Imputation of the 1989 Survey of Consumer Finances: Stochastic Relaxation and Multiple Imputation

Arthur B. Kennickell
Senior Economist and Project Director Survey of Consumer Finances
Board of Governors of the Federal Reserve System
Mail Stop 180, Washington, DC 20551

October 1991

Prepared for the Annual Meetings of the American Statistical Association
Atlanta, Georgia, August 18 to 21, 1991

KEY WORDS: Multiple imputation, Gibbs sampling, stochastic relaxation, Survey of Consumer Finances (SCF)

Opinions expressed in this paper are the responsibility of the author alone and do not necessarily reflect the views of the Board of Governors or the Federal Reserve System. The author gratefully acknowledges the support for this work by staff in the Division of Research and Statistics including Edward Ettin, David Jones, Myron Kwast, and Michael Prell, the advice and support of Fritz Scheuren, comments by Roderick Little and Donald Rubin on an earlier description of the imputation model presented here, the patience and creativity of Louise Woodburn in serving as sounding board for many of the ideas, Gerhard Fries for valuable help in debugging the software, and Robert Denk for his careful work in managing the SCF database. Finally, thanks are due also to Robert Avery, the author's predecessor as SCF project director, who nourished an environment where data collection and imputation are taken seriously.

The Survey of Consumer Finances (SCF) is designed to gather detailed information on the financial and demographic characteristics of U.S households. Inevitably in such a survey, some respondents are unwilling or unable to provide all of the information requested of them. In waves of the SCF before 1989, imputations of missing values were made on an ad hoc basis. A decision was made for the 1989 survey to build systematic imputation and editing software that reflects the current state of knowledge in the area and that would be substantially reusable in future waves of the survey.

This paper describes the Federal Reserve Imputation Technique Zeta (FRITZ) developed for the 1989 SCF. In the next section of this paper, I give a description of the structure of the 1989 SCF and evaluate the incidence of missing data. In the third section, I review some of the theory underlying the procedures applied. In the fourth section, I give an outline of the structure of the FRITZ model. The next section provides some statistics on the performance of the model. In the final section, I provide a brief summary and discuss areas for further research.

1989 SURVEY OF CONSUMER FINANCES

History and Purpose of the Survey

The SCF as a wealth survey traces its origins to the landmark 1963 Survey of Financial Characteristics of Consumers conducted by Dorothy Projector and Gertrude Weiss [1966]. Other surveys known as the SCF were conducted in the 1970s. However, it was only in 1983 that the SCF was revived as a wealth survey. The 1989 SCF is the third of the new series of SCFs. Our hope is that there will be support to continue the survey on a triennial basis.

The 1989 SCF was sponsored by the Board of Governors of the Federal Reserve System, the Statistics of Income Division of the Internal Revenue Service (SOI), the Congressional Joint

Committee on Taxation, the Office of the Assistant Secretary for Planning and Evaluation in the Department of Health and Human Services, the National Institute on Aging, the General Accounting Office, the Small Business Administration, the Office of the Comptroller of the Currency, and the Social Security Administration. The data from the project are used broadly for policy analysis and for more academic research in economics and other areas.

The purpose of the SCF is to provide a comprehensive and detailed view of the financial behavior of households. Altogether well over 1,500 variables were collected.[1] Detailed information was gathered on all assets and liabilities attributable to the primary economic unit in the household.[2] Liabilities include credit card debts, installment loans, mortgages, lines of credit, and other loans. Assets include the principal residence, all types of loans made to others, real estate assets, businesses, various types of accounts, including checking, saving, money market, IRA, Keogh, and brokerage accounts, stocks, mutual funds, bonds, and other assets. Detailed information was also collected on the current job of respondents and their spouses, their current and future pension rights, and other demographic characteristics. A supplementary survey was conducted of the pension providers of currently employed respondents and their spouses in order to obtain more comprehensive and accurate information than could reasonably be expected of households.

---

[1]This figure is adjusted for cases where the same figure could be reported in more than one place.

[2]The "primary economic unit" is defined as the person within a household who holds primary title to the dwelling, who is the principal lessee, or who is otherwise economically dominant, together with the spouse and minor children of that person and any other persons who are economically dependent on that group. A brief section of the interview was devoted to summary information about other economic units within the household.

Questionnaire Design

To accommodate the many types of information requested from a range of socio-economic groups, the questionnaire for the SCF is quite long and highly structured. Typically, the design of the instrument is such that questions about dollar amounts are preceded by one or more filter questions. For example, before asking the amount in a respondent's fifth checking account, the person is asked first whether the unit has any checking accounts, and then how many accounts the unit has. Sometimes respondents are asked different sequences of questions depending on the answer to such filter questions.

Sample Design

The sample design of the 1989 SCF is also complex.[3] The two major parts of the sample are the overlapping panel cross-section based on the 1983 SCF sample (1803 cases), and a new independent cross-section sample (2000 cases). Without loss of information essential for this paper, in this section I will review only the new cross-section sample design. This design is based on a dual-frame design. One part of this sample was drawn from a standard national area- probability frame and the remainder was selected from a list frame developed from administrative files maintained by SOI. A "wealth index" was constructed for each list frame element using income flows reported on tax returns capitalized at various rates of return.[4] Elements were then selected by stratified probability sampling to over-represent units with higher values of the wealth index. The motivation

---

[3]A more complete description of the sample design is available in Heeringa and Woodburn [1991].

[4]For example, taxable interest income was assumed to be supported by a stock of interest-bearing assets equal to ten times the interest income (implicitly this assumes a rate of return of ten percent). See Heeringa, Juster, and Woodburn [1991] for more details.

for the dual frame design was two-fold. First, since an important mission of the survey is to characterize the distribution of financial assets, which are highly concentrated in a small part of the population, an efficient design should over-represent that group.[5] Second, common survey folklore and on-going analyses of the 1989 SCF support the claim that nonresponse tends to be higher for households with higher levels of wealth.[6] In most area-probability samples, there is no means of making systematic adjustment for this differential nonresponse. The advantage of the list sample used for the 1989 SCF is that nonresponse adjustments can be made based on extensive income information contained in the administrative records that served as the sample frame.

Data Collection

Data for the survey were collected between the months of July 1989 and March 1990 by the Survey Research Center at the University of Michigan. Interviews were largely conducted in person and averaged about 75 minutes. Some panel interviews were conducted by telephone. In addition, some other interviews were conducted at least in part by telephone at the request of the respondent. In general, the telephone and in-person interviews appear identical in terms of the proportion of missing information, amount of marginal comments, and other such information.

Data Editing

Before the data were punched, the questionnaires went through several stages of editing. Interviewers performed a consistency check as soon as possible after the interview. In the field office,

---

[5]It was estimated from the 1983 and 1986 SCFs that about a half of one percent of households in the U.S. owns about a quarter of total household net worth. The degree of concentration was even greater for assets such as stocks and bonds. See Avery, Elliehausen and Kennickell [1988] and Avery and Kennickell [forthcoming, 1991].

[6]Preliminary information on unit nonresponse in the 1989 SCF is given in Woodburn [1991].

the questionnaires were examined more closely for evidence of interviewer or respondent error -- with particular attention to the possibility of double-counting of assets and liabilities. Further machine editing was performed on the punched data for more complicated logical problems. One area in which the survey is quite weak is in the retrieval of information missing due to interviewer or respondent errors.

