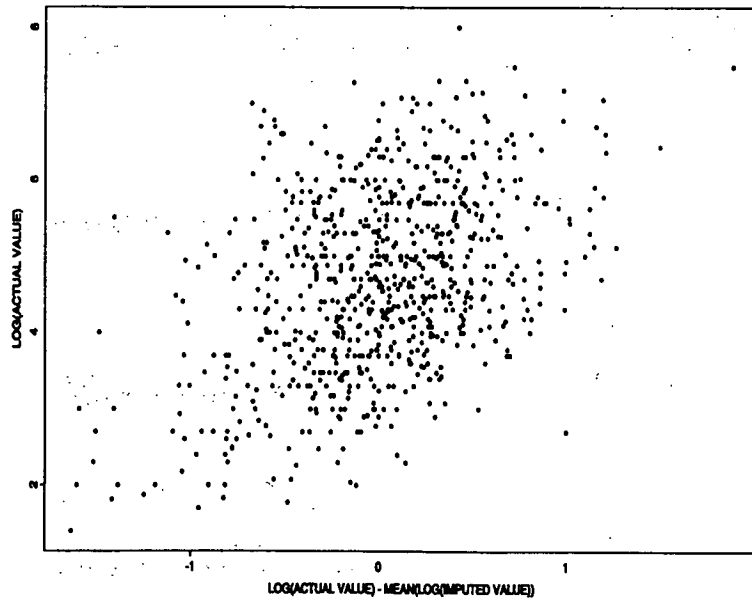**Scatterplot of Actual Values vs. "Bias," Publicly-Traded Stock, Experiment 2**

■ Appendix--continued

Scatterplot of Actual Values vs. "Bias," Total Household Income, Experiment 3
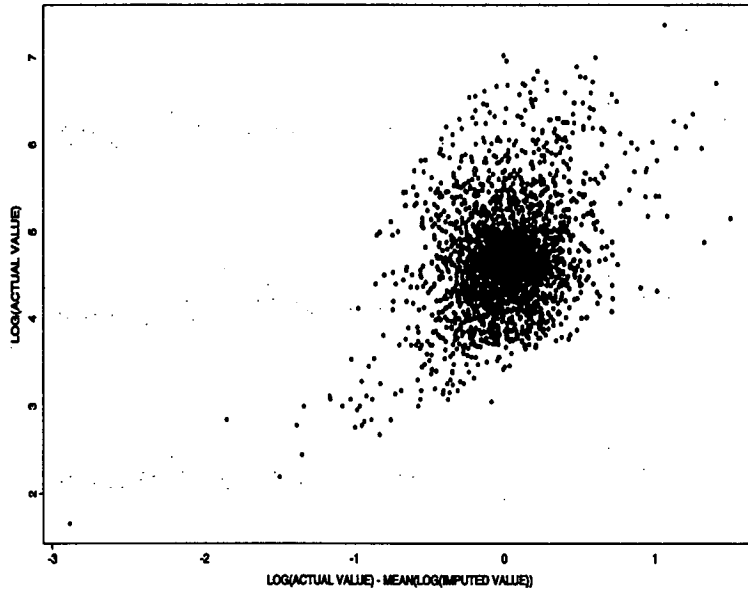

Scatterplot of Actual Values vs. "Bias," Publicly-Traded Stock, Experiment 3

■ Appendix--continued

Scatterplot of Actual Income vs. "Bias," Total Household Income, Experiment 4

Scatterplot of Actual Values vs. "Bias," Publicly-Traded Stock, Experiment 4