Data changes at all stages of editing represent something very close to, if not identical to, imputation. Generally, a conservative approach was taken to changing data in editing. However, when missing pieces of information were obvious in the context of other information in the questionnaire, the data were filled in at this stage. Records were kept of major changes to the data. As one might expect of an interview that was administered to households of all ranges of financial sophistication, editing was substantial and important for the quality of the final product. Many lessons have been learned in the effort both about how to avoid a number of the more serious problems through improved questionnaire design in the future and about what types of additional information are most useful in resolving inconsistencies that cannot be eliminated entirely.

Unit and Item Nonresponse

The achieved sample for the entire 1989 survey includes 3,803 households. Of this number 3,134 have cross-section representation and 1,479 have panel representation.[7] Of the 2000 new cross- section cases, 866 derive from the SOI list frame. Area-probability and list cases were treated slightly differently in the field. Area- probability cases were approached directly by interviewers, and about 69 percent of these cases were eventually interviewed. The list cases were given a prior

[7]This figure includes both cases from the new cross-section and from the part of the overlapping panel cross-section with cross- section representation.

opportunity to refuse participation by returning a postpaid card. About a third of the list cases refused participation at this stage. The remainder were approached by interviewers, yielding an overall interview rate for the list sample of about 34 percent. While the interview rate for the list cases is not high according to usual criteria, this figure merely makes explicit the differential nonresponse with respect to income that is hidden in other surveys that have insufficient frame information to reveal the problem. Moreover, in the SCF, we have at least the hope of making systematic adjustments to the sample by estimation of response models using the universe data under the assumption that units are conditionally missing at random.[8]

Every observation in the survey contains at least one piece of missing information -- often a very trivial item such as the interviewer ID number. Partial information was available for many items. Respondents who were reluctant to provide dollar values directly were offered a card containing dollar ranges labeled with letters. For total income a more directed "tree" approach was taken to bound income more tightly. Excluding range-card responses, the mean number of missing values per case is 21.6, the median is 11, the 90th percentile is 37, and the total number of missing values to be imputed is 82,125.[9] The mean number of range responses was 3.4 per interview and the total number of such responses was 3,477. For comparison, the maximum possible number of missing values is about 6 million.[10] However, all pieces of missing information are not of equal value in terms of the

---

[8]See Woodburn [1991] for a review of our recent efforts in this direction.

[9]For purposes of this count, if a branch variable is missing, all subsequent variables within the branch are also taken as missing.

[10]If one looks only at dollar amounts of financial assets (checking, money market, savings, and other such accounts, certificates of deposits, stocks, mutual funds, bonds, and trusts), out of a maximum of 136,908 data items, 3350 are missing, the mean number missing per case is 0.9 and the median number is zero.

overall objectives of the survey -- e.g., the amount a respondent has in a sixth checking account is usually less important that the total amount of corporate stocks. Another gauge of severity of the problem, the proportion of missing dollar amounts based on the imputed values, is given below in the discussion of the results of the model.

The structure of missing values is quite complicated. As noted above, the questionnaire is designed so that respondents are led down many question paths with several conditional branches. In addition, a very great number of patterns of missing data appear in the data. For all practical purposes, it is a safe assumption that the overall pattern of missingness for each case is unique. Thus, the imputation of the missing values cannot be addressed routinely using techniques developed for "monotone" patterns of missingness without sacrificing substantial information to achieve monotonicity for subgroups.[11]

Table 1 provides response rates for a nonrandom selection of survey variables for the panel and cross-section observations taken together. As shown in the table, item nonresponse rates vary widely, but generally within a range that is typical of other economic surveys. One exception is 1988 adjusted gross income, which was missing in over 28 percent of cases. I suspect that this very high level of nonresponse had two inportant sources. The field period began later than expected after April 15th and ran longer than expected, and respondents were not encouraged very strongly to look up data where appropriate.

[11]For a discussion of monotonicity and techniques for imputation where missing data patterns are monotone, see Rubin and Little [1987].

Shadow Variables

We have attempted to incorporate in the dataset as much information as possible about what went into the determination of each data value. For example, a piece of information might be provided exactly by the respondent, provided by the respondent as a range, refused by the respondent, unknown by the respondent, inferred from other information at the stage of editing and imputation, etc. Every variable in the survey has a shadow variable that reflects the information content of the primary variable. In all, each of these shadow variables can take on 63 values. The function of these variables is twofold. First, for purposes of automated imputation, these variables serve as a convenient flag for imputation. Such flags become even more important in the context of longitudinal imputation. Second, it is important for many analytic purposes to know exactly which items were not reported by the respondent. For example, some analysts may wish to use different methods of imputation, or to use other techniques to allow for the possibility of specific models of nonignorable nonresponse.

REVIEW OF IMPUTATION THEORY

There are numerous ancestors of the missing value techniques reviewed in this section. For a more complete history, I refer the reader to the detailed references in the landmark National Academy volumes (Madow, Olin, and Rubin [1983]), Little and Rubin [1987], and Rubin [1987].

Three strands of literature are particularly relevant to the work reported in this paper: the EM algorithm, multiple imputation, and Gibbs sampling, or stochastic relaxation. All of the methods discussed here are strongly influenced by Bayesian thinking. A more complete overview of this literature is given in Rubin [1990] and Gelfand and Smith [1990].

The EM algorithm presented as a distinct procedure first appeared in Demster, Laird and Rubin [1977]. That model is intended to estimate parameters in a dataset where some information is only partially observed and direct estimation in the presence of missing information is difficult, but estimation with complete data would be easier. Using observed information, starting estimates of the parameters are computed. These estimates are then used to simulate the missing information. The original information is used along with the simulated information as a basis for maximum likelihood estimation of the parameters of interest. This process continues iteratively until the parameter estimates are sufficiently close to a fixed point. The intuition of this landmark paper underlies all that is reported here.

Rubin's work on multiple imputation (see particularly Rubin [1987] and references therein) serves as a bridge between EM and the later simulation techniques that involve a structure similar to EM. Briefly, multiple imputation simulates the distribution of missing data and, thus, allows a more realistic assessment of variances and a more efficient representation of first moments.

A paper by Tanner and Wong [1987] follows from the methods of EM and ideas of multiple imputation and offers a clear framework for understanding the usefulness of iterative simulation methods in imputation. Tanner and Wong focus on the estimation of a set of parameters where some potential conditioning information is unobserved, but as shown below, it is easy to extend the argument to estimation of missing data. A brief review of part of their arguments may help in understanding the development of this paper.

Let X be unobserved values of a larger set X, let $X_o = \{X - X_u\}$, and let θ be a set of parameter values to be estimated. Using notation similar to Tanner and Wong, one may write

(1) $f(\theta|X_o) = \int_{X_u} f(\theta|X_o,Z)\, f(Z|X_o)\, dZ$ and

(2) $\quad f(X_u|X_o) = \int_\theta f(X_u|\phi,X_o)\, f(\phi|X_o)\, d\phi$.

By substitution and rearrangement of terms, one may write (as do Tanner and Wong)

(3) $\quad f(\theta|X_o) = \int_\theta \{\int_{X_u} f(\theta|X_o,Z) f(Z|\phi,X_o) dZ\,\}\, f(\phi|X_o)\, d\phi$.

Similarly, one may write

(4) $\quad f(X_u|X_o) = \int_{Xu} \{\int_\theta f(\theta|X_o,Z)\, f(X_u|\phi,X_o) d\phi\}\, f(Z|X_o)\, dZ$.

Both (3) and (4) are easily seen to be a recursive relationship that might be solved by iterative substitutions. Tanner and Wong prove that under regularity conditions, (3) (equivalently (4) by simple change of notation) converges uniformly to a unique fixed point.

Equation (4) has a simple interpretation in the imputation framework with the parameters θ serving as an intermediate step using simulation techniques and multiple imputation. Given some starting values, one could draw a number of replicates of θ and X_u in turn until convergence of the posterior distribution of θ or X_u is reached. In order for this method to be practical, one must be able to compute several conditional distributions to be able to draw samples of all of X_u and all of θ simultaneously. In some complex sets of data this constraint is not practical.

Papers by Geman and Geman [1984] and Li [1988] provide useful approaches for dealing with more complex data structures. These papers describe an iterative Markovian procedure of successive simulation of the distribution of variables conditioned on both observed data and distributions of variables previously simulated in the same iteration. The method is typically referred to as stochastic relaxation or Gibbs sampling. The procedure has had extensive applications in the area of image processing. The iterative nature of the procedure is similar to the model of Tanner and Wong with the following exception. If X_u above is partitioned into elements X_{ui}, where i=1 to U, the procedure can be described as a successive simulation of the distribution of the separate elements of θ and X_u conditioned on all available information, where "available" is taken to mean nonmissing information as well as simulated missing data. For example, in iteration I for variable V, one draws X_{uv} from

$$f(X_{uv}|X_o, \chi(I)_{u1},...,\chi(I)_{u\,v-1},\beta_v) \quad ,$$

where $\chi(I)_{uj}$ denotes simulated data on missing variable j in the Ith iteration. Moreover, the set of conditioning variables need not be the entire set of possible variables if it is know that some local structure (or "clique" in the terminology of Geman and Geman) can be assumed for each variable. A variation of this procedure is applied in the construction of the FRITZ model described below. Although convergence is reported to be slow for large numbers of variables, Geman and Geman show that under regularity conditions, the process converges and that the simulated distribution of X_u moves closer to the true latent distribution with each iteration.

DESCRIPTION OF FRITZ

After a review of the literature and of existing procedures, we decided to build comprehensive new imputation software for statistical imputation of the 1989 SCF. Fellegi and Holt [1976] and their proposal for an automated system of edit and imputation sets an imposing standard. Two software packages represent important extensions of the ideas in that paper. Statistics Canada maintains very interesting edit-imputation software, the Generalized Edit and Imputation System (GEIS). A review of this model is given in Giles [1987]. This is a very impressive model. However, for our purposes, the system is too limited in the types of imputation models available. In addition, it appears that it would be cumbersome to implement the multiply-imputed Gibbs sampling direction taken here. The Structured Program for Economic Editing and Referrals (SPEER) developed at the Census Bureau (Greenberg and Surdi [1984]) offers an excellent environment for the implementation of the types of complicated algebraic constraints important in the imputation of the SCF. However, given the nature of the SCF data and the theoretical direction taken here, SPEER appeared too difficult to adapt for our purposes.

In the past, imputation had been performed on an ad hoc basis, with significant and very frequent intervention by analysts at the level of individual imputations well beyond the editing stage. While the effort involved in the development of FRITZ has been great, we believe that much of the core set of procedures can be reused for future SCFs as well as for other purposes.

In designing the imputation procedures for the SCF, we were constrained in a number of ways. First, "reasonable" estimates of the missing data for a subset of financial variables needed to be available very quickly for pressing policy needs. Second, for several reasons we were limited to about a year from the time the first complete data tape was received in the fall of 1990. Third, the system was required to allow the imposition of prior information both in the form of edit rules and specific information about individual cases. Fourth, the procedure had to accommodate any possible pattern of missing values. Finally, the work had to be performed with limited computer resources (storage and CPU).

There is a continuum of changes to the respondents' answers from the point of interviewer recording, through primary data editing, to statistical imputation. Virtually all imputations made after the primary editing stage are model-based, though a small number of documented cases have been imputed judgmentally — typically variables that would be quite cumbersome to impute, but which are resolved with very high probability upon inspection.[12] Judgmentally imputed variables are flagged

[12] Fewer than 500 variables were imputed judgmentally. One frequent imputation of this type is the case where a respondent has reported an amount of a payment or a type of income, but the frequency of that amount is missing (e.g., a respondents who reports receiving $450 in Social

as such in the shadow variables and a file of these decisions was created as a part of the survey documentation.

FRITZ was designed to handle the great majority of statistical imputations. Although the procedure is iterative and involves multiple imputations, for relative transparency of exposition, it will be convenient to act at first as though the model were the more usual case of single imputations computed without iterations. The general procedures applied in the first iteration are used in all later iterations. Special problems induced by the mixture of panel and cross-section data will only be presented later in the discussion.

Basic Procedures in the First Iteration

Let the potential set of variables collected for a given case r (r=1 to R) be denoted by X_r where X_r is a vector of N variables.[13] Additionally, let X_g (of rank N) and X_m (of rank $N_m = N - N_r$) denote, respectively, the partitioning of X_r into variables that are available and those missing for some reason. The goal of the imputation process is to obtain a good estimate of $F(X_m | X_g)$. Multiple imputation allows the dataset itself to stand as a proxy for that distribution.

Using a variation on the technique of Gibbs sampling or stochastic relaxation described above, FRITZ proceeds through the variables to be imputed in a predetermined sequence making imputations variable-by-variable, and for a given variable, independently for each observation.[14] In the process,

---

Security payments, but reports no time interval over which the payments are received). Almost all other imputations of this sort are similarly obvious.

[13]In fact, N is not a constant for all cases. In particular, panel cases were asked different questions in a number of areas. Allowing for this distinction would only complicate the notation here.

[14]In principle, the sequencing of the variables for imputation should not be important. Every imputation should be conditioned on every possible bit of relevant information. Practically,

the information set available for imputation of each case expands as imputation proceeds through the sequence of variables. Imputed variables are treated exactly like reported variables within each iteration. That is, in the first iteration we estimate

$F(\beta_1 | X_g)$

$F(X | X_g, \beta_1)$

.....

$F(\beta_n | X_g \cup X_{m<n})$

$F(X_n | X_g \cup X_{m<n}, \beta_n)$

.....

$F(\beta_N | X_g \cup X_{m<n})$,

$F(X_N | X_g \cup X_{m<n}, \beta_N)$,

where $X_{m<n}$ denotes the missing values imputed in the sequence before variable n and where the parameters of the distribution are estimated from reported and simulated data in the previous iteration, and where β_j is an intermediate parameter vector corresponding to the "M" stage of EM.

In the FRITZ system, there are four types of model-based imputations: imputation of continuous variables, binary variables, and polychotomous variables, and nonparametric imputation. Unfortunately, theory does not offer much help in finding the "true" functional form of F. In the case

---

this is not possible both because of degrees of freedom problems and because of the time required to invert enormous matrices in the application. For example, suppose for an observation that the number of business the household owns is missing and the only information known is that for the second business the household loaned $500,000 to the business. A model could be specified especially for this case, but that would be quite cumbersome. Alternatively, a very large model could be built, but that would likely exhaust the degrees of freedom. It turns out that careful sequencing of the imputation of variables often allows the use of summary variables that appear to be a reasonable proxy for more detailed data (e.g., the total amount of money in all checking accounts instead of the amount in each of as many as six accounts).

of most continuous-variable imputations, it is assumed implicitly that the variables with missing values can be taken to have a conditional distribution of the form

$F(G(a)|H(b)) \sim \text{Normal}(m\ ,\ \sigma^2_a)$, where a is a variable with missing values, b is a set of conditioning variables, and G and H are transformations of a and b, respectively. This assumptions amounts to assuming that

$$G(a) = H(b) + \epsilon_a\ ,\ \text{ where } \epsilon_a \sim \text{Normal}(0,\sigma^2_a)\ .$$

Typically H is assumed to be multiplicative in b and the transformations G and H are taken as log transforms, implicitly yielding the linear model,

$$A = \text{constant} + \beta_1 B_1 + ... + \epsilon_A\ ,$$

where the capital letters indicate the log transform. The great benefit of this assumption is that a relatively simple covariance matrix of the variables forms a sufficient statistic for imputation and the simulation of A is straightforward.

In practice, we can be almost certain that the variables we observe are a subset of the appropriate vector B. At the least there are likely idiosyncratic factors for every observation that would be extremely difficult to represent as a reasonably small set of variables even in principle. Once we face the fact that all of B is not known, a potential problem of nonignorable nonresponse arises — that is, conditional on the observed variables the set of nonrespondents for a given item may be a nonrandom subset of the whole sample.[15]

In FRITZ an agnostic approach is taken to the set of observed variables chosen to proxy for B. In principle, it might be desirable to take the conditioning set as a series expansion of the function G involving all variables available for each observation. In practice, degrees of freedom limit the

---

[15]See Little [1983].

number of variables, interaction terms, and higher order terms that can feasibly be included. In any event, no attempt is made to exclude variables that have no obvious "structural" interpretation -- the underlying model is a pure reduced form. Most often, the maximal set of conditioning variables for a given case is on the order of 200 or more variables, frequently including a number of recoded variables particularly relevant for a given imputation. Typically included in the set of variables used is a group of interviewer observations on respondents' level of suspicion before and after the interview, their level of interest, etc. The data indicate a reasonable variation in the amount of information reported for all levels of these variables. The hope is that these variables will be correlated with unobserved characteristics of item nonrespondents and, thus, mitigate the potential for nonignorable nonresponse bias.

While there is no guarantee that such an approach eliminates — or even reduces — possible response bias, such a strategy may be the best practical insurance against bias. Our means for testing this assumption are very limited. One possibility may be to compare the distribution of variables available for the list sample in the administrative records with the distribution of comparable variables in the survey. While confidentiality concerns strictly limit direct comparison of cases in the two files, it may be possible to look at such distributions within some sub-groups.

Operationally, FRITZ looks at a given case, determines whether the current variable in the sequence should be imputed, determines which variables in the conditioning set are available either as reported values or previously imputed values, and computes a randomized imputation. As noted earlier, the combinations of missing values varies widely over all cases so that virtually every case involves a different "regression." Thus, $\hat{A}$, the imputed value of variable A for observation j is drawn according to

$\beta_A = [B_{g(i)}\ B_{g(i)}]^{-1}\ [B_{g(i)}\ A]$ and

$\hat{A}_j \sim F(A|B_{gj},\beta_A)$,

where $B_{g(i)}$ denotes the set of values of all observations for variables included in B_{gj}, the set of all available (reported and already imputed within the iteration) values for case j.

In the first iteration, an improper imputation is made by drawing a value from the distribution implied taking the model coefficients β to be fixed and assuming that ϵ_A is distributed normally with mean zero and variance given by $A'A - A'B_{g(i)}\ [B_{g(i)}'\ B_{g(i)}]^{-1}\ B_{g(i)}\ 'A$, where the relevant moments are computed as described below. The allowed distribution of ϵ_A may be truncated or otherwise altered using prior information or editing rules. Because the inversion of a large matrix is usually involved for each such imputation, this method is quite time-consuming.

The moment matrix for the continuous and binary imputations is computed for the appropriate sub-population — e.g., the covariance matrix needed for the imputation of the amount of holdings of certificates of deposit is computed using only households that actually have such instruments. Conveniently, a moment matrix computed using the maximal set of conditioning variables allowed will suffice for every case. The software automatically selects a submatrix for each case corresponding to the conditioning variables available. In the first iteration, the covariance matrix for the imputations is computed without weights and using all non-missing pairs of variables for each observation.[16] As is well-known, this method of calculation allows the possibility that the covariance matrix may no longer be positive definite, implying a negative value for ϵ^2_A. In practice ϵ^2_A is rarely estimated to be negative. For convenience at the first stage, ϵ^2_A is given a floor of zero. The

---

[16]Both cross-section and panel observations are pooled for estimations for variables that are common in the two parts of the survey. Special problems, which are noted later in this paper, arise when we begin to use prior-wave information in the estimation.

alternative of using only cases with full information would usually too drastically reduce the number of observations available for the calculation.

A more serious problem in the covariance estimation is that induced by the presence of very influential cases. Typically this has been a problem in cases where there are coefficients of conditioning variables that are identified by a very small number of observations. In such cases as have been detected, the set of conditioning variables has been reduced to exclude the offending variables. Unfortunately, I have not had either computer power or staff resources to explore this dimension systematically. FRITZ writes out information about imputations as it proceeds and such problems detected to date have been found through inspection of the model output. One sign of problems is the frequent inability of a given model to impute a value strictly within the bounds imposed by the constraints (either determined through edit rules, or from range card estimates). The most desirable approach would be to use robust estimation techniques for the covariance matrix. This will be an important line of research for this project in the future.

There appears to be another — perhaps related — class of problems with covariance matrices estimated in this way. Initially, it would happen occasionally that the model would impute values that were clearly absurd. Although a sweep algorithm with a facility for elimination of near-singularities is used in FRITZ, decomposition of the covariance matrix indicated a situation normally corresponding to near-collinearity (i.e., very large condition numbers). Moreover, the problem disappears once a completely imputed file is available for covariance estimation after the first iteration. Thus, the problem seems to stem from a characteristic of using all non-missing pairs for variables in the first iteration. Although I have not been able to resolve the problem analytically, I have implemented a numerical patch in the first iteration only that is related to principal component regression.

For binary variables, it is assumed that the same model holds as in the continuous case. This amounts to the somewhat more suspect assumption that the linear probability model applies. Problems with the linear probability model are well-known.[17] The model fails to account for the information implied by the fact that probabilities must be in the closed interval [0,1] and, because the model is heteroskedastic, produces inefficient estimates of the model parameters. Much better from a theoretical point of view would be to pose the relationship as a probit, logit or other such explicit probability model.[18] As it turns out here, however, the informational requirements of such models are too large to be practical. First, such models must be estimated iteratively, requiring an additional pass through all of the data at each iteration of that probability model. In addition, as is the case with the continuous variable imputations, patterns of missing data are such that virtually every observation has a different information set. Because there is no low-dimensional set of summary statistics that would apply to all subsets of conditioning variables, virtually every observation would require a different model and additional passes through the data. Goldberger [1964] has suggested that one use the estimates from the linear probability model to create weights for a second iteration which, as Rubin has pointed out in conversation, would amount to a first Newton step in the maximum

[17]See, for example, Judge et. al. [1985], p. 756 ff.

[18]As Roderick Little has pointed out in converstaion, the discriminate model uses the same set of input statistics as the linear probability model, but has the advantage that outcomes are contrained to lie between zero and one. In the on-going revision of the FRITZ model, the discriminant function approach is being explored.

likelihood estimation. Unfortunately, the time required for even that refinement is prohibitive given the current speed of the computers available to the project.

Given an estimated probability from the linear probability model, a draw is made from the implied binomial distribution to determine the outcome. Some key polychotomous imputations are structured as the sequential prediction of binary choices. The input covariance matrix is computed exactly as in the continuous variable case above.

Less critical polychotomous variables are imputed using a type of randomized hotdeck. Cases are arrayed as a multidimensional frequency table using a number of classifying variables. The imputation procedure randomly selects a value for the missing variable from the appropriate conditional cell. A minimum number of cases is required in each cell. If that minimum is not achieved, there are rules for collapsing adjacent cells. Very closely related to this simple hotdeck procedure is a nonparametric regression technique. Essentially, the difference is that continuous variables are allowed in the frequency table and the collapsing procedures select a slice of specified size from the joint distribution of the variables.

Higher-Order Iterations

In the first iteration, the goal is to estimate a reasonable set of starting values for further iterations. At the end of the first iteration, we have one copy of the dataset with all missing values filled in. From the second iteration and on, the initial dataset containing missing values is replicated 3 to 5 times, and the missing values are filled in based on statistics computed using the completed dataset from the prior iteration. In the second iteration, the covariance matrices and other such basic statistics needed for imputation are estimated from the reported data and the single imputations from the first iteration. In higher-order iterations, these statistics are pooled across the imputation replicates.

Following the example of Tanner and Wong, the number of replicates is allowed to vary over iterations. The first iteration involves one replicate, the second iteration three replicates, and later iterations five replicates. The primary justification for the varying number of replicates is the severe constraints on disk storage.[19] Given the complex tree structure of the data, it is an open question how many replicates may be needed to reflect adequately the variation due to imputation.

If the assumptions we have made do not move us too far from the requirements of the underlying theory, at each iteration, FRITZ will move closer toward the true latent posterior distribution of the data. For convenience, we define convergence in terms of changes in the implied distribution of wealth, rather than as a function of all possible variables. In many applications, Gibbs sampling is known to converge slowly. Unfortunately, this may be a severe limitation in this application. The first iteration of FRITZ requires at least 11 days — largely due to the number of matrices that must be inverted -- on a fairly fast Solbourne minicomputer computer dedicated to the project. Subsequent iterations can take 3 weeks or longer. The amount of time required places particularly strains on our ability to debug such complex software. For this paper, only the output of the first two iterations is available.

Structure of Software

FRITZ is written in the SAS language (version 6.03) and makes extensive use of the SAS MACRO language to create a subroutine structure. The entire model comprises about 100,000 lines of code. The majority of the basic computer code is written in PROC IML. The advantages of SAS

[19]An efficiently written replicate requires about 120M of disk space.

are that it is a known and closely-monitored product with excellent data management facilities. Among the important disadvantages are that it can be slow and there are bugs in PROC IML in the version of SAS used.

Generally, FRITZ requires four types of statements for each variable to be imputed. First, a set of rules is specified to impose editing rules and other prior information about the feasible distribution of imputations. Second, a set of transformations is specified given the imputed value. Third, a set of statistics is computed for the imputation for the appropriate sub-population. Finally, there is a call to the central imputation driver. An annotated version of a simple set of such routines for the imputation of directly-held corporate stocks is provided as appendix A.[20]

As in the example in the appendix, the edit rules generally are posed in terms of feasible upper and lower bounds for imputations. In the example there may be prior information from a range card, from information on the subset of stock held in the company where someone in the household works, from the number of companies in which the household owns stocks, or from the combination of legal requirements and the amount borrowed on a stock margin account. Often in other cases the constraints derive from complex algebraic relationships between variables. In the recode section, an attempt is made to create all related variables for which there is complete information at the time of the imputation, including constructed variables that may be needed in further imputations in the same program. In the example, the program may fill in the missing amount or percent that the stock had changed in value since its purchase and other items.

The call to the module that computes the moment matrix generally specifies a list of variables and a sub-population on which these estimates should be made. In the appendix, the calculation is

[20]The main imputation program is discussed further in Kennickell [1991].

made only for the population that has already been determined to have stocks. Finally, the call to the main imputation program specifies a type of model, the input covariance matrix, conditions under which the variable should be imputed, and so forth.

The larger program is constructed as a series of six large SAS MACROs. Each of these modules in turn creates a working dataset containing all variables to be imputed and all condtioning variables needed within the module, calls a large number of smaller modules for the imputation of individual variables (such as that described above), and then stores the basic variables that have changed in the main dataset. For the interested reader, Appendix B provides a sketch of how the overall code is arranged. More detailed information is available from the author upon request.

Panel Imputations

The work reported in this paper is primarily concerned with the imputation of data for the 1989 SCF cross-section, rather than the panel part of that project. With the exception of marital and employment history variables, which were constructed from updated historical information for the panel, the questions in the survey for the cross-section cases are a proper subset of the questions asked of the panel cases. In order to expand the basis for covariance estimation, the panel observations have been included with the cross- section variables in the imputations along with a dummy variable indicating that a case is a panel case.

Explicit recognition of the longitudinal dimension of the panel cases increases the complexity of the imputation and editing problem by an order of magnitude. Patterns of missing data are even more complicated when the longitudinal dimension is added: a respondent may report one data item in 1989, but have refused the same item in 1983, or data items may be only partially comparable. In principle, one should use all information in all periods to model the joint distribution of all variables and impute -- or re-impute -- all data missing in any period.

Panel imputation is an enormously complicated task that grows in complexity with the number of panel observations.[21] The amount of time needed for modeling and for computation would be extremely large. Moreover, degrees of freedom in modeling would quickly become very limited. The strategy I expect to follow is a compromise. A vector of key panel financial and other variables will be created along with indicators of the degree of missingness for each variable. If the amount of missing information exceeds a certain percent, a variable will be treated as missing. A single pass will then be made through the data, augmenting the sets of conditioning variables used for cross-section imputation to include the constructed variables and adding imputation modules for the newly included variables. Since the historical variables will only exist for the panel cases, problems may arise in the covariance estimation similar to those encountered in the first iteration of the imputations discussed above. Alternatively (and preferably), one could simply treat the panel variables as missing from the cross-section cases and proceed as before. Unfortunately, the very large number of variables that would then be "missing" may render such an approach infeasible with current resources. In light of the importance of panel data for research, further work in panel imputation will have a high priority in the SCF project, and I hope in the work of others.

SOME RESULTS FROM THE MODEL

As noted earlier, short of adding up all missing values equally, it is difficult to find a universally applicable single measure of the information missing due to nonresponse. After

[21]The task is further complicated here by the fact that the data structure of earlier SCFs makes it very difficult to identify which values were imputed.

imputation, other metrics are available. One such compelling measure is the proportion of dollars imputed for various items. Table 2 provides an estimate of the unweighted percentage of dollars that were imputed for selected items for the panel and cross-section cases together. Weighted percentages might be more informative here, but sampling weights are at such a stage that I do not believe such estimates would be reliable. Weighted and unweighted estimates will be provided later for the panel and cross-section separately.

An estimated 19 percent of total net worth in the sample was imputed, with 4.9 percentage points of that amount imputed using range information. In the case of total income, 35.2 percent of dollars were imputed with an amazing 30.5 percentage points of this amount constrained by range estimates. Most of the other figures reported lie somewhere between these cases.

Table 2 also displays the coefficient of variation due to imputation for components of household net worth and other variables based on data from the second iteration of FRITZ. As might be expected, the model performs better in terms of predicting higher-order aggregates than in terms of individual assets. For example, while the variation for money market accounts is 7.6 percent, the total variation in net worth is only 0.3 percent.

Because only the first two iterations of the model are currently available, it is impossible to say very much about the empirical convergence properties of FRITZ. However, as shown in figure 1, it does appear from the data that are currently available that the difference in the cumulative distribution of financial assets (a key variable) is virtually unchanged between the first two iterations.

SUMMARY AND FUTURE RESEARCH

The FRITZ model was developed to provide a coherent framework for the imputation of the 1989 SCF with the expectation that the model could be incrementally adjusted for use in future SCFs.

An attempt has been made to utilize the most current research in imputation. To our knowledge, this effort represents the first attempt to apply multiple imputation or methods of stochastic relaxation to a large social science survey.

In addition to the question of panel imputation noted above, there are many areas in need of further research. One of the most pressing concerns in the imputation of the SCF is to modify FRITZ to take advantage of obvious opportunities for parallel processing of the data. Although the software modifications would be complex, in principle on our UNIX system, it would be possible to farm the work out to a large number of independent processors with a central coordinator. While saving time is a reasonable goal alone, it is also the case that it is only by speeding up the processing that we can have a hope of implementing significant improvements in FRITZ. Of particular interest are changes related to improved robustness of the imputation and improved nonparametric imputation techniques.

Currently, the software used for nonparametric imputation is limited in the number of conditioning variables that can be used. It is possible to "trick" the software by creating complex index numbers to be used as conditioning variables. The difficulty in allowing a larger number of variables is in devising reliable classes of rules for grouping observations to create high-dimensional cells with a sufficient number of observations.

At the time this paper was completed, only the first two iterations of FRITZ were available. As we progress, it will be important to study the convergence properties of the model. If the model converges as slowly as Gibbs sampling appears to converge in some application, it is unlikely that in the near future there will be sufficient computer power to allow calculation to a near neighborhood of convergence. A related problem is the sensitivity of the model to starting values. Wu [1983] has noted that the convergence of EM to a global maximum is not always guaranteed. Since the Gibbs sampling approach is in a sense logically subordinate to EM, FRITZ might be expected to have similar problems.

Finally, we plan to examine how our estimates of imputation variance change as the number of replicates increases. Because the survey variables have a complicated hierarchical structure, it seems plausible that a larger number of replicates might be necessary to allow the data to express possible variations in that structure due to imputation. However, additional replicates are very expensive in terms of time required for imputation, amount of storage required for the data, and time required at the analysis stage. As in many other applied statistical exercises, greater computer power will eventually solve alot of problems.

BIBLIOGRAPHY

Avery, Robert B., Gregory E. Elliehausen, and Arthur B. Kennickell [1988] "Measuring Wealth with Survey Data: An Evaluation of the 1983 Survey of Consumer Finances," Review of Income and Wealth, Series 34, No. 4 (December), pp. 339-369.

__________ and Arthur B. Kennickell [1991, forthcoming] "Household Saving in the U.S" Review of Income and Wealth.

__________ and __________ [1991, forthcoming] "Changes in the Distribution of Wealth 1983 to 1986 based on the Surveys of Consumer Finances," mimeo, Cornell University.

Demster, A.P, N.M Laird, and D.B. Rubin [1977] "Maximum Likelihood From Incomplete Data Via the EM Algorithm," Journal of the Royal Statistical Society, Ser. B, v. 39, pp. 1-38.

Felligi, I.P. and D. Holt [1976] "A Systematic Approach to Automatic Edit and Imputation," Journal of the American Statistical Association, Vol. 71, No. 353, pp. 17-34.

Gelfand, Alan E. and Adrian F.M. Smith [1990] "Sampling-Based Approaches to Calculating Marginal Densities," Journal of the American Statistical Association, Vol. 85, No. 410, pp. 398-409.

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.

Giles, Philip [1987] "Towards the Development of a Generalized Edit and Imputation System," Proceeding of the Third Annual Research Conference, Bureau of the Census, pp. 185-193.

Goldberger, Arthur S. [1964] Econometric Theory. New York: Wiley.

Greenberg, B. and R. Surdi [1984] "A Flexible and Interactive Edit and Imputation System," Proceeding of the Section on Survey Research Methods, American Statistical Association, pp. 421-436.

Heeringa, Steven G. and R. Louise Woodburn [1991] "The 1989 Surveys of Consumer Finances Sample Design Documentation," mimeo, Survey Research Center, University of Michigan, Ann Arbor.

__________, F. Thomas Juster, and R. Louise Woodburn [1991] "The 1989 Survey of Consumer Finances: A Survey Design for Wealth Estimation," forthcoming in the Review of Income and Wealth.

Herzog, Thomas N. and Donald B. Rubin [1983] "Using Multiple Imputations to Handle Nonresponse in Sample Surveys" (Chapter 15), in Incomplete Data in Sample Surveys, New York: Academic Press.

Judge, George G., W.E. Griffiths, R. Carter Hill, Helmut Lutkepohl, and Tsoung-Chao Lee [1985] The Theory and Practice of Econometrics. New York: Wiley.

Kennickell, Arthur B. [1991] "A Manual for the MACRO IMPUTE1," mimeo, Board of Governors of the Federal Reserve System.

Li, Kim-Hung [1988] "Imputation Using Markov Chains," Journal of Statistical Computing and Simulation, Vol. 30, pp. 57-79.

Little, Roderick J.A. [1983] "The Ignorable Case" (Chapter 21) and "The Nonignorable Case" (Chapter 22), in Incomplete Data in Sample Surveys, New York: Academic Press.

__________ and Donald B. Rubin [1987] Statistical Analysis with Missing Data, New York: Wiley.

Madow, William G, Olin, and Donald B. Rubin (editors) [1983] Incomplete Data in Sample Surveys, New York: Academic Press.

Projector, Dorothy S. and Gertrude S. Weiss [1966] Survey of Financial Characteristics of Consumers, Washington: Board of Governors of the Federal Reserve System.

Rubin, Donald B. [1987] Multiple Imputation for Nonresponse in Surveys, Wiley: New York.

___________ [1990] "EM and Beyond," mimeo Department of Statistics, Harvard University (forthcoming in Psychometrica).

Tanner, Martin A. and Wing Hung Wong [1987] "The Calculation of Posterior Distributions by Data Augmentation,"(with comments) Journal of the American Statistical Association, Vol 82, No. 398, pp. 528-550.

Woodburn, R. Louise [1991] "Using Auxiliary Information to Investigate Nonresponse Bias," paper presented at the annual meetings of the American Statistical Association, Atlanta.

Wu, C.F. Jeff [1983] "On the Convergence Properties of the EM Algorithm," Annals of Statistics, Vol. 11, No. 1, pp. 91-103.

APPENDIX A
Imputation Modules for Amount of Directly-Held and Publicly-Traded Stocks

```
*   this MACRO defines constraints on the imputation of value of directly-held publicly-traded
    corporate stocks;
*   this MACRO and the following one are written in IML code and are called in the processing of
    the IMPUTE1 MACRO below;

    %MACRO TR1ISTK3;

*       define default bounds;
*       assume have at least $10 in each company where own stock;
        LB=10*MAX(1,NCOSTK);
        UB=9999999999;

*       use information on amount in margin account + legal requirements to set LB;
        IF (AMARGIN>0 & JAMARGIN<24) THEN LB=AMARGIN*4;
        ELSE IF (JAMARGIN>=24 & JAMARGIN<=45) THEN DO;
*           extract information from range card (JMARGIN is shadow variable for amount
            borrowed on margin account);
            %CARDBB(J=JAMARGIN,UB=MUB,LB=MLB);
            LB=MAX(LB,MLB*4);
        END;

*       use information from range card for stocks (JASTK is shadow variable for amount of
        stock);
        IF (JASTK>=24 & JASTK<=45) THEN DO;
            %CARDBB(J=JASTK,UB=SUB,LB=SLB);
            UB=MIN(UB,SUB);
            LB=MAX(LB,SLB);
        END;

*       use information on amount of stock in place where work + $10/company (JASTKWRK is
        shadow variable for amount of stock in company where household member works);
        WLB=0; WUB=0;
        IF (JASTKWRK>=24 & JASTKWRK<=45) THEN DO;
            %CARDBB(J=JASTKWRK,UB=WUB,LB=WLB);
        END;
        IF (WLB>0) THEN LB=MAX(LB,WLB+MAX(0,(NCOSTK-1)*10));
        ELSE IF (ASTKWRK>0) THEN
            LB=MAX(LB,ASTKWRK+MAX(0,(NCOSTK-1)*10));
        IF (NCOSTK=1 & WUB>0) THEN UB=MIN(UB,WUB);
```

```
*       put bounds in log form;
        UB=LOG(MAX(LB,UB));
        LB=LOG(MAX(10,LB));

    %MEND TR1ISTK3;

*   this MACRO sets recodes using imputation of log(stock);

 %MACRO TR2ISTK3;

*   compute level value of stock from log;
    ASTK=INT(EXP(LASTK)+.5);

*   compute percentage/amount of gain/loss since bought all stock;
    IF (GAINSTK=1 & PGSTK>.Z & AGSTK<=.Z) THEN AGSTK=
        MAX(1,INT(.5+ASTK*(1-1/(1+PGSTK/10000))));
    IF (GAINSTK=1 & PGSTK<=.Z & AGSTK>.Z) THEN PGSTK=
        MAX(1,INT(.5+((ASTK/(ASTK-AGSTK))-1)*10000));
    IF (GAINSTK=5 & PLSTK>.Z & ALSTK<=.Z) THEN ALSTK=
        MAX(1,INT(.5+ASTK*(1-1/(1+PLSTK/10000))));
    IF (GAINSTK=5 & PLSTK<=.Z & ALSTK>.Z) THEN PLSTK=
        MAX(1,INT(((ASTK/(ASTK-ALSTK))-1)*10000));

*   try to compute total financial assets;
    AFIN=ACHKG+AIRA+AMMA+ACD+ASAV+AMUTF+ASAVB+ABOND+ASTK;
    IF (AFIN>.Z) THEN LAFIN=LOG(MAX(1,AFIN));

*   if only stock in one company and have stock in business where work, the value of stock same
    as value of stock in business where work;
    IF (NCOSTK=1 & STKWORK=1) THEN DO;
        ASTKWRK=ASTK;
        LASTKWRK=LASTK;
    END;

*   create interaction term from log(stock) and log(number brokerage transactions in past year);
    LNBTSTK=LNBRTRA*LASTK;

%MEND TR2ISTK3;
```

* this MACRO computes covariance matrix for imputation using standard input set (%INCVARS2) and variables specific to variable — using only population with stocks for calculation;

```
%SSCPMISS(VAR=%INCVARS2 NCSTK GAINSTK LAGSTK LALSTK STKWORK
    LASTKWRK GAINMF NCMUTF LNBRTRA,DATA=&OLDI,OUT=_TAB,
    WHERE=%STR(DSTOCK=1));
```

* call to the main imputation MACRO;
* specify continuous variable model, dependent variable is log of holdings of corporate stock, JASTK is the name of the shadow variable, _TAB contains the covariance matrix estimated above, the dataset containing the values to be imputed is given by &NEWI, the MACROs TR1ISTK3 and TR2STK3 are called, imputation is restricted to cases that own stock and have a current missing value (or have a temporary value based on a range card), TOLER specifies a variance decomposition routine in the first iteration to stabilize the model, AUX specifies variables that are needed for the imputation, and KEEP specifies variables to be kept in the working dataset;

```
%IMPUTE1(TYPE=CONTIN,DEP=LASTK,MISS=JASTK,TABLE=_TAB,DATA=&NEWI,
    TRANS1=TR1ISTK3,TRANS2=TR2ISTK3,WHEREV=DSTOCK ASTK JASTK,
    WHERE=(DSTOCK=1 & (ASTK<=.Z | JASTK<45)),TOLER=YES, AUX=STKWORK
    NCOSTK JASTK ASTK ACHKG AIRA AMMA ACD ASAV AMUTF LNBRTRA JAFIN
    ASAVB ABOND ASTKWRK JASTKWRK AMARGIN, KEEP=ASTK AFIN LAFIN
    ASTKWRK LASTKWRK LNBTSTK AGSTK ALSTK PGSTK PLSTK GAINSTK);
```

APPENDIX B
Overall Organization of FRITZ

- Control file for FRITZ (Federal Reserve Imputation Technique Zeta);
- * Designed and implemented for the 1989 SCF;
- Arthur B. Kennickell;
- * original version December 20, 1989;
- Current version August 2, 1991

```
*   set and define all FILENAMES here;
FILENAME IMPUTE1 '/...';
FILENAME INCOME1 '/...';
FILENAME RESPROP1 '/...';
FILENAME INSTIT1 '/...';
FILENAME MORTDEB1 '/...';
FILENAME CONDEB1 '/...';
FILENAME BUS1 '/...';
FILENAME LABOR1 '/...';
FILENAME DEMOG1 '/...';
FILENAME SSCP '/...';
FILENAME CARDB '/...';
FILENAME CONVERGE '/...';
FILENAME BACKUP '/...';
LIBNAME LOUISE '/...';
LIBNAME LITTLE '/...';
LIBNAME RUBIN '/...';

•   all include statements here for MACROs;
%INCLUDE CARDB;
%INCLUDE SSCP;
%INCLUDE IMPUTE1;
%INCLUDE INCOME1;
%INCLUDE RESPROP1;
%INCLUDE INSTIT1;
%INCLUDE MORTDEB1;
%INCLUDE CONDEB1;
%INCLUDE BUS1;
%INCLUDE LABOR1;
%INCLUDE DEMOG1;
%INCLUDE CONVERGE;
%INCLUDE BACKUP;
```

```
* begin imputation control code;
%MACRO FRITZ;

* set-up variables;
    %LET ITERNUM=1;
    %LET CNVRG=NO;
    %LET NOPRINT=YES;
    %LET SEED=1111111;
    %GLOBAL ITERNUM NOPRINT SEED;

*   FRITZ1 is always the name of the dataset used to compute the statistics for imputation;
•   FRITZ2 is always the name of the dataset that contains the imputed values;
•   at the 1st iteration, the input dataset is the original dataset and the output dataset is a single
    replicate of the same dataset;
    %LET FRITZ1=%QUOTE(LOUISE.SCFR);
    %LET FRITZ2=%QUOTE(RUBIN.SCFR1);

•   begin iteration loop;
    %DO %UNTIL (&CNVRG EQ YES OR &ITERNUM EQ 100);

*       create (new) replicates of the original RECODES dataset to contain the imputations;
        DATA &FRITZ2;
            SET LOUISE.SCFR;

*           set number of replicates;
            %IF (&ITERNUM=1) THEN %LET NREPL=1;
            %ELSE %IF (&ITERNUM=2) THEN %LET NREPL=3;
            %ELSE %LET NREPL=5;

*           alter original ID number (XX1) to reflect replicate number;
            DO I=1 TO &NREPL;
                X1=XX1*10+I;
                OUTPUT %QUOTE(&FRITZ2);
            END;

        RUN;

*       invoke the MACROS that contain the imputation modules for each variable;

*       total income for the PEU, AGI, principal branch variables, financial assets;
        %INCOME1(OLDDATA=%STR(&FRITZ1),NEWDATA=%STR(&FRITZ2));

*       home value, vehicles, loans made, total value of investment properties;
```

```
        %RESPROP1(OLDDATA=%STR(&FRITZ1),NEWDATA=%STR(&FRITZ2));

*       financial institutional relationships;
        %INSTIT1(OLDDATA=%STR(&FRITZ1),NEWDATA=%STR(&FRITZ2));

*       impute mortgages;
        %MORTDEB1(OLDDATA=%STR(&FRITZ1),NEWDATA=%STR(&FRITZ2));

*       impute terms of all consumer loans and all non-mortgage loans for home purchase and
        home improvement;
        %CONDEB1(OLDDATA=%STR(&FRITZ1),NEWDATA=%STR(&FRITZ2));

*       businesses, credit cards, lines of credit, misc. properties, working or not, misc. attitudinal
        questions;
        %BUS1(OLDDATA=%STR(&FRITZ1),NEWDATA=%STR(&FRITZ2));

*       impute labor force participation, current job pensions, and employment history;
        %LABOR1(OLDDATA=%STR(&FRITZ1),NEWDATA=%STR(&FRITZ2));

*       current pension/SS, other future pensions, past settlements, inheritances, Section Y
        demographics unimputed at this point (including non-PEU finances), and misc. income,
        etc.;
        %DEMOG1(OLDDATA=%STR(&FRITZ1),NEWDATA=%STR(&FRITZ2));

*       determine convergence (CNVRG=YES/NO);
        %CONVERGE;

*       after the first iteration, back up imputed dataset (FRITZ1) on tape;
        %IF (&ITERNUM GT 1) %THEN %DO;
            %BACKUP(&FRITZ1);
        %END;

*       after first iteration, delete imputed dataset from previous iteration;
        %IF (&ITERNUM GT 1) %THEN %DO;
            PROC DATASETS;
                DELETE %STR(&FRITZ1);
            RUN;
        %END;
```

```
*       determine location of files for next iteration;
        %IF (%EVAL(MOD(&ITERNUM,2) EQ 0) %THEN %DO;
            %LET TAG1=RUBIN;
            %LET TAG2=LITTLE;
        %END;
        %ELSE %DO;
            %LET TAG1=LITTLE;
            %LET TAG2=RUBIN;
        %END;
        %LET FRITZ1=%QUOTE(%UNQUOTE(
            &TAG1.%QUOTE(.)%UNQUOTE(SCFRR%EVAL(&ITERNUM+1))));
        %LET FRITZ2=%QUOTE(%UNQUOTE(
            &TAG2.%QUOTE(.)%UNQUOTE(SCFRR&ITERNUM)));

*       increment iteration number;
        %LET ITERNUM=%EVAL(&ITERNUM+1);

    %END;

%MEND FRITZ;
%FRITZ;
```

Table1
Item Nonresponse Rates, Selected Items, Percent *
1989 Survey of Consumer Finances, Panel and Cross-Section, Unweighted

Item	*Don't know*	*Not avail.*	*Unknown whether have item*	*Range resp.*	*Memo item: % all cases inap.*
Balance on bank credit cards	0.6	1.2	0.0	0.8	30.9
Value of own home, excl. mobile homes	1.6	1.2	0.0	0.6	29.6
Amount outstanding on mortgage on home	3.2	2.0	0.1	1.2	58.5
Have any owned cars	0.0	0.0	0.0	0.0	0.0
Number of owned cars	0.0	0.2	0.0	0.0	11.5
Value of 1st business with mgt. role	15.0	3.1	1.1	4.8	73.7
Have checking acct.	0.0	0.2	0.0	0.0	0.0
Number of chkg. accts.	0.0	0.2	0.3	0.0	11.6
Amt. in 1st chkg. acct.	1.4	4.6	0.3	2.5	11.6
Amount of CDs	3.1	7.8	1.8	5.7	73.9
Amt. of savings bonds	4.9	2.3	2.5	2.5	76.3
Amount of stock, excl. mutual funds	5.4	5.1	1.5	5.4	65.4
Cash value of life insurance	31.7	1.6	4.5	1.9	53.0
Wage for respondent currently working	0.9	4.3	1.1	2.1	42.3
Balance in 1st defined contribution pension plan for respondent	16.7	1.8	6.2	2.7	84.2
Total family income	2.1	1.7	14.6	0.0	0.0
Filed 1988 tax return	0.2	1.0	0.0	0.0	0.0
Amount of 1988 adj. gross income	29.0	6.3	1.4	5.2	13.4
Amount of 1st inheritance	5.9	4.4	3.3	3.6	68.1
Amount of 1988 charitable contrib.	1.6	1.9	2.5	3.0	48.9
Wage income for non-primary unit members	30.3	2.9	5.1	5.9	90.1

* Computed as a percent of cases either where response was appropriate or where it was unknown whether response is appropriate.

Table 2
Proportion of Total Dollar Value Imputed,
Coefficient of Variation Due to Imputation, Various Items.
1989 Survey of Consumer Finances, Panel and Cross-Section, Unweighted.

Item	*% total $ imputed with range info.*	*% total $ imputed w/o range info.*	*Coeff. of var. due to imputation*
Checking accounts	3.1	11.8	0.039
IRA and Keogh accounts	10.9	4.2	0.013
Money market accounts	4.2	16.3	0.076
Savings accounts	3.6	13.8	0.056
Certificates of deposit	5.4	8.0	0.014
Corporate stock	13.2	15.5	0.056
Mutual funds	7.5	15.6	0.087
Savings bonds	3.6	41.7	0.026
Other bonds	3.9	8.3	0.042
Trust assets and annuities	7.5	6.0	0.024
Cash value of life insurance	1.8	19.0	0.033
Notes held	0.8	15.4	0.037
All financial assets	7.1	12.0	0.005
Principal residence	3.3	2.2	0.003
Other real estate	5.5	2.9	0.016
All businesses	22.2	6.3	0.066
Vehicles	2.6	0.3	0.001
Misc. assets	9.8	5.0	0.011
Total assets	5.3	12.9	0.005
Credit card debt	6.0	4.2	0.012
Consumer debt	0.1	4.2	0.000
Principal residence mortgage	0.7	6.3	0.002
Other mortgages	4.4	5.8	0.036
Lines of credit outstanding	0.9	3.4	0.007
Misc. debt	12.2	7.1	0.028
Total debt	3.8	6.3	0.030
Net worth	4.9	14.1	0.003
Total income	30.5	4.7	0.010
Adjusted gross income	15.6	38.6	0.036
Total inheritances received	6.6	19.5	0.117
Total charitable contributions	4.3	2.6	0.009

Figure 1: Cumulative Distribution of Total Financial Assets, Iterations 1 and 2,, FRITZ Model, 1989 SCF